UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO DE TECNOLÓGICO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Adriano Luiz de Souza Lima

Automated Assessment of the Visual Aesthetics of App Inventor User Interfaces with Deep Learning

Florianópolis
2023

Adriano Luiz de Souza Lima

Automated Assessment of the Visual Aesthetics of App Inventor User Interfaces with Deep Learning

Tese submetida ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Santa Catarina para a obtenção do título de doutor em Ciência da Computação.
Orientadora: Profa. Dr.a rer. nat. Christiane Gresse von Wangenheim, PMP

Florianópolis
2023

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Lima, Adriano Luiz de Souza
Automated Assessment of the Visual Aesthetics of App Inventor User Interfaces with Deep Learning / Adriano Luiz de Souza Lima ; orientadora, Christiane Gresse von Wangenheim, 2023.
167 p.

Tese (doutorado) - Universidade Federal de Santa Catarina, Centro Socioeconômico, Programa de Pós-Graduação em Ciência da Computação, Florianópolis, 2023.

Inclui referências.

1. Ciência da Computação. 2. estética visual. 3. design de GUI. 4. valiação de qualidade. 5. deep learning. I. Gresse von Wangenheim, Christiane. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Ciência da Computação. III. Título.

Adriano Luiz de Souza Lima

Automated Assessment of the Visual Aesthetics of App Inventor User Interfaces with Deep Learning

O presente trabalho em nível de doutorado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Mario de Noronha Neto, Dr.
Instituto Federal de Santa Catarina

Prof. Renan Vinicius Aranha, Dr.
Instituto Federal de Mato Grosso

Prof. Rafael de Santiago, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de doutor em Ciência da Computação.

GOV.BR
Documento assinado digitalmente
PATRICIA DELLA MEA PLENTZ
Data: 20/06/2023 07:23:29-0300
CPF: ***.243.700-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de
Pós-Graduação

ICP-Edu
Documento assinado digitalmente
Christiane Anneliese Gres...Wangenheim
Data: 19/06/2023 13:34:09-0300
CPF: ***.218.819-**
Verifique as assinaturas em https://v.ufsc.br

Profa. Dr.a rer. nat. Christiane Gresse von
Wangenheim, PMP
Orientadora

Florianópolis, 2023.

ACKNOWLEDGEMENTS

I would like to thank:

- My supervisor, Profa. Dr.a rer. nat. Christiane Gresse von Wangenheim, PMP, for her invaluable guidance throughout the whole process;
- My colleagues, for their help and support during the mandatory classes;
- The members of the GQS group, students and professors alike, for being part in the experiments and helping me out with the technical difficulties;
- My family, for their never ending patience, love and support.

RESUMO

A estética visual está sendo considerada cada vez mais como um fator essencial para o sucesso das aplicações móveis, afetando a experiência e a percepção dos usuários, o que torna a sua avaliação crucial no processo de design de interfaces. Recentemente, as abordagens de aprendizado de máquina têm apresentado resultados bastante promissores na previsão da estética visual. No entanto, até o presente momento, essas soluções propostas avaliam apenas interfaces de usuário baseadas na web. Portanto, neste trabalho foi desenvolvido um modelo de deep learning para quantificar a estética visual de interfaces de usuário móveis Android. Um modelo de rede neural convolucional (CNN) com um corpus de screenshots de interfaces de aplicativos Android foi treinado, adotando-se uma abordagem de aprendizado supervisionado baseado em regressão. Após o treinamento, o modelo prevê a distribuição das avaliações de estética visual para as GUIs de aplicativos Android, a partir das quais é possível calcular suas pontuações de estética visual e o grau de concordância entre os avaliadores. O seu desempenho foi medido como o erro quadrático médio entre o grau de estética visual previsto e o atribuído por avaliadores humanos. Também foi avaliada a saída do modelo analisando-se sua correlação e concordância com a avaliação humana. Entre as contribuições desta pesquisa estão um modelo de aprendizado profundo que pode automatizar a avaliação do aspecto estético de aplicativos móveis e um conjunto de dados com 820 imagens de interfaces de usuários desenvolvidas com o App Inventor e rotuladas. Com esse modelo, espera-se reduzir o custo e o tempo desse tipo de avaliação, permitindo sua execução a qualquer momento durante o processo de desenvolvimento de software. Ele pode estar disponível para organizações de software com poucos recursos alocados para design de interface do usuário, contribuindo para a melhoria da qualidade do software e processo de desenvolvimento. Outro uso possível é no contexto educacional. Espera-se que a automatização da avaliação da estética visual apoie o ensino do design visual, diminuindo o esforço de avaliação e resolvendo outros problemas, como o favoritismo.

Palavras-chave: estética visual; avaliação de qualidade; design de GUI; *deep learning.*

RESUMO EXPANDIDO

INTRODUÇÃO

Diversas limitações dos dispositivos móveis podem afetar a qualidade de seus aplicativos, principalmente a usabilidade. A diferença mais notável entre as interfaces gráficas de usuário (GUIs) de dispositivos móveis e desktops é o tamanho pequeno que limita o posicionamento dos elementos e a multiplicidade de contextos em que esses dispositivos são usados (RAHMAT et al., 2018).

Como parte da qualidade do produto de software, a estética visual refere-se à beleza das GUIs de sistemas de software interativos e tem sido cada vez mais reconhecida como um fator essencial em sua usabilidade percebida, credibilidade e avaliação geral (HAMBORG; HÜLSMANN; KASPAR, 2014). GUIs atraentes dão aos usuários a impressão imediata de que são úteis e fáceis de usar (TUCH et al., 2012a), mesmo quando oferecem usabilidade ruim (BHANDARI; CHANG, K.; NEBEN, 2019). E devido à sua importância, as avaliações de estética visual são cruciais para melhorar a forma como os usuários percebem a qualidade dos sistemas de software e aumentar suas chances de sucesso comercial (BHANDARI et al., 2017). Mas, embora a adesão a princípios objetivos de design ajude a criar GUIs atraentes (SCHLATTER; LEVINSON, 2013), a estética visual também é altamente subjetiva (PALMER; SCHLOSS; SAMMARTINO, 2013), o que significa que é preciso considerar como as pessoas o percebem para executar uma avaliação justa.

Embora as diretrizes de design sejam ferramentas apropriadas para ajudar a alcançar resultados amplamente aceitos dentro dos contextos-alvo, elas exigem treinamento extensivo e experiência prática, tornando difícil para não profissionais interpretá-las adequadamente (MINIUKOVICH; DE ANGELI, 2015a). A dificuldade de entregar GUIs com alto grau de estética reside na má compreensão das preferências estéticas dos usuários (WANG, C.; REN, 2018). O julgamento estético é altamente subjetivo e influenciado por gênero, valores culturais ou gosto pessoal, o que significa que as respostas estéticas podem diferir de pessoa para pessoa (PALMER; SCHLOSS; SAMMARTINO, 2013). No entanto, parece que quando um determinado grupo de usuários considera uma GUI atraente, eles compartilham uma experiência estética semelhante, alcançando algum grau de concordância intersubjetiva (ZEN; VANDERDONCKT, 2016). Essa concordância indica que eles apresentarão respostas semelhantes, como classificar ou classificar igualmente objetos (PALMER; SCHLOSS; SAMMARTINO, 2013).

A avaliação típica da estética visual é fazer com que os usuários-alvo indiquem manualmente sua percepção das GUIs (MOSHAGEN; THIELSCH, 2010), um método caro e demorado que demanda recursos consideráveis (MINIUKOVICH; DE ANGELI, 2015a). Uma forma de minimizar esse esforço é automatizando a avaliação da estética visual para detectar e visualizar prontamente aspectos problemáticos do projeto (MINIUKOVICH; DE ANGELI, 2014a). As avaliações automáticas podem beneficiar designers e desenvolvedores não profissionais, especialmente porque hoje em dia há uma tendência significativa em que cada vez mais pessoas com experiência em outros domínios estão criando aplicativos móveis para resolver problemas relacionados às suas áreas (PATERNÒ, 2013).

Mais recentemente, técnicas de deep learnig foram aplicadas para quantificar a estética visual de GUIs, como páginas da web (DOU et al., 2019) ou designs de GUI (XING et al., 2021), seguindo o sucesso de avaliações estéticas de fotografias (DENG; LOY; TANG, 2017). Embora existam pesquisas investigando a medição da estética visual de interfaces de desktop ou web, poucas pesquisas abordaram GUIs móveis (LIMA; GRESSE VON WANGENHEIM, 2021).

Nesta pesquisa será analisada a seguinte questão de pesquisa: É possível automatizar a avaliação da estética visual das interfaces de usuário dos aplicativos do App Inventor com desempenho, confiabilidade e validade aceitáveis?

OBJETIVOS

O objetivo geral deste trabalho é desenvolver, aplicar e avaliar um modelo para a avaliação estética visual automatizada de interfaces de usuário de aplicativos App Inventor usando técnicas de deep learning para automatizar essa avaliação. Os objetivos específicos são: sintetizar a fundamentação teórica sobre qualidade de software, especialmente no que diz respeito à estética visual de aplicativos móveis e deep learning; analisar o estado da arte em matéria de avaliação estética visual de interfaces de aplicações móveis; desenvolver um modelo usando deep learning para avaliar a estética visual das interfaces de usuário do App Inventor; e avaliar o modelo desenvolvido quanto ao seu desempenho, confiabilidade e validade.

METODOLOGIA

Este trabalho é classificado como pesquisa aplicada, visando gerar conhecimento para aplicações práticas voltadas para a solução de problemas específicos (SILVA; MENEZES, 2001), envolvendo o conhecimento necessário para o desenvolvimento de um modelo de deep learning para avaliar a estética visual de GUIs de aplicativos App Inventor. Esta pesquisa aborda quantitativamente o problema, buscando traduzir informações em números para classificá-los e analisá-los estatisticamente (SILVA; MENEZES, 2001). Quanto aos seus objetivos, a presente pesquisa é exploratória, pois visa abordar e conhecer o problema para explicitá-lo ou construir hipóteses (GIL, 2008). Foram adotados os seguintes passos: 1 - Fundamentação teórica, com a síntese de conceitos relevantes relacionados ao design visual e à estética visual em aplicativos móveis como parte da usabilidade da GUI e sua avaliação. Os conceitos relevantes para o trabalho foram sintetizados por meio de uma análise bibliográfica (GIL, 2008); 2 - Identificação do estado da arte, com pesquisa e revisão de estudos e trabalhos relacionados para estabelecer suas possíveis contribuições e definir o estado da arte da avaliação da estética visual da GUI. Nesta etapa, realizamos dois estudos de mapeamento sistemático seguindo o procedimento definido por Petersen et al. (2008): um para obter o estado da arte da avaliação da estética visual da GUI e outro para descobrir como a estética visual das GUIs é subjetivamente avaliada com base em avaliações humanas; 3 - Projeto e desenvolvimento do modelo de avaliação automatizada; 4: Avaliação do modelo. Avaliação de desempenho do modelo em relação ao uso do erro quadrático médio (MSE) (diferença quadrática média entre previsões e rótulos) como métrica de desempenho. Por fim, foram comparados os resultados do modelo com os de outros modelos, analisando sua correlação (BONETT; WRIGHT, 2000) e concordância (BLAND; ALTMAN, 1986) com as respostas humanas.

RESULTADOS E DISCUSSÃO

O estudo da estética é um desafio não só pela sua componente subjetiva, mas também pela forma como as respostas subjetivas se relacionam com as propriedades reais dos objetos observados. A pesquisa sobre o estado da arte em relação à avaliação estética visual de GUIs móveis revelou que poucos estudos propõem um método para esse tipo de avaliação.

E embora alguns outros trabalhos avaliem a estética visual de GUIs usando deep learning, nenhum deles é aplicado em GUIs móveis considerando as características que as diferem de outras GUIs. Como propriedade adicional desse tipo de avaliação, não foi encontrada nenhuma pesquisa que considerasse a concordância dos usuários a respeito da estética visual de qualquer GUI. Esses resultados demonstram a originalidade dos resultados desta pesquisa.

Embora esta pesquisa lide com a avaliação automática da estética visual, uma tarefa preliminar foi garantir que os humanos pudessem avaliar adequadamente cada GUI para que seus rótulos correspondessem corretamente às respostas estéticas dos humanos. Assim, foi realizado um mapeamento sistemático para entender como é feita essa avaliação subjetiva das GUIs e também quais instrumentos são mais adequados para a tarefa. Antes da definição do instrumento de classificação, foi realizado um estudo exploratório para garantir que ele tivesse alta confiabilidade e validade quando comparado com outro instrumento bem estabelecido.

Para mitigar o risco de comprometer essa atividade com avaliadores cansados e entediados, foi desenvolvido um aplicativo móvel, chamado GUI Labeler, usando o App Inventor com a intenção de reduzir o número de toques na tela para cada GUI e simplificar a tarefa. O GUI Labeler também permitiu que os participantes parassem a qualquer momento em que se sentissem cansados e retomassem facilmente seu trabalho. Este aplicativo pode ser facilmente adaptado para rotular qualquer imagem com uma escala de classificação.

Antes do treinamento dos primeiros modelos, foi criado um conjunto de dados com capturas de tela das GUIs que geraram um grau suficiente de concordância entre os avaliadores humanos em relação à sua estética visual. Após a criação do primeiro conjunto de dados, modelos ResNets com diferentes profundidades foram treinados para comparar seus desempenhos. Quando analisado em relação às avaliações humanas em um novo conjunto de dados contendo apenas GUIs não vistas, o ResNet50 mostrou excelente correlação e concordância, indicando que era um forte candidato para automatizar avaliações estéticas visuais.

Após os primeiros resultados, o conjunto de dados foi expandido com novas GUIs, buscando equilibrar as GUIs bonitas com as feias. Esta iteração mostrou a dificuldade de encontrar belas GUIs do App Inventor, apesar dos esforços para selecionar as mais bonitas. Novamente diferentes modelos ResNets treinados. Seu desempenho também foi superior ao de outras arquiteturas (VGG19 e EfficientNet B0) treinadas com o mesmo conjunto de dados. Com um conjunto de dados maior, as previsões mantiveram a forte correlação e concordância com os rótulos, embora um pouco abaixo da iteração anterior.

Como iteração final, as camadas de entrada e saída do modelo foram adaptadas para receber um vetor 5-dimensional, representando a distribuição de classificações, como os rótulos da GUI. Além de permitir o cálculo do grau de estética visual das GUIs, também foi possível extrair o grau de concordância entre os avaliadores humanos. Mais uma vez, o ResNet50 demonstrou ser superior quando comparado ao VGG19 e ao EfficientNet B0 e apresentou desempenho muito semelhante ao modelo treinado para prever a estética visual como um único escore, mostrando-se adequado para automatizar esse tipo de avaliação.

CONSIDERAÇÕES FINAIS

Neste trabalho apresentamos dois modelos para a avaliação automática da estética visual de GUIs móveis por meio de deep learning. O primeiro modelo apresenta como saída o grau de estética visual em um único valor numérico, enquanto que o segundo resulta na predição da distribuição das avaliações recebidas pelos avaliadores humanos, da qual é

possível calcular tanto o grau de estética visual quanto o grau de concordância entre os avaliadores. Outra contribuição é um conjunto de dados contendo 820 screenshots de GUIs móveis, rotuladas de acordo com a sua estética visual, e que está disponível para pesquisas futuras. Por fim, como trabalhos futuros, planeja-se fazer a integração desses modelos com o Codemaster para oferecer a estudantes de computação uma ferramenta para a avaliação de suas interfaces.

Palavras-chave: estética visual; avaliação de qualidade; design de GUI; *deep learning*.

ABSTRACT

Visual aesthetics is increasingly seen as an essential success factor for mobile applications, affecting users' experience and perception, making its assessment crucial in the interface design process. Recently, machine learning approaches have shown great promise in predicting visual aesthetics. Yet so far, these proposed solutions only evaluate web-based user interfaces. Therefore, we have developed a deep learning model to quantify the visual aesthetics of Android mobile user interfaces. We trained a convolutional neural network (CNN) model with a corpus of screenshots of Android app interfaces, adopting a regression-based supervised learning approach. After training, the model predicts the distribution of visual aesthetics ratings to Android app GUIs, from which it is possible to compute their visual aesthetics scores and the degree of agreement among raters. We measured its performance as the mean squared error between the predicted visual aesthetics degree and that assigned by human raters. We also evaluated the model output by analyzing its correlation and agreement with the ground truth. A contribution from our research is a deep learning model that can automate the assessment of the aesthetic aspect of mobile apps and a dataset with 820 labeled GUIs developed with App Inventor. With this model, we expect to reduce the cost and time of this type of assessment, allowing its execution at any time during the software development process. It can be available to software organizations with few resources allotted for UI design, contributing to the software quality improvement and development process. Another possible use is in the educational context. Automating the assessment of visual aesthetics is expected to support the teaching of visual design by decreasing the evaluation effort and solving other problems, such as favoritism.

Keywords: visual aesthetics; quality assessment; GUI design; deep learning

LIST OF FIGURES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

1.1 CONTEXTUALIZATION OF THE PROBLEM

Mobile devices are increasingly present in consumers' daily lives, available at any time and place (DOMOFF et al., 2019), with 250 million active devices in Brazil (ANATEL, 2023) and 17 billion worldwide (STATISTA, 2023) in 2023. Mobile devices offer different experiences from desktops (FLORA; WANG, X.; CHANDE, 2014), and their quality has gained importance, as they are present in many aspects of consumers' routines (MEHRA; PAUL; KAURAV, 2020). However, several limitations of mobile devices can affect app quality, especially usability. The most notable difference between mobile and desktop graphical user interfaces (GUIs) is their small size limiting the placement of elements and the multitude of contexts in which these devices are used (RAHMAT et al., 2018). Mobile GUIs also vary in the use of color and the vertical alignment of elements (MINIUKOVICH; DE ANGELI, 2015a) or the way they interact with users, using, for example, virtual keyboards partially covering GUIs and touch movements to enlarge or reduce visualization (RAHMAT et al., 2018).

Software quality is the degree to which the software product meets explicit and implicit needs when used under specified conditions (ISO, 2011). As part of the software product quality, visual aesthetics refers to the beauty of GUIs of interactive software systems and has been increasingly recognized as an essential factor in their perceived usability, credibility, and overall appraisal (HAMBORG; HÜLSMANN; KASPAR, 2014; TUCH et al., 2012b). Considering that most human-computer communication is visual, appealing GUIs give users the immediate impression that they are helpful and easy to use (NORMAN, 2002; TUCH et al., 2012b), even when they offer poor usability (ANDERSON, 2011; BHANDARI; CHANG, K.; NEBEN, 2019; ZEN; VANDERDONCKT, 2016). On the other hand, unattractive interfaces obscure the intent and meaning, slow down, and confuse users (ISMAIL, 2002). Furthermore, visual aesthetics has been reported as one of the strongest determinants of credibility (FOGG et al., 2003; OYIBO et al., 2018; ROBINS; HOLMES, 2008). And due to its importance, visual aesthetics assessments are crucial to improve how users perceive the quality of software systems and increase their chances for commercial success (BHANDARI et al., 2017). When it comes to mobile apps, visual aesthetics becomes essential because the first impression of its appearance is often decisive in the choice between downloading an app or not from the millions available at each app store (BHANDARI; CHANG, K.; NEBEN, 2019). That way, not only are the most popular mobile apps intuitive and easy to use, but they also feature beautiful GUIs (MORAN et al., 2018). But although adherence to objective design principles helps create appealing GUIs (SCHLATTER; LEVINSON, 2013; STONE et al., 2005), visual aesthetics is also highly subjective (PALMER; SCHLOSS; SAMMARTINO, 2013), which means that one needs to consider how people perceive it to execute a fair assessment.

Designing decisions based on personal taste are not a guarantee of visually pleasing GUIs (MINIUKOVICH; DE ANGELI, 2014a). And although design guidelines are appropriate tools to help achieve widely accepted results within target contexts, they require extensive training and practical experience, making it hard for non-professionals to interpret them accordingly (MINIUKOVICH; DE ANGELI, 2015a). Also, the difficulty of delivering GUIs with a high degree of aesthetics lies in the poor understanding of users' aesthetic preferences to meet their expectations (WANG, C.; REN, 2018). Aesthetic judgment is highly subjective and strongly influenced by gender, cultural values, or personal taste, meaning that aesthetic responses may significantly differ from person to person (PALMER; SCHLOSS; SAMMARTINO, 2013). Even among people with similar backgrounds, agreement on aesthetic matters is hard to achieve (LIMA; GRESSE VON WANGENHEIM, 2021). Yet, it seems that when a particular group of users considers a GUI attractive, they share a similar aesthetic experience, reaching some degree of intersubjective agreement (ZEN; VANDERDONCKT, 2016). That agreement indicates that they will present similar responses, such as equally rating or ranking objects (PALMER; SCHLOSS; SAMMARTINO, 2013). Therefore, the average response can indicate the visual aesthetics perception among a particular population (PALMER; SCHLOSS; SAMMARTINO, 2013). The challenge is to determine whether some representative set of people will judge an object as attractive or unattractive (PALMER; SCHLOSS; SAMMARTINO, 2013). Thus, visual aesthetics assessments need to be valid and reliable to avoid this dependency on personal preferences and guide and corroborate design choices (MINIUKOVICH; DE ANGELI, 2015a).

A typical approach to assess visual aesthetics is to have target users manually indicate their perception of GUIs' overall appearance (LAVIE; TRACTINSKY, 2004; MOSHAGEN; THIELSCH, 2010). Yet, this is an expensive and time-consuming method that demands considerable resources, which might be unavailable for small companies or individual professionals (MINIUKOVICH; DE ANGELI, 2015a). A way to minimize this effort is to automate the GUI visual aesthetic assessment to promptly detect and visualize problematic design aspects (MINIUKOVICH; DE ANGELI, 2014a). Automatic assessments are helpful in the early stages of mobile application development or at any time the process requires as they demand less effort than traditional assessment methods (BATTINA, 2019). Automatic assessments can also benefit non-professional designers and developers, especially as nowadays there is a significant trend where more and more people with a background in other domains are creating mobile applications to solve problems related to their areas (PATERNÒ, 2013), using, for example, visual programming languages (ALVES; GRESSE VON WANGENHEIM; HAUCK, 2019). With the popularization of computing, the importance of this type of programming language has increased as it facilitates end-user programming as well as teaching computing in a playful and easy-to-understand way already in K-12 (MONTIEL; GOMEZ-ZERMEÑO, 2021). Specifically for the development of mobile applications, App Inventor is an intuitive block-based programming environment that allows everyone to create fully functional mobile apps for smartphones and tablets as well as the design of their user interfaces (PATTON; TISSENBAUM; HARUNANI, 2019).

However, analyzing the quality of the user interface design of apps created with App Inventor a general low visual aesthetics is observed (SOLECKI et al., 2020a). Taking into consideration the importance of visual aesthetics for the success of an app, this points out the need for visual aesthetics assessments in order to provide feedback to end-users as well as students to help them to improve the interface design. And, in order to provide access to such assessments at any time to anyone it would be important to provide automated assessments available online that offer instant feedback. However, current automated solutions for code analysis created with App Inventor are limited to evaluating computational thinking concepts, such as by CodeMaster (GRESSE VON WANGENHEIM et al., 2018a) used for the assessment of learning in k-12 education (ALVES; GRESSE VON WANGENHEIM; HAUCK, 2019). An automated method that analyzes GUI visual aesthetics in conjunction with other usability aspects agrees with the idea that applications can be beautiful and also efficient and effective (SCHLATTER; LEVINSON, 2013).

Currently, there seems to exist a wide variety of strategies to automate the assessments of GUI visual aesthetics (LIMA; GRESSE VON WANGENHEIM, 2021). They can be divided into subjective approaches, based on human responses, or objective ones, based on measuring some GUI properties (SECKLER; OPWIS; TUCH, 2015). Subjective assessment approaches collect users' perception of visual aesthetics with questionnaire-based instruments (ALTABOLI; LIN, Y., 2011b; SECKLER; OPWIS; TUCH, 2015) or sensors that detect how users react upon seeing GUIs. Some sensors for this task are eye-tracking devices collecting users' gaze points over GUIs (GU et al., 2020; PAPPAS et al., 2020) and dermal electrodes to collect neurophysiological responses related to visual aesthetics perception (BHANDARI et al., 2017). Objective approaches, however, study how GUI properties associate with the users' perception of visual aesthetics (SECKLER; OPWIS; TUCH, 2015). They analyze elements organization to measure GUI layout properties (PURCHASE et al., 2011; ZEN; VANDERDONCKT, 2014) and extract image features, such as color, texture, and layout features, from GUI screenshots (MINIUKOVICH; DE ANGELI, 2015a; WU, O. et al., 2016). This variety has led several studies to compare these existing approaches. For example, Altaboli and Lin (2011a), Mõttus et al. (MÕTTUS et al., 2013), and Seckler et al. (SECKLER; OPWIS; TUCH, 2015) investigated the relationship between objective and subjective assessments. Altaboli and Lin (ALTABOLI; LIN, Y., 2011b) compared a subset of the metrics proposed by Ngo and Byrne (2001) with different element counts based on the Classical/Expressive Aesthetic Questionnaire (LAVIE; TRACTINSKY, 2004) and the Visual Aesthetics of Website Inventory questionnaire (VisAWI) (MOSHAGEN; THIELSCH, 2010) for the assessment of web pages. Mõttus and Lamas (2015) conducted a literature review to map studies on the aesthetics of interaction, which relates to the beauty of a product when in use (DJAJADININGRAT et al., 2004). And although this study investigated aesthetics assessment and its relation with interaction design, it presents no general overview of existing assessment methods in a broader context.

More recently, deep learning techniques have been applied to quantify the visual aesthetics of GUIs such as web pages (DOU et al., 2019; KHANI et al., 2016) or GUI designs (XING et al., 2021), following the success of aesthetics assessments of photographs (DENG; LOY; TANG, 2017; LU et al., 2014; MALU; BAPI; INDURKHYA, 2017). These deep-learning approaches can extract high-level features from raw input data to predict the image's visual aesthetics (LECUN; BENGIO; HINTON, G., 2015; POLYZOTIS et al., 2017). As a result, they can categorize visual aesthetics with discrete values, such as ugly/neutral/beautiful, or gauge it as a numerical value ranging within [0..1], assuming the assessment task as either classification or regression (KIRCHNER; HEBERLE; LÖWE, 2015).

Yet, although there exists research investigating the measurement of visual aesthetics of desktop or web interfaces, little research has addressed mobile GUIs (LIMA; GRESSE VON WANGENHEIM, 2021). Miniukovich and De Angeli (2014b, 2015a) assess Android and iOS apps using feature extraction techniques to analyze GUI properties and associate them with their visual aesthetics scores. Solecki et al. present a rubric and tool for the automated assessment of the conformity of the user interface design of App Inventor apps with style guides via code analysis (SOLECKI et al., 2020b). But so far there does not exist a solution for the automated assessment of the visual aesthetics of the user interface design of App Inventor apps which has also been evaluated in terms of reliability and validity.

Thus, in this research we analyze the following **research question**: Is it possible to automate the assessment of the visual aesthetics of user interfaces of App Inventor applications with acceptable performance, reliability, and validity?

1.2 OBJECTIVES

General Objective. The general objective of this work is to develop, apply, and evaluate a model for the automated visual aesthetics assessment of user interfaces of App Inventor applications using deep learning techniques to automate that assessment.

Specific Objectives:

- O1: Synthesize the theoretical foundation on software quality, especially regarding the visual aesthetics of mobile applications and deep learning;
- O2: Analyze the state-of-the-art concerning the visual aesthetics assessment of

mobile application interfaces;

- O3: Develop a model using deep learning to assess the visual aesthetics of App Inventor user interfaces;
- O4: Evaluate the model developed regarding its performance, reliability, and validity.

1.3 ADHERENCE TO THE GRADUATE PROGRAM IN COMPUTER SCIENCE

The present work is part of the Software Engineering research line of the Graduate Program in Computer Science (PPGCC), within the topics of Software Quality, following the definition of the knowledge area of Software Engineering by SBC (ZORZO et al., 2017). Following ISO/IEC 25010 (2011), usability is one software product characteristic that categorizes quality. GUI visual aesthetics is presented as a sub-characteristic of usability, defined as the “degree to which a user interface enables pleasing and satisfying interaction for the user” (ISO, 2011). Focusing on the assessment of GUI visual aesthetics, this work is situated within the software engineering area, more specifically software quality, considering that visual aesthetics is a software quality factor (ISO, 2011). Furthermore, this work aims to evolve the state-of-the-art regarding interface design assessments, as part of Software Engineering, by applying, but not evolving, Artificial Intelligence techniques.

1.4 METHODOLOGY

The nature of this work is classified as applied research since it aims to generate knowledge for practical application directed to the solution of specific problems (SILVA; MENEZES, 2001) that, in the present case, involves the necessary knowledge for the development of a deep learning model to assess the visual aesthetics of GUIs of App Inventor app. This research quantitatively addresses the problem, as it seeks to translate information into numbers to classify and analyze them using resources and statistical techniques (SILVA; MENEZES, 2001). Regarding its objectives, the present research is exploratory since it aims to approach and become familiar with the problem to make it explicit or build hypotheses (GIL, 2008). We executed the research using a multi-method methodology:

Step 1: Background. Synthesis of relevant concepts related to visual design and visual aesthetics on mobile applications as part of GUI usability and its assessment. The relevant concepts for the work were synthesized through a bibliographic analysis (GIL, 2008).

Step 2: Identification of the state-of-the-art. Research and review of studies and related works to establish their possible contributions and define the state-of-the-art of GUI visual aesthetics assessment. In this step, we performed two systematic mapping studies following the procedure defined by Petersen et al. (PETERSEN et al., 2008): one to elicit the state-of-the-art of the assessment of GUI visual aesthetics and another one to find out how the visual aesthetics of GUIs is subjectively assessed based on human ratings. Each of these systematic mappings was conducted by a) Planning: definition of research questions, which aim to meet the research objectives, and review protocol, containing search string, selection of databases, and criteria for inclusion and exclusion of the results; b) Execution: search based on the defined protocol, selecting relevant works and excluding those that distance themselves from the focus, following the selection criteria established in the protocol; c) Analysis and interpretation: extracting relevant information from the selected works to facilitate their analysis and discussion, mapping the studies currently existing and identifying their contribution to the current state-of-the-art.

Step 3: Design and development of the automated assessment model. The automated assessment model was developed by adopting deep learning techniques executing the process proposed by (AMERSHI et al., 2019; POLYZOTIS et al., 2017): a) Requirements analysis: specifying the requirements for the DL model and which types of models are most appropriate for the problem; b) Data collection: data collection to create a dataset; c) Dataset preparation: removal of inaccurate or noisy records from the dataset; d) Data labeling: definition of rating scale through an empiric case study and the assignment of ground truth labels to each data record and evaluation of the data quality (inter-rater reliability and agreement); e) Training of the neural network: training the the models.

Step 4: Model evaluation. Performance evaluation of the model regarding using the mean squared error (MSE) (mean squared difference between predictions and labels) as the metric for performance.

The performance quality of a deep learning model indicates how well its predictions match up against the ground truth. A typical quality measure for regression tasks is the mean squared error (MSE). MSE is the average of the squares of the errors. i.e., the average squared difference between the data labels (human rating scores) and the value of the deep learning model. MSE values are always non-negative values, with the lower, the better. As a quadratic function, it heavily penalizes outliers, which are common in visual aesthetics assessments:

$$MSE = \frac{1}{n}\sum_{n=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

We also compared the model results with those from other models by analyzing their correlation (BONETT; WRIGHT, 2000) and agreement (BLAND; ALTMAN, 1986) with human responses.

1.5 CONTRIBUTIONS

One of the scientific contributions of the work presented here is the creation of an automated model for the visual aesthetics assessment of mobile app GUIs using deep learning. An additional outcome is the elaboration of a dataset of screenshots of App Inventor app interfaces labeled by the degree of visual aesthetics by humans. Thus, our work creates an unprecedented contribution in the area of Software Quality as part of Software Engineering in which there is a growing interest in solutions with Artificial Intelligence/Machine Learning.

The following scientific contributions have been achieved:

- Systematic mapping on the state-of-the-art to provide an overview on how the assessments of the visual aesthetics of GUIs have been made (LIMA; GRESSE VON WANGENHEIM, 2021);
- Systematic mapping on the state-of-the-art to understand how the visual aesthetics of GUIs is subjectively assessed based on human ratings (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022a);
- A study comparing the reliability and validity of scales for the assessment of visual aesthetics of mobile guis through human judgment (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022b);
- A deep learning model to automate the visual aesthetics assessment of the GUIs of App Inventor apps on a single score (LIMA et al., n.d.);
- A deep learning model to predict the rating distribution of the visual aesthetics of the GUIs of App Inventor apps (LIMA; GRESSE VON WANGENHEIM, n.d.);
- A dataset with 820 screenshots of apps designed with App Inventor labeled according to their visual aesthetics[1];
- A mobile application for the labeling of screenshots[2].

As a technological contribution, a mobile application for the labeling of screenshots has been developed with the use of App Inventor. The application can be easily adapted to label any image with a 5-point scale.

The social contribution of the present work is the availability of a deep learning model to automatically assess the visual aesthetics of App Inventor app GUIs. This model is expected to reduce the cost, effort and time of such assessments, allowing their execution at any time during the development process. In this way, this assessment can be available to developers in contexts with few resources available for UI design or end-users, contributing to improving software quality and the development process. Another possible use is in the

[1] `https://bit.ly/app-inventor-dataset-v2`

[2] `https://bit.ly/gui-labeler`

educational context providing instantaneous feedback to students as part of computing education. As a result, automating visual aesthetics assessment is expected to support the teaching of visual design, which could decrease the evaluation effort and solve other problems, such as favoritism.

2 BACKGROUND

2.1 USABILITY AND AESTHETICS

An important quality aspect of a GUI is usability (ISO, 2011), which is "the extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (ISO, 2018). Following this characterization, not only should a software system allow its users to complete their tasks within the lowest time and effort possible, but it should also provide satisfaction, making GUIs easy to use, efficient, and enjoyable (PREECE; SHARP; ROGERS, 2015). The visual design of a GUI has a significant influence on its usability (SCHLATTER; LEVINSON, 2013). It can impact effectiveness and efficiency by guiding users to avoid errors and complete their tasks (SCHLATTER; LEVINSON, 2013), and improve user satisfaction when dealing with beautiful GUIs (LINDGAARD; DUDEK, 2002). Moreover, the overall user impression appears to increase user satisfaction to a greater degree than usability experience and content (HARTMANN; SUTCLIFFE; DE ANGELI, 2007). However, despite being the parameter intended to assess the qualitative aspects of user experience, such as pleasantness, comfort, and overall appreciation, satisfaction is frequently overlooked, while the other parameters, effectiveness and efficiency, are well-known, evaluated, and researched (BOLLINI, 2017).

GUIs need to draw users' attention before conveying a message that will be easily accessed or understood (ZEN; VANDERDONCKT, 2014). Visual aesthetics has a significant role in attracting users to visit a website or download a mobile app because it is the first thing that they experience after looking at GUIs (KHANI et al., 2016), and it is what primarily determines user preferences (ULRICH, 2006). Among software products with similar features, users tend to favor those they consider to be more visually attractive (KHANI et al., 2016; SCHENKMAN; JÖNSSON, 2000; TRACTINSKY, 2013). For that reason, aesthetics is an additional factor when comparing computer products or services (WU, O. et al., 2016).

The understanding of beauty has varied throughout history (LAVIE; TRACTINSKY, 2004). For some it was a property of objects that caused pleasure in the perceiver, but for others it was a function of the subject's personal qualities (MOSHAGEN; THIELSCH, 2010). For example, ancient and medieval theorists advocated the objectivist view, claiming that beauty is "located outside of anyone's particular experiences," and many philosophical accounts took beauty that way until the eighteenth century (SARTWELL, 2017). On the other hand, the subjectivist view is as old as the sophists, leading Hume to say that beauty "exists merely in the mind which contemplates" the objects, and Kant to state that the aesthetic judgment "can be no other than subjective" (SARTWELL, 2017). Nonetheless, conceiving beauty as purely subjective seems implausible, for two observers can reveal to each other their aesthetic judgments on the same object and agree or disagree about it.

Moreover, an entirely objectivist idea of beauty would mean that things could be beautiful or ugly even if they were never perceived (SARTWELL, 2017).

More recently, both views have been combined so that beauty has been conceived as arising from the relation between the object's properties and the perceiver's qualities (MOSHAGEN; THIELSCH, 2010). Santayana (2012) declares that "beauty is pleasure regarded as the quality of a thing," claiming that an object can only be beautiful if it entails some emotion in the observer. Even Hume and Kant, although assuming that the aesthetic experience is fundamentally subjective, also recognize that the aesthetic judgment claims inter-subjective validity, that everyone similarly situated should reach the same understanding (SARTWELL, 2017).

The view of beauty as deriving from the observed object and, at the same time, being the result of the observer's emotions is consonant with a widely accepted definition in the context of usability, in which visual aesthetics is "the degree to which it enables pleasing and satisfying interaction for the user" (ISO, 2011). Following this definition, more than only adhering to principles that contribute to visual aesthetics, GUI designs should also be meaningful for their users (SOUI et al., 2020). Yet, GUI visual aesthetics assessments focus on either GUIs and their properties or how users perceive them, but not both (LIMA; GRESSE VON WANGENHEIM, 2021; SECKLER; TUCH, 2012), even recognizing both the subjective and the objective aspects of visual aesthetics. That way, assessments that combine objective factors with different facets of subjective perception may help better understand how design elements influence perceived aesthetics (SECKLER; OPWIS; TUCH, 2015).

2.2 USABILITY FOR MOBILE APPLICATIONS

Mobile GUIs allow users to interact with their devices in novel ways that traditional desktop interactions do not, presenting new usability challenges (KUMAR; MOHITE, 2018). These devices have introduced a different paradigm such as widgets, touch, physical motion, and on-screen keyboards, including both the user input and the associated sensor information (FLORA; WANG, X.; CHANDE, 2014). The size and portability requirements of mobile devices limit the number of elements that can be presented at a time to avoid cluttering or extend on several screens, thus requiring the use of scrolling (RAHMAT et al., 2018). And although touch screen capabilities may facilitate some actions, they also pose new challenges through the lack of tactile feedback, button size, and varied movement possibilities, such as swiping, pinching, spreading, and flicking. Also, since fingertips are usually bigger than mouse arrows, touchable elements on screen also need to be big enough to avoid misselection (BALAGTAS-FERNANDEZ; FORRAI; HUSSMANN, 2009). These factors directly influence the interaction and interface design on these devices, demanding even greater attention in the development of the user interface design and its assessments (FLORA; WANG, X.; CHANDE, 2014). And, mobile device users might value usability and visual aesthetics even more as the on-the-go usage implies multiple external distractions and short and intensive interaction periods (CHOI, J. H.; LEE, H.-J., 2012).

Considering the different ways of interaction that mobile devices offer, the direct application of traditional usability guidelines and assessment methods might not work for mobile apps (KUMAR; MOHITE, 2018). Nonetheless, the usability definition presented by the International Organization for Standardization (ISO, 2018) has reached such a widespread adoption that today it is applied almost unchanged in a variety of contexts despite some resistance to a generally accepted definition of usability (WEICHBROTH, 2020). Usability still poses a challenge for developers and a point of concern for users even though mobile devices are constantly undergoing technological progress making the shift from the desktop environment smoother (SALMAN; WAN AHMAD; SULAIMAN, 2018; WEICHBROTH, 2020). But besides converging to the same usability requirements, the introduction of more advanced and sophisticated devices every year demands ever growing research on usability for mobile devices (KUMAR; MOHITE, 2018). Thus, the unique features of mobile devices and wireless networks which influence the usability of mobile applications, including mobile context, multimodality, connectivity, small screen size, different display resolutions, limited processing capability and power, and restrictive data entry methods (WEICHBROTH, 2020).

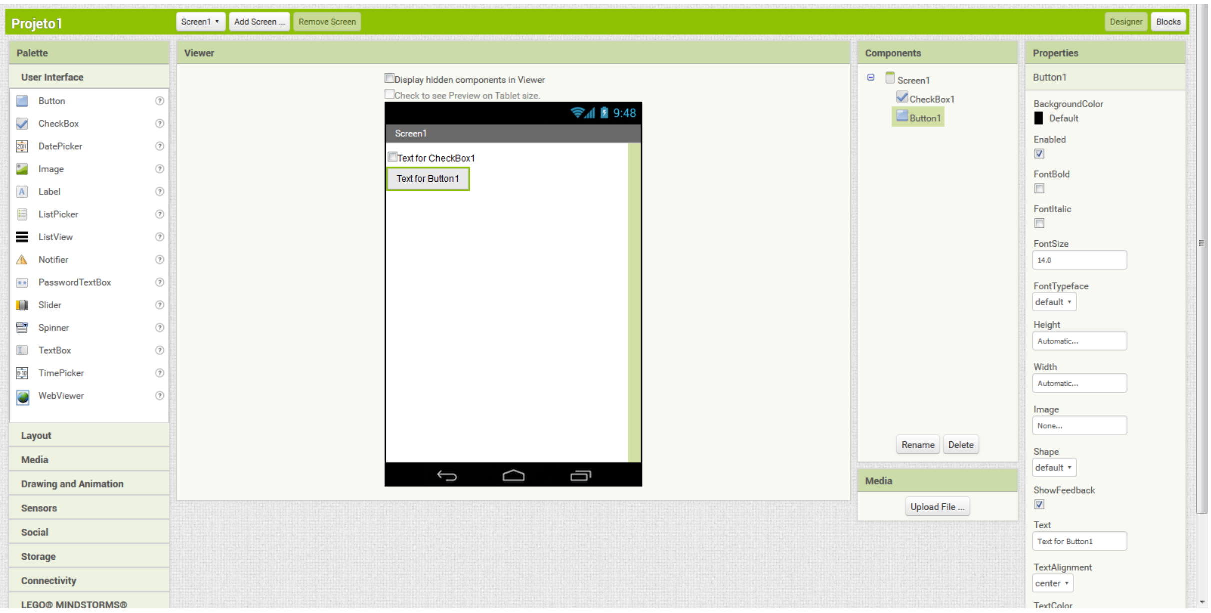
2.2.1 Applications developed with app inventor

Observing current trends of end-user development of mobile applications (BARRICELLI et al., 2019) as well as computing education in K-12 through the development of mobile apps (MEDEIROS; GRESSE VON WANGENHEIM; HAUCK, 2021), visual tools are becoming popular (GRESSE VON WANGENHEIM et al., 2021). In this context, App Inventor is a block-based programming environment that enables anyone to develop applications for Android devices. App Inventor has been around for over 12 years and has more than 18 million users worldwide (MIT APP INVENTOR, 2022).

Besides providing support for programming the functionality of mobile apps, App Inventor also provides an editor (Figure 1) to design user interfaces (PATTON; TISSENBAUM; HARUNANI, 2019). User interfaces are designed by drag and drop visual components such as buttons, text boxes, and images. Depending on the type of component, it is possible to configure properties such as height, width, fonts, color, among others.

App Inventor also provides an app gallery (Figure 2) where users can publish their complete app codes for sharing with the community or to be part of the App of the Month program (PATTON; TISSENBAUM; HARUNANI, 2019). They can also download other apps from the gallery for their own use or to adapt according to their interests (PATTON; TISSENBAUM; HARUNANI, 2019).

Figure 1 – App Inventor GUI editor.



Properties
Button1
BackgroundColor ■ Default
Enabled ☑
FontBold □
FontItalic □
FontSize 14.0
FontTypeface default ▾
Height Automatic...
Width Automatic...
Image None...
Shape default ▾
ShowFeedback ☑
Text Text for Button1
TextAlignment center ▾
TextColor

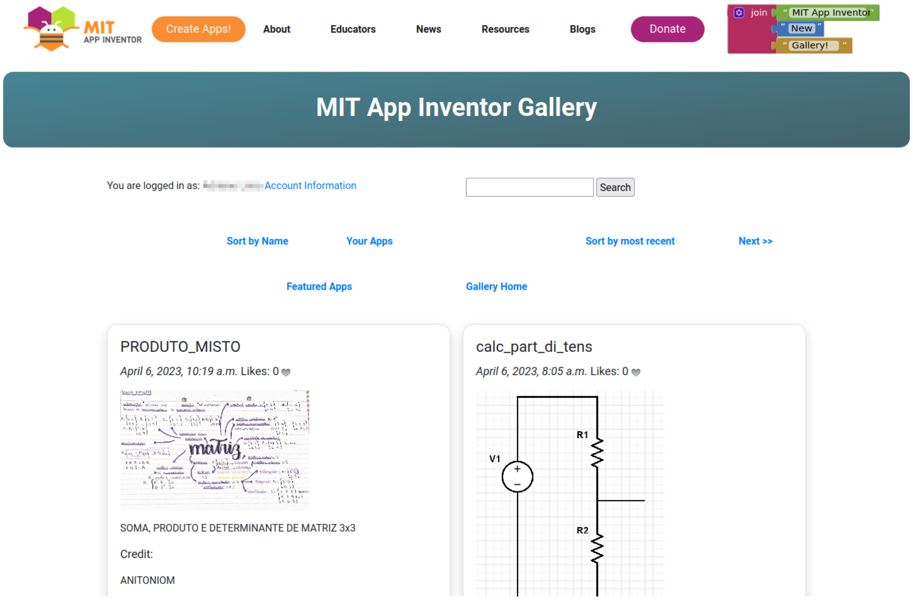
Source: MIT Media Lab, APP Inventor - ai2.appinventor.mit.edu, CC BY 3.0.

2.3 ASSESSING THE VISUAL AESTHETICS OF USER INTERFACES

Assessments guide the interface design and evaluate whether GUIs meet quality requirements (SHARMA, T.; MISHRA; TIWARI, 2016), like usability and visual aesthetics. When automated, GUI assessments such as PLAIN (SOUI et al., 2017) or GUIEvaluator and GUIExaminer (ALEMERIEN; MAGEL, 2015, 2014) can quickly detect and visualize problematic design aspects (SOUI et al., 2021). Behavioral methods, such as rating the visual aesthetics degree of GUIs, ranking them according to their degree of aesthetics, or comparing GUI pairs to indicate an aesthetic preference, are used to capture the subjective aspect of visual aesthetics (PALMER; SCHLOSS; SAMMARTINO, 2013). When related to behavioral response, a physiological reaction may also express the degree of aesthetics perceived, e.g., through electrodes on the skin (BHANDARI et al., 2017) or eye movements (GU et al., 2020; PAPPAS et al., 2020). When investigating object properties, design principles and elements have been found to influence GUI visual aesthetics (SECKLER; OPWIS; TUCH, 2015). Layout properties, such as symmetry, help establish regular structures and meaningful forms (BAUERLY; LIU, Yili, 2008). Color invokes many different emotions and influences general aesthetics (REINECKE et al., 2013; SECKLER; OPWIS; TUCH, 2015). Typography, imagery, controls, and affordance form a significant part of most GUIs (PURCHASE et al., 2011), and white space makes them distinguishable and helps compose the layout (SCHLATTER; LEVINSON, 2013).

There exists a wide variety of approaches for the visual aesthetics assessment of GUIs (Figure 4). Although most of them recognize that design principles and elements

Figure 2 – App Inventor gallery.



You are logged in as: [redacted] Account Information

PRODUTO_MISTO

April 6, 2023, 10:19 a.m. Likes: 0

SOMA, PRODUTO E DETERMINANTE DE MATRIZ 3x3

Credit:

ANITONIOM

calc_part_di_tens

April 6, 2023, 8:05 a.m. Likes: 0

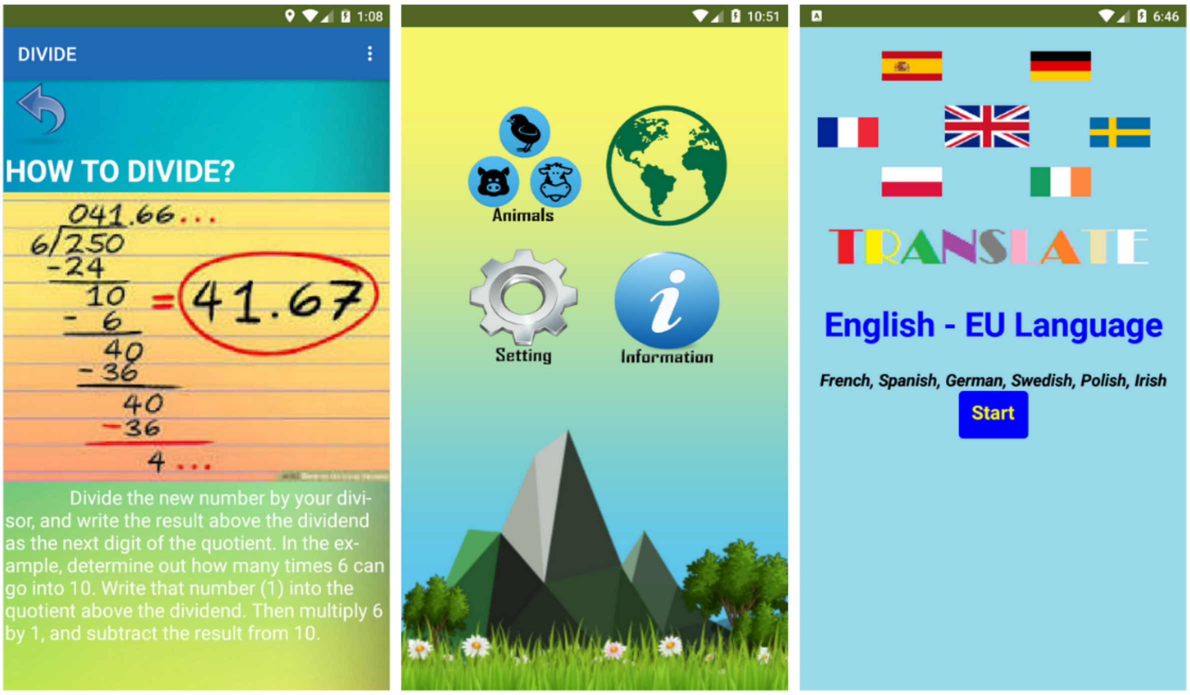
Source: MIT Media Lab, APP Inventor Gallery.

influence the GUI visual aesthetics and that the user's aesthetic reaction is based on idiosyncratic values, they usually focus on the GUI properties or the user's reactions (LAVIE; TRACTINSKY, 2004; SECKLER; OPWIS; TUCH, 2015). For that reason, visual aesthetics assessment approaches are classified as either subjective or objective (ALTABOLI; LIN, Y., 2011b; MÕTTUS et al., 2013; SECKLER; OPWIS; TUCH, 2015).

Subjective approaches consider visual aesthetics as a matter of how users perceive GUIs (ALTABOLI; LIN, Y., 2011b; MÕTTUS et al., 2013; SECKLER; OPWIS; TUCH, 2015). These approaches see a connection between beauty and emotion (LAVIE; TRACTINSKY, 2004) and assume that assessments should represent how users perceive visual aesthetics instead of looking for beauty in GUIs or some of their properties (SECKLER; OPWIS; TUCH, 2015). That way, potential users should take part in the assessment of every GUI, making it difficult to automatize. Subjective approaches are typically operationalized by using questionnaire-based instruments to measure users' perception of visual aesthetics (ALTABOLI; LIN, Y., 2011b; SECKLER; OPWIS; TUCH, 2015). An alternative is to use sensors to measure their response upon seeing GUIs, like eye-tracking devices collecting GUI points of interest (GU et al., 2020) or neurophysiological sensors collecting emotions, to relate them with the perception of visual aesthetics (BHANDARI et al., 2017).

Objective assessments recognize that some image properties, like order, proportion, and symmetry, trigger emotions in the viewer while looking at them (LAVIE; TRACTINSKY, 2004). Therefore, by measuring one or more GUI visual properties, these approaches

Figure 3 – Examples of apps available at the App Inventor gallery.



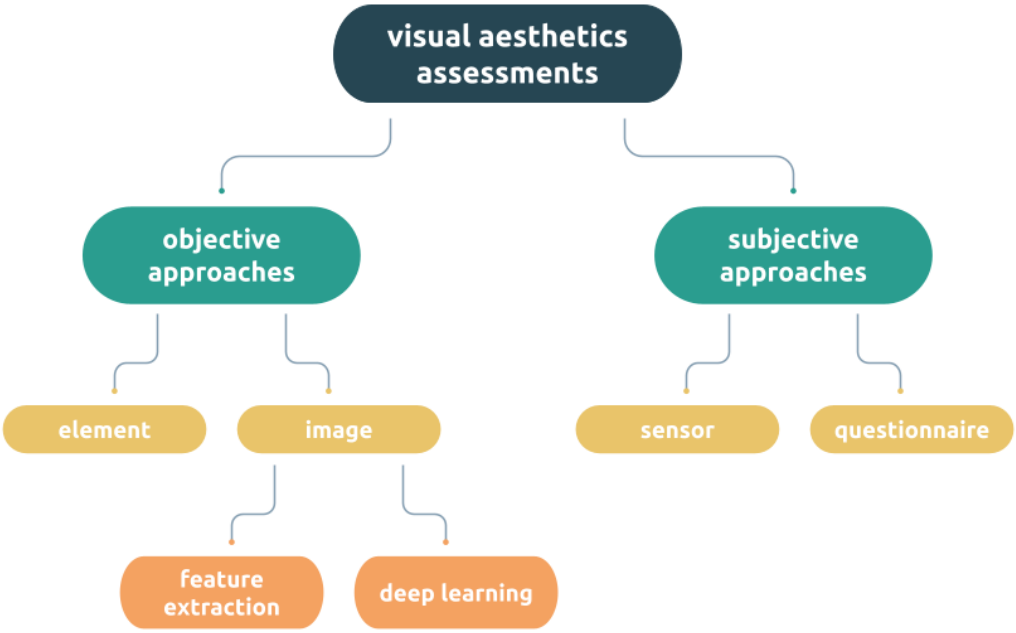
Divide the new number by your divisor, and write the result above the dividend as the next digit of the quotient. In the example, determine out how many times 6 can go into 10. Write that number (1) into the quotient above the dividend. Then multiply 6 by 1, and subtract the result from 10.

Source: MIT Media Lab, APP Inventor Gallery.

aim at indirectly estimating the perceived visual aesthetics. The properties are analyzed using objective measures (MÕTTUS et al., 2013), focusing on GUI elements and how they are organized or based on the whole GUI as a single image (MINIUKOVICH; DE ANGELI, 2015a). Element-based techniques apply metrics that either count or measure individual GUI element properties (e.g., size or aspect ratio) and layout properties (e.g., symmetry or density) (ALTABOLI; LIN, Y., 2011b). On the other hand, image-based techniques examine GUI screenshots to detect and extract features that might be associated with visual aesthetics (MINIUKOVICH; DE ANGELI, 2015a). These features can be manually engineered (called handcrafted features) before being extracted and measured using computer vision algorithms or automatically acquired by adopting deep learning approaches using GUI screenshots as input (NANNI; GHIDONI; BRAHNAM, 2017).

Following the trend of using deep learning techniques to assess the visual aesthetics of photographs (DENG; LOY; TANG, 2017; GAO et al., 2020; KARAYEV et al., 2014; LU et al., 2014; MALU; BAPI; INDURKHYA, 2017; SUCHECKI; TRZCISKI, 2017) they can also be applied to estimate the aesthetics of GUIs. In that context, especially convolutional neural networks (CNNs) are well-suited for analyzing images (ANDREARCZYK; WHELAN, 2017). CNNs use a supervised learning model that needs to be trained with the input data and the expected output, differently from the unsupervised learning model, that can cluster the input based on their statistical properties without being provided

Figure 4 – Common approaches to assess the visual aesthetics of GUIs.



Source: the author.

with the correct answer during training (**LI et al.**; **2016**). When used for assessing visual aesthetics, CNNs are trained with a large dataset of previously labeled GUI images. The label represents the expected output (visual aesthetics assessment) that can either be a category expressing an ordinal scale (ugly, neutral, beautiful) or some numerical value expressing a degree of aesthetics. Models that determine which category an input belongs to, based on its properties, are called classification models, whereas those models that predict numerical values within an interval are called regression models (KIRCHNER; HEBERLE; LÖWE, 2015). CNNs can learn those features that correlate with the UI visual aesthetics and assess new ones.

2.3.1 Measuring visual aesthetics based on human perception

Like other subjective constructs, visual aesthetics can be considered a unidimensional or multidimensional construct (BHATTACHERJEE, 2012). When regarded as a unidimensional construct, it is represented by a single underlying dimension, like "beauty" (ALTABOLI; LIN, Y., 2011a), "attractiveness" (SCHAIK; LING, 2009), or "appeal" (BAUGHAN et al., 2020), which is measured with a single scale or test. On the other hand, multidimensional constructs are composed of multiple dimensions analyzed separately (BHATTACHERJEE, 2012). For example, Lavie and Tractinsky (2004) decompose the construct of visual aesthetics into two dimensions: "classical aesthetics," which is related to orderly and clean design, and "expressive aesthetics," which is manifested by creativity, originality, and the ability to break design conventions. Others often decompose visual aesthetics into several facets, such as simplicity, colorfulness, diversity, and craftsmanship (MOSHAGEN; THIELSCH, 2010). Here, “simplicity” refers to the aspects of a layout that contribute to perception, “diversity” refers to creativity and novelty, “colorfulness” relates to color perception, and “craftsmanship” refers to the skill and care employed in the design of GUIs.

Multidimensional constructs are commonly assessed via multiple-item questionnaires, in which one or more items are used to rate each dimension. Lavie and Tractinsky (2004) presented a questionnaire to rate classical and expressive aesthetics using ten items (five items for each dimension) on a 7-point Likert scale. Users indicate their degree of agreement (from “1” - “strongly disagree” to “7” - “strongly agree”) to affirmations representing classical (e.g., “I feel the design of this website is clean.”) or expressive aesthetics (e.g., “I feel the design of this web site is creative.”) to assess GUIs. Another instrument is the Visual Aesthetics of Website Inventory (VisAWI), a questionnaire composed of 18 items to rate the simplicity/diversity/colorfulness/craftsmanship dimension of visual aesthetics (MOSHAGEN; THIELSCH, 2010). They also use 7-point Likert scales. Although these well-established questionnaires assessing visual aesthetics subdimensions are suitable to provide a detailed understanding of what contributes to the visual aesthetics of GUIs, they are lengthy and demand considerable effort, which becomes impractical when assessing a large number of GUIs. In this regard, the short version of VisAWI (VisAWI-S) aims to minimize effort by being composed of only four items, one for each dimension (Table 1) (MOSHAGEN; THIELSCH, 2013). However, although shorter than its original version, it is still four times longer than any unidimensional measure using a single scale.

Table 1 – VisAWI-S Items.

Dimension	**Item**
simplicity	“Everything goes together on this site”
diversity	“The layout is pleasantly varied”
colorfulness	“The color composition is attractive”
craftsmanship	“The layout appears professionally designed”

Source: adapted from (MOSHAGEN; THIELSCH, 2013).

There are many possible ways to measure the subjective perception of visual aesthetics as a unidimensional construct (PALMER; SCHLOSS; SAMMARTINO, 2013). Rank-ordering is a simple measuring technique in which subjects order GUIs from the most to the least beautiful. In a complete ranking design, all GUIs are presented at once before subjects can order them (PURCHASE et al., 2011). Despite its simplicity, that might be a disadvantage when assessing many GUIs. An alternative is the balanced incomplete block design (BIBD), in which subjects rank subsets (blocks), always with the same number of GUIs. With smaller sets, participants make fewer comparisons each time, reducing perceptual and memory load and resulting in more reliable outcomes (BAUERLY; LIU, Yili, 2006). A typical example of the BIBD is the 2-design, using blocks of two elements to allow pairwise comparisons.

Another technique for assessing the visual aesthetics of GUIs is the use of rating scales that represent their mapping to measurement values (WOHLIN et al., 2012). Different rating scales are typically used to assess visual aesthetics (Table 2).

Table 2 – Rating scales.

Scale	Type	Description
Binary	nominal	Respondents choose between only two mutually exclusive values (BHATTACHERJEE, 2012)
Likert	ordinal	Statements to which respondents express their degree of agreement (MURPHY; LIKERT, 1938)
Magnitude estimation	interval	Respondents assign the stimuli with numerical values proportional to the reference stimulus (STEVENS; MARKS, 2017)
Mean opinion score	ordinal	Respondents estimate the quality of stimuli on a 5-point ordinal scale (bad, poor, fair, good, excellent) (HUYNH-THU et al., 2011)
Semantic differential	ordinal	Respondents express their opinions or feelings toward a statement using pairs of opposing adjectives (BHATTACHERJEE, 2012)
Visual analog	interval	A 10-cm line with the marked position converted into a 101-point scale (0–100) (COUPER et al., 2006)

Source: the author.

When considering visual aesthetics a unidimensional construct, its subjective assessment can be done with a single rating scale while, on the other hand, multidimensional constructs demand measurement with multiple scales to rate each of their dimensions (BHATTACHERJEE, 2012).

Idiosyncratic properties strongly influence aesthetic preferences and People with different backgrounds tend to disagree when assessing the same stimulus (PALMER; SCHLOSS; SAMMARTINO, 2013). A lack of equivalence between assessments lowers the inter-rater reliability and agreement, which reduces their overall quality. Some demographic differences that might influence how subjects assess GUI visual aesthetics are:

- **Age**: Reinecke and Gajos (2013) found that subjects aged 31 to 40 prefer GUIs with few colors while older participants consider colorless GUIs visually less appealing than any other age group.
- **Gender**: Aesthetic preferences change for females and males. For instance, Moss and Gunn (2009) indicate that users have a statistically significant tendency to prefer GUIs produced by people of the same gender. Also, females rate colorful websites higher on visual aesthetics than males (REINECKE; GAJOS, 2014).
- **Education**: Subjects with pre-high school education tend to rate the visual aesthetics of colorful and visually complex websites higher (REINECKE; GAJOS,

2014), while preference for such GUIs decreases with higher educational stages.

Yet, as it is not feasible to study entire populations in most empirical research, samples must be representative and large enough to minimize error when generalizing the results (WOHLIN et al., 2012). Sampling techniques can either be categorized into probability or non-probability sampling (BHATTACHERJEE, 2012). When using probability sampling, the chance of selecting each subject is known and higher than zero (WOHLIN et al., 2012). Examples are simple random sampling and systematic sampling. Non-probability sampling is adopted when it is impossible to estimate the probability of selection or when some subjects have zero chance of being selected. Some of these techniques are convenience, expert, volunteering, and quota sampling. Convenience sampling is the selection of those subjects that are readily available as the students in a class or the researcher's friends. Expert sampling is adopted when some attributes of the individuals are relevant to the study, such as some specific knowledge that influences how they rate. It is also possible to have subjects volunteer themselves by responding to the research advertising or enrolling in a crowdsourcing service. And in quota sampling, subjects are selected until each group of interest reaches a predefined number of individuals.

2.3.2 Assessment quality

Reliability and validity are indispensable indicators of the quality of assessments (KIMBERLIN; WINTERSTEIN, 2008). They express the rigor of research processes and the credibility of research findings (ROBERTS; PRIEST; TRAYNOR, 2006). Reliability indicates how consistent an assessment is, whereas validity expresses how well it represents the construct one intends to measure (THORNDIKE; THORNDIKE-CHRIST, 2014). In this way, visual aesthetics assessments are reliable if the most beautiful GUIs consistently receive higher scores than the ugliest ones and are valid if they precisely measure visual aesthetics and not something else (e.g., usability).

As there is not one single direct measure of reliability, it is analyzed through its attributes of homogeneity, equivalence, and stability (HEALE; TWYCROSS, 2015). Homogeneity, also called internal consistency, indicates the consistency between items of a data collection instrument, equivalence is consistency between different raters, and stability is consistency over time (KIMBERLIN; WINTERSTEIN, 2008). Different reliability types analyze each of these attributes, as shown in Table 3.

Internal consistency indicates how homogeneous a data collection instrument is (HEALE; TWYCROSS, 2015). An instrument to assess GUI visual aesthetics should consist of items that examine the construct in its wholeness but do not include items that measure other constructs (STREINER, 2003). Inter-rater reliability and inter-rater agreement express how equivalent the responses from different observers are (KOTTNER et al., 2011). Although they are often used interchangeably in the literature, they are different in what they represent (VET et al., 2006; GISEV; BELL; CHEN, 2013). The

Table 3 – Overview on types of reliability.

Consistency	Type	Description	Examples of analysis
Homogeneity	Internal consistency	The consistency between different items of the same data collection instrument (BHATTACHERJEE, 2012).	• Cronbach's alpha; • Composite reliability
Equivalence	Inter-rater reliability and agreement	The extent to which raters can consistently distinguish between different items (reliability) or different raters assign the same value for each item (agreement) (GISEV; BELL; CHEN, 2013).	• Correlation analysis; • Kendall's coefficient of concordance; • Intraclass correlation coefficient (ICC)
Stability	Intra-rater reliability	The consistency between two ratings of the same construct assigned by the same rater at two different points in time.	• Correlation analysis; • Intraclass correlation coefficient (ICC)

Source: the author.

inter-rater agreement shows how identical scores are, while inter-rater reliability analyzes the general trend in ratings, not the absolute score each rater assigns, analyzing if raters rank GUIs the same way (GISEV; BELL; CHEN, 2013). The intra-rater reliability analyzes stability over time by the same subject under similar circumstances at different moments (KOTTNER et al., 2011).

Validity refers to the extent to which an instrument correctly represents the construct it purports to measure, including content, construct, and criterion-related validity (PUNCH, 1998), as shown in Table 4.

Table 4 – Overview on types of validity.

Type	Description	Examples of analysis
Content	The extent to which an assessment represents the full content of the assessed construct (PUNCH, 1998).	Judgment by an expert panel.
Construct	The extent to which an assessment actually represents the construct to be measured(PUNCH, 1998).	• Internal consistency analysis; • Factor analysis; • Convergent/discriminant validity.
Criterion-related	The extent to which the scores of an instrument correlate with other measures of the same construct accepted as ground truth (KIMBERLIN; WINTERSTEIN, 2008).	Correlation analysis with present (concurrent) or future (predictive) criteria.

Source: the author.

The content validity of instruments is determined by whether they sufficiently comprise everything they should assess (HEALE; TWYCROSS, 2015). It is the lowest level and regards how relevant and representative the instrument items are (ROBERTS; PRIEST; TRAYNOR, 2006). Content validity is usually analyzed by an expert panel

(KIMBERLIN; WINTERSTEIN, 2008). Construct validity includes demonstrating the relationship between what the instrument measures and the construct under research (ROBERTS; PRIEST; TRAYNOR, 2006). The confirmatory factor analysis is a way of confirming how well the measured variables represent that construct (ROBERTS; PRIEST; TRAYNOR, 2006). It should also be highly correlated to other instruments assessing similar constructs (convergent validity) and poorly correlated to those that measure something else (discriminant validity) (HEALE; TWYCROSS, 2015). Criterion-related validity involves analyzing how much the outcome of an assessment instrument correlates with those of other already validated instruments, known as criteria, to assess the same construct (ROBERTS; PRIEST; TRAYNOR, 2006). Concurrent validity is given when the criterion exists in the present or predictive if it will exist in the future (PUNCH, 1998).

3 STATE-OF-THE-ART

To provide a comprehensive overview on the topic of this thesis, we performed two systematic mapping studies. We analyzed the state-of-the-art with regard to the assessment of the visual aesthetics of GUIs (section 3.1), and as well as the state-of-the-art on how the visual aesthetics of GUIs is subjectively assessed based on human ratings (section 3.2).

3.1 ASSESSMENT OF THE VISUAL AESTHETICS OF GUIS

3.1.1 Definition of the review protocol

The **research question** is “What studies exist on the assessment of the visual aesthetics of GUIs?” focusing on the following analysis questions:

- AQ1. What are the characteristics of the assessed objects?
- AQ2. What techniques are used for assessment?
- AQ3. How have the approaches been evaluated?

Data source. We conducted the search on Scopus, the largest abstract and citation database of peer-reviewed literature, including publications from ACM, Elsevier, IEEE, and Springer, considering articles with free access through the Capes Portal.

Inclusion/exclusion criteria. We included only articles that presented visual aesthetics assessments of graphical user interfaces, including web and mobile applications. We searched for peer-reviewed articles in English published in the ten-year period before this research, from January 2010 to August 2020. Although attempts to assess GUI aesthetics have been made since the beginning of the 2000s, e.g., Ngo et al. (2000), we chose to examine articles published in the last decade considering the significant differences in the visual design of GUIs to earlier software systems (PUNCHOOJIT; HONGWARITTORRN, 2017). The search was updated in March 2023, when new articles were included. We excluded articles that do not present assessment approaches themselves but rather study or compare existing ones. Moreover, we considered only articles that presented significant information on the assessment to enable relevant information extraction regarding the analysis questions and, therefore, excluded abstract-only or one-page articles.

Definition of the search string. Pursuing the research objective, we defined the search string to identify articles dealing with core concepts and their synonyms, as stated in Table 5, also based on several informal searches for the calibration of the search string. The term “visual aesthetics” represents the object of assessment. Although some articles use synonyms when dealing with the visual aesthetics of GUIs, like “visual appeal” (LINDGAARD, 2007), “visual quality” (WU, O. et al., 2011), or “beauty” (MOSHAGEN; THIELSCH, 2010), adding these expressions to the string did not minimize the risk of omission and, therefore, were left out. We also included the key term “assessment”, including common synonyms in the GUI analysis context such as "evaluation" and "measurement." Focusing on GUIs, we chose to use the core concept "interface" as a broader term. As some articles refer rather to the entire software systems, like "application" or "website," although targeting their GUIs, we also included these as synonyms, as well as "mobile," "android," and "ios."

Table 5 – Keywords.

Core concepts	**Keywords and synonyms**
visual aesthetics	aesthetics
assessment	assess*, evaluat*, measure*
interface	ui, gui, web*, mobile, app*, android, ios

Source: the author.

We defined the search string in conformance with the specific syntax of the data source using wildcard characters to cover as many variations of the terms as possible:

```
TITLE-ABS-KEY(aesthetics AND (assess* OR evaluat* OR measure*) AND
(interface OR ui OR gui OR web* OR mobile OR app OR application OR
android OR ios)) AND PUBYEAR>2009 AND (LIMIT-TO (LANGUAGE, "English")
```

3.1.2 Execution of the search

The first author executed the search in August 2020, which was reviewed by the second author. The initial search returned 1,114 articles. We applied Scopus filters to exclude works on unrelated fields such as Medicine (166), Dentistry (69), Business, Management and Accounting (61), Agricultural, Biological Sciences (47), Health Professions (24), Biochemistry, Genetics, and Molecular Biology (21), Earth and Planetary Sciences (17), Nursing (13), Economics, Econometrics, and Finance (10), Pharmacology, Toxicology, and Pharmaceutics (8), and Immunology and Microbiology (1). We quickly reviewed titles, abstracts, and keywords of all filtered search results (759 articles) to identify those articles matching the exclusion criteria during the first analysis. After the removal of irrelevant and duplicates, we identified 49 potentially relevant articles. In the next step, we analyzed the full texts and excluded irrelevant ones, following the inclusion/exclusion criteria. As a result, we selected 27 articles that present approaches to assessing GUI visual aesthetics (Table 6). In March 2023, the research was updated using the same criteria to search from articles that were published from 2020, when 11 new articles were included.

We excluded many articles that focus on visual aesthetics assessments of other objects, such as photographs. On the other hand, we included one that assessed mobile game apps because the presented method can be extended for GUI elements (JYLHÄ; HAMARI, 2020). Other articles were excluded as they do not present substantial information on the assessment approach to enable the extraction of relevant information (BOURGUET, 2018;

Table 6 – Number of articles per selection stage.

	Initial search result	**Filtered search result**	**Potentially relevant**	**Relevant articles**
original	1,114	759	49	27
update	2,134	883	25	12

Source: the author.

WEN et al., 2018; KO; LIU, Yuchun, 2019) or presented a comparative analysis of different approaches instead of introducing a specific assessment approach (ALTABOLI; LIN, Y., 2011a; MÕTTUS et al., 2013). We also excluded one article that was not available for the full-text analysis (KONG; GUO, 2019). On the other hand, we included the approach presented by Maity et al. (2016) that focuses on the visual aesthetics assessment of only one specific element (text), different from the others that aim to assess the visual aesthetics of the GUIs as a whole. Evolving their research, the authors Maity and Bhattacharya (2019) complete the text assessment later on, together with images and white space, as part of a model to assess the whole GUI visual aesthetics. As a result, we identified 38 relevant articles that present different approaches to visual aesthetics assessment (Table 7).

Table 7 – Relevant articles.

(To be continued)

Title	**Reference**
An Arabic Version of the Visual Aesthetics of Websites Inventory (AR-VisAWI): Translation and Psychometric Properties	(ABBAS; HIRSCHFELD; THIELSCH, 2022)
Modeling and Evaluating User Interface Aesthetics Employing ISO 25010 Quality Standard	(ABBASI et al., 2012)
Investigating Effects of Screen Layout Elements on Interface and Screen Design Aesthetics	(ALTABOLI; LIN, Y., 2011a)
I Don't Have That Much Data! Reusing User Behavior Models for Websites from Different Domains	(BAKAEV et al., 2020)
Benchmarking Neural Networks-Based Approaches for Predicting Visual Perception of User Interfaces	(BAKAEV et al., 2022)
Webthetics: Quantifying Webpage Aesthetics with Deep Learning	(DOU et al., 2019)
Predicting Webpage Aesthetics with Heatmap Entropy	(GU et al., 2020)
Development of Measurement Instrument for Visual Qualities of Graphical User Interface Elements (VISQUAL): A Test in the Context of Mobile Game Icons	(JYLHÄ; HAMARI, 2020)
A Novel Approach for Website Aesthetic Evaluation Based on Convolutional Neural Networks	(KHANI et al., 2016)

(Continuation of Table 7)

Title	Reference
Aesthetic assessment of website design based on multimodal fusion	(LIU, X.; JIANG, 2021)
A Computational Model to Predict Aesthetic Quality of Text Elements of GUI	(MAITY; MADROSIYA; BHATTACHARYA, 2016)
Is My Interface Beautiful? - A Computational Model-based Approach	(MAITY; BHATTACHARYA, 2019)
A Quantitative Approach to Measure Webpage Aesthetics	(MAITY; BHATTACHARYA, 2020)
Computing Aesthetics of Concrete User Interfaces	(MBENZA; BURNY, 2020)
Quantification of Interface Visual Complexity	(MINIUKOVICH; DE ANGELI, 2014a)
Visual Impressions of Mobile App Interfaces	(MINIUKOVICH; DE ANGELI, 2014b)
Computation of Interface Aesthetics	(MINIUKOVICH; DE ANGELI, 2015a)
Visual Diversity and User Interface Quality	(MINIUKOVICH; DE ANGELI, 2015b)
Visual Complexity of Graphical User Interfaces	(MINIUKOVICH; SULPIZIO; DE ANGELI, 2018)
Facets of Visual Aesthetics	(MOSHAGEN; THIELSCH, 2010)
A Short Version of the Visual Aesthetics of Websites Inventory	(MOSHAGEN; THIELSCH, 2013)
How Quickly Can We Predict Users' Ratings on Aesthetic Evaluations of Websites? Employing Machine Learning on Eye-tracking Data	(PAPPAS et al., 2020)
Investigating Objective Measures of Web Page Aesthetics and Usability	(PURCHASE et al., 2011)
Predicting Users' First Impressions of Website Aesthetics with a Quantification of Perceived Visual Complexity and Colorfulness	(REINECKE et al., 2013)
Quantifying Visual Preferences Around the World	(REINECKE; GAJOS, 2014)
An Indonesian Adaption of Visual Aesthetics of Website Inventory (VisAWI) Questionnaire for Evaluating Video Game User Interface	(SADITA et al., 2022)
Farsi Version of Visual Aesthetics of Website Inventory (FV-VisAWI): Translation and Psychometric Evaluation	(SAREMI et al., 2022)
Assessing the quality of mobile graphical user interfaces using multi-objective optimization	(SOUI et al., 2020)

(Continuation of Table 7)

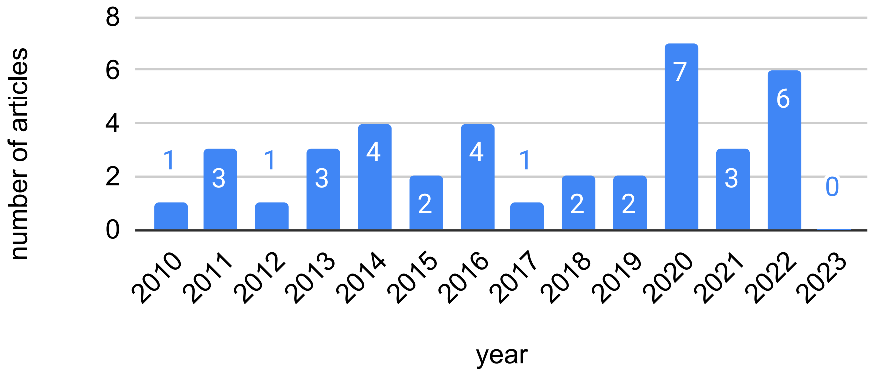
Title	Reference
User's Web Page Aesthetics Opinion: A Matter of Low-level Image Descriptors Based on MPEG-7	(URIBE; ÁLVAREZ; MENÉNDEZ, 2017)
A novel webpage layout aesthetic evaluation model for quantifying webpage layout design	(WAN et al., 2021)
An Entropy-based Approach for Computing the Aesthetics of Interfaces	(WANG, C.; REN, 2018)
Evaluating the Visual Quality of Web Pages Using a Computational Aesthetic Approach	(WU, O. et al., 2011)
Multimodal Web Aesthetics Assessment Based on Structural SVM and Multitask Fusion Learning	(WU, O. et al., 2016)
Computational model for predicting user aesthetic preference for GUI using DCNNs	(XING et al., 2021)
AI-driven user aesthetics preference prediction for UI layouts via deep convolutional neural networks	(XING et al., 2022)
Evaluation of Digital Twin Interface Based on Aesthetics	(YANG et al., 2022)
Metric-based Evaluation of Graphical User Interfaces: Model, Method, and Software Support	(ZEN, 2013)
Towards an Evaluation of Graphical User Interfaces Aesthetics Based on Metrics	(ZEN; VANDERDONCKT, 2014)
Assessing User Interface Aesthetics Based on the Inter-subjectivity of Judgment	(ZEN; VANDERDONCKT, 2016)

Source: the author.

Over the period of analysis, the number of published articles has varied from one to four. In the last few years, the number of studies has risen to seven in 2020 and six in 2022 (Figure 5). Most studies on questionnaire-based assessments were published earlier, from 2010 to 2013, with translated versions in 2022. Objective approaches have been published throughout the last ten years, with those based on deep learning techniques being published more recently, especially from 2019 on.

For each article that met the inclusion and quality criteria, we extracted information that characterized and classified the assessment approaches and techniques following the analysis questions. Data has been extracted by the first author and revised by the second author until consensus has been achieved. Data extraction was hindered in several cases due to a lack of a common format to describe these approaches. Many papers lack sufficient detail about research quality, such as validity and reliability. And, even when available, the evaluation descriptions often lack details, such as a clear definition of the adopted research design. Thus, the authors inferred some information based on the information

Figure 5 – Distribution of relevant articles per year of publication.



Source: the author.

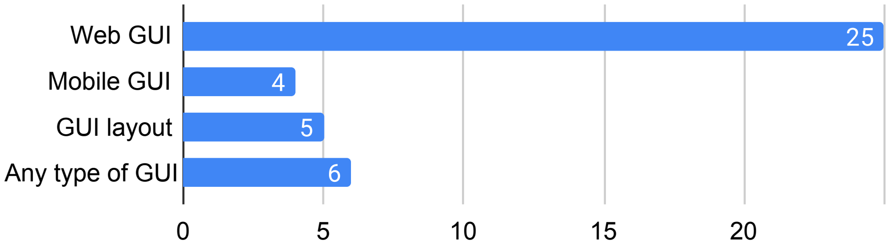
reported.

3.1.3 Data analysis

3.1.3.1 What are the characteristics of the assessed objects?

Most articles describe a method to assess the visual aesthetics of GUIs or some of their elements, while six papers present a study on visual aesthetics components, such as visual complexity (MINIUKOVICH; DE ANGELI, 2014b; MINIUKOVICH; SULPIZIO; DE ANGELI, 2018), visual diversity (MINIUKOVICH; DE ANGELI, 2015b), or layout features that correlate with visual aesthetics (ZEN, 2013; ZEN; VANDERDONCKT, 2014). One article measures 15 visual qualities of single GUI items (JYLHÄ; HAMARI, 2020).

Figure 6 – Distribution of assessments studies per type of GUI.



Source: the author.

About two-thirds of the studies (64.1%) focus on assessing web GUIs, while only four articles present the assessment of mobile GUIs (JYLHÄ; HAMARI, 2020; MINIUKOVICH; DE ANGELI, 2014b, 2015a; SOUI et al., 2020). Another research investigated mobile and web GUIs, but the mobile apps dataset was discarded because the connection with visual aesthetics was not clear (MINIUKOVICH; DE ANGELI, 2015b). Five papers analyze the visual aesthetics assessment of artificially constructed GUI models (blank screens with black squares representing the elements) designed by the authors (ALTABOLI; LIN, Y., 2011a) or by other users (WAN et al., 2021; XING et al., 2021, 2022; YANG et al., 2022). About 15% of the articles do not focus on a specific GUI type, presenting generic assessment approaches instead (Figure 6 and Table 8).

Table 8 – Classification of studies per type of GUI.

Type of GUI	References
Web	(ABBAS; HIRSCHFELD; THIELSCH, 2022; ABBASI et al., 2012; BAKAEV et al., 2020; DOU et al., 2019; GU et al., 2020; KHANI et al., 2016; LIU, X.; JIANG, 2021; MAITY; BHATTACHARYA, 2019, 2020; MINIUKOVICH; DE ANGELI, 2014a, 2015a, 2015b; MINIUKOVICH; SULPIZIO; DE ANGELI, 2018; MOSHAGEN; THIELSCH, 2010, 2013; PAPPAS et al., 2020; PURCHASE et al., 2011; REINECKE; GAJOS, 2014; REINECKE et al., 2013; SADITA et al., 2022; SAREMI et al., 2022; URIBE; ÁLVAREZ; MENÉNDEZ, 2017; WU, O. et al., 2011, 2016)
Mobile	(JYLHÄ; HAMARI, 2020; MINIUKOVICH; DE ANGELI, 2014b, 2015a; SOUI et al., 2020)
GUI model	(ALTABOLI; LIN, Y., 2011a; WAN et al., 2021; XING et al., 2021, 2022; YANG et al., 2022)
Any type	(ZEN, 2013; ZEN; VANDERDONCKT, 2014; MAITY; MADROSIYA; BHATTACHARYA, 2016; ZEN; VANDERDONCKT, 2016; WANG, C.; REN, 2018; MBENZA; BURNY, 2020)

Source: the author.

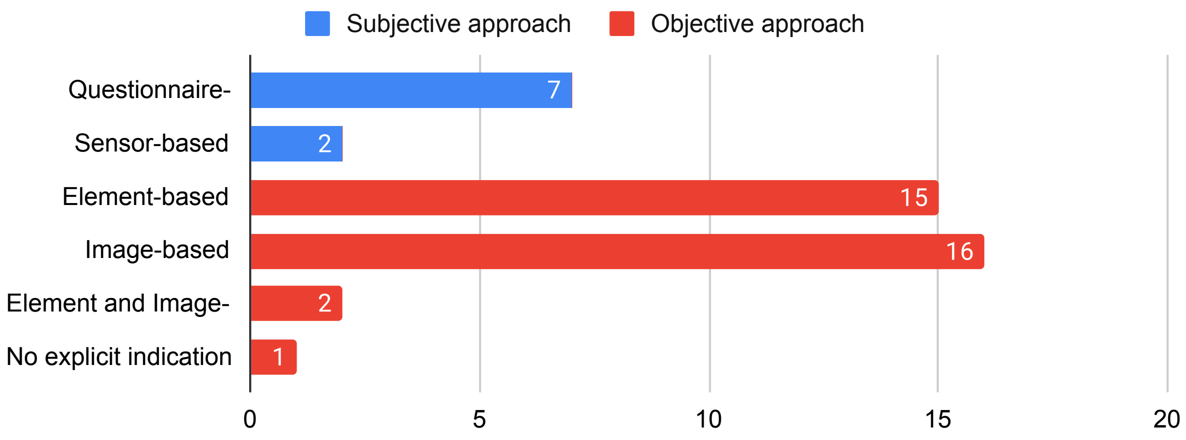
3.1.3.2 What techniques are used for assessment?

The vast majority of the studies adopt some objective approach (77%), with fewer using subjective ones. Among the subjective ones, seven studies capture the perceived visual aesthetics through questionnaire-based assessments, with four developing their own instruments, e.g.: (MOSHAGEN; THIELSCH, 2010), and three adapting the VisAWI questionnaire to foreign languages, e.g.: (ABBAS; HIRSCHFELD; THIELSCH, 2022). Also, two articles use a sensor-based technique using eye-tracking to measure the interest of users in websites (GU et al., 2020; PAPPAS et al., 2020). Among the objective approaches, the majority uses either element-based techniques to assess GUI visual aesthetics or image-based techniques, or a combination of both. While Zen (2013) does not explicitly indicate the used technique, the assessment seems to employ an objective approach.

3.1.3.2.1 Questionnaire-based approaches

The questionnaire-based approaches aim at capturing the perceived visual aesthetics based on the responses of the users. Two out of four studies using questionnaire-based

Figure 7 – Distribution of studies per assessment approaches.



Source: the author.

Table 9 – Classification of studies per type of assessment approach.

Approach	References
Questionnaire-based	(ABBAS; HIRSCHFELD; THIELSCH, 2022; BAKAEV et al., 2022; MOSHAGEN; THIELSCH, 2010; ABBASI et al., 2012; MOSHAGEN; THIELSCH, 2013; JYLHÄ; HAMARI, 2020; SADITA et al., 2022; SAREMI et al., 2022)
Sensor-based	(GU et al., 2020; PAPPAS et al., 2020)
Element-based	(ALTABOLI; LIN, Y., 2011a; PURCHASE et al., 2011; WU, O. et al., 2011; ZEN; VANDERDONCKT, 2014, 2016; MAITY; MADROSIYA; BHATTACHARYA, 2016; WU, O. et al., 2016; WANG, C.; REN, 2018; MAITY; BHATTACHARYA, 2019, 2020; MBENZA; BURNY, 2020; SOUI et al., 2020; LIU, X.; JIANG, 2021; YANG et al., 2022; WAN et al., 2021)
Image-based	(WU, O. et al., 2011; REINECKE et al., 2013; MINIUKOVICH; DE ANGELI, 2014a, 2014b, 2015a, 2015b; KHANI et al., 2016; REINECKE; GAJOS, 2014; WU, O. et al., 2016; URIBE; ÁLVAREZ; MENÉNDEZ, 2017; MINIUKOVICH; SULPIZIO; DE ANGELI, 2018; DOU et al., 2019; BAKAEV et al., 2020, 2022; XING et al., 2021, 2022)
Not informed	(ZEN, 2013)

Source: the author.

techniques aim at measuring the visual aesthetics of websites in terms of simplicity, diversity, color, and craftsmanship. The Visual Aesthetics of Website Inventory questionnaire (VisAWI) (MOSHAGEN; THIELSCH, 2010) is composed of 18 items on a 7-point Likert scale to which respondents assessing websites indicate their level of agreement. Simplicity and diversity are measured by five items each, while color and craftsmanship are measured by four. Based on the responses, each factor is calculated as the mean value of the corresponding items, and the higher their final score, as the mean value of all responses, the higher the contribution to the GUI visual aesthetics. In the same way, the visual aesthetics final score, which is the general second-order factor comprising the four factors, is computed as the mean value of all 18 items on the 7-point Likert scale. Moshagen and Thielsch (2013) also propose a short version of this questionnaire (VisAWI-S) with only four items that can provide a brief visual aesthetics assessment of websites. Despite having one item for each factor, the final score computation follows the same procedure as the full questionnaire. Three other studies adapted and validated the VisAWI to foreign languages (Arabic, Farsi, and Indonesian) (ABBAS; HIRSCHFELD; THIELSCH, 2022; SADITA et al., 2022; SAREMI et al., 2022).

Another questionnaire targets the visual aesthetics measure as a requirement of the "quality in use" aspect as defined by ISO/IEC 25010 (ABBASI et al., 2012). Here, the GUI aesthetics is represented through a factor of "pleasure," a sub-characteristic of "satisfaction" as part of quality in use, considering that visual aesthetics is a strong determinant of user satisfaction. The questionnaire is composed of two items on a 7-point Likert scale for each of the four sub-characteristics (attractiveness, enjoyment, admirability, engaging) to capture the visual aesthetics of GUIs. The visual aesthetic final score is obtained from the mean value of the eight items converted into a percentage value (0 - 100%).

Table 10 – Overview of the questionnaire-based studies.

Reference	**Assessed object**	**Assessed factors**	**N. of items**	**Scale**	**Final score**
(ABBASI et al., 2012)	whole website	attractiveness, enjoyment, admirability, engaging	8	7-point Likert	percentual mean score
(MOSHAGEN; THIELSCH, 2010; ABBAS; HIRSCHFELD; THIELSCH, 2022; SADITA et al., 2022; SAREMI et al., 2022)	whole website	simplicity, diversity, color, craftsmanship	18	7-point Likert	mean value
(MOSHAGEN; THIELSCH, 2013)	whole website	simplicity, diversity, color, craftsmanship	4	7-point Likert	mean value
(JYLHÄ; HAMARI, 2020)	mobile app icons	excellence/inferiority, graciousness/harshness, idleness/liveliness, normalness/bizarreness, complexity/simplicity	15	7-point semantic differential	not computed

Source: the author.

Jylhä and Hamari (2020) present a questionnaire (VISQUAL) to assess five dimensions of GUI elements visual quality (excellence/inferiority, graciousness/harshness, idleness/liveliness, normalness/bizarreness, and complexity/simplicity). When testing the method, users responded to the questionnaire with 22 adjective pairs (beautiful/ugly, calm/exciting, colorful/colorless, complex/simple, and so on) on a 7-point semantic differential scale to assess four icons out of pre-selected 68 mobile game icons. After analyzing the results using exploratory factor analysis and confirmatory factor analysis, they adjusted the questionnaire to 15 items. No final score for visual aesthetics is presented.

3.1.3.2.2 Sensor-based approaches

An alternative to subjective questionnaires is to collect areas of interest from the users when they look at web GUIs via eye-tracking (GU et al., 2020; PAPPAS et al., 2020). That is done by collecting data about the eye reaction when looking at something, which includes the pupil diameter, fixations of the eyes at some point, saccades (movement of both eyes between fixations), and other events. Gu et al. (2020) measure the gaze points and the eye movement speed to identify areas of interest over GUIs. They calculate an entropy value (rVAE) through a derivation of Shannon entropy from each GUI heatmap, i. e., the probability distribution of those areas of interest (SHANNON, 1948). Since the rVAE negatively correlates with human assessments as “good” or “bad”, a low rVAE score indicates high visual aesthetics.

Table 11 – Overview of the sensor-based studies.

Reference	**Metrics**	**Technique**
(GU et al., 2020)	gaze points, eye movement speed, areas of interest	Computation of relative visual attention entropy (rVAE) from the probability distribution of areas of interest
(PAPPAS et al., 2020)	pupil diameter (mean, median, max, sd); fixation duration (mean, median, max, sd); fixation dispersion (mean, median, max, sd); skewness of fixation duration histogram; ratio of forward saccades to total saccades; ratio of global and local saccades with a threshold on sac. vel.; skewness of saccade velocity histogram; saccade velocity (mean, median, max, sd); saccade amplitude (mean, median, max, sd); saccade duration (mean, median, max, sd); number of fixations, number of saccades, fixation to saccade ratio	Random Forest regression with metrics as independent variables to estimate simplicity, diversity, color, and craftsmanship

Source: the author.

Pappas et al. (2020) collect 31 metrics, such as mean, variance, minimum, maximum, and median of pupil diameter, eye fixation details, and saccades details. Using the collected data as input for a Random Forest regression algorithm, they estimate the simplicity, diversity, color, and craftsmanship of GUIs that have been previously assessed using the VisAWI questionnaire (MOSHAGEN; THIELSCH, 2010). In this way, each factor’s final score ranges from 1 to 7, as the questionnaire is responded to on a 7-point Likest scale. No general visual aesthetics score is computed.

3.1.3.2.3 Element-based approaches

Element-based approaches consider GUI visual aesthetics as being composed of measurable characteristics (metrics) aiming at quantitatively measuring layout properties. Analyzing the proposed metrics, we can observe that they vary considerably in complexity. One of the simplest metrics is the number of GUI elements, their size, and aspect ratio (WU, O. et al., 2011, 2016; MAITY; MADROSIYA; BHATTACHARYA, 2016; MAITY; BHATTACHARYA, 2019). Another study also considers the white space area a relevant metric (MAITY; BHATTACHARYA, 2019). Metrics related to text elements include the number of text blocks (WU, O. et al., 2011), text blocks aspect ratio (WU, O. et al., 2016), relative area to the whole GUI (WU, O. et al., 2016; MAITY; BHATTACHARYA, 2019), word and letter spacing, character density, font size, and line-height (MAITY; MADROSIYA; BHATTACHARYA, 2016; MAITY; BHATTACHARYA, 2019).

Seven studies (PURCHASE et al., 2011; ALTABOLI; LIN, Y., 2011a; ZEN; VANDERDONCKT, 2014, 2016; MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; WAN et al., 2021) used a subset of the 14 metrics presented by Ngo (NGO; SAMSUDIN; ABDULLAH, 2000; NGO, 2001; NGO; TEO; BYRNE, 2002, 2003) to quantify different layout aspects of GUIs. These metrics are calculated based on properties of the GUI elements, such as area, distance from the central line of the GUI, the number of elements, and others, to result in values ranging from 0 to 1, indicating the presence of aesthetic factors on the GUI (PURCHASE et al., 2011). None of the studies calculate all 14 metrics, and Mbenza and Burny (2020) combine six of Ngo's metrics with three metrics proposed by Vanderdonckt and Gillo (1994) in addition to one originally defined by the authors (grouping). Wang and Ren (2018) propose Shannon entropy calculation to measure GUI disorder and uncertainty, i.e., their visual complexity. The entropy is computed based on manually formed element blocks and the proportion of the area occupied by each block.

In general, the studies do not apply different treatments for the different types of GUI elements. However, few exceptions adopt distinct approaches to discriminate against them (WU, O. et al., 2011, 2016; PURCHASE et al., 2011; MAITY; BHATTACHARYA, 2019; WAN et al., 2021). Wu et al. (2011) detect text blocks as well as the blocks that have no text, applying specific text metrics (number of text blocks, the ratio of the area of all text blocks to the whole GUI area, and character density) to analyze their influence on the general visual aesthetics. Wu et al. (2016) use a similar procedure calculating the ratio of the area of all text blocks to the whole GUI area and the proportion of the textual pixels in the entire region. Maity and Bhattacharya (MAITY; BHATTACHARYA, 2019) primarily study text, image, and white space properties from which they calculate an aesthetics score based on the weighted average of each value. Purchase et al. (2011) calculate an overall score using Ngo's formulas based on the ten individual metrics and the order and complexity metric. The order and complexity metric is computed separately for the set of text, images, control components, and also for all elements, resulting in 14 metrics.

Most element-based approaches consider at least ten metrics to calculate visual aesthetics score, although some are based on a smaller number (ALTABOLI; LIN, Y., 2011a; MAITY; BHATTACHARYA, 2020; WAN et al., 2021; WANG, C.; REN, 2018; ZEN; VANDERDONCKT, 2016). Altaboli and Lin (2011) considered only three metrics (balance, unity, and sequence) (NGO; TEO; BYRNE, 2002) to simplify their experiment in which they manipulate each of these factors to assume either a high or low level to yield different results. Zen and Vanderdonckt (2016) implemented four metrics (balance, equilibrium, density, and economy) (NGO; SAMSUDIN; ABDULLAH, 2000) to identify relationships between the automatically computed and human scores. Wang and Ren (2018) proposed the computation of the element blocks entropy to measure the visual complexity as representative of the visual aesthetics of the GUI.

Table 12 – Description of the metrics.

(To be continued)

Metric	Description	Reference
alignment	as in (VANDERDONCKT; GILLO, 1994)	(MBENZA; BURNY, 2020)
aspect ratio	height/width	(WU, O. et al., 2011, 2016; MAITY; BHATTACHARYA, 2019)
balance	as in (NGO; TEO; BYRNE, 2000)	(ALTABOLI; LIN, Y., 2011a; MAITY; BHATTACHARYA, 2020; PURCHASE et al., 2011; ZEN; VANDERDONCKT, 2014; WAN et al., 2021)
	as in (VANDERDONCKT; GILLO, 1994)	(MBENZA; BURNY, 2020)
	adapted from (GALITZ, 2007)	(YANG et al., 2022)
cohesion	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; PURCHASE et al., 2011; WAN et al., 2021)
complexity	as in (SOUI et al., 2017)	(SOUI et al., 2020)
composition	as in (SOUI et al., 2017)	(SOUI et al., 2020)
continuity	adapted from (GALITZ, 2007)	(YANG et al., 2022)
density	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
	as in (SOUI et al., 2017)	(SOUI et al., 2020)

(Continuation of Table 12)

Metric	Description	Reference
economy	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
element ratio	ratio of text blocks area to the whole GUI area	(WU, O. et al., 2011, 2016)
equilibrium	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
grouping	similar sizes, shapes, colors, or positions for related information and keeping them close to each other	(MBENZA; BURNY, 2020)
homogeneity	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
integrality	as in (SOUI et al., 2017)	(SOUI et al., 2020)
intensity	adapted from (GALITZ, 2007)	(YANG et al., 2022)
number of elements	number of layout blocks number of text blocks number of layers of the block tree number of leaf nodes	(WU, O. et al., 2011, 2016) (WU, O. et al., 2011) (WU, O. et al., 2011, 2016) (WU, O. et al., 2011, 2016)
order and complexity	as in (NGO; TEO; BYRNE, 2000)	(PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
proportion	as in (NGO; TEO; BYRNE, 2000) as in (VANDERDONCKT; GILLO, 1994) adapted from (GALITZ, 2007)	(MAITY; BHATTACHARYA, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014) (MBENZA; BURNY, 2020) (YANG et al., 2022)
regularity	as in (NGO; TEO; BYRNE, 2000) as in (SOUI et al., 2017)	(MAITY; BHATTACHARYA, 2020; ZEN; VANDERDONCKT, 2014; WAN et al., 2021) (SOUI et al., 2020)
repartition	as in (SOUI et al., 2017)	(SOUI et al., 2020)

(Continuation of Table 12)

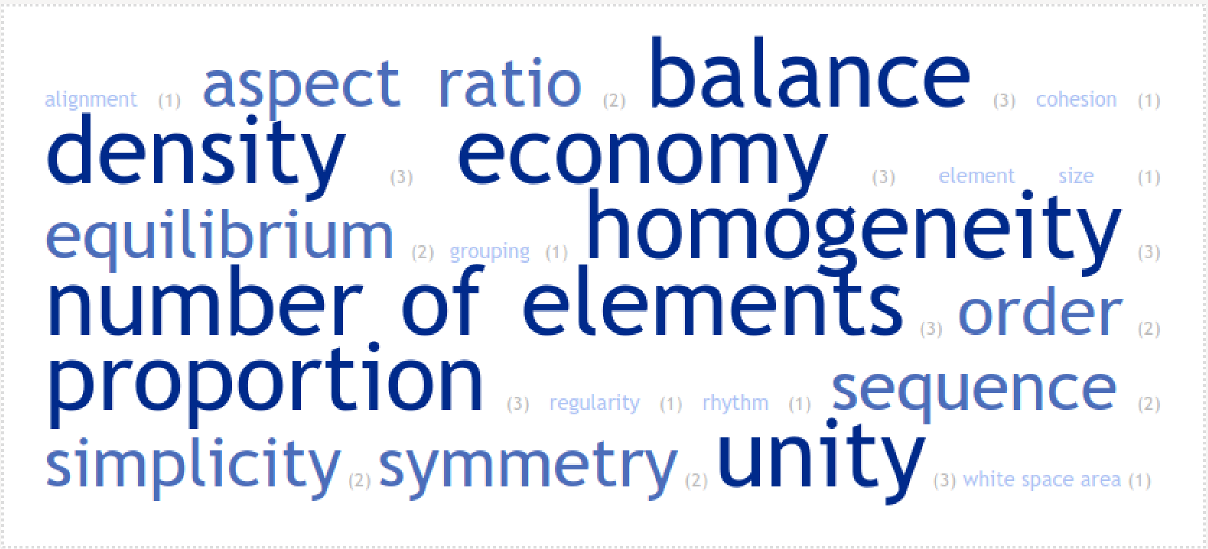
Metric	Description	Reference
rhythm	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
sequence	as in (NGO; TEO; BYRNE, 2000)	(ALTABOLI; LIN, Y., 2011a; MAITY; BHATTACHARYA, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
simplicity	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
sorting	as in (SOUI et al., 2017)	(SOUI et al., 2020)
symmetry	as in (NGO; TEO; BYRNE, 2000)	(MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
	as in (SOUI et al., 2017)	(SOUI et al., 2020)
	adapted from (GALITZ, 2007)	(YANG et al., 2022)
unity	as in (NGO; TEO; BYRNE, 2000)	(ALTABOLI; LIN, Y., 2011a; MAITY; BHATTACHARYA, 2020; MBENZA; BURNY, 2020; PURCHASE et al., 2011; WAN et al., 2021; ZEN; VANDERDONCKT, 2014)
visual complexity	ratio of JPEG size to the whole GUI image area	(WU, O. et al., 2011, 2016)
	Shannon entropy (SHANNON, 1948)	(WANG, C.; REN, 2018)
white space	white space area	(MAITY; BHATTACHARYA, 2019)
width and height of elements	sum of width and height of all layout blocks	(WU, O. et al., 2011, 2016)

Source: the author.

Figure 8 illustrates the frequencies of use of the metrics. Balance, unity, and number of elements are the most common metrics, being considered in four studies. Metrics related to the aspect ratio, density, economy, homogeneity, proportion, sequence, and simplicity are used in three studies. Equilibrium, order and complexity, and symmetry appear in two studies. The least frequent metrics are alignment, cohesion, grouping, regularity, rhythm, white space area, and entropy.

Some studies automated the calculation of the metrics. Purchase et al. (2011) developed a Firefox extension in JavaScript to identify the size and the position of the

Figure 8 – Measured factors (frequency of appearance shown in parenthesis).



Source: the author.

GUI elements directly from the web-page HTML code. The script calculates the values of the 11 metrics. Instead of measuring single elements, Zen and Vanderdonckt (2016; 2014) present a web application that analyzes regions that users manually define over the GUI. Those regions are not limited by the number of elements but are arbitrarily defined by users. Any GUI type can be loaded into the application by providing its image file. The calculation of the metrics is based on the regions defined by the users. Mbenza and Burny (2020) developed a RESTful web service in Java to compute the metrics from GUIs accessible via URI or as an image file. Once the metrics are calculated, they are presented to the user in HTML format.

3.1.3.2.4 Image-based approaches

The majority of the approaches adopt an image-based approach, in which they assess visual aesthetics based on features of screenshots of the GUIs extracting and measuring features associated with visual aesthetics. Therefore, the majority use computer vision techniques to extract handcrafted features that had been previously selected. Six approaches use neural networks, directly learning these features associated with visual aesthetics.

In image-based approaches, the use of color is, by far, the most popular feature examined for assessing GUI visual aesthetics. All image-based approaches extracting handcrafted features analyze color variability, i.e., the number of different colors perceived in GUIs. Some studies divide color variability into two sub-features, e.g., the number of dominant colors and color range (MINIUKOVICH; DE ANGELI, 2014a, 2014b, 2015a), or three sub-features, such as the number of dominant colors, variance among different colors, and variance among different tonalities of the same color (URIBE; ÁLVAREZ; MENÉNDEZ, 2017). Other studies compute color variability as a single feature, using Hasler's metric of colorfulness (HASLER; SUESSTRUNK, 2003) as one of its components (WU, O. et al., 2011, 2016; REINECKE et al., 2013; REINECKE; GAJOS, 2014). Miniukovich and De Angeli (2015a), Miniukovich et al. (2018), and Maity and Bhattacharya (2019) analyze only the number of dominant colors.

Besides color, some studies also extract texture features from GUI screenshots (MINIUKOVICH; DE ANGELI, 2015b; URIBE; ÁLVAREZ; MENÉNDEZ, 2017), imagery (MAITY; BHATTACHARYA, 2019), or their visual blocks (WU, O. et al., 2011, 2016) or layout features including symmetry (MINIUKOVICH; DE ANGELI, 2014a, 2014b), grid quality (MINIUKOVICH; DE ANGELI, 2015a), white space (MINIUKOVICH; DE ANGELI, 2015a), and the number of alignment points (MINIUKOVICH; SULPIZIO; DE ANGELI, 2018).

Table 13 – Overview on the analyzed texture features.

Texture feature	Description	Reference
coarseness	as in (TAMURA; MORI; YAMAWAKI, 1978) - average, average contrast, and variance	(WU, O. et al., 2011, 2016)
contrast	as in (TAMURA; MORI; YAMAWAKI, 1978) - average, average contrast, and variance	(WU, O. et al., 2011, 2016)
directionality	as in (TAMURA; MORI; YAMAWAKI, 1978) - average, average contrast, and variance	(WU, O. et al., 2011, 2016)
edge histogram (EH)	as in (IAKOVIDOU et al., 2014)	(MINIUKOVICH; DE ANGELI, 2015b)
color and edge directivity histogram (CEDD)	as in (IAKOVIDOU et al., 2014)	(MINIUKOVICH; DE ANGELI, 2015b)
fuzzy color and texture histogram (FCTH)	as in (IAKOVIDOU et al., 2014)	(MINIUKOVICH; DE ANGELI, 2015b)
line energy	from edge histogram descriptor (MANJUNATH; SALEMBIER; SIKORA, 2002)	(URIBE; ÁLVAREZ; MENÉNDEZ, 2017)
line homogeneity	from edge histogram descriptor (MANJUNATH; SALEMBIER; SIKORA, 2002)	(URIBE; ÁLVAREZ; MENÉNDEZ, 2017)
smoothness	as in (DAUBECHIES, 1992)	(MAITY; BHATTACHARYA, 2019)

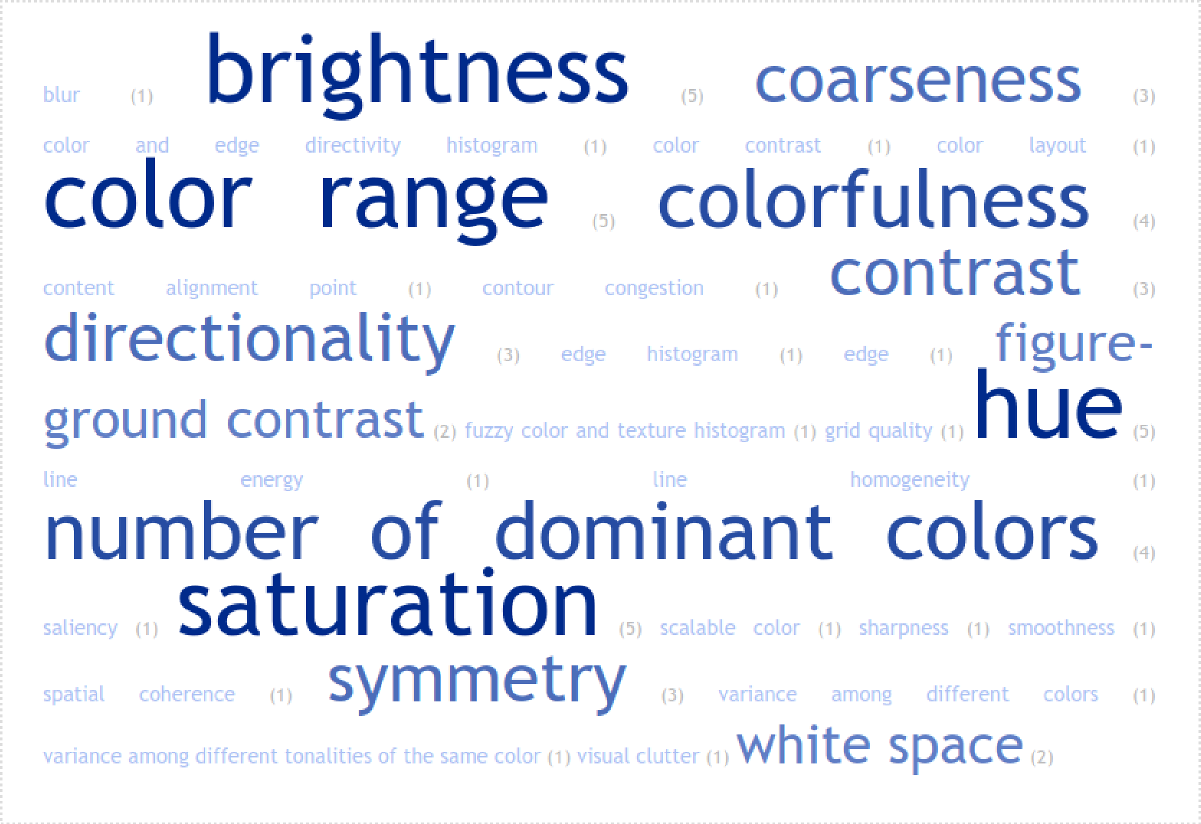
Source: the author.

The layout features include symmetry, grid quality, white space, and the number of alignment points. The symmetry feature indicates how much the left side of an image matches with the right side. It can be measured through the ratio of matching contour pixels across the central vertical axis to all GUI contour pixels (MINIUKOVICH; DE ANGELI, 2014a, 2014b). Another measurement technique segments the GUI into visual blocks (CAO; MAO; LUO, 2010) and analyzes their shift from the central vertical axis when it crosses the blocks or from a matching block across that axis (MINIUKOVICH; DE ANGELI, 2015a; MINIUKOVICH; SULPIZIO; DE ANGELI, 2018). The GUI segmentation into visual blocks also allows measuring the grid quality and white space. Grid quality is based on the number of visual blocks, the number of alignment points of blocks, the number of block sizes, and the proportion of the GUI that is covered by same-size blocks (BALINSKY, 2006) when applied to web GUIs (MINIUKOVICH; DE ANGELI, 2015a). On the other hand, the grid quality of mobile GUIs is based on the number of vertical block sizes instead of the number of visual blocks and the number of alignment points of blocks, considering that mobile GUIs often display elements in a single-aligned column. As a layout component, white space is the area that is not covered by visual blocks (MINIUKOVICH; DE ANGELI, 2015a). When the segmentation of GUIs uses the Quadtree decomposition, which is the recursive division of GUI images into four sub-images until they are sufficiently uniform, it results in large content-free areas and also content-dense squares (MINIUKOVICH; SULPIZIO; DE ANGELI, 2018). With this method, white space is considered the number of 64-pixel and 128-pixel squares, where the number of alignment points is the count of vertically aligned block points in a GUI (MINIUKOVICH; SULPIZIO; DE ANGELI, 2018).

Six approaches also adopt neural networks to assess the GUI visual aesthetics. While only two of them were found in the first search (KHANI et al., 2016; DOU et al., 2019), the other five were published more recently and we only found them after updating this research (BAKAEV et al., 2020, 2022; XING et al., 2021, 2022). Deep learning networks are multi-layered structures that can automatically extract a complex data representation and identify more complicated relationships between the system input and output, so there is no need to extract features manually (LI, X. et al., 2016).

Except for Bakaev et al. (2020), all other studies use convolutional neural networks (CNNs) to analyze the GUIs. CNNs comprise multiple layers that learn to make linear and non-linear operations (MAHMOOD et al., 2017). They are commonly composed of three layer types: convolution, pooling, and fully connected layers (AGGARWAL, 2018). Convolutional layers contain sets of filters to detect image features, from simple features in the first layer to larger and more complex features in the subsequent ones (ANDREARCZYK; WHELAN, 2017). Pooling layers reduce the dimensionality of the feature maps generated by a convolutional layer. That makes features detectable anywhere in the image (AGGARWAL, 2018). Fully connected layers use convolution/pooling process results to classify the image into a category, such as “good” or “bad.” That was the case in Khani et al. (2016). Because all other models adopted regression to assign the image with a

Figure 9 – Features extracted and the frequency of appearance (in parenthesis).



Source: the author.

numerical score, they added a single neuron as the final layer. The synthesis of the models are in Table 14.

3.1.3.3 How have the approaches been evaluated?

Most of the approaches evaluate models' validity using the statistical significance of the relationship between GUI factors and human ratings. One study (DOU et al., 2019), however, performs a correlation analysis between their deep learning model and human ratings, while a few studies measure the error between the assessments of their models and GUI actual scores (WU, O. et al., 2011; URIBE; ÁLVAREZ; MENÉNDEZ, 2017; MAITY; BHATTACHARYA, 2019; PAPPAS et al., 2020). When analyzing the human assessment quality, few studies evaluated inter-rater reliability (REINECKE et al., 2013; REINECKE; GAJOS, 2014; MINIUKOVICH; DE ANGELI, 2014a, 2014b, 2015a; MINIUKOVICH;

[1] (KRIZHEVSKY; SUTSKEVER; HINTON, G. E., 2017)
[2] (KARAYEV et al., 2014)
[3] (KONONENKO; KUKAR, 2007)
[4] (HU et al., 2019)
[5] (SZEGEDY et al., 2015)
[6] (TAN; LE, 2020)

Table 14 – Characteristics of the deep learning models.

Ref	Framework	Input	Deep layers	Output layers	Learning algorithm	Dataset separation	Error rate
(KHANI et al., 2016)	AlexNet [1]	227 x 227 (pixels) x 3 (colors)	5 conv, 3 max-pool	2 fully-connected	SVM with Gaussian radial basis function kernel	90% training; 10% validation	34.15%
(DOU et al., 2019)	CaffeNet [2]	256 x 192 (pixels) x 3 (colors)	5 conv, 2 max-pool	2 fully-connected, 1 neuron	Backprop	75.4% training; 24.6% validation	20.41%
(BAKAEV et al., 2020, 2022)	ANN [3]	32 metrics	2 fully-connected	1 neuron	Backprop	80% training; 20% validation	87.86%
(XING et al., 2021)	SE-VGG19 [4]	224 x 224 (pixels) x 3 (colors)	5 conv, 2 max-pool	1 SE block, 1 fully-connected, 1 neuron	Backprop	90% training; 10% validation	14.9% (collection) 25.38% (likes)
(BAKAEV et al., 2022)	GoogLeNet [5]	224 x 224 (pixels) x 3 (colors)	2 conv, 9 inception, 4 max-pool	1 avg-pool, 1 dropout, 1 linear, 1 neuron	Backprop	80% training; 20% validation	98.38%
(XING et al., 2022)	EfficientNet [6]	224 x 224 (pixels) x 3 (colors)	5 conv, 2 max-pool	1 fully-connected, 1 neuron	Backprop	60% training; 20% validation; 20% testing	1.46%

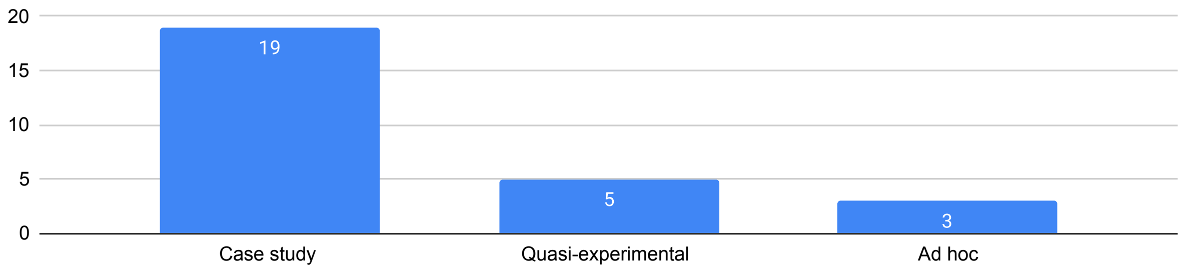
Source: the author.

SULPIZIO; DE ANGELI, 2018) or inter-rater agreement (ZEN; VANDERDONCKT, 2016).

The majority of the approaches conduct case studies for the evaluation. In these non-experimental studies, the assessment method is systematically defined, and the results of the proposed approaches are compared (via correlation or linear regression) with participants' perception of visual aesthetics collected through questionnaires. Five studies applied a quasi-experimental strategy, in which they compare their automated assessment with other approaches acting as the control group. It differs from an experimental design because it does not apply a random allocation of participants to the experimental or control group. Three studies were conducted in an *ad hoc* manner, describing their observations in pilot studies only informally.

Most of the studies collected GUI samples from the internet yet used different meth-

Figure 10 – Distribution of evaluations per study type.



Source: the author.

ods. In some cases, students collected a fixed number of samples (WU, O. et al., 2016), manually choosing beautiful and ugly GUIs from specific repositories (MINIUKOVICH; DE ANGELI, 2014b, 2015a, 2015b; GU et al., 2020) or using a crowdsourcing platform (MINIUKOVICH; DE ANGELI, 2015a, 2015b). Some articles do not detail the dataset acquisition, only mentioning restrictions applied to the process (REINECKE et al., 2013; PURCHASE et al., 2011; MINIUKOVICH; SULPIZIO; DE ANGELI, 2018; MINIUKOVICH; DE ANGELI, 2014a). Few studies use samples designed specifically for their research (ALTABOLI; LIN, Y., 2011a; MAITY; MADROSIYA; BHATTACHARYA, 2016; MAITY; BHATTACHARYA, 2019), while others use the dataset collected by Reinecke and Gajos (2014) that was later made available as a public dataset (KHANI et al., 2016; WU, O. et al., 2016; URIBE; ÁLVAREZ; MENÉNDEZ, 2017; DOU et al., 2019). That is the same set Reinecke et al. (REINECKE et al., 2013) use, excluding 20 grayscale GUIs. URIBE et al. (2017) and Altaboli and Lin (ALTABOLI; LIN, Y., 2011a) also use one of the datasets prepared by Moshagen and Thielsch (2010) to validate their findings. Similarly, one of the sets collected by Miniukovich and De Angeli (2015a) is reused in another work (MINIUKOVICH; DE ANGELI, 2015b).

Table 15 – Overview on the datasets used for evaluation.

(To be continued)

#	Reference	N. of samples	N. of human raters	N. of ratings by sample
1	(MOSHAGEN; THIELSCH, 2010)	2	78	39
2	(MOSHAGEN; THIELSCH, 2010)	2	375	188
3	(ABBASI et al., 2012)	2	31	31
4	(ZEN; VANDERDONCKT, 2014)	4	25	25
5	(MBENZA; BURNY, 2020)	5	16	16
6	(URIBE; ÁLVAREZ; MENÉNDEZ, 2017)	6	110	110

(Continuation of Table 15)

#	Reference	N. of samples	N. of human raters	N. of ratings by sample
7	(ALTABOLI; LIN, Y., 2011a)	8	13	13
8	(PAPPAS et al., 2020)	9	23	23
9	(MOSHAGEN; THIELSCH, 2013)	10	305	31
10	(ZEN; VANDERDONCKT, 2016)	10	15	15
11	(PURCHASE et al., 2011)	15	21	21
12	(MOSHAGEN; THIELSCH, 2013)	24	764	32
13	(MAITY; MADROSIYA; BHATTACHARYA, 2016)	30	30	100
14	(GU et al., 2020)	40	40	30
15	(MOSHAGEN; THIELSCH, 2010; ALTABOLI; LIN, Y., 2011a; URIBE; ÁLVAREZ; MENÉNDEZ, 2017)	42	512	512
16	(MOSHAGEN; THIELSCH, 2013)	50	604	12
17	(MINIUKOVICH; SULPIZIO; DE ANGELI, 2018)	55	26	26
18	(JYLHÄ; HAMARI, 2020)	68	569	33
19	(MAITY; BHATTACHARYA, 2019)	95	185	37
20	(MINIUKOVICH; DE ANGELI, 2014b)	99	20	20
21	(MOSHAGEN; THIELSCH, 2010)	100	300	21
22	(MOSHAGEN; THIELSCH, 2010)	100	506	101
23	(MINIUKOVICH; DE ANGELI, 2014a)	140	10	10
24	(MINIUKOVICH; DE ANGELI, 2015b)	150	45	5
25	(MAITY; BHATTACHARYA, 2019)	150	130	130
26	(WU, O. et al., 2011)	154	7	7
27	(MAITY; BHATTACHARYA, 2019)	250	83	83
28	(MINIUKOVICH; DE ANGELI, 2015a, 2015b)	300	60	10
29	(MINIUKOVICH; DE ANGELI, 2015a)	300	51	8
30	(REINECKE; GAJOS, 2014; KHANI et al., 2016; WU, O. et al., 2016; URIBE; ÁLVAREZ; MENÉNDEZ, 2017; DOU et al., 2019)	430	32,222	1,800
31	(REINECKE et al., 2013)	450	424	8
32	(WU, O. et al., 2011)	500	7	7

(Continuation of Table 15)

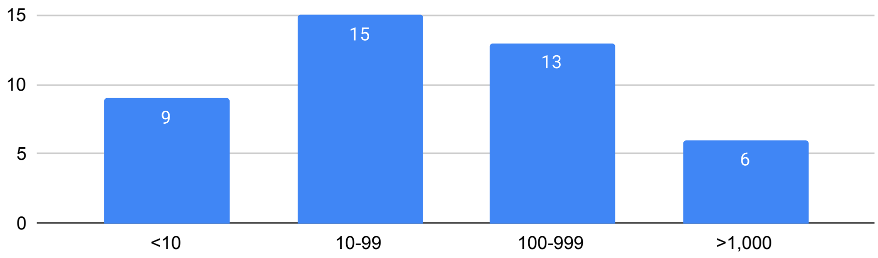
#	Reference	N. of samples	N. of human raters	N. of ratings by sample
33	(WU, O. et al., 2016)	1,000	10	10
34	(BAKAEV et al., 2020)	3,249	137	5
35	(SOUI et al., 2020)	24	20	20
36	(MAITY; BHATTACHARYA, 2020)	209	100	100
37	(XING et al., 2021)	38,423	-	-
38	(LIU, X.; JIANG, 2021)	11,000	110	1
39	(YANG et al., 2022)	8	10	10
40	(ABBAS; HIRSCHFELD; THIELSCH, 2022)	31	223	223
41	(BAKAEV et al., 2022)	2,932	137	5
42	(WAN et al., 2021)	50	90	90
43	(XING et al., 2022)	12,186	-	-
44	(SAREMI et al., 2022)	1	200	200
45	(SADITA et al., 2022)	1	56	56

Source: the author.

The studies used sample sets of varied sizes (Table 15). The smallest samples include only two (MOSHAGEN; THIELSCH, 2010; ABBASI et al., 2012) or less than ten GUIs (ALTABOLI; LIN, Y., 2011a; ZEN; VANDERDONCKT, 2014; URIBE; ÁLVAREZ; MENÉNDEZ, 2017; MBENZA; BURNY, 2020; PAPPAS et al., 2020). The largest samples, with more than 1,000 GUIs, are those that use machine learning techniques (BAKAEV et al., 2020, 2022; LIU, X.; JIANG, 2021; XING et al., 2021, 2022; WU, O. et al., 2016). Another set containing 430 web GUI screenshots is used in that same study to validate the same model (REINECKE; GAJOS, 2014). Several studies use datasets that vary from 10 to 99 samples, whereas most of them use datasets of considerable size, with at least 100 but less than 1,000 samples.

The studies collected assessments from different numbers of human raters to compare the results of the proposed approaches with their ratings. Not all studies used fully crossed research designs. One study collected ratings from more than 32,000 participants, each rating 30 GUIs, out of 430, twice (REINECKE; GAJOS, 2014), calculating the GUI's score as the mean of both ratings. The study reports a total of around 771,000 pairs of ratings, which makes an average of nearly 1,800 ratings for each GUI. On the other hand, some studies collected less than ten human ratings for each sample (REINECKE et al., 2013; MINIUKOVICH; DE ANGELI, 2015a, 2015b; WU, O. et al., 2011), while the average among the studies is about 110 ratings per GUI.

Figure 11 – Distribution of datasets per size.



Source: the author.

Very few studies evaluate the quality of human assessments. To confirm the inter-rater reliability, studies measured the standard deviation between participants' first and second ratings (REINECKE et al., 2013; REINECKE; GAJOS, 2014) or the intraclass correlation coefficient (MINIUKOVICH; DE ANGELI, 2014a, 2014b, 2015a; MINIUKOVICH; SULPIZIO; DE ANGELI, 2018) confirming high reliability among scores. Only Zen and Vanderdonckt (2016) presented the interrater agreement analysis using Randolph's Kappa (RANDOLPH, 2005), finding a moderate agreement between raters.

The studies also analyzed mean scores to detect sample skewing, confirming an acceptable balance between GUIs with high and low aesthetics (MINIUKOVICH; DE ANGELI, 2014a, 2014b, 2015a). However, sometimes a negative skewness has been observed, which shows bias to high visual aesthetics GUIs, indicating an unbalanced sample, which might pose some threat to the research validity. Zen and Vanderdonckt (2014) apply the Kolmogorov-Smirnov and the Shapiro-Wilk test to each GUI score to verify data normality. Since both tests brought a negative conclusion, they proceed to use non-parametric tests to investigate whether there is a similarity between their metric scores and user ratings. The one-sample Wilcoxon signed-rank test and the match-paired Wilcoxon signed-rank test help them conclude that, although metrics and rating values are not significantly similar, four out of twelve metrics allow the ranking of GUIs similarly to the way humans do.

Several studies present a Pearson correlation analysis between the results of their approaches and the human ratings (PURCHASE et al., 2011; ALTABOLI; LIN, Y., 2011a; MINIUKOVICH; DE ANGELI, 2015b; GU et al., 2020; DOU et al., 2019). In parts, these studies vary concerning the GUI properties they compare, following the way their respective approach is defined. The correlation values vary considerably from $r = .52$ (MINIUKOVICH; DE ANGELI, 2015b) to $r = .85$ (DOU et al., 2019), with a typical value of .69, showing a moderate to strong correlation (Table 16). One study (GU et al., 2020) presented a negative correlation, showing that the higher the visual attention entropy, the lowest the humans rated the GUI visual aesthetics.

Table 16 – Overview of Pearson correlation between results and human ratings.

Reference	Result of the proposed approach	Human rating score	*r*
(PURCHASE et al., 2011)	visual aesthetics rank of color GUIs	mean rank on perceived visual aesthetics of color GUIs	$.66^{*}$
(PURCHASE et al., 2011)	visual aesthetics rank of black and white GUIs	mean rank of perceived visual aesthetics of black and white GUIs	$.71^{*}$
(ALTABOLI; LIN, Y., 2011a)	visual aesthetics of GUI models	mean value of perceived visual aesthetics rated on a 10-point Likert scale	$.84^{**}$
(MINIUKOVICH; DE ANGELI, 2015b)	visual diversity metrics (average) - study 1	mean value of perceived visual aesthetics rated on a 7-point scale	$.52^{**}$
(MINIUKOVICH; DE ANGELI, 2015b)	visual diversity metrics (average) - study 2	mean value of perceived visual aesthetics rated on a 7-point scale	$.56^{*}$
(GU et al., 2020)	relative visual attention entropy (rVAE)	mean value of perceived visual aesthetics on a nominal scale rated as “good” (1) or “bad” (0)	$-.66^{**}$
(DOU et al., 2019)	deep learning model prediction	mean value of perceived visual aesthetics rated on 9-point scale	$.85^{**}$

* $p < .01$; ** $p < .001$

Source: the author.

Other approaches aim at establishing the relationship between several metrics or GUI features and visual aesthetics by conducting multiple regression analyses (Table 17) (MINIUKOVICH; DE ANGELI, 2014a, 2014a, 2015a; REINECKE et al., 2013; URIBE; ÁLVAREZ; MENÉNDEZ, 2017). Different from correlation, regression analysis aims to establish a causal relationship between the independent variables (metrics or features) and the dependent variable (visual aesthetics), expressed with a coefficient of determination (R^2). The studies show coefficients of determination varying from $R^2 = .13$ (MINIUKOVICH; DE ANGELI, 2015a) to $R^2 = .51$ (MINIUKOVICH; DE ANGELI, 2014a), indicating that the variability of visual aesthetics can be explained by up to 51% by GUI features.

Maity and Bhattacharya (2019) apply support vector regression (SVR) to model the visual aesthetics of text, images, which, combined with the white space area, are part of the metrics to assess the visual aesthetics of the whole GUIs. For the presented models, the predicted values are the assessments of the visual aesthetics of the samples in the interval [1, 5], while the observed values for text and images are the mean values, and for the whole GUIs are the statistical modes of the human ratings on a five-point Likert scale. The RMSE of each model (text, images, and GUI) is .59, .67, and .79, respectively. These results show that the RMSE in the whole GUI model is greater than in other models since

Table 17 – Regression between automatic assessment results and human ratings.

Reference	Coefficients	Human rating score	Model	R^2
(MINIUKOVICH; DE ANGELI, 2014a)	6 features of web GUI screenshots	mean value of perceived visual aesthetics rated on a 5-point Likert scale	linear regression model	.51
(MINIUKOVICH; DE ANGELI, 2014b)	6 features of Android app GUI screenshots	mean value of perceived visual aesthetics rated on a 7-point Likert scale	linear regression model	.36
(MINIUKOVICH; DE ANGELI, 2015a)	8 features of web GUI screenshots (150ms) - study 1	mean value of perceived visual aesthetics rated on a 7-point Likert scale	linear regression model	.49
(MINIUKOVICH; DE ANGELI, 2015a)	8 features of web GUI screenshots (4s) - study 1	mean value of perceived visual aesthetics rated on a 7-point Likert scale	linear regression model	.43
(MINIUKOVICH; DE ANGELI, 2015a)	8 features of iOS app GUI screenshots (150ms) - study 2	mean value of perceived visual aesthetics rated on a 7-point Likert scale	linear regression model	.13
(MINIUKOVICH; DE ANGELI, 2015a)	8 features of iOS app GUI screenshots (4s) - study 2	mean value of perceived visual aesthetics rated on a 7-point Likert scale	linear regression model	.18
(REINECKE et al., 2013)	colorfulness, visual complexity, and demographic variables	mean value of participants' perceived visual aesthetics rated on a 9-point Likert scale	linear mixed-effects model	.48
(URIBE; ÁLVAREZ; MENÉNDEZ, 2017)	10 low-level characteristics of GUIs	mean value of perceived visual aesthetics rated on a 9-point Likert scale	stepwise multiple regression algorithm	.404

Source: the author.

it relies on the individual text, images, and white space assessments to derive its predicted values.

Uribe et al. (2017) evaluate the results using RMSE on a regression model developed using 5-fold cross-validation to avoid overfitting. Applying the model on the complete dataset used for its development, comparing the results to human ratings on a 9-point Likert scale, results in an RMSE value of .86. They compare it with another study that uses the same dataset and presents an RMSE = .90 (REINECKE et al., 2013). The model validation uses two different datasets, a generic set composed of web GUIs of various types and another specific set of six distinct search engine GUIs. The samples in both datasets were rated using the VisAWI questionnaire on a 7-point Likert scale (MOSHAGEN; THIELSCH, 2010), yet without further information on how the scores were converted to the model scale. The results show an RMSE = 1.28 for the generic dataset and an RMSE = 1.18 for the specific dataset indicating a loss of precision when assessing previously unseen GUIs.

Pappas et al. (PAPPAS et al., 2020) use the normalized RMSE (NRMSE) to test the adopted Random Forest regression algorithm. They establish 25 seconds of eye-tracking data as the baseline for the prediction of participants' assessments and compare the NRMSE for the baseline with the NRMSE of three other pipelines (10s, 15s, and 20s). For each aesthetic dimension assessed with the VisAWI questionnaire (MOSHAGEN; THIELSCH, 2010), they pick the pipeline that is not significantly different from the baseline for each aesthetic facet. That way, they verify a NRMSE = .11 for simplicity and a NRMSE = .12 for colorfulness in 15s, and a NRMSE = .10 for diversity and a NRMSE = .13 for craftsmanship in 20s. These values indicate that users need to observe a GUI between 15 and 20 seconds before forming their visual aesthetics impressions.

Wu et al. (2011) apply two models to classify the visual aesthetics of web GUIs into "high" or "low" and to assign numerical scores. The classification task employs a cost-sensitive SVM, which uses the misclassified cost instead of the misclassified error rate to evaluate the model, and considers the average cost of each rated GUI as its standard cost. The model is evaluated by the average misclassified cost (AMC), which is the average of all average costs, assuming that the lower the AMC the better. They compute the AMC for each feature separately, for all features combined, and for feature subsets of different sizes. A subset of six extracted features classified the GUIs with an AMC = .18, indicating that the combination of these features performs better than each feature alone or combining all features. The assignment of visual aesthetics scores is done by adopting a support vector regression model. Here, the predicted values are the same features used for classification, and the lowest RSSE = .54 is achieved with the same feature subset, confirming their close relation with visual aesthetics.

A typical way to evaluate the performance of deep learning models is by measuring how well they predict the output based on the input. Depending on the kind of model, accuracy, precision, recall, F1 score for classification, and mean squared error and mean absolute error for regression are popular metrics for its evaluation (SOKOLOVA; LAPALME, 2009). Khani et al. (2016) developed a neural network to classify the visual aesthetics of GUIs into "good" or "bad." Their system can assign the label of GUIs right with a test error of 34.15%. Dou et al. (2019) present a CNN that handles the GUI visual aesthetics assessment as a regression problem instead of dealing with the task using a classification approach. They evaluate their model performance in the same way as Khani et al. and achieve a lower teste error rate of 20.41% error, obtaining better results.

3.1.4 Discussion

Despite the recognition of the importance of visual aesthetics for the quality of software systems, only a relatively small number of research investigating how to assess visual aesthetics has been encountered.

Considering the GUI types assessed in the studies, it seems reasonable that most of them are web GUIs due to the ever-growing importance of online services. What is surprising, though, is the small amount of research addressing mobile GUIs, considering that in 2023 there were more than 17 billion worldwide (STATISTA, 2023). The current relevance of mobile apps and the lack of research on their visual aesthetics assessment indicate the need for further research that considers mobile GUI-specific characteristics.

The variety of different techniques used for the assessment task and the lack of consensus on which features best represent a GUI beauty also indicate that the aesthetics concept is not easy to grasp. Most studies aim to assess visual aesthetics objectively, while a few types of research propose a subjective approach for measuring the perceived visual aesthetics via questionnaires or sensors. And although questionnaires may be a reliable tool for GUI visual aesthetics assessments, they might not be suitable when the same user needs to assess multiple GUIs. Furthermore, the manual collection of user perception is time-consuming and represents a considerable effort. Sensor-based techniques seem to be among the least adopted, despite representing novel possibilities to understand how users react to the experience of seeing GUIs.

Considering objective approaches, we found a quite balanced number of studies using element- or image-based techniques (10 and 12). Although both analyze what users see, the first measures mostly layout properties of the GUIs, like symmetry, balance, and unity, while image-based techniques investigate more low-level features, like color and its variations, texture, and contrast. Element-based approaches use a large variety of metrics. And, although half of the encountered studies used the metrics proposed by Ngo (NGO; SAMSUDIN; ABDULLAH, 2000; NGO, 2001; NGO; TEO; BYRNE, 2002, 2003), none of them used the complete set of metrics or the same subset as another study. The number of elements, balance, and unity are the most common ones, present in 40% of the papers. Besides, the number of metrics measured in each study varies from as little as one (WANG, C.; REN, 2018) to fourteen (PURCHASE et al., 2011). That variety regarding the quantity and factors reflects a lack of consensus on assessing visual aesthetics based on GUI properties.

A similar variety of factors can be observed for image-based approaches, revealing a lack of consensus on which features best reflect the GUI visual aesthetics. The color was by far the most frequent feature associated with visual aesthetics, being analyzed by various aspects, like the number of dominant colors, color range, or HSV color space. Layout and texture features are also commonly observed, but each study extracted a distinct set of features. An exception is Wu et al. (2011; 2016), which extracted Tamura's texture features in both studies. In contrast, deep learning approaches directly learn patterns to assess the GUI visual aesthetics, not requiring manual feature engineering. Therefore, it may even recognize unnoticed patterns, not limited to already known features. But despite the current trend of deep learning and its promising results for the visual aesthetics assessments, we found only two recently published studies aiming specifically at assessing GUIs, indicating that such approaches seem to be only emerging. Yet, we have encountered no deep learning approaches for mobile applications so far.

Regarding the reliability and validity of the approaches proposed, we observed that most of the studies present some evaluation. Most adopt a case study or quasi-experimental approach, comparing their results with similar previously conducted research. One issue that complicates these evaluations is the preparation of a dataset with a balanced set of visually attractive and unattractive GUIs and their reliable labeling by human raters, often requiring several ratings for each GUI to deal with inter-rater agreement and reliability issues. While the number of ratings for each GUI varied widely from 5 to 1,800, different research scopes justify that variation. The studies that collected fewer ratings for each GUI had specific purposes (websites of London-based civil-engineering companies), whereas the study that collected 1,800 ratings per GUI intended to understand the assessment variation under demographic variables within a much larger scope context. We also observed that some research types overcame the difficulty of preparing a proper dataset using datasets made available by other researchers. The availability of a well-founded and reliably labeled public dataset, such as the one Reinecke and Gajos (2014) prepared, is a significant contribution to research in this area. Yet, as that dataset was created in 2014, the GUI designs may be outdated and, thus, no longer valid to represent today's interfaces. Extending the set with more recent GUIs would therefore be essential. Some studies used highly reduced sample sets, in some cases considering only four GUIs, and, therefore, the results may not be generalizable.

Although some of the studies ran very rigorous evaluations, some did not evaluate the proposed approaches at all. Thus, in general, it remains questionable whether these approaches allow the reliable and valid assessment of the visual aesthetics of GUI.

These results indicate the need for further research to support the assessment of visual aesthetics of graphical user interfaces in a more effective way that can be easily adopted in practice, and the need for more robust assessment models focusing on more current types of interfaces such as mobile applications.

3.2 ASSESSMENT OF VISUAL AESTHETICS OF GUIS BASED ON HUMAN JUDGMENTS

3.2.1 Definition of the review protocol

Research question. This research aims to elicit how the visual aesthetics of GUIs is subjectively assessed based on human ratings. To achieve that goal, we focused on the following analysis questions:

- AQ1: What visual aesthetic factors are assessed through human ratings?
- AQ2: How are visual aesthetics assessments designed?
- AQ3: What are the characteristics of the raters?

- AQ4: What data collection instruments are used?
- AQ5: How are the assessments evaluated in terms of reliability and validity?

Data source. We search on Scopus, the largest abstract and citation database of peer-reviewed literature, including peer-reviewed publications in English from the principal scientific portals in computing such as ACM, Elsevier, IEEE, and Springer, considering articles with free access through the Portal CAPES.

Inclusion/exclusion criteria. We included only articles that presented subjective assessments of the visual aesthetics of GUIs. Aiming at a comprehensive overview, we also included articles that describe subjective assessments, even when they are part of a usability evaluation and not specifically proposed as new assessment approaches. We excluded articles not focusing on the assessment of visual aesthetics or not related specifically to graphical user interfaces. We also included only articles considering the entire GUI, excluding articles that focus only on individual GUI elements such as icons, text, or imagery.

Quality criteria. We considered only articles with considerable information to enable the extraction of relevant data regarding the analysis questions. Therefore, we excluded abstract-only or one-page articles.

Definition of the search string. Pursuing the research objective, we defined the search string to identify articles dealing with core concepts and their synonyms, also based on several informal searches for the calibration of the search string:

```
TITLE-ABS-KEY (((aesthetics OR beaut* OR "visual appeal" OR "visual
quality") AND (assess* OR evaluat* OR measure* OR rating) AND
(interface OR ui OR gui OR web* OR mobile OR app OR application OR
android OR ios) AND (human OR people OR person OR volunteer OR
participant OR subject OR user))) AND (EXCLUDE (DOCTYPE, "ch") OR
EXCLUDE (DOCTYPE, "bk") OR EXCLUDE (DOCTYPE, "er" ) OR EXCLUDE
(DOCTYPE, "le") OR EXCLUDE (DOCTYPE, "ed")) AND (LIMIT-TO (LANGUAGE,
"English"))
```

We applied Scopus filters to exclude works from unrelated fields: Medicine, Dentistry, Biochemistry, Genetics, and Molecular Biology, Agricultural, Biological Sciences, Health Professions, Pharmacology, Toxicology, and Pharmaceutics, Earth and Planetary Sciences, Nursing, Economics, Econometrics, and Finance, and Immunology and Microbiology.

3.2.2 Execution of the search

The first author executed the search in March 2021, which was reviewed by the second author. The initial search returned 1,456 articles. We quickly reviewed titles, abstracts, and keywords of all search results to identify those articles matching the exclusion criteria during the first analysis. After the removal of irrelevant articles, we identified 230 potentially relevant articles. In the next step, we analyzed the full texts and excluded irrelevant ones following the inclusion/exclusion and quality criteria. As a result, we selected 121 articles that present some kind of human assessment of the visual aesthetics of a GUI. The selection process has been performed by both authors together, discussing the selection until a consensus was reached (Table 18).

Table 18 – Selected articles.

(To be continued)

Title	Reference
Modeling and Evaluating User Interface Aesthetics Employing ISO 25010 Quality Standard	(ABBASI et al., 2012)
The Impact of Web Page Usability Guideline Implementation on Aesthetics and Perceptions of the E-Retailer	(AGARWAL; HEDGE, 2008)
Investigating a Multi-faceted View of User Experience	(AL-SHAMAILEH; SUTCLIFFE, 2012a)
The Effect of Website Interactivity and Repeated Exposure on User Experience	(AL-SHAMAILEH; SUTCLIFFE, 2012b)
Experimental Investigation of Effects of Balance, Unity, and Sequence on Interface and Screen Design Aesthetics	(ALTABOLI; LIN, Y., 2010)
Effects of Unity of Form and Symmetry on Visual Aesthetics of Website Interface Design	(ALTABOLI; LIN, Y., 2012)
Quality Attributes Analysis in a Crowdsourcing-based Emergency Management System	(AMORIM et al., 2017)
I Don't Have That Much Data! Reusing User Behavior Models for Websites from Different Domains	(BAKAEV et al., 2020)
Computational Modeling and Experimental Investigation of Effects of Compositional Elements on Interface and Design Aesthetics	(BAUERLY; LIU, Yili, 2006)
Effects of Symmetry and Number of Compositional Elements on Interface and Design Aesthetics	(BAUERLY; LIU, Yili, 2008)
Evaluation and Improvement of Interface Aesthetics with an Interactive Genetic Algorithm	(BAUERLY; LIU, Yili, 2009)
Keep it Simple: How Visual Complexity and Preferences Impact Search Efficiency on Websites	(BAUGHAN et al., 2020)
Understanding Visual Appeal and Quality Perceptions of Mobile Apps: An Emotional Perspective	(BHANDARI; NEBEN; CHANG, K., 2015)
Effects of Interface Design Factors on Affective Responses and Quality Evaluations in Mobile Applications	(BHANDARI et al., 2017)

(Continuation of Table 18)

Title	Reference
Understanding the Impact of Perceived Visual Aesthetics on User Evaluations: An Emotional Perspective	(BHANDARI; CHANG, K.; NEBEN, 2019)
Electrophysiological Correlates of Aesthetic Processing of Webpages: A Comparison of Experts and Laypersons	(BÖLTE et al., 2017)
Entropy and Compression Based Analysis of Web User Interfaces	(BOYCHUK; BAKAEV, 2019)
A User-centered Evaluation and Redesign Approach for E-Government APP	(CHANG, D.; LI, F.; HUANG, L., 2020)
The Influence of the Search Complexity and the Familiarity with the Website on the Subjective Appraisal of Aesthetics, Mental Effort and Usability	(CHEVALIER; MAURY; FOUQUEREAU, 2014)
I2Evaluator: An Aesthetic Metric-Tool for Evaluating the Usability of Adaptive User Interfaces	(CHETTAOUI; BOUHLEL, 2017)
The Effects of Aesthetics and Cognitive, Style on Perceived Usability	(CONKLIN et al., 2006)
Searching vs. Browsing—The Influence of Consumers' Goal Directedness on Website Evaluations	(DAMES et al., 2019)
Interaction, Usability and Aesthetics: What Influences Users' Preferences?	(DE ANGELI; SUTCLIFFE; HARTMANN, 2006)
Efficiency, Trust, and Visual Appeal: Usability Testing through Eye Tracking	(DJAMASBI et al., 2010)
Effects of Different Website Designs on First Impressions, Aesthetic Judgements and Memory Performance after Short Presentation	(DOUNEVA; JARON; THIELSCH, 2016)
Considering Aesthetics and Usability Temporalities in a Model Based Development Process	(DUPUY-CHESSA; LAURILLAU; CÉRET, 2016)
Online Viewing and Aesthetic Preferences of Generation Y and the Baby Boom Generation: Testing User Web Site Experience Through Eye Tracking	(DJAMASBI et al., 2011)
Usability and Aesthetics: The Case of Architectural Websites	(FALIAGKA et al., 2015)
Youth Matters: Philly (YMP: Development, usability, usefulness, & accessibility of a mobile web-based app for homeless and unstably housed youth	(GREESON et al., 2020)
Predicting Webpage Aesthetics with Heatmap Entropy	(GU et al., 2020)
The Impact of Web Page Text-Background Colour Combinations on Readability, Retention, Aesthetics and Behavioural Intention	(HALL; HANNA, 2004)

(Continuation of Table 18)

Title	Reference
Assessing the Attractiveness of Interactive Systems	(HARTMANN, 2006)
Investigating Attractiveness in Web User Interfaces	(HARTMANN; SUTCLIFFE; DE ANGELI, 2007)
Towards a Theory of User Judgment of Aesthetics and User Interface Quality	(HARTMANN; SUTCLIFFE; DE ANGELI, 2008)
The Interplay of Beauty, Goodness, and Usability in Interactive Products	(HASSENZAHL, 2004)
The Inference of Perceived Usability from Beauty	(HASSENZAHL; MONK, 2010)
A Study of Affective Meanings Predicting Aesthetic Preferences of Interactive Skins	(HUANG, S.-M., 2013)
LEMtool: Measuring Emotions in Visual Interfaces	(HUISMAN et al., 2013)
Aesthetics in Context—The Role of Aesthetics and Usage Mode for a Website's Success	(ITEN; TROENDLE; OPWIS, 2018)
The Effects of Perceived Visual Aesthetics on Process Satisfaction in GSS Use	(IVANOV; SCHNEIDER, 2010)
Perceived Website Aesthetics by Users and Designers: Implications for Evaluation Practice	(KOUTSABASIS; ISTIKOPOULOU, 2013)
Are You Willing to Donate?: Relationship Between Perceived Website Design, Trust and Donation Decisions Online	(KÜCHLER; HERTEL; THIELSCH, 2020)
Underlying Quality Factors in Spanish Language Apps for People with Disabilities	(LARCO et al., 2018)
Objective Design to Subjective Evaluations: Connecting Visual Complexity to Aesthetic and Usability Assessments of eHealth	(LAZARD; KING, 2020)
What is This Evasive Beast We Call User Satisfaction?	(LINDGAARD; DUDEK, 2003)
Attention Web Designers: You Have 50 Milliseconds to Make a Good First Impression!	(LINDGAARD et al., 2006)
Judging Web Page Visual Appeal: Do East and West Really Differ?	(LINDGAARD; LITWINSKA; DUDEK, 2008)
Evaluating Localized MOOCs: The Role of Culture on Interface Design and User Experience	(LIU, S. et al., 2020)
A Model to Compute Webpage Aesthetics Quality Based on Wireframe Geometry	(MAITY; BHATTACHARYA, 2017)
Relating Aesthetics of the GUI Text Elements with Readability using Font Family	(MAITY; BHATTACHARYA, 2018)

(Continuation of Table 18)

Title	Reference
Is My Interface Beautiful?—A Computational Model-Based Approach	(MAITY; BHATTACHARYA, 2019)
Affective Graphs: The Visual Appeal of Linked Data	(MAZUMDAR et al., 2015)
Computing Aesthetics of Concrete User Interfaces	(MBENZA; BURNY, 2020)
An Investigation of Visual Appeal and Trust in Websites	(MCDONNELL; LEE, A., 2016)
The Effect of Interaction on Visual Appeal and Trust in Online Health Information	(MCDONNELL; O'REILLY, 2017)
Quantification of Interface Visual Complexity	(MINIUKOVICH; DE ANGELI, 2014a)
Visual Impressions of Mobile App Interfaces	(MINIUKOVICH; DE ANGELI, 2014b)
Computation of Interface Aesthetics	(MINIUKOVICH; DE ANGELI, 2015a)
Visual Diversity and User Interface Quality	(MINIUKOVICH; DE ANGELI, 2015b)
Relationship Between Visual Complexity and Aesthetics of Webpages	(MINIUKOVICH; MARCHESE, 2020)
Comparison of Three Digital Library Interfaces: Open Library, Google Books, and Hathi Trust	(MILLER; CHOI, G.; CHELL, 2012)
Facets of Visual Aesthetics	(MOSHAGEN; THIELSCH, 2010)
A Short Version of the Visual Aesthetics of Websites Inventory	(MOSHAGEN; THIELSCH, 2013)
Gender Differences in Website Production and Preference Aesthetics: Preliminary Implications for ICT in Education and Beyond	(MOSS; GUNN, 2009)
Aesthetic Measures for Screen Design	(NGO; BYRNE, 1998)
Aesthetic Measures for Assessing Graphic Screens	(NGO; SAMSUDIN; ABDULLAH, 2000)
Formalising Guidelines for the Design of Screen Layouts	(NGO; TEO; BYRNE, 2000)
Measuring the Aesthetic Elements of Screen Designs	(NGO, 2001)
A Review of Mindfulness-Based Apps for Children	(NUNES; CASTRO; LIMPO, 2020)
Visual Clarity as Mediator Between Usability and Aesthetics	(OTTEN; SCHREPP; THOMASCHEWSKI, 2020)
Homepage Aesthetics: The Search for Preference Factors and the Challenges of Subjectivity	(PANDIR; KNIGHT, 2006)

(Continuation of Table 18)

Title	Reference
The Subjective and Objective Nature of Website Aesthetic Impressions	(PAPACHRISTOS; AVOURIS, 2009)
Are First Impressions about Websites Only Related to Visual Appeal?	(PAPACHRISTOS; AVOURIS, 2011)
The Influence of Website Category on Aesthetic Preferences	(PAPACHRISTOS; AVOURIS, 2013)
A Comparison of Gaze Behavior of Experts and Novices to Explain Website Visual Appeal	(PAPPAS et al., 2018)
How Quickly Can We Predict Users' Ratings on Aesthetic Evaluations of Websites? Employing Machine Learning on Eye-Tracking Data	(PAPPAS et al., 2020)
The Influence of Aesthetic and Usability Web Design Elements on Viewing Patterns and User Response: An Eye-tracking Study	(PAVLAS; LUM; SALAS, 2010)
An Experimental Investigation of the Influence of Website Emotional Design Features on Trust in Unfamiliar Online Vendors	(PENGNATE; SARATHY, 2017)
The Influence of the Centrality of Visual Website Aesthetics on Online User Responses: Measure Development and Empirical Investigation	(PENGNATE; SARATHY; ARNOLD, 2021)
The Engagement of Website Initial Aesthetic Impressions: An Experimental Investigation	(PENGNATE; SARATHY; LEE, J., 2019)
From the Ground-Up: Role of Usability and Aesthetics Evaluation in Creating a Knowledge-Based Website for the U.S. Army Corps of Engineers	(PROPST et al., 2013)
Investigating Objective Measures of Web Page Aesthetics and Usability	(PURCHASE et al., 2011)
Analyzing the Effects of Visual Aesthetic of Web Pages on Users' Responses in Online Retailing Using the VisAWI Method	(RAMEZANI NIA; SHOKOUHYAR, 2020)
Improving Performance, Perceived Usability, and Aesthetics with Culturally Adaptive User Interfaces	(REINECKE; BERNSTEIN, A., 2011)
Predicting Users' First Impressions of Website Aesthetics With a Quantification of Perceived Visual Complexity and Colorfulness	(REINECKE et al., 2013)
Quantifying Visual Preferences Around the World	(REINECKE; GAJOS, 2014)
The Effect of Aesthetically Pleasing Composition on Visual Search Performance	(SALIMUN et al., 2010b)

(Continuation of Table 18)

Title	Reference
Preference Ranking of Screen Layout Principles	(SALIMUN et al., 2010a)
Implementing Recommendations From Web Accessibility Guidelines: Would They Also Provide Benefits to Nondisabled Users	(SCHMUTZ; SONDEREGGER; SAUER, 2017)
The Influence of Hedonic Quality on the Attractiveness of User Interfaces of Business Management Software	(SCHREPP; HELD; LAUGWITZ, 2006)
Stay Present with Your Phone: A Systematic Review and Standardized Rating of Mindfulness Apps in European App Stores	(SCHULTCHEN et al., 2021)
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	(SECKLER; OPWIS; TUCH, 2015)
Users' Emotional Valence, Arousal, and Engagement Based on Perceived Usability and Aesthetics for Web Sites	(SEO et al., 2015)
A GA-based Approach to Improve Web Page Aesthetics	(SINGH; BHATTACHARYA, 2011)
The Negative Impact of Saturation on Website Trustworthiness and Appeal: A Temporal Model of Aesthetic Website Perception	(SKULMOWSKI et al., 2016)
Expressive and Classical Aesthetics: Two Distinct Concepts with Highly Similar Effect Patterns in User–Artefact Interaction	(SONDEREGGER; SAUER; EICHENBERGER, 2014)
Aesthetics on the Web: Effects on Approach and Avoidance Behaviour	(STREBE, 2016)
Assessing Interaction Styles in Web User Interfaces	(SUTCLIFFE; DE ANGELI, 2005)
Getting the Message Across: Visual Attention, Aesthetic Design and What Users Remember	(SUTCLIFFE; NAMOUNE, 2008)
Neuroanatomical Correlates of Perceived Usability	(THANH VI; HORNBÆK; SUBRAMANIAN, 2017)
User Evaluation of Websites: From First Impression to Recommendation	(THIELSCH; BLOTENBERG; JARON, 2014)
Expected Usability Is Not a Valid Indicator of Experienced Usability	(THIELSCH; ENGEL; HIRSCHFELD, 2015)
Evaluating the Consistency of Immediate Aesthetic Perceptions of Web Pages	(TRACTINSKY et al., 2006)

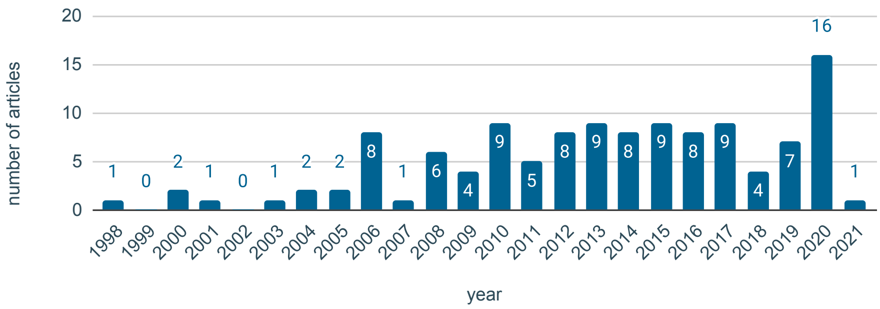
(Continuation of Table 18)

Title	Reference
The Role of Visual Complexity and Prototypicality Regarding First Impression of Websites: Working Towards Understanding Aesthetic Judgments	(TUCH et al., 2012a)
Is Beautiful Really Usable? Toward Understanding the Relation Between Usability, Aesthetics, and Affect in HCI	(TUCH et al., 2012b)
User's Web Page Aesthetics Opinion: A Matter of Low-Level Image Descriptors Based on MPEG-7	(URIBE; ÁLVAREZ; MENÉNDEZ, 2017)
Five Psychometric Scales for Online Measurement of the Quality of Human-Computer Interaction in Web Sites	(SCHAIK; LING, 2005)
Modelling User Experience with Web Sites: Usability, Hedonic Value, Beauty and Goodness	(SCHAIK; LING, 2008)
The Role of Context in Perceptions of the Aesthetics of Web Pages over Time	(SCHAIK; LING, 2009)
User-Experience from an Inference Perspective	(SCHAIK; HASSENZAHL; LING, 2012)
Towards an Understanding of Visual Appeal in Website Design	(VARELA, Martín et al., 2013)
QoE in the Web: A Dance of Design and Performance	(VARELA, Martin et al., 2015)
The Impact of Symmetric Web-Design: A Pilot Study	(VASSEUR; LÉGER; SÉNÉCAL, 2020)
Aesthetics in Hypermedia: Impact of Colour Harmony on Implicit Memory and User Experience	(VENNI; BÉTRANCOURT, 2020)
Temporal Evaluation of Aesthetics of User Interfaces as one Component of User Experience	(VOGEL, 2013)
Evaluating the Visual Quality of Web Pages Using a Computational Aesthetic Approach	(WU, O. et al., 2011)
Multimodal Web Aesthetics Assessment Based on Structural SVM and Multitask Fusion Learning	(WU, O. et al., 2016)
The Visual Aesthetics Measurement on Interface Design Education	(WU, C. M.; LI, P., 2019)
Towards an Evaluation of Graphical User Interfaces Aesthetics based on Metrics	(ZEN; VANDERDONCKT, 2014)
Assessing User Interface Aesthetics Based on the Intersubjectivity of Judgment	(ZEN; VANDERDONCKT, 2016)

Source: the author.

The first relevant article within the scope of our study was published by Ngo and Byrne in 1998 (Figure 12). Then, from 2006 on, the number of publications increased, maintaining a steady number of about 7-9 publications per year. Yet, in 2020 there was a peak, almost doubling the number of publications in the year before. At the beginning of 2021 (until March 2021), we found one publication so far.

Figure 12 – Distribution of relevant articles per year.



Source: the author.

3.2.3 Data analysis

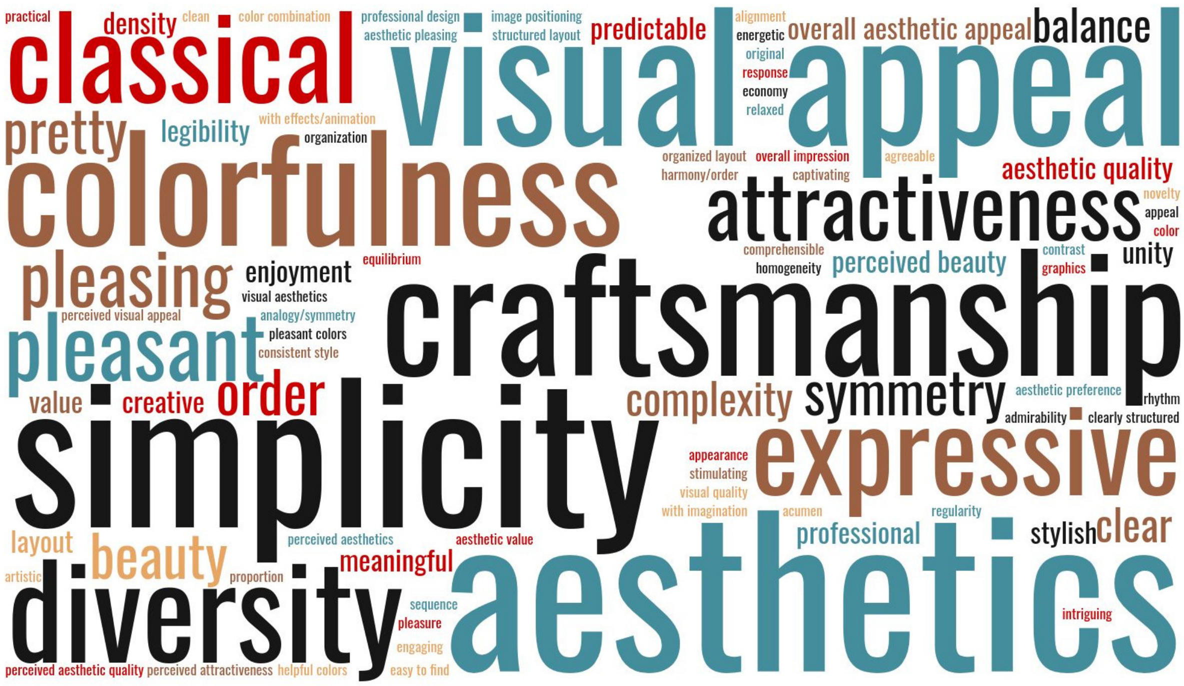
We present the results of the analysis of the 121 relevant articles. As some of them introduce more than one study, we extracted the information for each of the studies separately, analyzing a total of 148 studies. Some articles do not focus on the visual aesthetics assessment as their primary goal, but as part of a broader evaluation, like user satisfaction (LINDGAARD; DUDEK, 2003), user preference (DE ANGELI; SUTCLIFFE; HARTMANN, 2006), or usability and aesthetics at the same time (CHANG, D.; LI, F.; HUANG, L., 2020). From those articles, we only extracted data related to the assessment of visual aesthetics. For all articles that met the inclusion/exclusion and quality criteria, we extracted data that characterized the visual aesthetics assessments to answer the analysis questions. Several papers, however, lack explicit information or details on specific characteristics about their assessments, such as the number of ratings for each GUI received or how they compute a final score. In those cases, we could infer some information based on their reported information or cited references. When that was not possible, we indicated the lack of data as "NI" (not informed).

3.2.3.1 What visual aesthetic factors are assessed through human ratings?

We observed that there does not exist a consensus regarding the construct of visual aesthetics. Just over half of the studies (53.9%) consider it a multidimensional construct, decomposing visual aesthetics into several sub-dimensions. Twenty-two of these studies divide the construct into the simplicity/diversity/colorfulness/craftsmanship dimensions (MOSHAGEN; THIELSCH, 2013, 2010). Other studies consider visual aesthetics a composition of the classical/expressive dimensions (LAVIE; TRACTINSKY, 2004)

(N=22; 15.9%), while two of these focus on classical aesthetics only (MILLER; CHOI, G.; CHELL, 2012; SCHMUTZ; SONDEREGGER; SAUER, 2017) and two others on the expressive dimension (AL-SHAMAILEH; SUTCLIFFE, 2012b, 2012a). In some studies, respondents rate the subjective dimensions of visual aesthetics like the hedonic qualities pleasant, pretty, appealing, inviting, good, pleasing, and motivating (HASSENZAHL; BURMESTER; KOLLER, 2003). But others assess design elements such as color (AGARWAL; HEDGE, 2008), symmetry (KOUTSABASIS; ISTIKOPOULOU, 2013), simplicity (SCHAIK; LING, 2005), or balance (PAPACHRISTOS; AVOURIS, 2009).

Figure 13 – Assessed factors.



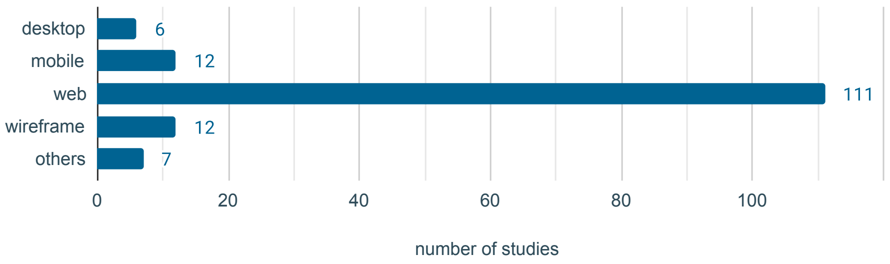
Source: the author.

On the other hand, a considerable number understand visual aesthetics as a unidimensional construct. In these studies, participants rate, e.g., how “beautiful” (ALTABOLI; LIN, Y., 2010), “attractive” (SCHAIK; LING, 2009), or “appealing” (BAUGHAN et al., 2020) they perceive the GUIs. The majority of these studies assess visual aesthetics on single-item scales, with few using multiple-item questionnaires, with three (PAPACHRISTOS; AVOURIS, 2009), seven (THIELSCH; BLOTENBERG; JARON, 2014), or ten items (CONKLIN et al., 2006), for example.

3.2.3.2 How are visual aesthetics assessments designed?

Focusing on GUIs, we observed that the vast majority of the studies assessed web GUIs (N=102; 73.9%) (Figure 14). Most of these were either screenshots or live pages from existing websites, while a few used mockups specifically designed for the study. A small number of the studies manipulated existing websites so that raters could assess the same content with different degrees of aesthetics. Application GUIs (mobile and desktop) accounted for 11.6% of the studies, whereas 8% of the studies use GUI wireframes to focus on specific layout elements such as symmetry and balance (BAUERLY; LIU, Yili, 2006). The other studies either do not explicitly inform the specific GUI type (MAITY; BHATTACHARYA, 2018) or select GUIs from diverse multimedia systems (NGO, 2001).

Figure 14 – Distribution of studies per assessed GUI type.



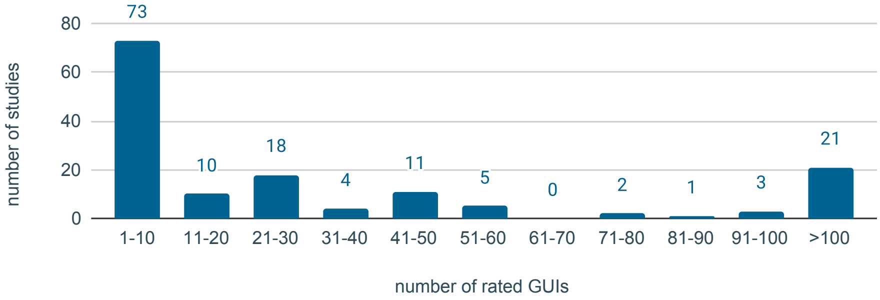
Source: the author.

An important characteristic of this type of study is the number of assessed GUIs (Figure 15). Some studies do not report the total number of GUIs as they assess the visual aesthetics of a specific website or mobile app as a whole and not individual screens of the computational system. In these cases, we considered each system as one GUI as long as it is rated as a unity. In general, the number of assessed GUIs varied substantially, but about half of the studies assess less than ten GUIs (N=66; 47.8%). Several studies rate only one GUI (N=7; 5.1%). On the other hand, there is a significant number of studies assessing more than one hundred GUIs (14.5%).

On the other end, there is a significant number of studies assessing more than one hundred GUIs (14.2%). Among these, three articles present studies with much larger datasets. Wu et al. (2016) selected 1,000 website screenshots to train a machine learning model to assess visual aesthetics. Miniukovich and Marchese (2020) assembled six previously used datasets totaling 1,506 GUIs to replicate past results and test additional hypotheses for the divergence on the relationship of aesthetics with complexity. Bakaev et al. (BAKAEV et al., 2020) captured 3,429 web GUI screenshots from different domains to train and test 21 artificial neural network models to predict subjective assessments for websites of other categories.

We also analyzed the number of GUIs rated by each participant (Figure 17 17). In about half of the studies (N=75; 54.3%), each subject assesses less than ten GUIs. Although most of these studies have respondents rating the whole set of GUIs, others choose to assign parts of the set to avoid fatigue (LARCO et al., 2018) or to ask the subjects to rate

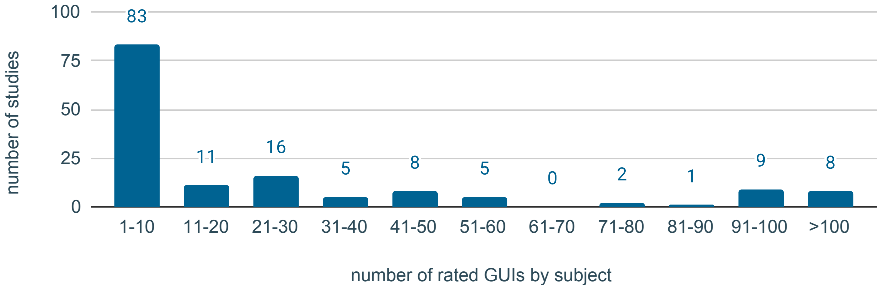
Figure 15 – Distribution of studies per total number of rated GUIs.



Source: the author.

only one of the different design conditions within the dataset (BHANDARI; CHANG, K.; NEBEN, 2019).

Figure 16 – Distribution of studies per number of rated GUIs by subject.

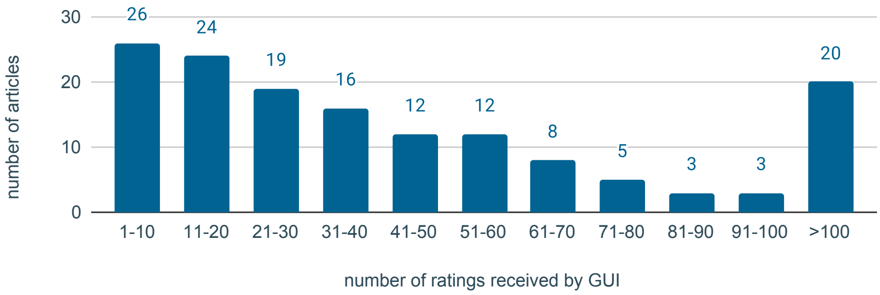


Source: the author.

In the majority of the studies (N=95; 68.8%), participants rated the entire set of GUIs. On the other hand, analyzing how many ratings each GUI received, we observed that this information is not reported explicitly in several articles. In those cases, we inferred this information from the total number of GUIs and the number of GUIs each participant rated. Although not always providing a precise number, it allows us to understand the magnitude of the rating sample used in this type of assessment (Figure 17). Some studies collect less than ten ratings (N=24; 17.4%), with some as little as only one (NGO; BYRNE, 1998) or two (BAUERLY; LIU, Yili, 2009) per GUI.

Very few articles explicitly report sampling details (Figure 18). Almost half of the studies (N=58; 42%) do not provide details about how subjects were selected, making it impossible to determine the sampling technique used. In general, studies choose their subjects by applying non-probability sampling techniques. More than a quarter of the studies adopt volunteering sampling in which they accept volunteers from crowdsourcing platforms or people that replied to advertisements at college or Facebook. Studies using

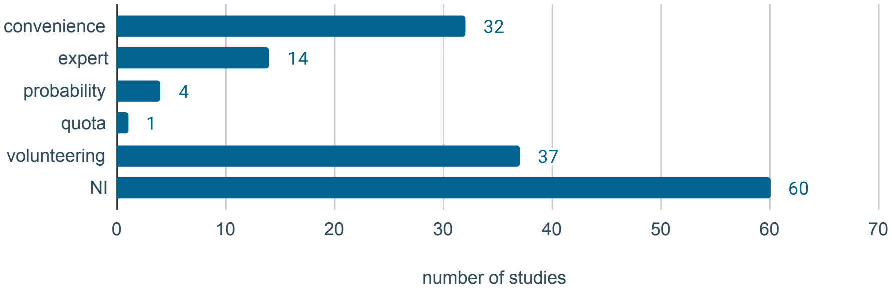
Figure 17 – Distribution of studies per number of ratings received by GUI.



Source: the author.

convenience sampling select subjects that are the students of one or more specific courses (KOUTSABASIS; ISTIKOPOULOU, 2013), members of a research group (WU, O. et al., 2011), or friends and family of members involved in the study (PURCHASE et al., 2011; SONDEREGGER; SAUER; EICHENBERGER, 2014). Several studies adopt expert sampling based on the subjects' knowledge (MCDONNELL; O'REILLY, 2017). One study selects respondents from 25 different nationalities, aiming at a sample group with varied cultural backgrounds (REINECKE; BERNSTEIN, A., 2011). Therefore, they use quota sampling, including up to four individuals of the same nationality in the study.

Figure 18 – Distribution of studies per sampling technique.



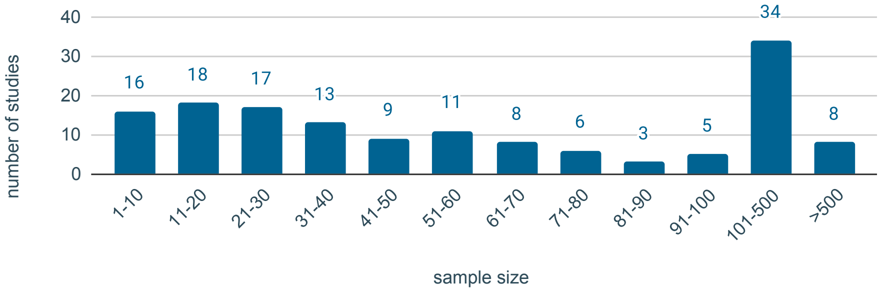
Source: the author.

Only three studies use probability sampling, selecting subjects in a random (HARTMANN; SUTCLIFFE; DE ANGELI, 2008; MOSS; GUNN, 2009) or semi-random fashion (LINDGAARD; DUDEK, 2003), without providing further information on the population or sampling frame from which the subjects come.

3.2.3.3 What are the characteristics of the raters?

Based on the reported information, we also analyze the considered characteristics of the raters in the studies. As some studies indicate removing invalid assessments from the original set (VARELA, Martín et al., 2013), our analysis is based on the valid responses as reported. Half of the studies (N=69; 50%) involve up to 50 subjects assessing GUIs (Figure 19). The number of studies gradually decreases when the groups of subjects get larger. Nonetheless, more than a quarter of the studies employ more than 100 subjects (N=40; 29%), of which the largest groups of raters involve 2,265 (KÜCHLER; HERTEL; THIELSCH, 2020) and 32,222 individuals (REINECKE; GAJOS, 2014) with valid responses. On the other end, one study involves only two expert participants assessing 192 GUIs (SCHULTCHEN et al., 2021).

Figure 19 – Distribution of studies per sample size.

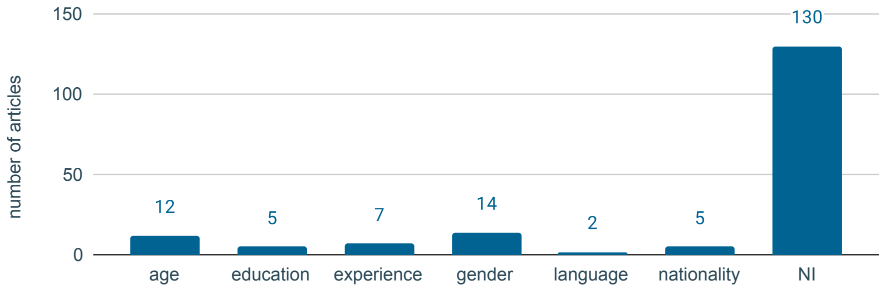


Source: the author.

Although every article reports some demographic information about their subjects, very few studies (N=18; 13%) consider the influence of demographic factors on the aesthetics assessment (Figure 20). Among those that do, they examine age, education, experience, language, nationality, and gender. In general, the studies that consider demographics analyze if these factors have any statistically significant influence on visual aesthetics. Three studies, however, are particularly interested in how demographic factors affect the perception of visual aesthetics. One study compares how groups with different cultural backgrounds, represented by Canadian and Taiwanese respondents, rate visual aesthetics (LINDGAARD; LITWINSKA; DUDEK, 2008). But, although Taiwanese subjects rate Taiwanese and Chinese GUIs higher than Canadians, no significant difference was observed when both groups rate North American GUIs, thus, not leading to a conclusive result about cultural differences. Another study investigates differences due to gender assessing 30 male-designed and 30 female-designed GUIs, concluding that subjects prefer GUIs typical of their gender (MOSS; GUNN, 2009). Djamasbi et al. (2011) analyze the aesthetic preferences of two different generations. They found that subjects from the baby boom generation, born from 1946 to 1964, and subjects from Generation Y, born from 1977 to 1990, rate GUI visual aesthetics similarly even though eye-tracking devices show they have different browsing behaviors.

One factor that can bias a visual aesthetics assessment is the gender distribution within the sample group (MOSS; GUNN, 2009). Nonetheless, almost one-third of the

Figure 20 – Distribution of studies per demographic data.



Source: the author.

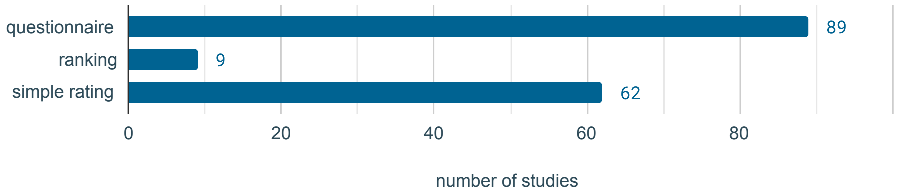
studies (N=41; 30%) do not report the proportion of females and males among their subjects. Among those reporting this information, half of the studies (N=69; 50%) have a reasonably balanced group of raters, with the proportion of female participants ranging between one-third and two-thirds. Nine studies report a perfectly balanced rater group, with the same number of female and male raters. Yet, about a quarter of these studies present less equitable rater groups, with two-thirds of their members are either females (N=19; 14%) or males (N=15; 11%). One study involves only male raters in their sample of ten subjects (SINGH; BHATTACHARYA, 2011).

In terms of education, more than half of the studies (N=77; 54%) involve students, including undergraduate students in 61 (44%) studies and graduate students in 19 (14%) studies. Other studies report their participants' degrees (N=15; 11%). Participants have at least a high school diploma in six studies (4%), a college degree in 11 (8%), and a postgraduate degree in two (1%). In 30 studies (22%), the education degree of the participants is not reported.

3.2.3.4 What data collection instruments are used?

We observed that diverse instruments are used to collect subjects' assessments of the visual aesthetics of GUIs (Figure 21). Most of them either apply questionnaires with multiple items (57.2%) or simple ratings (44.2%), in which subjects rate visual aesthetics on a single scale (BAKAEV et al., 2020). In most of the studies that apply simple ratings, subjects rate each GUI once (N=49; 35.5%), but in others (N=12; 8.7%), they are required to rate them twice to compare different exposure times (MINIUKOVICH; DE ANGELI, 2014b) or to evaluate consistency between ratings (REINECKE et al., 2013). Other studies use ranking scales of GUIs relative to their visual aesthetics (5.8%). In these studies, subjects either rank all GUIs at once (PURCHASE et al., 2011) or in groups of four (BAUERLY; LIU, Yili, 2006) or two (ZEN; VANDERDONCKT, 2016) in a balanced incomplete block design (BIBD).

Figure 21 – Distribution of studies per assessing instruments.



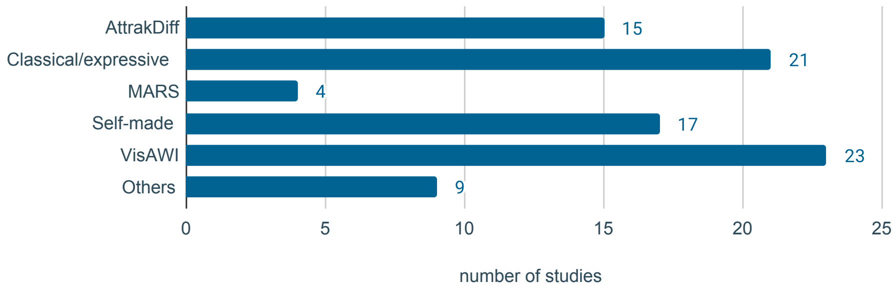
Source: the author.

Two questionnaires and their variations are adopted in many more studies than the others (Figure 22). The Classical/expressive questionnaire was originally designed to measure the classical and the expressive dimensions of visual aesthetics, each with five items (LAVIE; TRACTINSKY, 2004). Respondents indicate their degree of agreement on 7-point Likert scales to statements like “The website has a clean design” (classical aesthetics) and “The website has a creative design” (expressive aesthetics). Although most studies that used this questionnaire applied its full version with ten items (N=12; 8.7%) (CHEVALIER; MAURY; FOUQUEREAU, 2014), some adapted it either by reducing the number of items (TRACTINSKY et al., 2006) or by rating only the classical (MILLER; CHOI, G.; CHELL, 2012; SCHMUTZ; SONDEREGGER; SAUER, 2017) or the expressive dimension (AL-SHAMAILEH; SUTCLIFFE, 2012a, 2012b).

A considerable number of studies also apply the Visual Aesthetics of Website Inventory (VisAWI) questionnaire. Most of these studies use the full VisAWI (MOSHAGEN; THIELSCH, 2010), an 18-item questionnaire (N=12; 8.7%) in which respondents assess the simplicity, diversity, colorfulness, and craftsmanship dimensions of web GUIs (DOUNEVA; JARON; THIELSCH, 2016). Subjects use 7-point Likert scales to show their level of agreement to statements such as “Everything goes together on this site” (simplicity), “The layout is pleasantly varied” (diversity), “The color composition is attractive” (colorfulness), and “The layout appears professionally designed” (craftsmanship). Others choose the short version (VisAWI-S) (MOSHAGEN; THIELSCH, 2013) to assess the same dimensions but using four items instead (KÜCHLER; HERTEL; THIELSCH, 2020). Two studies, however, adapted the full version of the questionnaire by dropping items (RAMEZANI NIA; SHOKOUHYAR, 2020; URIBE; ÁLVAREZ; MENÉNDEZ, 2017).

The AttrakDiff is a 28-item questionnaire on 7-point semantic differential scales in which seven items are used to rate visual aesthetics (HASSENZAHL; BURMESTER; KOLLER, 2003). Although the six studies that applied the AttrakDiff are also interested in its full scope, including pragmatic and hedonic qualities of GUIs, we collected only data referring to visual aesthetics. The Mobile App Rating Scale (MARS) is an instrument originally developed for measuring the quality of mobile health apps, comprising visual aesthetics as one of its dimensions (STOYANOV et al., 2015). Four studies applied the

Figure 22 – Distribution of studies per questionnaire.

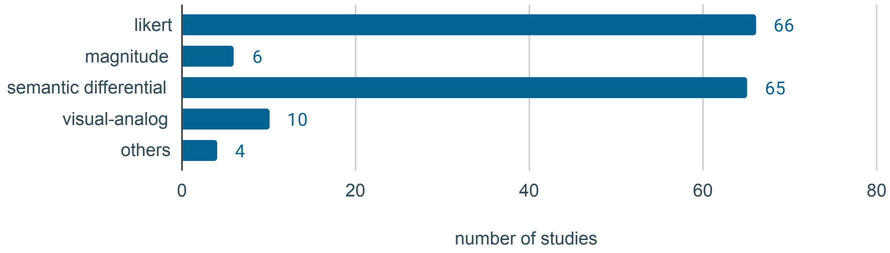


Source: the author.

MARS to evaluate apps for mindfulness promotion (NUNES; CASTRO; LIMPO, 2020; SCHULTCHEN et al., 2021) and accessibility (GREESON et al., 2020; LARCO et al., 2018).

The most used scales are the Likert (VENNI; BÉTRANCOURT, 2020), adopted in nearly half of the studies (47.1%), and the semantic differential scale (40.6%) (WU, C. M.; LI, P., 2019) (Figure 23). As we observed that semantic differential scales are, in some cases, inappropriately denoted as Likert, we adjusted the terminology as defined in section 2. One study used both scale types (TUCH et al., 2012b) on different instruments to rate visual aesthetics. The visual-analog scale, and unmarked line or slider, is used as an alternative to avoid the nonlinearity of Likert or semantic differential scales (LINDGAARD et al., 2006). Studies that apply the magnitude estimation method use its corresponding scale to rate GUIs compared to a benchmark (BAUERLY; LIU, Yili, 2006, 2008, 2009). Other applied alternatives include the interval scale [0..1] (NGO, 2001), the binary scale (GU et al., 2020), or the MOS scale (VARELA, Martín et al., 2013).

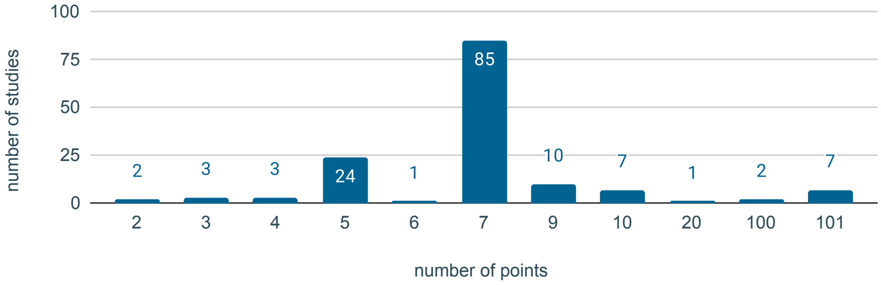
Figure 23 – Distribution of studies per rating scale.



Source: the author.

When analyzing the rating scales used, we can also observe that they differ in the number of points (Figure 24), including one study that applies binary scales (GU et al., 2020). The majority, however, uses ordinal scales with 3 to 101 points, with more than half of the studies applying 7-point scales (54.3%), followed by scales with five or nine points. Although interval scales, such as the visual-analog or the interval [0..1], can be virtually divided into infinite points, we could infer from the articles that it generates 101 points (from 0 to 100) (PAPACHRISTOS; AVOURIS, 2011), or 100 points when no midpoint is available (SCHAIK; LING, 2009). An exception is proposed by Tuch et al. (2012a), who map visual-analog responses to nine points.

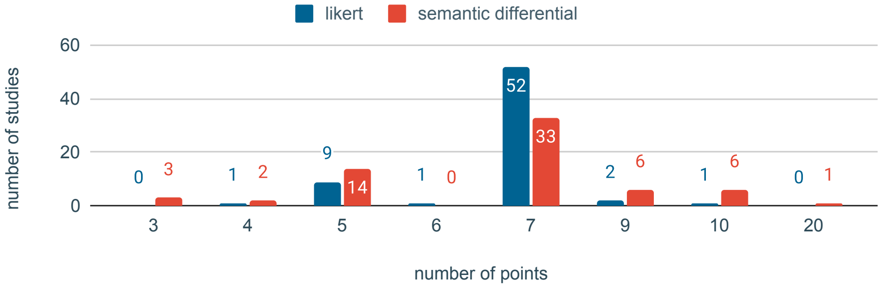
Figure 24 – Distribution of studies per number or points.



Source: the author.

When isolating the scales that are not number dependent (Figure 25), we can observe mostly Likert scales with seven or five points were used. Scales with an unusual number of points, like three, ten, or twenty, were more likely used in semantic differential scales. Another characteristic differentiating rating scales is the inclusion or not of a midpoint resulting in rating scales with an odd or even number of points. We observed that the vast majority of the studies (N=115; 83.3%) use odd-number-point scales, while 17 studies applied scales with an even number of points forcing respondents to choose either a more positive or more negative perception of the visual aesthetics of GUIs.

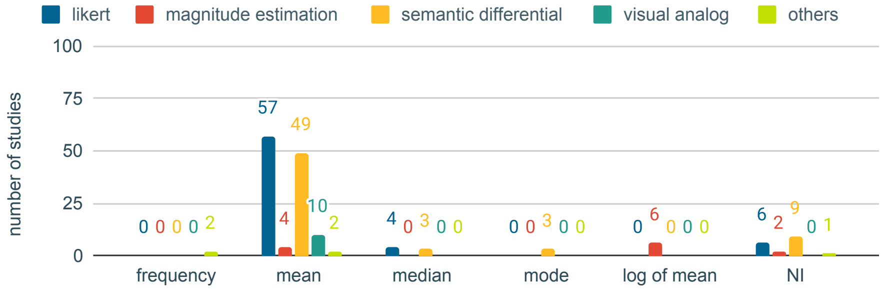
Figure 25 – Distribution of studies per number or points (Likert and semantic differential scales).



Source: the author.

To compile the responses of different raters into a single score, the majority of the studies (N=115; 83.3%) computes such a score on the visual aesthetics of each GUI as the mean score of all ratings received (Figure 26). This is the case for many studies that apply Likert or semantic differential scales, as these are generally considered to be ordinal scales for which the “distance” between each successive category may not be equivalent. However, ten studies (7.2%) compute the median score or the mode of all ratings to obtain the central tendency of visual aesthetics with ordinal scales. The studies that use the magnitude estimation scale calculate the log of the geometric mean of subject rating, as magnitude estimation data are log-normally distributed (BAUERLY; LIU, Yili, 2006, 2008, 2009). The study that applies binary scales computes the frequency of positive ratings (’beautiful’) (GU et al., 2020). The computation of a single score to indicate visual aesthetics was not reported in 16 studies (10.9%).

Figure 26 – Distribution of studies per single score computation.



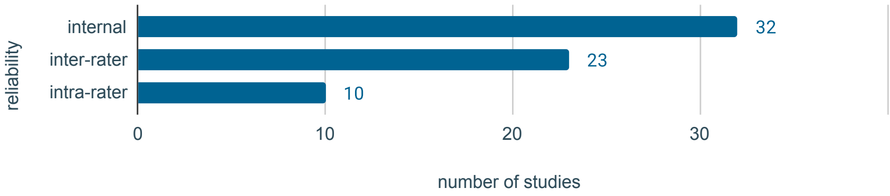
Source: the author.

3.2.3.5 How are the assessments evaluated in terms of reliability and validity?

Considering the importance of evaluating the assessment quality, we noted that more than half of the studies (N=78; 56.5%) do not present a reliability or validity evaluation. Among those that analyze the assessment quality, almost all studies evaluate reliability (91.7%), whereas only about a quarter examine validity (23.3%).

Regarding reliability (Figure27), 32 studies measure the internal consistency of their instruments, which is less than half of all studies in which subjects rate visual aesthetics with more than one item (43.8%). Considering the importance of evaluating reliability, especially in the case of newly designed questionnaires not evaluated before, as the case in 17 studies, that proportion is even lower (N=5; 29.3.6%), pointing out a lack of reliability evaluation of more than half of these questionnaires. To analyze the internal consistency of the data collection instrument, all of these studies but one used Cronbach’s alpha. An exception is a study by Ivanov and Schneider (2010), who compute composite reliability, which four other studies also calculate in combination with Cronbach’s alpha.

Figure 27 – Distribution of studies per reliability type.



Source: the author.

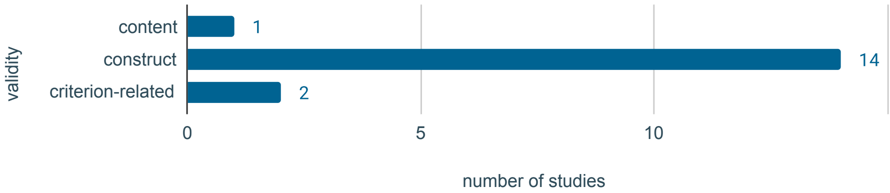
The lowest Cronbach's alpha scores ranged from .68 to .74 (ALTABOLI; LIN, Y., 2012) when applying the Classical/Expressive Questionnaire, a result below the recommended .8 for basic research instruments (STREINER, 2003). Other eight studies report Cronbach's alpha scores below .8 for previously built questionnaires. When analyzing custom-made questionnaires, the lowest alpha score is .7, which is above the recommended .50 to .60 for the early stages of research (STREINER, 2003). On the other hand, Ivanov and Schneider (2010) report a composite reliability score of .93 for their self-designed questionnaire, whereas Dupuy-Chessa et al. (2016) obtained a Cronbach's alpha score of .97.

We noticed that an even smaller number of studies analyze the equivalence among ratings (15.9%), of which 15 studies establish inter-rater reliability, and seven analyze the inter-rater agreement. Indices to indicate inter-rater reliability include the intraclass correlation coefficient (ICC) and the Pearson's correlation coefficient (N=8). Among the studies that use ICC, Huang (2013) reports a low inter-rater reliability (ICC = .305; no confidence interval reported). Two studies relate moderate but acceptable reliability among subjects rating GUIs viewing them for only 50ms (MINIUKOVICH; DE ANGELI, 2014b) or 150ms (MINIUKOVICH; DE ANGELI, 2015a). The other studies report high inter-rater reliability, with ICC above .76. None of the studies using Pearson's correlation in this type of analysis reported a coefficient below .5.

Three studies compute Kendall's coefficient of concordance to analyze inter-rater agreement (PANDIR; KNIGHT, 2006; PAPACHRISTOS; AVOURIS, 2009; SALIMUN et al., 2010b), and three others apply ICC (MCDONNELL; O'REILLY, 2017; SCHULTCHEN et al., 2021; SECKLER; OPWIS; TUCH, 2015). One study uses Randolph's Kappa coefficient (ZEN; VANDERDONCKT, 2016). Two studies indicate a low inter-rater agreement (PANDIR; KNIGHT, 2006; SALIMUN et al., 2010b), whereas the others report from moderate to high degrees of inter-rater agreement.

Ten studies have subjects rating the same GUI set twice to analyze the stability of ratings in time. All of them compute the correlation between ratings for intra-rater reliability. In one study, subjects rate 100 GUIs, from which 14 are rated twice, totaling 114 valid ratings (MINIUKOVICH; MARCHESE, 2020). The intra-rater reliability is computed using Pearson's and ICC coefficients over the rated-twice GUIs to exclude data deemed untrustworthy. Although also having subjects rate the same GUI set twice, two studies only compute the standard deviation of the rating difference (REINECKE et al., 2013; REINECKE; GAJOS, 2014). Another study does not calculate intra-rater reliability because different instruments are used in each rating (SCHAIK; LING, 2009).

Figure 28 – Distribution of studies per validity type.



Source: the author.

Fourteen studies (10.1%) analyze the construct validity of their assessment instruments, of which nine examine convergent/discriminant validity (Figure 28). They apply factor analysis (IVANOV; SCHNEIDER, 2010), average variance extracted (RAMEZANI NIA; SHOKOUHYAR, 2020), and correlation analysis (MOSHAGEN; THIELSCH, 2010) to verify if their instruments are highly related to other similar instruments (convergent validity) and poorly related to instruments that measure other constructs (divergent validity). Moshagen and Thielsch (2010; 2013) analyze criterion-related validity by correlating the VisAWI to participants' intention to revisit the website (concurrent validity) and assessing its ability to discriminate between ugly and beautiful GUIs (discriminative validity). One study surveys experts (university professors) to analyze content validity (RAMEZANI NIA; SHOKOUHYAR, 2020).

3.2.4 Discussion

In general, we encountered a significant amount of research on this topic which seems also increasing in 2020. Yet, the wide diversity of the presented assessment approaches also evidences a lack of consensus on how to measure the visual aesthetics of GUIs based on human judgment. This becomes evident concerning the definition as a unidimensional or multidimensional construct. And, continues when considering the variety of different terms, e.g., for the assessment as a unidimensional construct ("beautiful," "attractive," "visually appealing," or "aesthetically pleasing") or as a multidimensional construct, converging to either the simplicity/diversity/colorfulness/craftsmanship or the classical/expressive dimensions.

The choice of the assessment method depends on how visual aesthetics is understood. Those studies that treat it as a unidimensional construct tend to assess visual aesthetics on a single item scale with subjects indicating their general perception. This has the advantage of making the assessment process quick and simple by reducing the subjects' fatigue and allowing them to rate more GUIs in less time. On the other hand, approaches considering visual aesthetics as a multidimensional construct enable capturing the users' perception on a more detailed level.

Regarding scales, the use of ordinal scales, like Likert or semantic differentials, appears to be consonant with the assessment of GUI visual aesthetics because the distance between scale points (e.g., "beautiful" and "very beautiful") may differ for each respondent. Yet, some studies make use of interval scales such as the visual analog. The application of the magnitude estimation scale tries to overpass this issue by allowing subjects to rate each GUI freely. That way, they can reflect how distant they are from each other regarding visual aesthetics, supporting parametric statistics.

Regarding ordinal scales, there also seems to be no consensus so far on the optimal number of points (LEWIS; ERDINÇ, 2017). While some studies report that reliability and validity increase when respondents have more options to choose from (LOZANO; GARCÍA-CUETO; MUÑIZ, 2008; MAYDEU-OLIVARES et al., 2009), increasing the number beyond nine points may present low marginal returns (COX, 1980). Nonetheless, we can observe a trend for 7-point scales ±2 (five or nine points), while some works have concluded that this choice should be content-specific and a function of the measurement conditions (GARLAND, 1991).

A related issue is the inclusion of a midpoint representing a neutral choice concerning visual aesthetics. Literature shows some evidence that rating scales with mid-points might produce distortions in the results, arising from biased respondents that desire to please the interviewer or appear helpful (GARLAND, 1991). Besides, raters might choose the midpoint to represent their "no opinion" choice if questions are not evident due to language complexity (KLARE, 1950) or clarity (COOMBS, C. H.; COOMBS, L. C., 1976). Therefore, without the survey readability verification during the design phase, the inclusion of a midpoint could result in a systematic error, impacting both the reliability and validity of the collected data (VELEZ; ASHWORTH, 2007). However, most studies adopted a scale with a mid-point following general assessment trends and allowing subjects to express they neither perceive the GUI as beautiful or ugly.

Another issue is the computation of a final score based on the collected data to represent the visual aesthetics of a GUI. However, it seems that many approaches do not include this part, or in some cases, calculating the average value seems inappropriate in the context of certain types of scales. This misconception may indicate that they are still poorly understood, jeopardizing the results inferred from them.

In terms of the assessed GUI types, most studies analyze web GUIs, which is explained by the growing importance of online services. Nonetheless, there still seems to be a lack of studies aiming at mobile GUIs considering the current popularity of this type of device.

Regarding the research design, the studies also varied in terms of the number of GUIs assessed. By presenting the subjects with a high number of different GUIs, it is possible to expose them to greater variability in the stimuli, allowing the respondents to perceive GUIs with different levels of visual aesthetics. On the other hand, those studies that assess a small number of GUIs are either interested in only those GUIs or choose a few that they can manipulate to change some characteristics of interest.

Studies involving large numbers of GUIs tend to receive fewer ratings per GUI. Yet, we did not encounter a correlation between the total number of rated GUIs and the number of ratings each GUI received (r=.01). For example, in the study presented by Wu et al. (2016) using a set with 1,000 GUIs, each GUI received ten ratings, the same number each of the 30 GUIs received in the study by Singh and Bhattacharya (2010). That is also confirmed by analyzing studies from the ends of the range. For example, studies assessing only two GUIs collect from ten (LINDGAARD; DUDEK, 2003) to 481 ratings per GUI (DAMES et al., 2019), while studies that assess more than 300 GUIs receive from two (BAUERLY; LIU, Yili, 2009) to 1,793 ratings (REINECKE; GAJOS, 2014).

Another issue that has become evident is the lack of information about the target population. Many studies also do not mention the sampling technique used for the selection of the subjects. And among those stating this information, all use a non-probability sample. And none of the three exceptions that report using probability sampling make it clear the sampling technique they use or the target population at which they aim. Yet, as non-probability samples are not representative of wider groups, and, thus, findings cannot be generalized (COHEN; MANION; MORRISON, 2017), they may be more suitable for pilot studies that aim to test their instruments or validate their measurements (BHATTACHERJEE, 2012). Among these sampling techniques, we can see the trend of volunteering with the help of crowdsourcing platforms such as Microworkers[7] or Lab in the Wild[8], which considerably enlarges the number of sampled individuals.

And, despite assessing a construct that is highly influenced by subjective factors, very few studies analyze if demographic factors have any statistically significant influence on their results, even though almost all of them report some demographic data, like gender or age. Even so, some studies report quite unbalanced samples regarding gender and age, which might threaten their reliability, as indicated, e.g., by Moss and Gunn (2009) and Djamasbi et al. (2011). On the other hand, the composition of a more balanced group of raters may be complicated in practice, especially when aiming at a sample large enough to be representative of the target population in order to allow reliable generalizations.

Surprisingly we also observed a lack of information on the evaluation of the assessment methods. When assessing a construct so strongly affected by subjective factors like

[7] `https://www.microworkers.com/`

[8] `https://www.labinthewild.org/`

visual aesthetics, it is reasonable to expect a high variability in the responses. Notwithstanding, few studies analyze the inter-rater or intra-rater consistency of the subjects' responses to ensure that they systematically rate beautiful GUIs higher than the ugly ones. Regarding reliability, the majority of the studies only analyze internal consistency. Nonetheless, internal consistency is a function of the scores and, therefore, is not a property of the assessment instrument, making it sample-dependent (STREINER, 2003). For this reason, it should be established every time a study is conducted, regardless of if a new or a well-established instrument has been used. Yet, not even half of the studies in which subjects rate GUIs with multiple items report an analysis of the internal consistency of their instruments. And, information on the analysis of the assessment validity is even more scarce. For example, only two out of the 17 studies that designed their own questionnaire examine its validity.

These results, although demonstrating through the significant amount of research on this topic its relevance, that there is still a larger consensus lacking on how to assess the visual aesthetics of GUI based on human judgments.

4 MODEL DEVELOPMENT

As a solution we present a deep learning model to assess the visual aesthetics of GUIs designed with App Inventor. The model takes screenshots of App Inventor GUIs as input and delivers their visual aesthetics degrees as outputs, with performance measured as the difference to human perception.

The use of deep learning neural networks that can deal with numerous parameters is on par with the idea that beauty derives from every visual aspect of GUIs. Thus, instead of analyzing a handful of handcrafted features, such as color or density, deep learning neural networks can weigh how much image features contribute to the visual aesthetics of the GUI as a unity. We adopted a supervised learning approach training for convolution neural network (CNN) models, with images labeled according to their aesthetic degree. Following that approach, the model adjusts its parameters to reduce the difference between the inference to predict the labels and the actual labels representing the ground truth.

We adopted an exploratory approach testing different CNN architectures to find those that present the most promising results. An example is residual networks (ResNets) (HE et al., 2016), which are state-of-the-art in image recognition (DOSOVITSKIY et al., 2021; KOLESNIKOV et al., 2020). ResNets were proposed to alleviate the vanishing gradient phenomenon that makes deep neural networks difficult to train (IOFFE; SZEGEDY, 2015; ZHANG et al., 2019) and allow for neural networks with up to 152 layers. In this way, we can choose or design a network with adequate depth for the complexity of our problem. They also enable the use of hyperparameter optimization strategies specially developed for these CNN models, allowing faster training (SMITH, 2018; SMITH; TOPIN, 2019).

In the following sections, we present the results of our study to define the rating scale to label the screenshots, the preparation of the dataset and the sequence of iterations to develop the model.

4.1 REQUIREMENTS ANALYSIS

Our objective is to develop a deep learning model that learns from experience E for some class of tasks T and performance measure P, in which its performance at tasks in T, as measured by P, improves with E. In this research:

- A task in T is the assessment of the visual aesthetics of an App Inventor app screenshot with a numerical value within the interval [0..1];
- Experience E is a labeled dataset of App Inventor app screenshots, being each label a degree of visual aesthetics label within the interval [0..1] or a distribution of visual aesthetics ratings; and
- Performance P is the model loss, measured as the mean squared error (MSE)

between the predicted visual aesthetics degrees and the actual screenshot labels.

The model inputs are screenshots of Android applications developed with App Inventor and manually labeled by human raters. We adopted a supervised machine learning approach to deal with the visual aesthetics assessments and propose the aesthetics assessment task as a real-valued regression problem, rather than classification, following Dou et al. (2019). Our goal is to predict continuous scores for the screenshot visual aesthetics instead of discrete category labels. One reason for our choice is that we do not consider images as being beautiful or ugly (or any other category between these two) by themselves, but rather recognize that they provoke aesthetic experiences in different degrees from person to person. Second, regression has achieved better results than classification in similar works (DOU et al., 2019), which was corroborated in our first tests. Therefore, the output is a numerical value within [0..1], interpreted as the visual aesthetics degree, where 0 = "very ugly" and 1 = "very beautiful."

We selected a high-level CNN framework called fast.ai to develop our machine learning model (HOWARD; GUGGER, 2020). Fast.ai is based upon the PyTorch/Torch Python CNN framework, a good performing, flexible, and research-oriented CNN framework (FONNEGRA; BLAIR; DÍAZ, 2017). It offers ready-to-use and customizable functions to train models, making it suitable for practitioners mainly interested in applying pre-existing deep learning methods (HOWARD; GUGGER, 2020). Our choice for fast.ai relies on the fact that we are performing research on interface design assessments employing deep learning techniques rather than advancing state-of-the-art deep learning technologies.

Before training the final model to execute the proposed task, we tested different CNN architectures to identify the ones with better performance. In the first iteration, we used residual network architectures (ResNets), including ResNet18, ResNet34, ResNet50, and ResNet101 (HE et al., 2016). ResNets employ identity connections that act as shortcuts, bypassing several layers at a time, providing two parallel learning paths in several network sections, and avoiding the typical gradient loss of very-deep networks. This architecture allows for much deeper networks with up to 152 layers. Due to their design principle, we can choose or design a network with adequate depth for the complexity of the problem at hand. We chose ResNets mainly because they enable hyperparameter optimization strategies (HYPOs) especially developed for these CNN models, allowing faster training (SMITH, 2018; SMITH; TOPIN, 2019).

In the second iteration, we expanded the dataset with new screenshots and trained the best performing architecture in the previous one (ResNet50) and two other architectures to compare their performances, VGG19 and EfficientNet B0. The VGG19 architecture has achieved high accuracy with large-scale image recognition (SIMONYAN; ZISSERMAN, 2015) and significant results with the visual aesthetics assessments of photographs (LIN, R., 2022; SAKAGUCHI; TAKIMOTO; KANAGAWA, 2022). It uses small 3x3 convolution filters, the smallest possible size that still captures up/down and left/right.

All hidden layers use ReLU as the activation function. The EfficientNet B0 uses a fixed set of coefficients to scale up width, depth, and image resolution uniformly (TAN; LE, 2020) and overcomes the difficulty of randomly scaling up each of those dimensions by trial and error. The result is a performance improvement with less use of computational resources. We selected EfficientNet B0, which has shown a performance similar to ResNet50 (TAN; LE, 2020).

In the third iteration, we changed the labels from single scores representing visual aesthetics degrees to 5-dimensional vectors representing the distribution of ratings that the GUIs had received from the human raters. In addition to the visual aesthetics single score, it is also possible to compute the central tendency deviation from these vectors, representing the raters' degree of agreement towards the visual aesthetics of the GUI. We, thus, adapted the input and output layers to accept the vectors.

4.2 DATA PREPARATION

For the development of the model with sufficient labeled App Inventor GUI screenshots is required. To build the dataset, we took screenshots from apps available in the MIT App Inventor Gallery and apps developed in the context of the Software Quality Group/Computação na Escola initiative of the Universidade Federal de Santa Catarina (GQS/INE/UFSC). As part of the cleaning, we eliminated duplicates and images with unacceptable content (e.g., commercial, political, religious, ethical). We also pre-processed the screenshots to reduce and standardize the screenshot sizes (e.g., image downsampling).

All screenshots have been labeled with numeric values indicating their visual aesthetics resulting from a set of individual scores assigned by human raters. To compute the labels, we used a central tendency measure (e.g., mean, median, and mode) to indicate how that particular group of people perceive the GUIs. Labeling was done by ten researchers rating each GUI to achieve a representative score for that group. We conducted the labeling in multiple iterations to avoid fatigue.

4.2.1 Data collection

Screenshots of App Inventor apps were captured by 6 members of GQS/INE/UFSC following a pre-designed script. The apps were randomly selected from the App Inventor Gallery. However, as some apps are incomplete or look experimental, we selected only apps with at least one visible component to reduce the chance of getting screenshots of blank or unfinished screens. In addition, we also captured screenshots of apps developed in the context of Computing in School, an initiative of GQS/INE/UFSC.

After loading each project into the App Inventor IDE, we used its live testing feature to execute their GUIs. We manually took all screenshots using Genymotion v.3.1.2 to emulate a Google Pixel device running Android 8.0 - API 26 and saved them as PNG images with 1080x1920 pixels. The process resulted in 8,303 screenshots from different 1,552 apps. From this set, we selected 820 App Inventor screenshots to create a dataset for developing the neural network.

When preparing the dataset, we only included:

- screenshots in portrait mode;
- screenshots that display text in the latin alphabet; and
- screenshots containing acceptable images concerning ethical matters.

We excluded:

- screenshots of games, as their interface design significantly differs from other types of apps;
- screenshots of unfinished apps (e.g., blank screens and screens displaying only one element in the upper left corner or placeholder for text);
- screenshots very similar to others already selected (e.g., GUIs containing maps or variations of the same application); and
- screenshots displaying intense texting (e.g., 'about' screens) or advertisements.

Furthermore, aiming at creating a balanced dataset, the screenshots were fairly distributed on a 5-point scale, ranging from very ugly to very beautiful, according to the author's perception.

4.2.2 Definition of the labeling scale

Motivated by the need to assess large sets of mobile GUIs in the context of the development of Machine Learning-based solutions, we conducted an exploratory study aiming at analyzing different alternatives of scales. We selected the alternative scales based on a systematic mapping of the most common instruments (questionnaires, scales, and evaluation) to provide subjective ratings of GUI visual aesthetics (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022b). As a result, we identified that among the most used questionnaires is the Visual Aesthetics of Website Inventory questionnaire, in its full or short form (MOSHAGEN; THIELSCH, 2010, 2013). Regarding the unidimensional measurement of visual aesthetics, most studies adopt either the Likert or semantic differential scales, with five or seven points. Therefore, in this study, we adopted the short form of the VISAWI (VisAWI-S) as the golden standard, comparing the ratings with four alternatives of 5 or 7-point Likert and semantic differential scales.

4.2.2.1 Study definition

Using the GQM approach (BASILI; CALDIEIRA; ROMBACH, 1994), we defined the purpose of the study and systematically decomposed it into analysis questions and metrics.

Research question: Which is the most appropriate scale in terms of reliability and validity to assess GUI visual aesthetics based on human judgments?

Analysis questions:

- AQ1: Which scale alternative shows the highest inter-rater reliability or agreement?
- AQ2: Which scale alternative shows the highest intra-rater reliability or agreement?
- AQ3: How valid are the scale alternatives compared to the 7-point VisAWI-S questionnaire being considered a golden standard?

Rating scales: Participants expressed their perception of visual aesthetics on two different rating instruments for each GUI: a single-item scale and a four-item questionnaire. Each participant was randomly assigned to one out of the four groups that presented a different single-item scale. They varied in type (semantic differential or Likert) and the number of points/categories (five or seven) (Table 20). All respondents also indicated their agreement to four items on 7-point Likert scales as part of the VisAWI-S questionnaire (Table 1).

Table 19 – Rating instruments applied to each group of respondents.

Group	Rating instruments	
A	7-point semantic differential	VisAWI-S
B	5-point semantic differential	VisAWI-S
C	7-point Likert	VisAWI-S
D	5-point Likert	VisAWI-S

Source: the author.

Stimuli. We used nine different GUIs extracted from mobile apps developed with App Inventor (Figure 29). The apps were collected from the App Inventor Gallery or developed in the initiative Computação na Escola/INCoD/INE/UFSC context. We selected GUIs with Portuguese or well-known English words. We also aimed at creating a balanced dataset, including screenshots with interfaces ranging from very ugly to very beautiful, according to the authors' perception. The GUIs were individually assessed. GUI 2 was rated twice for intra-rater reliability/agreement analysis. For that reason, it was the first and last interface to be rated. All the other GUIs were presented in random order.

Procedure. The survey was conducted online, with participants using their own devices. Before starting the assessment, participants received on-screen information about the assessment objective, the rating process, and ethical concerns. They only started assessing the GUIs if they expressed their consent to participate. Next, they provided demographic background information (age, gender, education, experience, and mobile device system) and self-reported their vision condition concerning color-blindness (Tables 21 and

Table 20 – Semantic differential scale to rate unidimensional visual aesthetics (seven and five points).

Scale type	**Items**	**Response categories (n. of points)**
Semantic differential	How do you rate the visual aesthetics of this interface?	1 very ugly 2 ugly* 3 somewhat ugly 4 neither beautiful nor ugly 5 somewhat beautiful 6 beautiful* 7 very beautiful
Likert	I think this interface is beautiful.	1 strongly disagree 2 disagree* 3 somewhat disagree 4 neither agree nor disagree 5 somewhat agree 6 agree* 7 strongly agree

* Categories removed in 5-point scales

Source: the author.

22). Participants rated the whole set of nine GUIs, of which one GUI was rated twice. They expressed their perception of the visual aesthetics using one of the four uni-dimensional scale alternatives and responded to the VisAWI-S questionnaire for each GUI. We used LimeSurvey[1] to conduct the survey online. The assessment was conducted in Brazilian Portuguese.

Approval for conducting this research with human participants was granted by the Human Research Ethics Committee (CEPSH) at UFSC (Certificate No. 4.971.708).

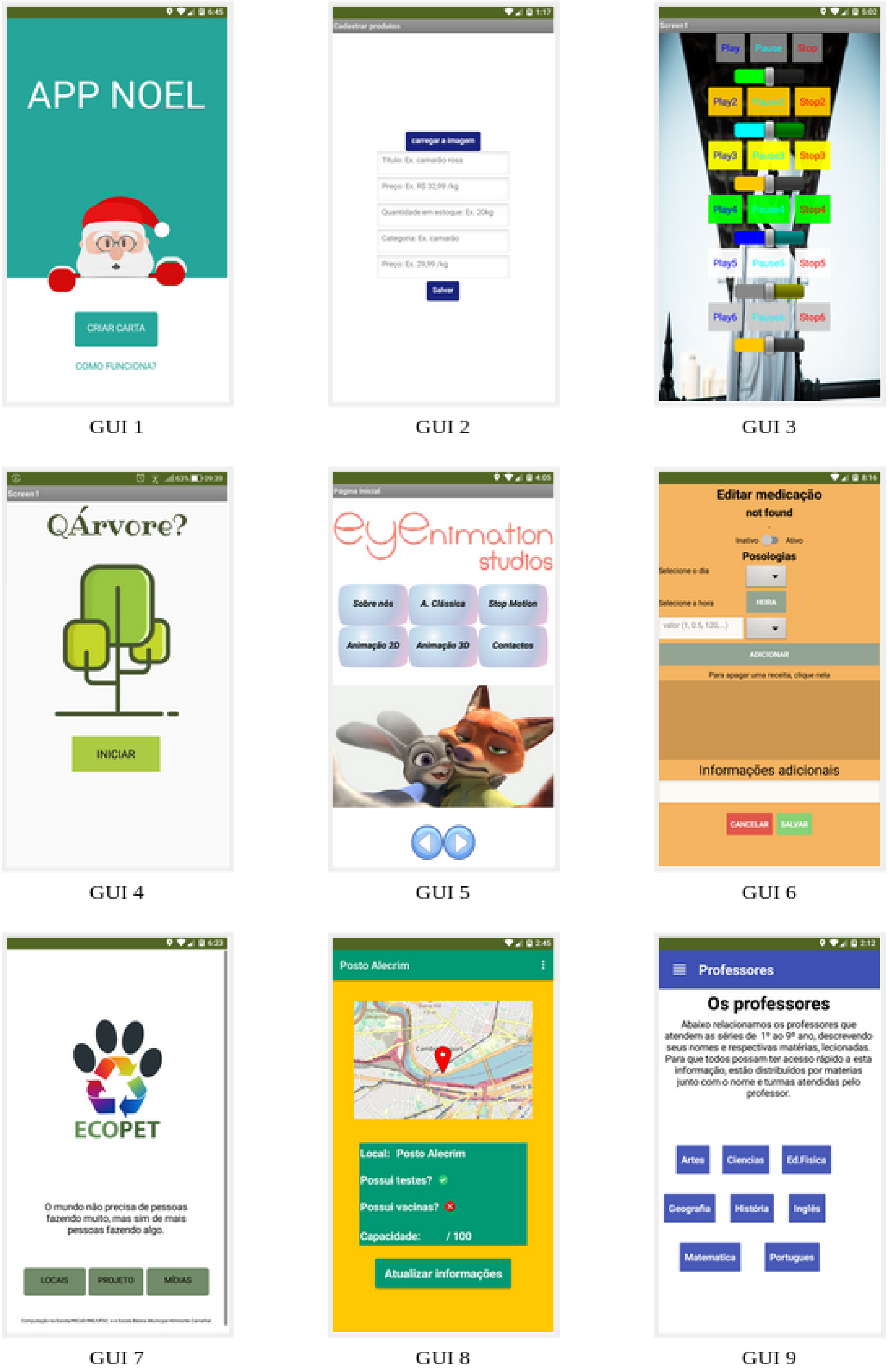
4.2.2.2 Execution of the study

We executed an explorative survey (WOHLIN et al., 2012) to collect evidence for future choices on subjective assessments of GUI visual aesthetics.

Participants. We used a convenience sample to select a total of 208 participants (84 female). Most were selected among students, faculty, and staff from UFSC, but we also invited family, friends, and acquaintances to participate in the study. Eleven participants that reported being color blind or not being sure about their condition had their responses removed. Participants completed the study anonymously and were not compensated for their participation. No time constraint was imposed on participants, and it took them an average of fewer than 10 minutes to complete the survey.

[1] `https://www.limesurvey.org`

Figure 29 – GUIs assessed in the study.



GUI 1 GUI 2 GUI 3

GUI 4 GUI 5 GUI 6

GUI 7 GUI 8 GUI 9

Source: the author.

4.2.2.3 Results

To evaluate the assessment instruments, we analyzed their reliability and validity. To analyze internal consistency, we computed Cronbach's alpha of the four VisAWI-S items. For the inter- and intra-rater reliability/agreement analysis, we calculated the intraclass correlation coefficient (ICC) for reliability and Kendall's coefficient of concordance for

Table 21 – Demographic questions.

Questions	Answer options
How old are you?	• 18 to 29 years old • 30 to 39 years old • 40 to 49 years old • 50 to 59 years old • 60+ years old
What is your biological gender at birth?	• male • female • no answer
What is your highest education level?	• middle school diploma • high school diploma • undergraduate degree • graduate degree • other
How many years of experience with interface design do you have?	• none • less than 1 year • less than 2 years • less than 3 years • over 3 years
What type of cell phone/mobile device do you use?	• iOS • Android • other
Are you color blind?	• yes • no • I don't know

Source: the author.

agreement. We analyzed the construct validity by calculating the Spearman correlation between each scale and the VisAWI-S (convergent validity).

We computed the visual aesthetics for each GUI according to the VisAWI-S following its proposed methodology (MOSHAGEN; THIELSCH, 2013), in which the final score is the mean score between all ratings received in the four subdimensions. Table 23 shows the rank of each GUI when compared to the others and the final aesthetic value in parenthesis. We first computed the degree of visual aesthetics by considering all responses from the four groups together and then considering each group separately. The results were fairly consistent for all GUIs. GUI 8 presented the highest difference between its lowest and highest scores (.87; 38%).

To compute the degree of visual aesthetics according to each scale type, we calculated the median, which is the indicated measure of central tendency for ordinal scales (NUNNALLY; BERNSTEIN, I. H., 1994) (Table 23). Comparing scales with the same number of points, we noted that the scores obtained with the Likert scales are equal to or lower than those with the semantic differential ones. An exception is GUI 4, which received a higher score from the 7-point Likert scale than from the 7-point semantic differential.

To analyze internal consistency, we computed Cronbach's alpha for the VisAWI-

Table 22 – Demographics about respondents.

Demographics	Frequency	(%)	Total (%)
Gender			
Male	124	59.6%	59.6%
Female	84	40.4%	100.0%
Age			
18 to 29 years old	112	53.8%	53.8%
30 to 39 years old	48	23.1%	76.9%
40 to 49 years old	25	12.0%	88.9%
50+ years old*	23	11.1%	100.0%
Highest education level			
Middle school diploma	2	1.0%	1.0%
High school diploma	33	15.9%	16.8%
Undergraduate degree	96	46.2%	63.0%
Graduate degree	77	37.0%	100.0%
Mobile system			
Android	154	74.0%	74.0%
iOS	51	24.5%	98.6%
other	3	1.4%	100.0%
Experience			
none	118	56.7%	56.7%
< 1 year	26	12.5%	69.2%
1 to 2 years	16	7.7%	76.9%
2 to 3 years	10	4.8%	81.7%
3+ years	38	18.3%	100.0%

* We merged the categories “50 to 59 years old” and “60+ years old” due to their small number of participants.

Source: the author.

S. Nunnally and Bernstein (1994) recommend an alpha of at least .80 for basic research instruments and above .90 if “important decisions” depend on the test scores. In its original study, the VisAWI-S showed a Cronbach’s alpha of .79 (MOSHAGEN; THIELSCH, 2013). As a result of our study, we observed an alpha of .94 when considering all nine GUIs. When computed individually for each GUI, it ranged between .85 (GUI 9) and .95 (GUI 3).

AQ1: Which scale alternative shows the highest inter-rater reliability/agreement?

To analyze inter-rater reliability/agreement, we computed the intraclass correlation coefficient (ICC) for the inter-rater reliability and Kendall’s coefficient of concordance with corrected ties (Wt) for agreement. The ICC can handle several rating situations, including the experience setting in our study (GISEV; BELL; CHEN, 2013). ICC scores range from 0 to 1, in which scores between .75 and .90 indicate good reliability, and above .90 indicate excellent reliability (KOO; LI, M. Y., 2016). As all respondents rated the

Table 23 – Visual aesthetics scores for each GUI according to VisAWI-S.

GUI	All responses rank (value)	Group A	Group B	Group C	Group D
1	1 (5.43)	1 (5.47)	1 (5.47)	1 (5.48)	1 (5.31)
2	7 (2.53)	8 (2.51)	7 (2.57)	5 (2.69)	6 (2.34)
3	9 (1.53)	9 (1.54)	9 (1.78)	9 (1.54)	9 (1.27)
4	2 (5.09)	2 (5.14)	2 (5.08)	2 (5.22)	2 (4,92)
5	5 (2.81)	6 (3.06)	5 (3.09)	5.5 (2.62)*	5 (2.50)
6	8 (2.44)	7 (2.54)	8 (2.51)	8 (2.58)	8 (2.14)
7	3 (4.42)	3 (4.67)	3 (4.33)	3 (4.46)	3 (4.20)
8	6 (2.72)	5 (3.15)	6 (2.85)	5.5 (2.62)*	7 (2.28)
9	4 (3.04)	4 (3.17)	4 (3.13)	4 (3.14)	4 (2.73)

* Average of ranks 5 a 6, following (PORTNEY, 2020).

Source: the author.

Table 24 – Visual aesthetics scores for each GUI by rating scale.

GUI	7-point semantic differential	5-point semantic differential	7-point Likert	5-point Likert
1	6	4	6	4
2	3	2	2	1
3	1	1	1	1
4	5	4	6	4
5	3	2	2	2
6	2	2	2	1
7	5	4	5	4
8	3	2	2	2
9	3	2	2	1

Source: the author.

same GUIs within each group, we applied the two-way random effects model for average measurements (MCGRAW; WONG, 1996). We used the "irr" package running on R version 4.1.1 to compute ICC estimates and their 95% confidence intervals.

Kendall's coefficient of concordance is suitable to express inter-rater agreement of multiple raters on ordinal scales (GISEV; BELL; CHEN, 2013). Wt ranges between 0 and 1, where 0 means no agreement and 1 means complete agreement. As the number of raters increases, it becomes harder to achieve a higher Wt. As a consequence, low Wt scores can also be significant (GISEV; BELL; CHEN, 2013).

We first computed the inter-rater reliability and agreement of the VisAWI-S, considering the responses of the four groups together (Table 25). Instead of analyzing the ratings of each subdimension, we computed the visual aesthetics score according to each respondent. Reliability was very high, with ICC = .997. Kendall's coefficient of concordance resulted in a significant Wt = .582, indicating that respondents did not achieve this degree of agreement by chance.

Table 25 – VisAWI-S inter-rater reliability and agreement considering all responses.

	Reliability (ICC)	**Agreement (Kendall)**
VisAWI-S	ICC(C,208) = .997; 95% CI [.993, .999]	Wt = .582

Source: the author.

When analyzing each group separately, inter-rater reliability is again very high. All ICC scores are above .97 (Table 26). The differences between the scales and the VisAWI-S are quite small, within and across groups. Also, scores are slightly higher for Likert scales than for semantic differentials. Inter-rater agreement varies around Wt = .6 among the groups.

Table 26 – Inter-rater reliability and agreement within each group.

Group	**Scale**	**Reliability (ICC)**	**Agreement (Kendall)**
A	VisAWI-S	ICC(C,53) = .984; 95% CI [.964, .996]	Wt = .594
	7-point semantic differential	ICC(C,53) = .986; 95% CI [.969, .996]	Wt = .594
B	VisAWI-S	ICC(C,48) = .983; 95% CI [.962, .995]	Wt = .543
	5-point semantic differential	ICC(C,48) = .983; 95% CI [.962, .995]	Wt = .539
C	VisAWI-S	ICC(C,54) = .988; 95% CI [.973, .997]	Wt = .588
	7-point Likert	ICC(C,54) = .987; 95% CI [.971, .996]	Wt = .615
D	VisAWI-S	ICC(C,53) = .989; 95% CI [.975, .997]	Wt = .616
	5-point Likert	ICC(C,53) = .987; 95% CI [.971, .996]	Wt = .637

Source: the author.

AQ2: Which scale alternative shows the highest intra-rater reliability/agreement?

To analyze intra-rater reliability and agreement, we computed the same indices (ICC and Wt) that we used for inter-rater analysis. In this case, however, we only considered the scores assigned to GUI 2, which the respondents rated twice during their participation. The ICC score was again very high, presenting a good to excellent intra-rater reliability. Kendall's coefficient of concordance was also very high, showing agreement between both ratings (Table 27).

Intra-rater reliability ranged from good to excellent within groups, except for the groups that used 5-point scales (Likert and semantic differential), achieving a moderate to good reliability (Table 28). Wt scores regarding intra-rater agreement were also high, yet again the 5-point Likert scale achieved the lowest value (Wt = .753).

Table 27 – VisAWI-S intra-rater reliability and agreement considering all responses.

	Reliability (ICC)	**Agreement (Kendall)**
VisAWI-S	ICC(C,2) = .913; 95% CI [.885, .934]	Wt = .904

Source: the author.

Table 28 – Intra-rater reliability and agreement within each group.

Group	**Scale**	**Reliability (ICC)**	**Agreement (Kendall)**
A	VisAWI-S	ICC(C,2) = .86; 95% CI [.757, .919]	Wt = .817
	7-point semantic differential	ICC(C,2) = .856; 95% CI [.751, .917]	Wt = .817
B	VisAWI-S	ICC(C,2) = .918; 95% CI [.853, .954]	Wt = .923
	5-point semantic differential	ICC(C,2) = .797; 95% CI [.639, .886]	Wt = .855
C	VisAWI-S	ICC(C,2) = .928; 95% CI [.876, .958]	Wt = .923
	7-point Likert	ICC(C,2) = .879; 95% CI [.792, .93]	Wt = .87
D	VisAWI-S	ICC(C,2) = .926; 95% CI [.872, .957]	Wt = .914
	5-point Likert	ICC(C,2) = .772; 95% CI [.606, .869]	Wt = .753

Source: the author.

AQ3. How valid are the scale alternatives compared to the 7-point VisAWI-S questionnaire being considered a golden standard?

We used Spearman's rank correlation coefficient (ρ) to compare how close the responses with each scale type are to those with the VisAWI-S. The choice for the Spearman's rank correlation coefficient, rather than the more common correlation test Pearson's r, is justified by the ordinal scales (BRYMAN; CRAMER, 1990). The coefficient ρ ranges between -1 and 1. A $\rho = 1$ indicates a perfect association between responses, a $\rho = 0$ indicates no association and a $\rho = -1$ indicates a perfect negative association. That way, ρ scores between .7 and .89 can be considered strong, and between .9 and 1, very strong (SCHOBER; BOER; SCHWARTE, 2018). The closer ρ is to zero, the weaker the association between both rating instruments. We also calculated the p-value (or probability) to verify how likely any observed correlation is given by chance. P-values range between 0 (0%) and 1 (100%), and a p-value close to 1 suggests no correlation except by chance and, therefore, the null hypothesis assumption is correct. The results of the analysis are presented in Table 29. We correlated the responses within each group.

We also correlated all the scales, two by two, to see how they converge. In this case, instead of correlating individual ratings, we used the final visual aesthetic score of the nine GUIs achieved with each scale (Table 30).

Table 29 – Spearman's correlation coefficient between each scale type and the VisAWI-S

Group	**Scale**	**Spearman***
A	7-point semantic differential x VisAWI-S	$\rho = .91$
B	5-point semantic differential x VisAWI-S	$\rho = .9$
C	7-point Likert x VisAWI-S	$\rho = .91$
D	5-point Likert x VisAWI-S	$\rho = .9$

* p < .001

Source: the author.

Table 30 – Spearman's correlation coefficient between the final scores achieved with each scale type and the VisAWI-S.

Scale	**VisAWI-S**	**7-point semantic differential**	**5-point semantic differential**	**7-point Likert**
7-point semantic differential	.93	-	-	-
5-point semantic differential	.87	.91	-	-
7-point Likert	.88	.91	.99	-
5-point Likert	.83*	.86*	.84	.83

* p < .01; all others: p < .001

Source: the author.

4.2.2.4 Discussion

The responses with each one of the four scales showed excellent inter-rater reliability based on the ICC. This consistency among responses indicates that, although raters might not agree on how beautiful each GUI is, they would probably agree on which ones are the most beautiful and the ugliest, ranking them similarly. A reason for such high inter-rater reliability might be the result of assessing a group of GUIs that are distinctly beautiful or ugly. Although we had not formally assessed the GUIs, we selected those that we assumed would respond to the complete extension of each scale. And although the most beautiful and the ugliest ones stand out, several GUIs in the middle range are not so clearly distinguishable.

None of the groups showed an inter-rater agreement that was preponderant over the others, as all varied around .6. Although this is near the middle of the scale, we consider this value as good, considering the difficulties of finding agreement regarding visual aesthetics. When assessing visual aesthetics, the agreement shows that judges have a similar opinion about the same object. A high inter-rater agreement is an indication that they share a similar background.

Our survey showed considerably higher inter-rater reliability when compared to previous studies that executed the same analysis. McDonnell and Lee (2016) found a good coefficient (ICC (2,16) = .865, 95% CI [.779, .928]), whereas it was poor in Huang (2013) (ICC = .305; no confidence interval reported). In terms of the agreement, our analysis indicates that respondents achieved a higher level of agreement than those in the works of Pandir and Knight (2006) (W = .193, $p < .01$) and Salimun et al. (2010a) (W = .1023), but much lower than Papachristos and Avouris (2009) (W = .886 to .971). But in this case, only three GUIs that were designed to look different from each other were rated.

The Likert scales show slightly higher inter-rater reliability than semantic differential ones. We can also observe that when comparing inter-rater agreements. There are even less observable differences when comparing scales with five points vs. seven points. Although the 7-point semantic differential scale presents higher reliability than its 5-point counterpart, both 5- and 7-point Likert scales result in equal scores. In terms of inter-rater agreement, the semantic differential scale with seven points displays a slightly higher value than that with five points, but, on the other hand, the 5-point Likert scale shows higher agreement than the 7-point one. Nonetheless, those differences are too small to indicate any preference.

Regarding intra-rater reliability/agreement we observed that, even though both measures were taken in the same session, many differed the second time. Nonetheless, intra-rater reliabilities were good to excellent for responses using almost all scales. Exceptions are the ICC values for the responses within the groups that used 5-point scales (groups B and D), which only achieved moderate to good intra-rater reliability. The same results can be observed intra-rater agreement. Yet, in this case, only responses with the 5-point Likert scales achieved a lower Wt than the others. One hypothesis concerns the number of points: a variation of one point from one rating to another would cause a greater impact on a 5-point scale than on a 7-point scale. Also, as the number of points grows, respondents tend to choose the same response from item to item (WEATHERS; SHARMA, S.; NIEDRICH, 2005). However, the responses on semantic differential scales with five points found a similar agreement to those using seven points. It is also possible that the other GUIs might have affected those respondents stronger than those in other groups, causing them to change their minds about this particular GUI. The reasons for that consistency/agreement reduction between ratings seem to be beyond the scope of this work.

The 7-point Likert scale shows the highest intra-rater reliability/agreement, although the difference from the 7-point semantic differential scale is small. However, those two scales present higher reliability than scales with five response options. Regarding the agreement, the 5-point semantic differential scale score is on par with those for 7-point scales, yet the Likert scale presents a lower score. Although this may not be a reason to discard this scale alternative, it should be used with caution.

Regarding the validity of the scale alternatives when compared to the 7-point VisAWI-S questionnaire as a golden standard, we observed that all scale types show a strong correlation with the VisAWI-S (> .9), indicating convergent validity. Thus, any scale alternative can provide a very close approximation to perceived visual aesthetics as measured by the VisAWI-S questionnaire. Also, the strong correlation between scales, when compared pairwise, suggests that any of them can be used to assess the same construct, i.e., the visual aesthetics of mobile GUIs. In this regard, our results are also consistent with other works concerning the number of points (LEWIS, 2021) or instrument types (BRÜHLMANN et al., 2020; CHYUNG et al., 2018; LEWIS, 2018).

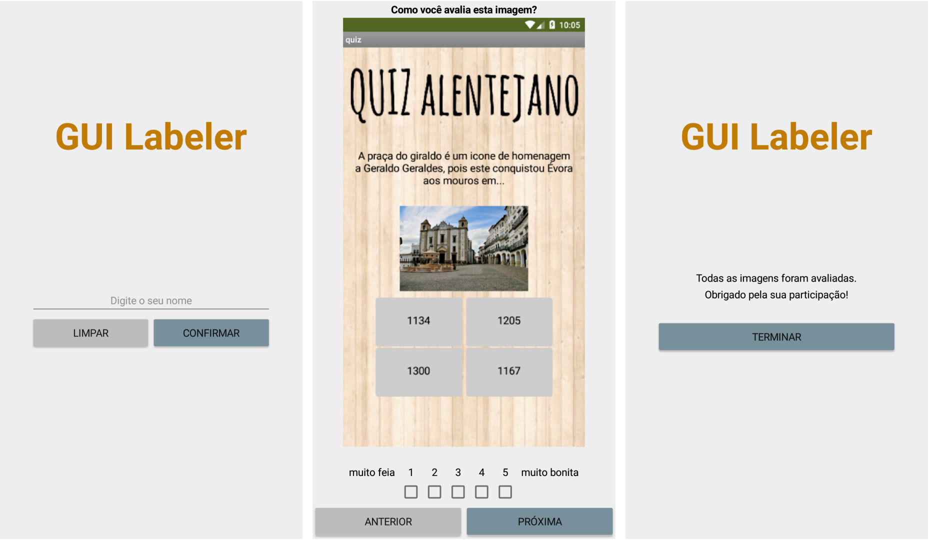
In summary, the results of our study provide evidence that single-item questionnaires can measure the perceived visual aesthetics of mobile app GUIs with a similar reliability/validity compared to well-established multiple-item questionnaires, such as the VisAWI-S. Comparing the four single-item questionnaire alternatives, no significant differences could be identified. However, no matter which instrument is chosen, a quality analysis (inter and/or intra-rater reliability/agreement) is required to make sure that the responses obtained are representative of the target group.

4.2.3 Labeling process

Ten GQS/INE/UFSC members participated in the labeling process (50% female), all with degrees in computing-related areas. None of them reported being colorblind. Four participants reported having no experience with interface design, and another one had less than one year of experience. The other five participants had at least two years of experience. Each participant used their own device to label the screenshots in three sessions. In the first two sessions, they labeled 600 screenshots that were used in the first training iteration (dataset 1). In the last session, they labeled another 220 screenshots that were added to the expanded dataset used in the second training iteration (dataset 2). Although three participants had iOS devices, only one was unfamiliar with Android applications.

Based on the results from the exploratory study, the participants labeled the screenshots on a 5-point semantic differential scale (“1” = “very ugly”; “5” = “very beautiful”). This scale type has shown high reliability and validity when used to rate mobile GUI visual aesthetics as result of our study presented in Section 4.2.2. It has also shown a good correlation with the short version of the Visual Aesthetics of Website Inventory (VisAWI-S), a widely used questionnaire to assess the visual aesthetics of websites. That way, participants could express their perception on only one scale instead of four in the VisAWI-S, considerably reducing their effort. We developed the application GUI Labeler with App Inventor to operationalize the rating process. The app enabled participants to assign a degree of perceived visual aesthetics to each of the screenshots (Figure 30) during the labeling process. The app shows each screenshot separately with the respective rating scale below. The rating process can be interrupted at any moment, allowing its continuation from where it stopped. This allowed participants to halt their assessments whenever needed or wanted, e.g., due to an external interruption or fatigue. In addition, it enabled participants to rate the apps anywhere and anytime, bringing them closer to real-life users. The responses from this rating process showed excellent inter-rater reliability (ICC(C,10) = .877; 95% CI [.862; .891]).

Figure 30 – Mobile application developed for the rating process.



Como você avalia esta imagem?

A praça do giraldo é um icone de homenagem a Geraldo Geraldes, pois este conquistou Évora aos mouros em...

Todas as imagens foram avaliadas.
Obrigado pela sua participação!

Source: the author.

To compute the final label of each screenshot, we calculated the median of all ratings as the indicated measure of central tendency for ordinal scales (NUNNALLY; BERNSTEIN, I. H., 1994). Besides, the median can provide a typical value that is not as skewed by extremely high or low scores. As each screenshot received an even number of ratings (10), the median could result in the intermediate value between two points on the rating scale. We decided to keep these values to minimize the loss of granularity from converting the scale values to the continuous interval [0..1]. Thus, although the screenshots received ratings on a 5-point scale, their labels could have nine possible values (five scale points and their intermediate values). We also computed the average absolute deviation (AAD) as a measure of dispersion among the responses (LEYS et al., 2013). A high deviation indicates that the ratings are spread along the rating scale, allowing us to interpret that participants disagree about the visual aesthetics of that particular screenshot. Therefore, we removed 97 screenshots that received ratings with a deviation equal to or greater than 1 from dataset 1 in order to avoid training our model with confusing data. In addition, another 22 screenshots (all with a deviation of .9) were removed to balance the set between beautiful and ugly images. Applying the same criteria. 17 screenshots were removed from dataset 2. Finally, we normalized the labels to the interval [0..1], where "0" = "very ugly" and "1" = "very beautiful."

As a result, dataset 1 contained 481 GUI screenshots and dataset 2 contained 684 screenshots. All images had labels indicating their visual aesthetics degrees within [0..1]. We randomly set aside screenshots with average absolute deviations of .5 or less to form the test set, fairly distributed along the rating scale. The test set had 15 screenshots in the first training iteration, and 20 in the second iteration. None of these screenshots was used in the training or validation steps. We randomly divided the other screenshots using 80% for training and 20% for validation. In this way, dataset 1 was split into a training set of 373 screenshots and a validation set of 93 screenshots, and dataset 2 into a training set of 532 screenshots and a validation set of 132 screenshots. For pre-processing, we downsampled the screenshots from 1080x1920 to 448x448 pixels. We performed no other transformations, such as cropping, to avoid distorting image features relevant to visual aesthetics perception. The dataset is available online[2].

4.3 SINGLE-POINT REGRESSION

4.3.1 Model training

To train our models, we adopted a transfer learning approach. Following this approach, we adapted the output layers of a pre-trained network to work with our evaluation metrics and train its input and final fully-connected output layers on our data, maintaining the pre-trained internal feature representation structure of the network. After this training, we unfroze its internal features, allowing all its layers to learn. We then performed a fine-tuning, with the same dataset of the specific domain, to adjust all internal features to our data (YOSINSKI et al., 2014). In addition to allowing faster training, transfer learning enables using small datasets, such as ours, without overfitting the model. We used a pre-trained model with ImageNet, one of the largest publicly available general-purpose datasets (RUSSAKOVSKY et al., 2015).

As an optimization training strategy, we chose an automatic hyperparameter (HYPO) called fit1cycle (SMITH, 2018; SMITH; TOPIN, 2019), developed for residual networks. It works with a varying adaptive learning rate and momentum, where the learning rate is automatically increased first and then decreased while the momentum rate follows an opposite strategy (SMITH; TOPIN, 2019).

Considering that we are dealing with our problem as a regression task, we adapted the input layer to the image resolution of our dataset represented by a vector of a screenshot and its numerical label. We also substituted the original output layers, which in the ImageNet pre-trained networks represent a categorical variable with 1,000 values (containing 1,000 neurons for the ImageNet dataset) for a regression layer with a single neuron. Also, the cross-entropy loss for classification is changed for the regression loss. Thus, the results for a predicted aesthetic score of a mobile GUI screenshot results range within [0..1]. Although the screenshots received ratings on an ordinal 5-point scale, we aimed to predict continuous scores for the screenshot visual aesthetics instead of discrete category labels. The reason is because the cross-entropy loss function of classification models would

[2] `https://bit.ly/app-inventor-dataset-v2`

not reflect the distances between different points on the rating scale. For example, “2” is closer to “3” than it is to “5,” but the cross-entropy loss would be the same no matter the distance. Also, regression has achieved better results than classification in similar research (DOU et al., 2019; XING et al., 2021). In this way, the output is a numerical value within [0..1], interpreted as the visual aesthetics degree, where “0” = “very ugly” and “1” = “very beautiful.”

4.3.1.1 First iteration

The output layers of the networks were transfer-trained until the validation error stopped improving. We trained two models for each architecture, one using standard training (fit) and another using automated hyperparameter optimization (fit1cycle), with transfer-learning and fine-tuning phases (SMITH, 2018; SMITH; TOPIN, 2019). This strategy works with a varying adaptive learning rate and momentum, where the learning rate is automatically increased first and then decreased while the momentum rate follows the opposite way (SMITH, 2018). We kept all default parameters (Table 31) and trained for no more than 100 epochs during transfer learning.

Table 31 – Summary of the compared models in the first iteration.

Architectures	ResNet18, ResNet34, ResNet50, ResNet101	
Input dimension of images	448 x 448 (pixels) x 3 (color channels)	
Predictive model	Regression [0..1]	
Learning algorithm	Backpropagation	
Dataset separation	373 screenshots for training (80%); 93 screenshots for validation (20%)	
Training strategy	fit, fit1cycle	
Phase	Transfer learning	Fine tuning
Epochs	100	20
Learning rate	.003	range of optimum learning rates[3]
Weight decay	No	No

Source: the author.

In the fine-tuning phase, we employed the same strategy as in the transfer learning, unfreezing and allowing for the adaptation of all weights in the network. We determined a range of optimum learning rates using the method suggested by Smith and Topin (2019). It resulted in a different range of rates for each network. After fine-tuning, all models slightly improved their performance. Given the results presented in Table 32, ResNet50 trained with the fit strategy performed best (Figure 31).

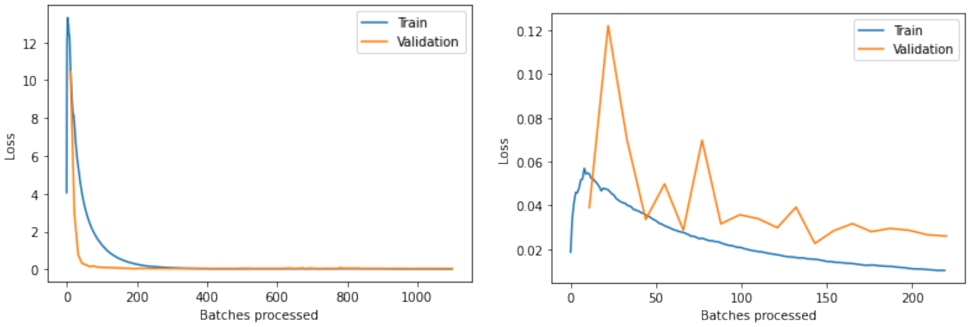
[3] (SMITH; TOPIN, 2019)

Table 32 – Best MSE for each model in the first iteration.

	Transfer learning		**Fine-tuning**	
Architecture	**Strategy**	**MSE**	**LR**	**MSE**
ResNet18	fit	.036627	6.31e-07; 3.31e-07	.035908
	fit1cycle	.034406	3.98e-06; 1.32e-03	.031995
ResNet34	fit	.031096	3.31e-06; 9.12e-06	.030115
	fit1cycle	.032325	6.31e-07; 7.59e-08	.032314
ResNet50	fit	.038136	1e-04; 1e-03	**.022649**
	fit1cycle	.034477	1e-05; 1e-04	.026813
ResNet101	fit	.030617	3e-06; 3e-07	.02787
	fit1cycle	.032051	2.75e-06; 2.09e-04	.027051

Source: the author.

Figure 31 – Train and validation losses for ResNet50 in the first iteration; transfer learning (left); fine-tuning (right).



Source: the author.

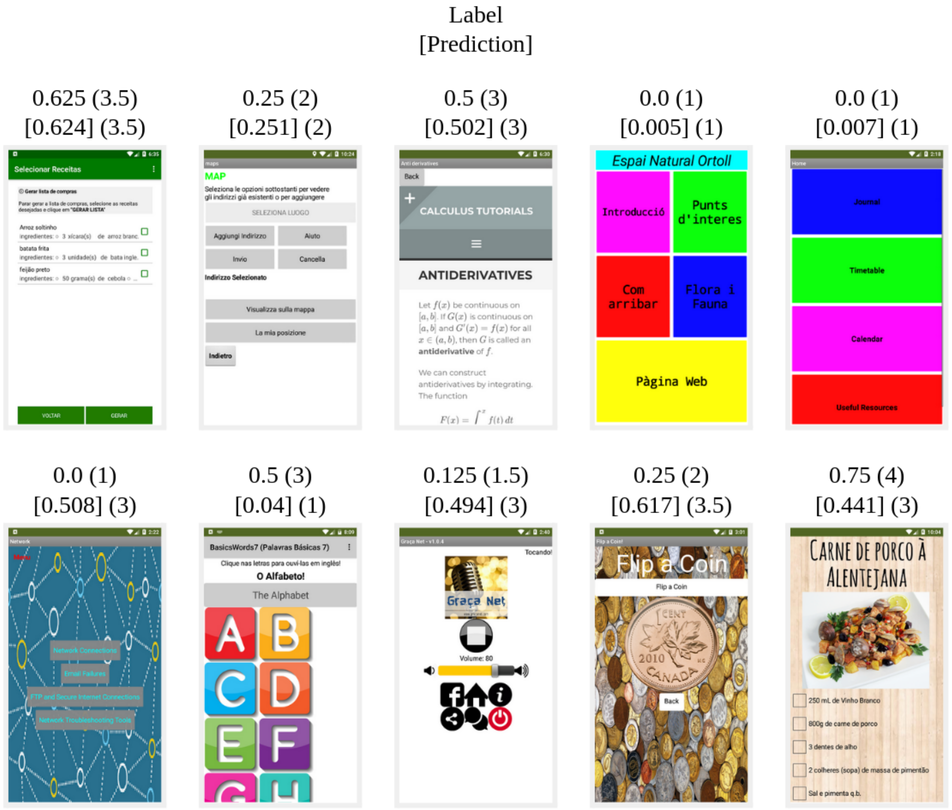
These results demonstrate that the model performed very well in classifying visual aesthetics. Most classifications (81.7%) differed by less than one point when converted back to a 5-point scale. That means that most of the time, it can classify a "beautiful" GUI somewhere between "very beautiful" and "neither beautiful nor ugly" but it does not classify such a GUI as "ugly," for example. That is an acceptable result, considering that even humans have trouble agreeing about visual aesthetics (GRESSE VON WANGENHEIM et al., 2018b).

These good performance results have also been observed when using the test set, containing only unseen screenshots without access to their labels for prediction (Figure 33). Considering only this set, the MSE was .0166 and for only two GUIs the prediction differed by one point when converted to a five-point scale. The model differed by just half a point on the visual aesthetics degree of eight GUIs and got the label right on another five.

4.3.1.2 Second iteration

In the second iteration, we expanded the training/validation dataset with 198 new screenshots and randomly selected 20 screenshots for the test set. We trained a ResNet50,

Figure 32 – Validation set: best (top) and worst (bottom) classifications (values converted to the 5-point scale in parenthesis).



Source: the author.

which was the best performing model in the previous iteration and a VGG19 and an EfficientNet B0. In this way, we could compare the performance of the ResNet with that of two architectures that are widely used for similar tasks. Here, we used the same training steps (transfer learning and fine tuning) and strategies used in iteration 1 (Table33).

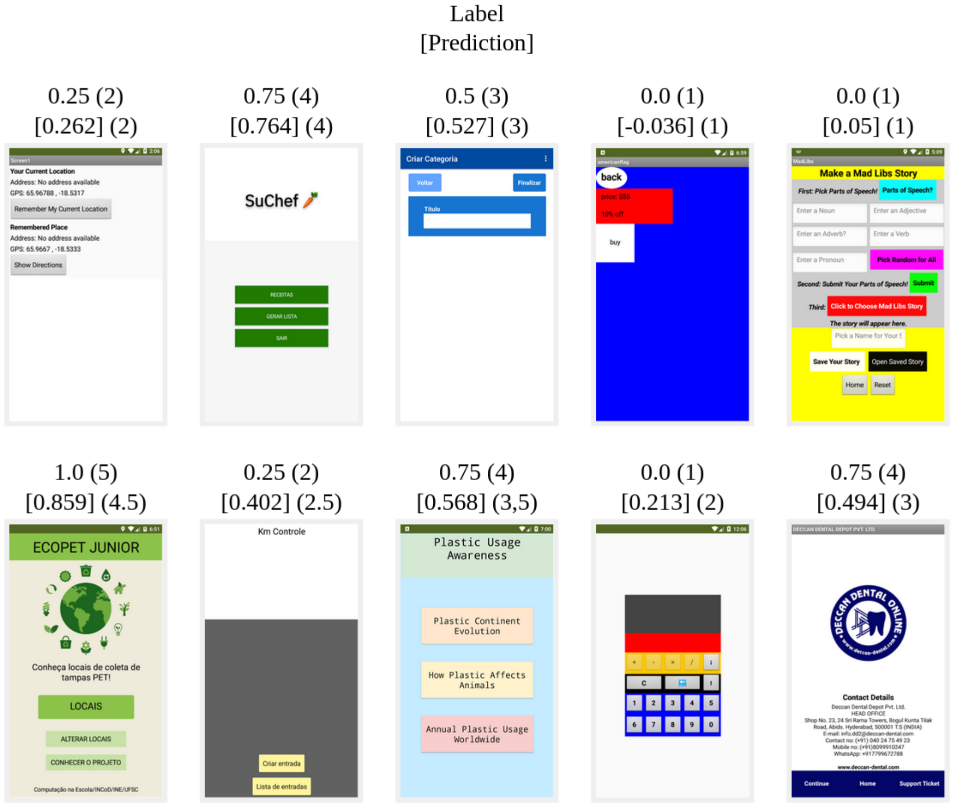
Again, the Resnet50 model presented the best performance when compared with the other architectures (Table 34). Although the VGG19 and the EfficientNet B0 showed superior performance when compared with other ResNets, in this iteration the ResNet50 showed a lower MSE than in the previous iteration.

The ResNet50 model kept good performance when classifying visual aesthetics. In this iteration, 79.5% of the classifications differed by less than one point when converted back to a 5-point scale, which is very close to the previous iteration. On the other hand, the classifications for only five screenshots (out of 132) differed by one and a half points from their labels (Figure 35, bottom).

The model also demonstrated very good performance with the test set (Figure 36). When converted to a five-point scale, predictions differed by at most one point and were correct for eight out of twenty GUIs (40%).

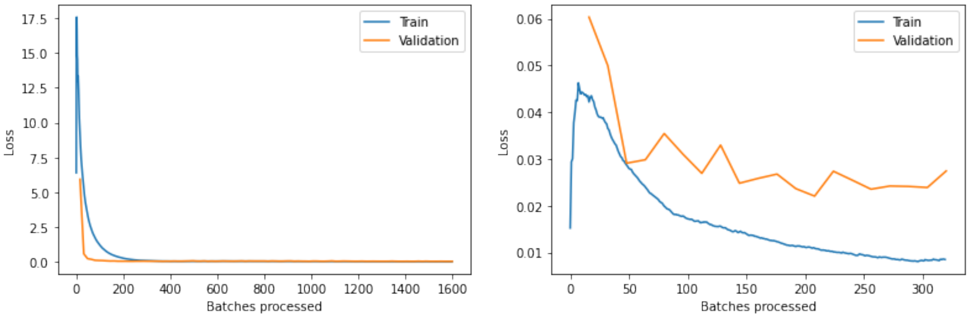
[4] (SMITH; TOPIN, 2019)

Figure 33 – Test set: best (top) and worst (bottom) predictions (values converted to the 5-point scale in parenthesis).



Source: the author.

Figure 34 – Train and validation losses for ResNet50 in the second iteration; transfer learning (left); fine-tuning (right).



Source: the author.

4.3.2 Model evaluation

Studies using deep learning models typically evaluate their performances 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 the GUIs separated from the dataset before training.

We evaluated the best performing model (ResNet50) to assess how close the visual

Table 33 – Summary of the compared models in the second iteration.

Architectures	VGG19, ResNet50, EfficientNet B0	
Input dimension of images	448 x 448 (pixels) x 3 (color channels)	
Predictive model	Regression [0..1]	
Learning algorithm	Backpropagation	
Dataset separation	532 screenshots for training (80%); 132 screenshots for validation (20%)	
Training strategy	fit, fit1cycle	
Phase	Transfer learning	Fine tuning
Epochs	100	20
Learning rate	.003	range of optimum learning rates[4]
Weight decay	No	No

Source: the author.

Table 34 – Best MSE for each model in the second iteration.

	Transfer learning		**Fine-tuning**	
Architecture	**Strategy**	**MSE**	**LR**	**MSE**
VGG19	fit	.026874	6.31e-07; 6.92e-05	.025613
	fit1cycle	.029634	6.31e-07; 5.75e-05	.028517
ResNet50	fit	.026767	1e-04; 1e-03	**.022049**
	fit1cycle	.029282	1e-05; 1e-04	.027709
EfficientNet B0	fit	.024025	1e-05; 1e-04	.023942
	fit1cycle	.023812	2e-04; 2e-05	.023651

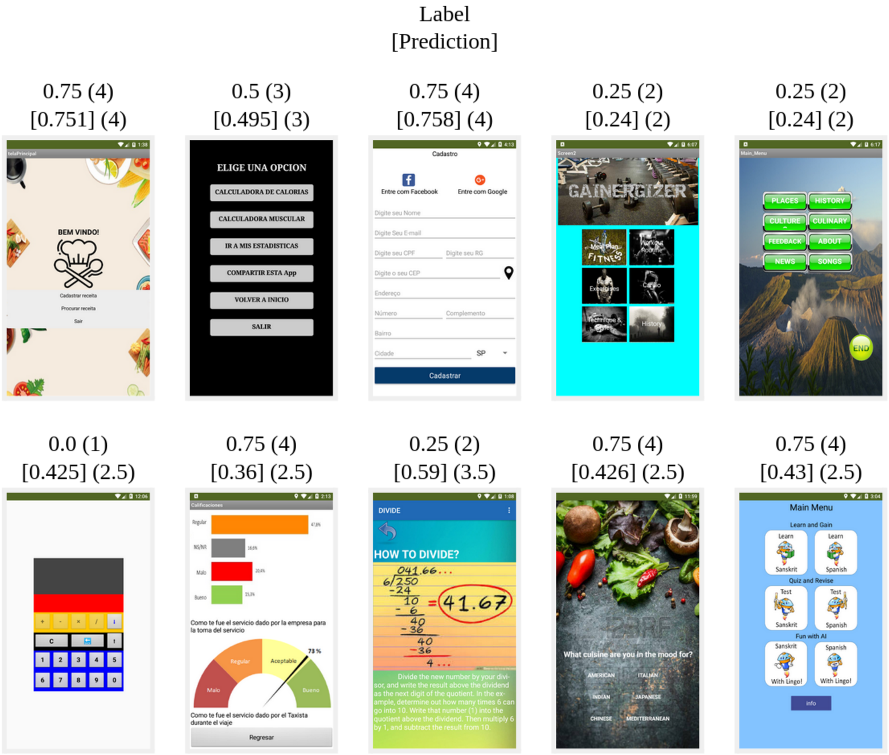
Source: the author.

aesthetics predicted in our models are to human ratings. That is typically done by conducting a correlation analysis, which we executed to enable the comparison of our results to a similar work (DOU et al., 2019). However, correlations quantify the degree to which two variables are related, not how much they agree. And as they only evaluate the linear association between two sets of observations, they can be inadequate and misleading when assessing their degree of agreement (GIAVARINA, 2015). Thus, we also used the Bland-Altman (B&A) plot analysis to measure the degree of agreement between the automatic assessment of our model and the human ratings (BLAND; ALTMAN, 1986).

4.3.2.1 Correlation analysis

To evaluate if the trained model performs well on previously unseen inputs, we analyzed the performance of the learned model against the test dataset. We measured the strength of the linear association between the results of the deep learning network and ground truth based on human assessments using the Spearman rank correlation. The choice

Figure 35 – Validation set: best (top) and worst (bottom) classifications (values converted to the 5-point scale in parenthesis).



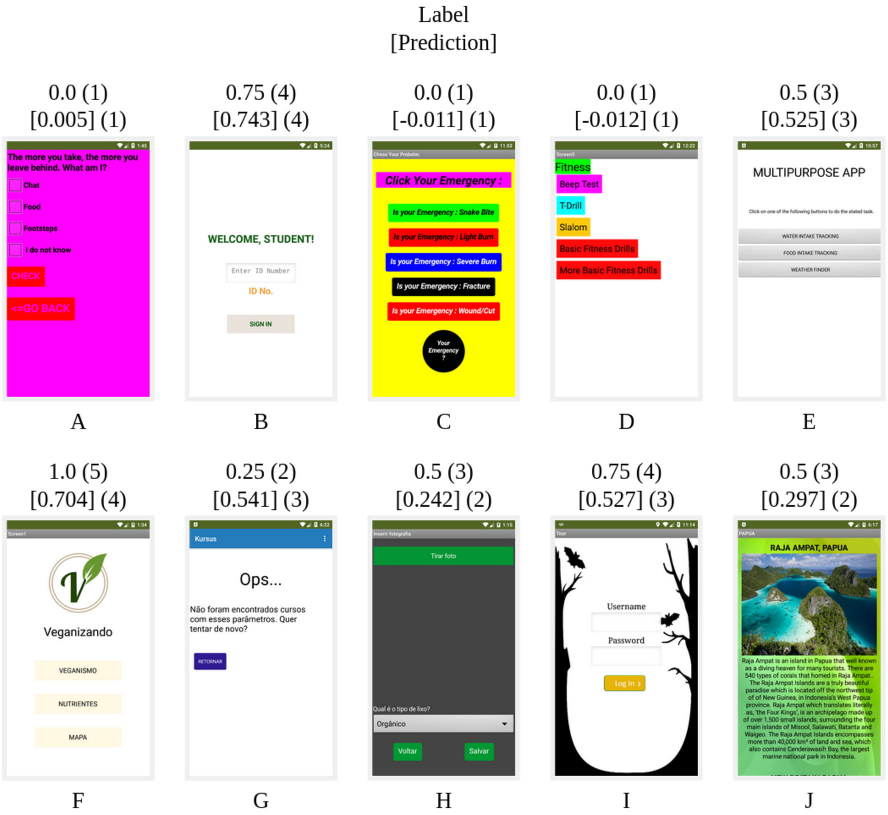
Source: the author.

for Spearman's rank correlation coefficient, rather than the more common correlation test Pearson's r, is justified by the non-normality of the data (BRYMAN; CRAMER, 1990). The numerical value of ρ ranges from -1 to +1. The closer the coefficients are to -1 or +1, the stronger the linear relationship is.

In the first iteration, the ResNet50 model trained with the fit strategy showed the best correlation between predictions and labels ($\rho = .87$) on the validation set (Figure 37, left). In the second iteration, that same model trained with a larger set (dataset 2) achieved a correlation of $\rho = .79$. When comparing the performance with a similar work, the correlation in the first iteration is on par with Dou et al. (2019), that report a correlation of $r = .85$ on the validation set. That work, however, used a dataset with a normal-like distribution, where most of the labels are close to the center of the scale and very few or no labels are close to its ends (Figure 38, left). Such a dataset can bias the model and improve the chance of getting the prediction right in the validation set, since it follows the same distribution and each label has a greater chance to be in the middle of the scale. We tried to use datasets that are as balanced as possible (Figure 38, right). The difficulty was finding samples on the upper end of the scale ("very beautiful" GUIs). Nonetheless, except for these GUIs, the model learned to classify all others with the same chance.

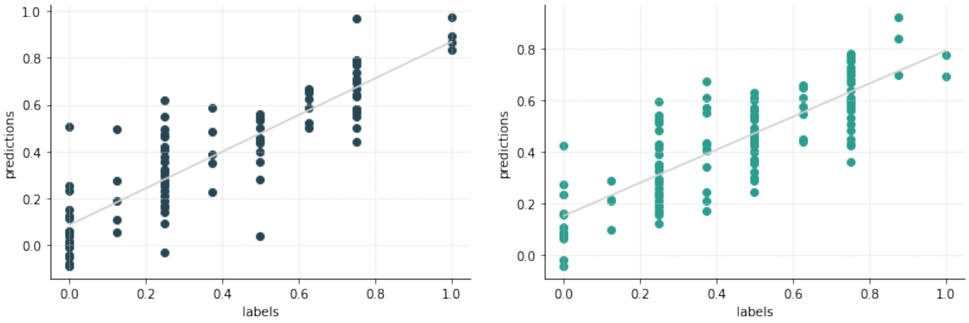
After both training iterations, our models performed very well predicting the visual esthetics of the screenshots in the test set (Figure 39). The ResNet50 model trained with

Figure 36 – Test set: best (top) and worst (bottom) predictions (values converted to the 5-point scale in parenthesis).



Source: the author.

Figure 37 – Correlation between labels and predictions in the first iteration (left) and in the second iteration (right) on the validation set.

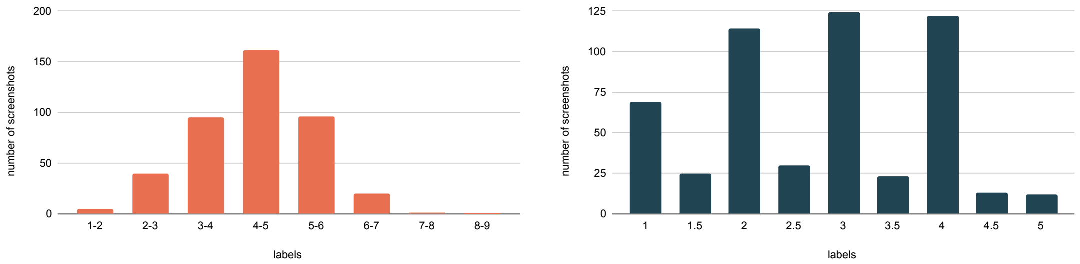


Source: the author.

the fit strategy showed a correlation of $\rho = .95$ in the first iteration and $\rho = .9$ in the second one ($\rho = .9$). This represents an excellent correlation, considering that the models were assessing images that they had not seen before.

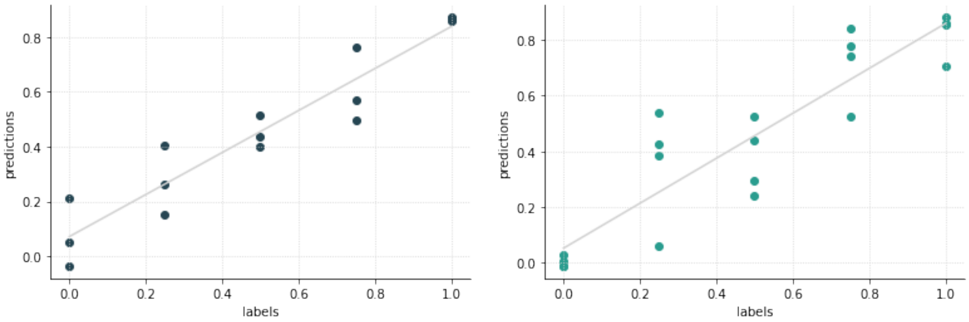
We also correlated the predictions with the labels on the set containing the removed screenshots (average absolute deviations equal to or greater than 1) in both training iterations. All models resulted in ρ between .50 and .66, indicating that our choice to remove those screenshots about which humans show a higher degree of disagreement on

Figure 38 – Distribution of samples on the validation set in Dou et al. (2019) (left) and in the second training iteration (right).



Source: the author.

Figure 39 – Correlation between labels and predictions in the first iteration (left) and in the second iteration (right) on the test set.



Source: the author.

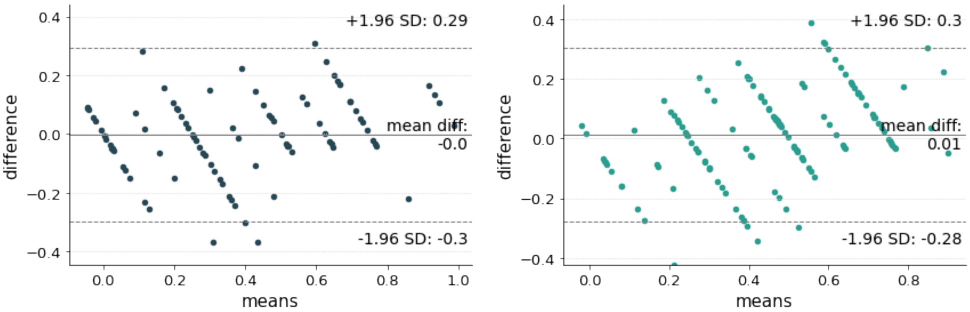
their visual aesthetics was correct.

4.3.2.2 Bland-Altman analysis

As correlation analysis shows the relationship between two variables, not their differences, we also performed a Bland-Altman (B&A) plot analysis to assess how they compare (GIAVARINA, 2015). The B&A plot analysis describes the agreement between two quantitative measurements by studying the mean difference and constructing limits of agreement. It allows us to evaluate a bias between the mean differences and estimate an agreement interval, within which 95% of the differences between the first and the second methods fall (BLAND; ALTMAN, 1986). This analysis does not indicate if the agreement between the predicted values and the human ratings is sufficient or if the automated assessment is suitable to replace the human one. It only quantifies the bias and a range of agreement, within which 95% of the differences between one measurement and the other are included (GIAVARINA, 2015). B&A recommends that 95% of the data points lie within ± 2 standard deviations of the mean difference.

Figure 40 shows the B&A charts. It can be seen that the average difference between labels and predictions is zero in the validation set in the first and .01 in the second iteration. That is an indication of bias absence predicting visual aesthetics in the validation sets.

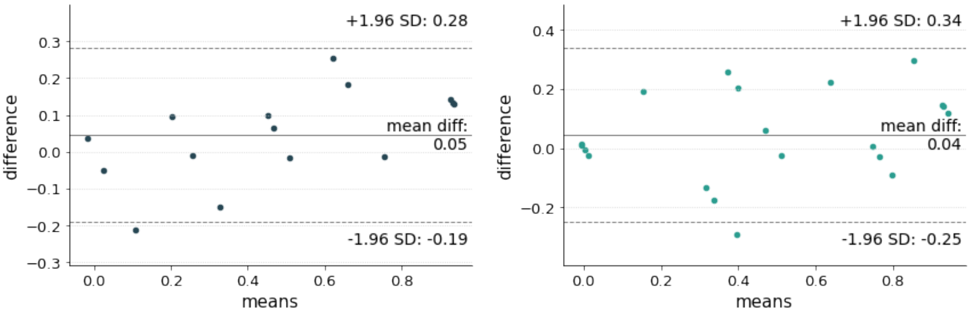
Figure 40 – B&A plot analysis on the validation set in the first iteration (left) in the second iteration (right)



Source: the author.

The confidence interval (CI) is within the expected range. In the first iteration, the CI varied from -.3 to .29, and in the second iteration, it ranged from -.28 to .3, showing that 95% of the predictions differ from the labels by just over one point or less on a five-point scale.

Figure 41 – B&A plot analysis on the test set in the first iteration (left) in the second iteration (right)



Source: the author.

For the test set, the labels are on average .04 larger than the predictions in the first iteration and .05 in the second one. This is less than half a point on a 5-point scale. However, it also shows a slight tendency of the model to assign a lower degree of visual aesthetics than humans. The CI in the first iteration (.-19 to .28) was smaller than in the second one (-.25 to .34). It shows that 95% of the predictions differ from the labels by just over one point or less on a 5-point scale.

4.4 MULTI-POINT REGRESSION

We executed a third training iteration so that we could compare the results of a model predicting a single score visual aesthetics degree with those of a model predicting a distribution of ratings received.

4.4.1 Model training

For this training, the input layer was adapted to the image resolution of our dataset, represented by the screenshot vector and its label. The original output layers representing a categorical variable with 1,000 values 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%)."

Again, we used the mean squared error (MSE) as the loss function. The output layers of the networks were transfer-trained for 100 epochs, which the previous iterations have shown to be enough for the validation error to stop improving. Each architecture was trained with two different training strategies, the 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 35).

Table 35 – Summary of the compared models for multi-point regression.

Architectures	VGG19, ResNet50, EfficientNet B0	
Input dimension of images	448 x 448 (pixels) x 3 (color channels)	
Predictive model	Multi-point regression (5-dimensional vector)	
Learning algorithm	Backpropagation	
Dataset separation	640 screenshots for training (80%); 160 screenshots for validation (20%)	
Training strategy	fit, fit1cycle	
Phase	Transfer learning	Fine tuning
Epochs	100	20
Learning rate	.003	range of optimum learning rates[5]
Weight decay	No	No

Source: the author.

After unfreezing the intermediate layers to allow all weights to adapt to our dataset, we trained each model for 20 more epochs using the same training strategy. The learning rates were defined with the method suggested by Smith and Topin (SMITH; TOPIN, 2019). All models improved their performances after the fine-tuning phase. The ResNet50 trained with the fit strategy achieved the best performance, even better than a ResNet50 model to predict the visual aesthetics score directly (Table 36). For this reason, we show the detailed results for this model only.

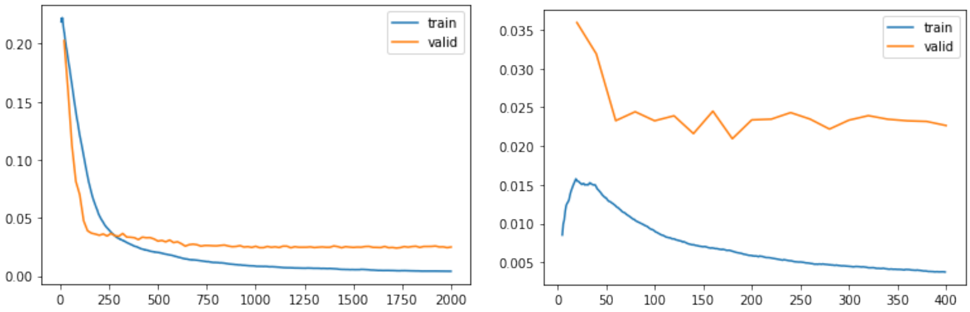
[5] (SMITH; TOPIN, 2019)

Table 36 – Best MSE for each model for multi-point regression.

		Transfer learning		Fine-tuning	
Architecture	**Prediction**	**Strategy**	**MSE**	**LR**	**MSE**
ResNet50	regression	fit	.026767	1e-04, 1e-03	.022049
		fit1cycle	.029282	1e-05, 1e-04	.027709
VGG19	histogram	fit	.02486	6.31e-07, 6.92e-05	.023841
		fit1cycle	.025081	6.31e-07, 5.75e-05	.022129
ResNet50	histogram	fit	.024029	1e-04, 1e-03	**.02095**
		fit1cycle	.025765	1e-05, 1e-04	.021359
EfficientNet B0	histogram	fit	.023992	1e-05, 1e-04	.023362
		fit1cycle	.024021	2e-04, 2e-05	.022814

Source: the author.

Figure 42 – Train and validation losses for ResNet50 transfer learning (left); fine-tuning (right)

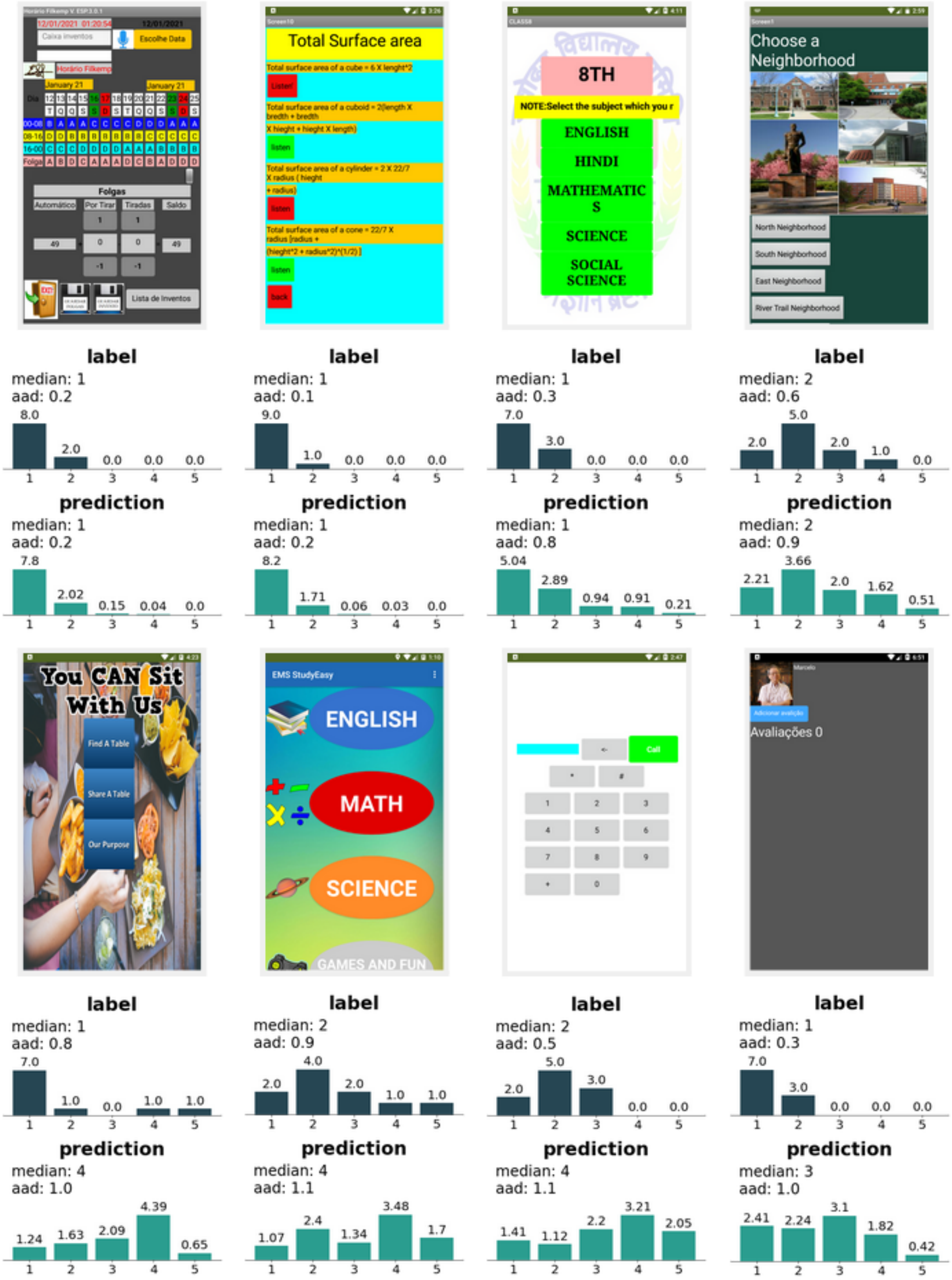


Source: the author.

These results demonstrate that the model performed well in predicting the visual aesthetics distributions. The medians from almost half of the predicted distributions (79 out of 190) were the same as those computed from the labels (Figure 43, top). 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 six distributions (3.2%) differed by one and a half points or more from their labels (Figure 43, bottom). The worst case was one GUI that was considered very ugly by most of the human raters, with seven ratings "1," but was assigned a median "4" from the predicted distribution, which is interpreted as "beautiful." Yet, observing such divergences in only a very small number of cases, these results can still be considered satisfactory since even humans strongly disagree about visual aesthetics sometimes (GRESSE VON WANGENHEIM et al., 2018b).

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

Figure 43 – Validation set: best (top) and worst (bottom) predictions.

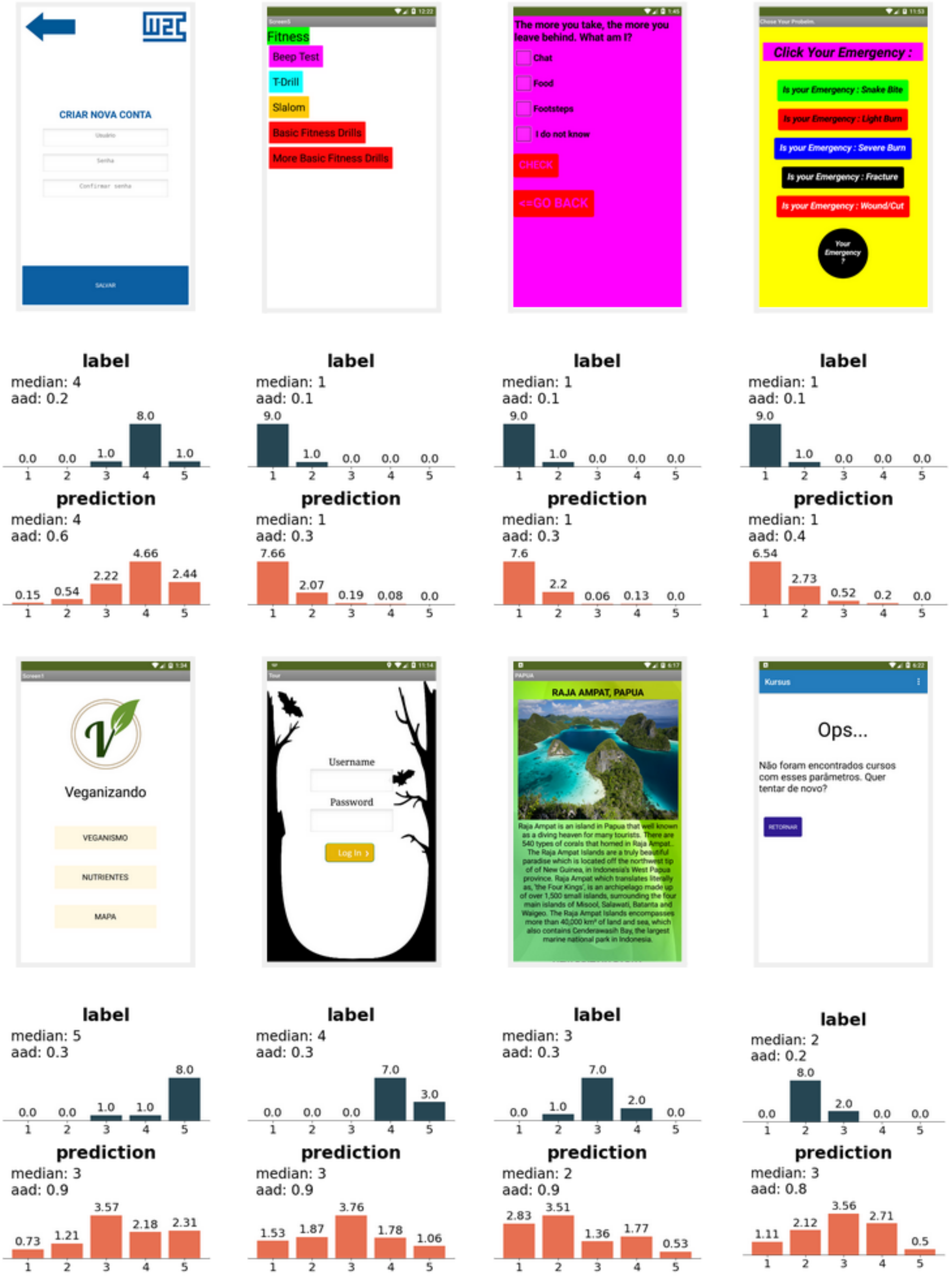


Source: the author.

the difficulty in assessing those GUIs.

A similar performance was observed with the test set, which contains new screenshots without labels (Figure 44). The median of the predicted distribution of one GUI was two points lower than its label (Figure 44, bottom). All other medians differed by one point or less, with half of the predictions resulting in the same 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.

Figure 44 – Test set: best (top) and worst (bottom) predictions.



RAJA AMPAT, PAPUA

Raja Ampat is an island in Papua that well known as a diving heaven for many touristts. There are 540 types of corals that homed in Raja Ampat.. The Raja Ampat Islands are a truly beautiful paradise which is located off the northwest tip of of New Guinea, in Indonesia's West Papua province. Raja Ampat which translates literally as, 'the Four Kings', is an archipelago made up of over 1,500 small islands, surrounding the four main islands of Misool, Salawati, Batanta and Waigeo. The Raja Ampat Islands encompasses more than 40,000 km² of land and sea, which also contains Cenderawasih Bay, the largest marine national park in Indonesia.

Ops...

Não foram encontrados cursos com esses parâmetros. Quer tentar de novo?

Source: the author.

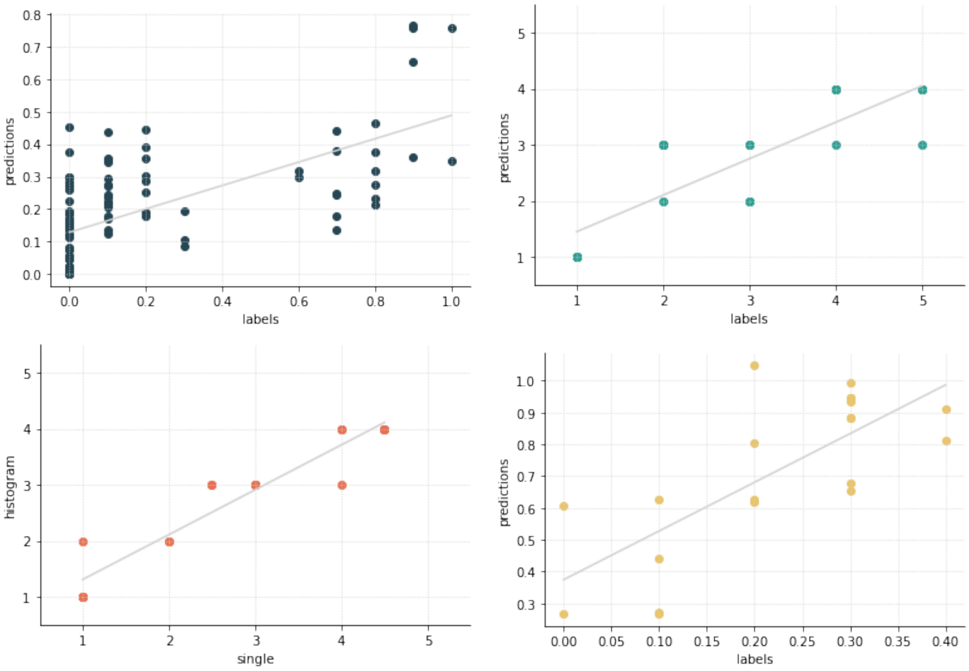
4.4.2 Model evaluation

We used the same methods, the Spearman rank correlation and the Bland-Altman plot analysis (B&A) to evaluate the performance of this model.

4.4.2.1 Correlation analysis

The correlation between the individual scores of the predicted distribution with those of the labels resulted in a moderate correlation ($\rho = .65$) (Figure 45, 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 45). The highest scores assigned by the model were .74 and .71.

Figure 45 – 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).

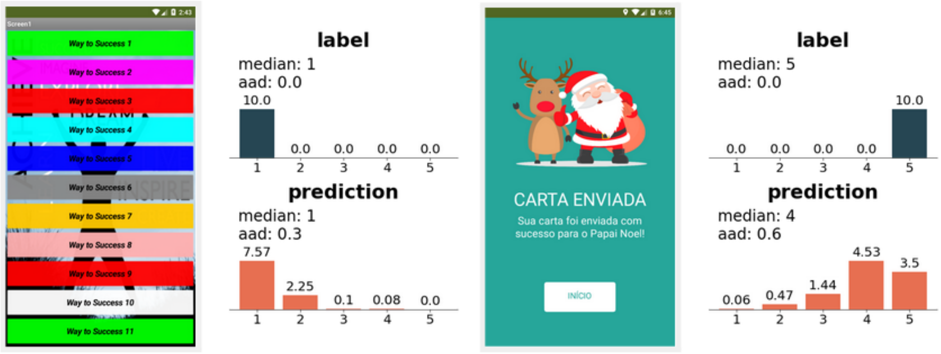


Source: the author.

We also compared the medians of predictions with the medians of labels (Figure 45, top right). The median is preferred against the mean score as a measure of central tendency for ordinal scales (NUNNALLY; BERNSTEIN, I. H., 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 (ρ = .85). In this case, the median of the predictions varied at most one point from the median of the labels. For example, all predictions for GUIs with visual aesthetics degree of “1” = “very ugly” were correct. On the other hand, all predictions for GUIs with visual aesthetics of “5” = “very beautiful” were wrong, 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 lower rates. This result is very similar to the model predicting the visual aesthetics directly (Figure 45, bottom left). When comparing the median from our results with the visual aesthetics degree directly predicted, the correlation was ρ = .97.

Comparing the AADs of labels and predictions (Figure 46, bottom right), again only a moderate correlation (ρ = .67) was obtained. Although this dataset contains GUIs with low AADs (.4 or lower), reflecting a high level of agreement between the human

Figure 46 – GUIs that generated consensus among raters.



Sua carta foi enviada com sucesso para o Papai Noel!

Source: the author.

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.

4.4.2.2 Bland-Altman analysis

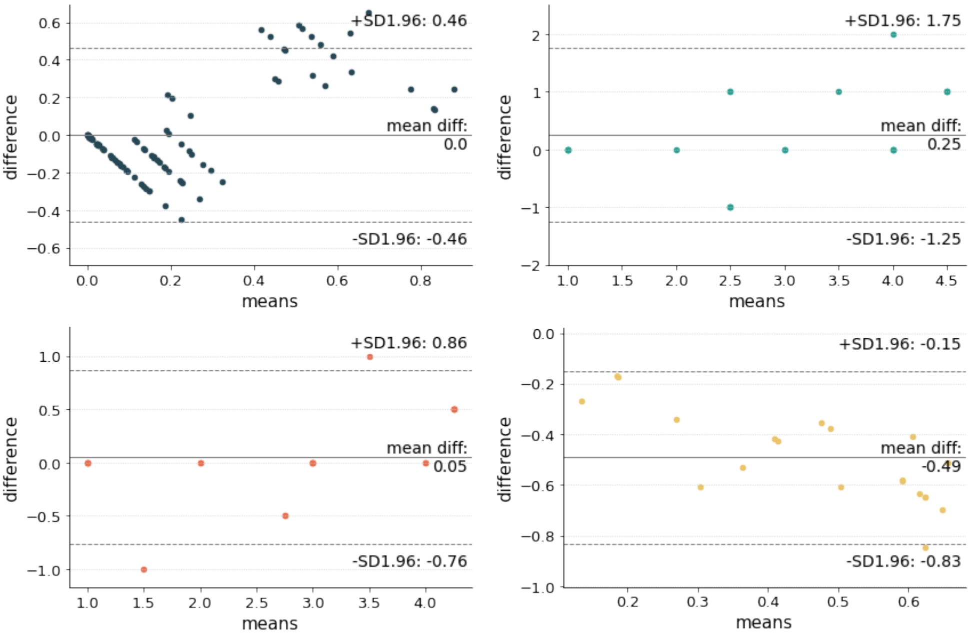
Analyzing the B&A plot for individual scores of the predicted distribution and those of the labels (Figure 47, top left), it is possible to observe that the mean difference between them is zero. This is an indication of bias absence in predicting the score distribution for the whole set. Nonetheless, 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 (Figure 48).

Figure 47 - 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)

Source: the author.

The B&A plot analysis for the medians of the predicted distributions and the medians of the labels shows a different scenario (Figure 47, 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 47, bottom left). The mean difference between them was half point on a 5-point scale and the CI was (-.76 to .86). For only two GUIs

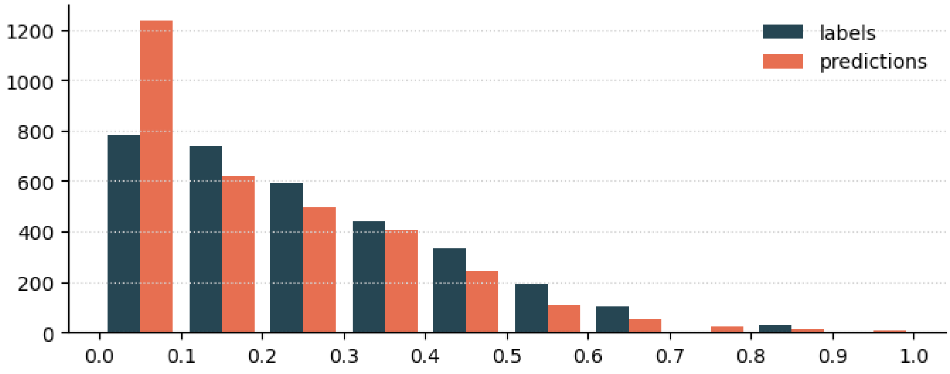
Figure 47 – B&A plot analysis: label distribution vs predicted distribution (top left); label median vs predicted median (top right); direct prediction vs predicted median (bottom left); label AAD vs predicted AAD (bottom right).



Source: the author.

the median for the predicted distribution was one point higher than directly predicted.

Figure 48 – Histogram of number of ratings received.



Source: the author.

The AAD is clearly biased on the B&A plot (Figure 47, bottom right). All differences are below zero, and the mean difference is -.49, confirming that the model output tends to have AADs higher than the labels.

5 DISCUSSION

Part of the challenge of assessing visual aesthetics lies in the subjective response of this type of judgment. Therefore, the optimal computational representation of visual aesthetics is far from obvious. Because people have different perceptions, even when they share a similar background, they tend to disagree about visual aesthetics (GRESSE VON WANGENHEIM et al., 2018b). 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"). Yet, analyzing the inter-rater agreement/reliability among labelers we observed a considerable high agreement and reliability among the human raters. And developing deep learning models using a single score the resulting model achieved an MSE below .022 predicting visual aesthetics in a single score, surpassing the precision of the assessment of web page GUIs (MSE = .042) reported by Dou et al. (2019). Our model also performed slightly better than the one assessing GUI designs (MSE = .0222) presented by Xing et al. (2021) that do not directly predict visual aesthetics, but indirect indicators of user aesthetic preference (number of likes received and number of collections to which the GUIs belong).

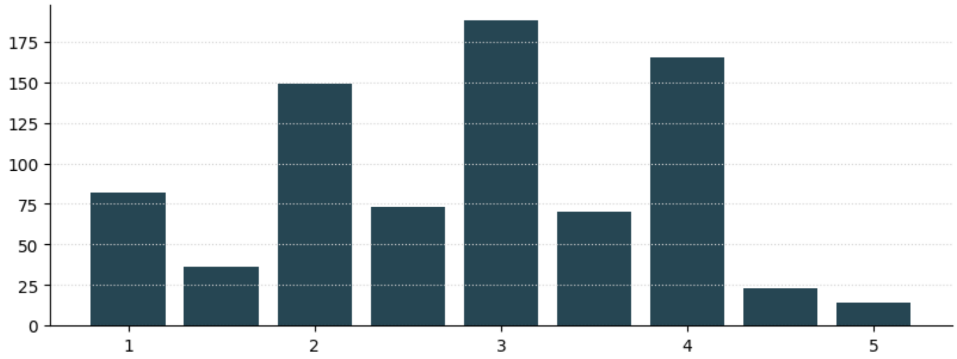
We agree with Dou et al. (2019) that formulating the problem as a regression task is a significant factor for this performance. Previous informal tests with classification models also yielded lower performance results. These evaluation results demonstrate that a convolutional neural network can learn to predict the visual aesthetics of mobile GUIs based on their screenshots. The predictions of our model showed an excellent correlation with the human ratings ($\rho = .9$), with the B&A plot analysis indicating that more than 95% of them agree, i.e., 19 out of the 20 outputs are within the 95% confidence interval. These results outperform other models assessing web GUIs (DOU et al., 2019; KHANI et al., 2016) in an unprecedented approach for mobile GUIs.

As a single score to represent visual aesthetics may hide how much disagreement a GUI can provoke before it is computed from all the individual assessments, we also experimented a deep learning model on visual aesthetics distributions. Using a distribution of received ratings not only allows 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. As a result, when predicting rating distributions, we achieved an even better performance (MSE = .0209) much above the similar ones reported on web GUIs in literature with the additional difference that we kept those samples with high AAD.

Yet, when 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 the labels. One possible reason for such a 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.

The model also seems to have trouble predicting the visual aesthetics of some GUIs on the higher end of the rating scale. For example, the model did not correctly predict the visual aesthetics scores for any of the four "very beautiful" GUIs in the test set. And as the B&A plot analyses indicate, the models tend to assign lower visual aesthetic scores to screenshots than humans. This might happen because there are very few "very beautiful" GUIs in the dataset (Figure 49). We detected this problem after the first training iterations and tried to mitigate it by adding more beautiful screenshots when composing dataset 2. Nonetheless, the human raters' preferences did not reflect the authors' and the new beautiful GUIs did not receive the highest ratings from the participants. 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 for 95% of screenshots on the five-point scale. This means that no GUI labeled as "4" ("beautiful") was predicted as "2" ("ugly") by 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") from the model predicting rating distributions.

Figure 49 – Distribution of the label medians in the dataset.



Source: the author.

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 assigning the visual aesthetics degree of mobile GUIs. Yet, the correlation between the medians of the predicted distributions and the label medians is below that of the model that directly assesses visual aesthetics as a single score.

It is also interesting to note the considerable difference in the correlations of the human assessments with the test set and with the one containing those GUIs removed from training in the first and second iterations due to the high mean absolute deviation in labeling. The test set consisted of GUIs on which human raters expressed a high degree of agreement as to their visual aesthetics. When assessing the screenshots from this group, the results showed a strong correlation with the human ratings ($\rho = .9$). The B&A analysis also indicated the agreement between the two assessment methods. On the other hand, the correlation between the model results and the set of GUIs excluded from the training was considerably lower ($\rho = .61$). In the third iteration, however, these GUIs were kept in the dataset because the labels were rating distributions. Differently from the labels used in the first two iterations that were derived from the individual ratings, the distributions represented the direct response from the participants and did not hide their degree of agreement. Anyway, these results show that just as humans have difficulty agreeing on the visual aesthetics of some GUIs, so does the deep learning model as expected. And although it represents an objective representation of visual aesthetics, it derives from the GUI properties and human intersubjectivity when rating them, composed of different subjective evaluations. This shows that the deep learning model is susceptible to the same difficulties humans face when assessing GUIs with conflicting or confusing aesthetic elements.

Examining the evaluation examples presented in Figure 36, we can also try to understand which design elements contribute to the visual aesthetics of GUIs. Those that received the lowest ratings, i.e., GUIs considered "very ugly," make heavy use of very saturated colors (A, C, and D). Long pieces of text also seem to reduce the visual aesthetics ratings (G and J). On the other hand, GUIs that received intermediate ratings have large areas of blank space (E, G, and H). Some higher-rated GUIs also use whitespace, with an additional contrast between colors much softer than on ugly screens (B, F, and I). We also observe that a lack of symmetry between the elements seems to contribute to lower visual aesthetics (A, D, and G). Finally, we noticed that GUIs with fewer large elements (B and F) receive better evaluations than GUIs with many small ones (A, C, and D). However, representing just a superficial analysis of this issue, the automatization of visual aesthetics can also support such an analysis in detail on a larger scale with reasonable effort.

Overall, both models present very similar results. On one hand, the model predicting rating distributions achieved lower MSE (.0209 vs. .022), while on the other hand, the model directly predicting the visual aesthetics as a single score presented a better correlation with the labels on unseen GUIs (.9 vs. .85). Nonetheless, the results from both models show a near-perfect correlation over the same test set ($\rho = .97$), indicating that no matter whether the visual aesthetics is predicted directly with a single score or computed from a predicted rating distribution, both models offer very similar outputs. Yet, from the rating distributions it is also possible to compute the degree of agreement of the raters, which is an additional information to identify GUIs that have conflicting characteristics (beautiful for some, ugly for others). For this reason, the model predicting score distributions is preferred over that predicting direct visual aesthetics degree.

Thus, although there already exist first proposals using deep learning to assess the visual aesthetics of web pages, with the results achieved in this research we evolve the current state of the art by focusing specifically on mobile GUIs, with different characteristics than web pages. This work also presents the first model to assess the visual aesthetics of GUIs (desktop, web, or mobile) as a distribution of individual visual aesthetic ratings for mobile GUIs. And, it also presents an innovative contribution in the education context of computing education enabling the automatic assessment of the visual aesthetics of GUIs of apps created with App Inventor.

Threats to validity. A potential threat to our results 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. Another threat comes from the subjective character of human classification during labeling. To reduce it, we analyzed the inter-rater agreement of the human responses and removed those screenshots about which the human raters disagreed on their visual aesthetics. 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. Also, based on related work, we chose well-tested CNN architectures that had been used 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 based on a test set (not previously used for training or validation) randomly chosen from the dataset.

6 CONCLUSION AND FUTURE WORK

The study of aesthetics is challenging not only because of its subjective component, but also because of the way that subjective responses are related with the actual properties of observed objects. This is a matter which we do not intend to solve as a philosophical question, but we expect to have contributed to the practical issue of assessing the visual aesthetics of GUIs in an educational context with our work.

The research on the state-of-the-art concerning the visual aesthetics assessment of mobile application interfaces revealed very few studies propose a method for this type of assessment. Those that do investigate visual aesthetics on mobile GUIs employ techniques for handcrafted feature extraction (ALEMERIEN; MAGEL, 2015; MINIUKOVICH; DE ANGELI, 2014b, 2015a; TABA et al., 2014). And although some other works assess visual aesthetics of GUIs using deep learning techniques, none of them are applied on mobile GUIs considering their unique features that differ from other types of GUIs. As an additional property of this type of assessment, no research that considered how much different users would agree on visual aesthetics of any GUI was found. These results demonstrate the originality of our research results.

Although this research deals with the automatic assessment of visual aesthetics, a preliminary task was to ensure that humans could adequately rate each GUI so that their labels would correctly correspond to the humans' aesthetic responses. Therefore, we executed a systematic mapping to understand how this subjective assessment of GUIs is typically done and also what instruments, such as questionnaires and scales, are most suitable for the task. We published these results in 2022 (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022a). Before the definition of the rating instrument, we executed an exploratory study to ensure it had high reliability and validity when compared with another well established instrument. In this way, the risk of imprecise subjective assessments when creating the datasets used for training the deep learning models was mitigated. We published that study in 2022 (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022b).

The subjective assessment of GUIs can be a long and tedious task depending on the number of samples. To mitigate the risk of jeopardizing this activity with fatigued and bored raters, we developed a mobile app, called GUI Labeler, using App Inventor with the intent of reducing the number of screen touches for each GUI and simplifying the task. The GUI Labeler also enabled the participants to stop at any time they felt tired and easily resume their work. This app can be easily adapted for the labeling of any image with a rating scale. Although the GUI Labeler did not present any type of error or crash during the labeling process, we expect to broaden its tests and present the results in a future article.

Before the training of the first models, we created a dataset with GUI screenshots that had generated a sufficient degree of agreement among the human raters regarding their visual aesthetics. The rationale behind this decision is that we could train these models only with GUIs that were clearly beautiful or ugly. This way, we could mitigate the difficulty of assessing GUIs with conflicting features. After the creation of the first dataset, we trained ResNets models with different depths to compare their performances. The ResNet50 consistently showed better performance when compared with the other models and also with similar works using deep learning to assess web GUIs. When analyzed against the human ratings on a new dataset containing only unseen GUIs, the ResNet50 showed excellent correlation and agreement, indicating that it was a strong candidate for automating visual aesthetics assessments. The results of our first models were published in 2022 (LIMA et al., 2022).

After the first results, we expanded the dataset with new GUIs, trying to balance the beautiful GUIS with the ugly ones. This iteration showed the difficulty of finding beautiful App Inventor GUIs despite our efforts to select the most beautiful ones. We again trained different ResNets models and confirmed the ResNet50 as the best performing model. Its performance was also superior to that of other architectures (VGG19 and EfficientNet B0) trained with the same dataset. With a larger dataset, the predictions kept the strong correlation and agreement with the labels, although it was a little lower. The results of this training iteration were submitted in 2023 (LIMA et al., n.d.), and the dataset is available online at `https://bit.ly/app-inventor-dataset-v2`.

As a final iteration, we adapted the input and output layers of our model to receive a 5-dimensional vector, representing the rating distribution, as the GUI labels. Besides allowing the computation of the visual aesthetics degree of GUIs, from the rating distribution it was also possible to extract the agreement degree among the human raters, supporting a richer analysis. Again, the ResNet50 demonstrated to be superior when compared to the VGG19 and the EfficientNet B0 and presented a performance that was very similar to the model trained to predict visual aesthetics as a single score, showing that it is suitable for automating this type of assessment. These results were submitted for publication in 2023 (LIMA; GRESSE VON WANGENHEIM, n.d.).

Now we are finally able to answer our research question. With a deep learning model that achieves an MSE = .021, a correlation between its results and unseen GUIs above .85 and more than 95% of the assessments differing in less than one point from the actual labels when automatically assessing mobile GUIs, we can state that **it is possible** to automate the assessment of the visual aesthetics of user interfaces of App Inventor applications with acceptable performance, reliability, and validity.

As future work, we expect to integrate the trained model with the tool Codemaster, a tool for the automated assessment of App Inventor applications (GRESSE VON WANGENHEIM et al., 2018a). With Codemaster, teachers and students can evaluate if computational and design principles are well implemented in their apps. After the integration, they will also be able to assess the visual aesthetics of their apps.

Table 37 – Summary of the research results.

Scientific contributions
Assessing the Visual Esthetics of User Interfaces: A Ten-Year Systematic Mapping (LIMA; GRESSE VON WANGENHEIM, 2021)
Assessment of Visual Aesthetics through Human Judgments: a Systematic Mapping (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022a)
Comparing Scales for the Assessment of Visual Aesthetics of Mobile GUIs Through Human Judgments (LIMA; GRESSE VON WANGENHEIM; BORGATTO, 2022b)
Automated Assessment of Visual aesthetics of Android User Interfaces with Deep Learning (LIMA et al., 2022)
A Deep Learning Model for the Assessment of the Visual Aesthetics of Mobile User Interfaces (LIMA et al., n.d.)
A Deep Learning Model for the Distribution of Visual Aesthetics Degree of Mobile User Interfaces (LIMA; GRESSE VON WANGENHEIM, n.d.)
Technological contributions
A mobile application for the labeling of screenshots `https://bit.ly/gui-labeler`
A dataset with 820 screenshots of apps designed with App Inventor labeled according to their visual aesthetics `https://bit.ly/app-inventor-dataset-v2`
A deep learning model for the direct assessment the visual aesthetics of App Inventor GUIs with a single score `https://bit.ly/appsthetics-code`
A deep learning model for visual aesthetics rating distribution of App Inventor GUIs `https://bit.ly/appsthetics-code`

Source: the author.

REFERÊNCIAS

ABBAS, Ali; HIRSCHFELD, Gerrit; THIELSCH, Meinald T. An Arabic Version of the Visual Aesthetics of Websites Inventory (AR-VisAWI): Translation and Psychometric Properties. **International Journal of Human–Computer Interaction**, v. 0, n. 0, p. 1–11, June 2022. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2085409. ISSN 1044-7318.

ABBASI, Maissom Qanber; WENG, Jingnong; WANG, Yunhong; RAFIQUE, Irfan; WANG, Xinran; LEW, Philip. Modeling and Evaluating User Interface Aesthetics Employing ISO 25010 Quality Standard. In: 2012 Eighth International Conference on the Quality of Information and Communications Technology. [S.l.: s.n.], Sept. 2012. P. 303–306.

AGARWAL, Anshu; HEDGE, Alan. The Impact of Web Page Usability Guideline Implementation on Aesthetics and Perceptions of the E-Retailer. en. **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, v. 52, n. 6, p. 528–532, Sept. 2008. Publisher: SAGE Publications Inc. ISSN 2169-5067.

AGGARWAL, Charu C. **Neural Networks and Deep Learning: A Textbook**. Cham: Springer International Publishing, 2018. ISBN 978-3-319-94462-3 978-3-319-94463-0.

ALEMERIEN, Khalid; MAGEL, Kenneth. GUIEvaluator: A Metric-tool for Evaluating the Complexity of Graphical User Interfaces. In: PROCEEDINGS of the Twenty-Sixth International Conference on Software Engineering & Knowledge Engineering. Vancouver, BC, Canada: [s.n.], 2014. P. 13–18.

ALEMERIEN, Khalid; MAGEL, Kenneth. SLC: a visual cohesion metric to predict the usability of graphical user interfaces. In: PROCEEDINGS of the 30th Annual ACM Symposium on Applied Computing. New York, NY, USA: Association for Computing Machinery, Apr. 2015. (SAC ’15), p. 1526–1533.

ALTABOLI, Ahamed; LIN, Yingzi. Effects of Unity of Form and Symmetry on Visual Aesthetics of Website Interface Design. en. **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, v. 56, n. 1, p. 728–732, Sept. 2012. Publisher: SAGE Publications Inc. ISSN 2169-5067.

ALTABOLI, Ahamed; LIN, Yingzi. Experimental Investigation of Effects of Balance, Unity, and Sequence on Interface and Screen Design Aesthetics. In: IADIS International

Conference Interfaces and Human Computer Interaction 2010 (part of MCCSIS 2010). [S.l.: s.n.], 2010. P. 243–250.

ALTABOLI, Ahamed; LIN, Yingzi. Investigating Effects of Screen Layout Elements on Interface and Screen Design Aesthetics. en. **Advances in Human-Computer Interaction**, v. 2011, p. 1–10, 2011. ISSN 1687-5893, 1687-5907.

ALTABOLI, Ahamed; LIN, Yingzi. Objective and Subjective Measures of Visual Aesthetics of Website Interface Design: The Two Sides of the Coin. en. In: JACKO, Julie A. (Ed.). **Human-Computer Interaction. Design and Development Approaches**. Berlin, Heidelberg: Springer, 2011. (Lecture Notes in Computer Science), p. 35–44.

ALVES, Nathalia da Cruz; GRESSE VON WANGENHEIM, Christiane; HAUCK, Jean Carlo Rossa. Approaches to Assess Computational Thinking Competences Based on Code Analysis in K-12 Education: A Systematic Mapping Study. en. **Informatics in Education**, v. 18, n. 1, p. 17–39, 2019. Publisher: Vilnius University Institute of Mathematics and Informatics, Lithuanian Academy of Sciences. ISSN 1648-5831.

AMERSHI, Saleema; BEGEL, Andrew; BIRD, Christian; DELINE, Robert; GALL, Harald; KAMAR, Ece; NAGAPPAN, Nachiappan; NUSHI, Besmira; ZIMMERMANN, Thomas. Software Engineering for Machine Learning: A Case Study. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). [S.l.: s.n.], May 2019. P. 291–300.

AMORIM, Ana Maria; BOECHAT, Glaucya; NOVAIS, Renato; VIEIRA, Vaninha; VILLELA, Karina. Quality Attributes Analysis in a Crowdsourcing-based Emergency Management System. In: p. 501–509.

ANATEL. **Painéis de Dados da Agência Nacional de Telecomunicações**. [S.l.: s.n.], 2023. Available from: `https://informacoes.anatel.gov.br/paineis/acessos/telefonia-movel`. Visited on: 4 Apr. 2023.

ANDERSON, Stephen P. **Seductive Interaction Design: Creating Playful, Fun, and Effective User Experiences**. 1st edition. Berkeley, CA: New Riders Pub, June 2011. ISBN 978-0-321-72552-3.

ANDREARCZYK, Vincent; WHELAN, Paul F. Chapter 4 - Deep Learning in Texture Analysis and Its Application to Tissue Image Classification. In: DEPEURSINGE, Adrien; S. AL-KADI, Omar; MITCHELL, J. Ross (Eds.). **Biomedical Texture Analysis**. [S.l.]: Academic Press, Jan. 2017. (The Elsevier and MICCAI Society Book Series). P. 95–129. ISBN 978-0-12-812133-7.

BAKAEV, Maxim; HEIL, Sebastian; CHIRKOV, Leonid; GAEDKE, Martin. Benchmarking Neural Networks-Based Approaches for Predicting Visual Perception of User Interfaces. en. In: DEGEN, Helmut; NTOA, Stavroula (Eds.). **Artificial Intelligence in HCI**. Cham: Springer International Publishing, 2022. (Lecture Notes in Computer Science), p. 217–231.

BAKAEV, Maxim; SPEICHER, Maximilian; HEIL, Sebastian; GAEDKE, Martin. I Don't Have That Much Data! Reusing User Behavior Models for Websites from Different Domains. en. In: BIELIKOVA, Maria; MIKKONEN, Tommi; PAUTASSO, Cesare (Eds.). **Web Engineering**. Cham: Springer International Publishing, 2020. (Lecture Notes in Computer Science), p. 146–162.

BALAGTAS-FERNANDEZ, Florence; FORRAI, Jenny; HUSSMANN, Heinrich. Evaluation of User Interface Design and Input Methods for Applications on Mobile Touch Screen Devices. en. In: GROSS, Tom; GULLIKSEN, Jan; KOTZÉ, Paula; OESTREICHER, Lars; PALANQUE, Philippe; PRATES, Raquel Oliveira; WINCKLER, Marco (Eds.). **Human-Computer Interaction – INTERACT 2009**. Berlin, Heidelberg: Springer, 2009. (Lecture Notes in Computer Science), p. 243–246.

BALINSKY, Helen. Evaluating interface aesthetics: measure of symmetry. In: DIGITAL Publishing. [S.l.]: SPIE, Feb. 2006. v. 6076, p. 52–63.

BARRICELLI, Barbara Rita; CASSANO, Fabio; FOGLI, Daniela; PICCINNO, Antonio. End-user development, end-user programming and end-user software engineering: A systematic mapping study. en. **Journal of Systems and Software**, v. 149, p. 101–137, Mar. 2019. ISSN 0164-1212.

BASILI, Victor R; CALDIEIRA, Gianluigi; ROMBACH, H Dieter. Goal Question Metric Paradigm. **Encyclopedia of software engineering**, p. 528–532, 1994. Publisher: John Weily and Sons.

BATTINA, Dhaya Sindhu. **Artificial Intelligence in Software Test Automation: A Systematic Literature Review**. en. Rochester, NY: [s.n.], Dec. 2019. Available from: `https://papers.ssrn.com/abstract=4004324`. Visited on: 6 Apr. 2023.

BAUERLY, Michael; LIU, Yili. Computational modeling and experimental investigation of effects of compositional elements on interface and design aesthetics. en. **International Journal of Human-Computer Studies**, v. 64, n. 8, p. 670–682, Aug. 2006. ISSN 1071-5819.

BAUERLY, Michael; LIU, Yili. Effects of Symmetry and Number of Compositional Elements on Interface and Design Aesthetics. **International Journal of Human–Computer Interaction**, v. 24, n. 3, p. 275–287, Mar. 2008. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447310801920508. ISSN 1044-7318.

BAUERLY, Michael; LIU, Yili. Evaluation and Improvement of Interface Aesthetics with an Interactive Genetic Algorithm. **International Journal of Human–Computer Interaction**, v. 25, n. 2, p. 155–166, 2009. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447310802629801. ISSN 1044-7318.

BAUGHAN, Amanda; AUGUST, Tal; YAMASHITA, Naomi; REINECKE, Katharina. Keep it Simple: How Visual Complexity and Preferences Impact Search Efficiency on Websites. In: PROCEEDINGS of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2020. P. 1–10. ISBN 978-1-4503-6708-0.

BHANDARI, Upasna; CHANG, Klarissa; NEBEN, Tillmann. Understanding the Impact of Perceived Visual Aesthetics on User Evaluations: An Emotional perspective. **Information & Management**, v. 56, n. 1, p. 85–93, 2019.

BHANDARI, Upasna; NEBEN, Tillmann; CHANG, Klarissa. Understanding Visual Appeal and Quality Perceptions of Mobile Apps: An Emotional Perspective. en. In: KUROSU, Masaaki (Ed.). **Human-Computer Interaction: Design and Evaluation**. Cham: Springer International Publishing, 2015. (Lecture Notes in Computer Science), p. 451–459.

BHANDARI, Upasna; NEBEN, Tillmann; CHANG, Klarissa; CHUA, Wen Yong. Effects of Interface Design Factors on Affective Responses and Quality Evaluations in Mobile Applications. **Computers in Human Behavior**, v. 72, p. 525–534, 2017. Place: Netherlands Publisher: Elsevier Science. ISSN 1873-7692(Electronic),0747-5632(Print).

BHATTACHERJEE, Anol. **Social Science Research: Principles, Methods, and Practices**. 2nd edition. [S.l.]: CreateSpace Independent Publishing Platform, Apr. 2012. ISBN 978-1-4751-4612-7.

BLAND, J. Martin; ALTMAN, Douglas G. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. en. **The Lancet**, v. 327, n. 8476, p. 307–310, 1986. ISSN 0140-6736.

BOLLINI, Letizia. Beautiful interfaces. From user experience to user interface design. **The Design Journal**, v. 20, sup1, s89–s101, July 2017. Publisher: Routledge _eprint: https://doi.org/10.1080/14606925.2017.1352649. ISSN 1460-6925.

BÖLTE, Jens; HÖSKER, Thomas M.; HIRSCHFELD, Gerrit; THIELSCH, Meinald T. Electrophysiological correlates of aesthetic processing of webpages: a comparison of experts and laypersons. en. **PeerJ**, v. 5, e3440, June 2017. Publisher: PeerJ Inc. ISSN 2167-8359.

BONETT, Douglas G.; WRIGHT, Thomas A. Sample size requirements for estimating pearson, kendall and spearman correlations. en. **Psychometrika**, v. 65, n. 1, p. 23–28, Mar. 2000. ISSN 1860-0980.

BOURGUET, Marie-Luce. Metrics-Based Evaluation of Graphical User Interface Aesthetics: The Segmentation Problem. In: PROCEEDINGS of the 2018 ACM Companion International Conference on Interactive Surfaces and Spaces. New York, NY, USA: Association for Computing Machinery, Nov. 2018. (ISS '18 Companion), p. 31–38.

BOYCHUK, Egor; BAKAEV, Maxim. Entropy and Compression Based Analysis of Web User Interfaces. en. In: BAKAEV, Maxim; FRASINCAR, Flavius; KO, In-Young (Eds.). **Web Engineering**. Cham: Springer International Publishing, 2019. (Lecture Notes in Computer Science), p. 253–261.

BRÜHLMANN, Florian; PETRALITO, Serge; RIESER, Denise C.; AESCHBACH, Lena F.; OPWIS, Klaus. TrustDiff: Development and Validation of a Semantic Differential for User Trust on the Web. **Journal of Usability Studies**, v. 16, n. 1, p. 29–48, Nov. 2020. ISSN 19313357.

BRYMAN, Alan; CRAMER, Duncan. **Quantitative Data Analysis for Social Scientists**. Florence, KY, US: Taylor & Francis/Routledge, 1990. (Quantitative data analysis for social scientists). Pages: xiv, 290. ISBN 978-0-415-02664-2 978-0-415-02665-9.

CAO, Jiuxin; MAO, Bo; LUO, Junzhou. A segmentation method for web page analysis using shrinking and dividing. **International Journal of Parallel, Emergent and Distributed Systems**, v. 25, n. 2, p. 93–104, Apr. 2010. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/17445760802429585. ISSN 1744-5760.

CHANG, D.; LI, F.; HUANG, L. A User-centered Evaluation and Redesign Approach for E-Government APP. In: 2020 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM). [S.l.: s.n.], Dec. 2020. P. 270–274.

CHETTAOUI, Neila; BOUHLEL, Med Salim. I2Evaluator: An Aesthetic Metric-Tool for Evaluating the Usability of Adaptive User Interfaces. en. In: HASSANIEN, Aboul Ella; SHAALAN, Khaled; GABER, Tarek; TOLBA, Mohamed F. (Eds.). **Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017**. Cham: Springer International Publishing, 2017. (Advances in Intelligent Systems and Computing), p. 374–383.

CHEVALIER, Aline; MAURY, Anne-Claire; FOUQUEREAU, Nicolas. The Influence of the Search Complexity and the Familiarity with the Website on the Subjective Appraisal of Aesthetics, Mental Effort and Usability. **Behaviour & Information Technology**, v. 33, n. 2, p. 117–132, 2014. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2013.819936. ISSN 0144-929X.

CHOI, Junho H.; LEE, Hye-Jin. Facets of Simplicity for the Smartphone Interface: A Structural Model. en. **International Journal of Human-Computer Studies**, v. 70, n. 2, p. 129–142, Feb. 2012. ISSN 1071-5819.

CHYUNG, Seung Youn (Yonnie); SWANSON, Ieva; ROBERTS, Katherine; HANKINSON, Andrea. Evidence-Based Survey Design: The Use of Continuous Rating Scales in Surveys. en. **Performance Improvement**, v. 57, n. 5, p. 38–48, 2018. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pfi.21763. ISSN 1930-8272.

COHEN, Louis; MANION, Lawrence; MORRISON, Keith. **Research Methods in Education**. 8th edition. London ; New York: Routledge, Nov. 2017. ISBN 978-1-138-20988-6.

CONKLIN, Sara M.; KOUBEK, Richard J.; THURMAN, James A.; NEWMAN, Leah C. The Effects of Aesthetics and Cognitive, Style on Perceived Usability. en. **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**, v. 50, n. 18, p. 2153–2157, Oct. 2006. Publisher: SAGE Publications Inc. ISSN 2169-5067.

COOMBS, Clyde H.; COOMBS, Lolagene C. "Don't Know": Item Ambiguity or Respondent Uncertainty? **Public Opinion Quarterly**, v. 40, n. 4, p. 497–514, Jan. 1976. ISSN 0033-362X.

COUPER, Mick P.; TOURANGEAU, Roger; CONRAD, Frederick G.; SINGER, Eleanor. Evaluating the Effectiveness of Visual Analog Scales: A Web Experiment. en. **Social Science Computer Review**, v. 24, n. 2, p. 227–245, May 2006. Publisher: SAGE Publications Inc. ISSN 0894-4393.

COX, Eli P. The Optimal Number of Response Alternatives for a Scale: A Review. en. **Journal of Marketing Research**, v. 17, n. 4, p. 407–422, Nov. 1980. Publisher: SAGE Publications Inc. ISSN 0022-2437.

DAMES, Hannah; HIRSCHFELD, Gerrit; SACKMANN, Timo; THIELSCH, Meinald T. Searching vs. Browsing—The Influence of Consumers' Goal Directedness on Website Evaluations. **Interacting with Computers**, v. 31, n. 1, p. 95–112, Jan. 2019. ISSN 0953-5438.

DAUBECHIES, Ingrid. **Ten lectures on wavelets**. [S.l.: s.n.], Jan. 1992. Publication Title: CBMS-NSF regional conference series in applied mathematics ADS Bibcode: 1992tlw..conf.....D.

DE ANGELI, Antonella; SUTCLIFFE, Alistair; HARTMANN, Jan. Interaction, usability and aesthetics: what influences users' preferences? In: PROCEEDINGS of the 6th conference on Designing Interactive systems. New York, NY, USA: Association for Computing Machinery, June 2006. (DIS '06), p. 271–280.

DENG, Yubin; LOY, Chen Change; TANG, Xiaoou. Image Aesthetic Assessment: An experimental survey. **IEEE Signal Processing Magazine**, v. 34, n. 4, p. 80–106, July 2017. Conference Name: IEEE Signal Processing Magazine. ISSN 1558-0792.

DJAJADININGRAT, Tom; WENSVEEN, Stephan; FRENS, Joep; OVERBEEKE, Kees. Tangible products: redressing the balance between appearance and action. en. **Personal and Ubiquitous Computing**, v. 8, n. 5, p. 294–309, Sept. 2004. ISSN 1617-4917.

DJAMASBI, Soussan; SIEGEL, Marisa; SKORINKO, Jeanine; TULLIS, Tom. Online Viewing and Aesthetic Preferences of Generation Y and the Baby Boom Generation: Testing User Web Site Experience Through Eye Tracking. **International Journal of**

Electronic Commerce, v. 15, n. 4, p. 121–158, 2011. Publisher: Routledge _eprint: https://doi.org/10.2753/JEC1086-4415150404. ISSN 1086-4415.

DJAMASBI, Soussan; SIEGEL, Marisa; TULLIS, Tom; DAI, Rui. Efficiency, Trust, and Visual Appeal: Usability Testing through Eye Tracking. In: 2010 43rd Hawaii International Conference on System Sciences. [S.l.: s.n.], Jan. 2010. P. 1–10. ISSN: 1530-1605.

DOMOFF, Sarah E.; RADESKY, Jenny S.; HARRISON, Kristen; RILEY, Hurley; LUMENG, Julie C.; MILLER, Alison L. A Naturalistic Study of Child and Family Screen Media and Mobile Device Use. en. **Journal of Child and Family Studies**, v. 28, n. 2, p. 401–410, Feb. 2019. ISSN 1573-2843.

DOSOVITSKIY, Alexey et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. **arXiv:2010.11929 [cs]**, June 2021. arXiv: 2010.11929.

DOU, Qi; ZHENG, Xianjun Sam; SUN, Tongfang; HENG, Pheng-Ann. Webthetics: Quantifying Webpage Aesthetics with Deep Learning. en. **International Journal of Human-Computer Studies**, v. 124, p. 56–66, Apr. 2019. ISSN 1071-5819.

DOUNEVA, Maria; JARON, Rafael; THIELSCH, Meinald T. Effects of Different Website Designs on First Impressions, Aesthetic Judgements and Memory Performance after Short Presentation. **Interacting with Computers**, v. 28, n. 4, p. 552–567, June 2016. ISSN 0953-5438.

DUPUY-CHESSA, Sophie; LAURILLAU, Yann; CÉRET, Eric. Considering aesthetics and usability temporalities in a model based development process. In: ACTES de la 28ième conference francophone sur l'Interaction Homme-Machine. New York, NY, USA: Association for Computing Machinery, Oct. 2016. (IHM '16), p. 25–35.

FALIAGKA, Evanthia; LALOU, Eleni; RIGOU, Maria; SIRMAKESSIS, Spiros. Usability and Aesthetics: The Case of Architectural Websites. en. In: KUROSU, Masaaki (Ed.). **Human-Computer Interaction: Users and Contexts**. Cham: Springer International Publishing, 2015. (Lecture Notes in Computer Science), p. 54–64.

FLORA, Harleen K.; WANG, Xiaofeng; CHANDE, Swati V. An Investigation into Mobile Application Development Processes: Challenges and Best Practices. en. **International Journal of Modern Education and Computer Science (IJMECS)**, v. 6, n. 6, p. 1, 2014.

FOGG, B. J.; SOOHOO, Cathy; DANIELSON, David R.; MARABLE, Leslie; STANFORD, Julianne; TAUBER, Ellen R. How do users evaluate the credibility of Web sites? a study with over 2,500 participants. In: PROCEEDINGS of the 2003 conference on Designing for user experiences. New York, NY, USA: Association for Computing Machinery, June 2003. (DUX '03), p. 1–15.

FONNEGRA, Rubén D.; BLAIR, Bryan; DÍAZ, Gloria M. Performance Comparison of Deep Learning Frameworks in Image Classification Problems Using Convolutional and Recurrent Networks. In: 2017 IEEE Colombian Conference on Communications and Computing (COLCOM). [S.l.: s.n.], Aug. 2017. P. 1–6.

GALITZ, Wilbert O. **The Essential Guide to User Interface Design: An Introduction to GUI Design Principles and Techniques**. [S.l.]: John Wiley & Sons, Apr. 2007. Google-Books-ID: Q3Xp_Awu49sC. ISBN 978-0-470-14622-4.

GAO, Fei; LI, Ziyun; YU, Jun; YU, Junze; HUANG, Qingming; TIAN, Qi. Style-adaptive Photo Aesthetic Rating Via Convolutional Neural Networks and Multi-task Learning. en. **Neurocomputing**, v. 395, p. 247–254, 2020. ISSN 0925-2312.

GARLAND, Ron. The Mid-Point on a Rating Scale: Is it Desirable? en. **Marketing Bulletin**, v. 2, n. 1, p. 66–70, 1991.

GIAVARINA, Davide. Understanding Bland Altman Analysis. eng. **Biochemia Medica**, v. 25, n. 2, p. 141–151, 2015. ISSN 1330-0962.

GIL, Antonio Carlos. **Métodos e Técnicas de Pesquisa Social**. 6ª edição. São Paulo: Atlas, July 2008. ISBN 978-85-224-5142-5.

GISEV, Natasa; BELL, J. Simon; CHEN, Timothy F. Interrater Agreement and Interrater Reliability: Key Concepts, Approaches, and Applications. en. **Research in Social and Administrative Pharmacy**, v. 9, n. 3, p. 330–338, May 2013. ISSN 1551-7411.

GREESON, Johanna K. P.; TREGLIA, Daniel; MORONES, Seth; HOPKINS, Marcia; MIKELL, Dominique. Youth Matters: Philly (YMP): Development, usability, usefulness, & accessibility of a mobile web-based app for homeless and unstably housed youth. en. **Children and Youth Services Review**, v. 108, p. 104586, Jan. 2020. ISSN 0190-7409.

GRESSE VON WANGENHEIM, Christiane; HAUCK, Jean Carlo Rossa; DEMETRIO, Matheus Faustino; PELLE, Rafael; ALVES, Nathalia da Cruz; AZEVEDO, Luiz Felipe; BARBOSA, Heliziane. CodeMaster - Automatic Assessment and Grading of App Inventor and Snap! Programs. English. **Informatics in Education - An International Journal**, v. 17, n. 1, p. 117–150, 2018. Publisher: Vilniaus Universiteto Leidykla. ISSN 1648-5831.

GRESSE VON WANGENHEIM, Christiane; HAUCK, Jean Carlo Rossa; PACHECO, Fernando S.; BERTONCELI BUENO, Matheus F. Visual tools for teaching machine learning in K-12: A ten-year systematic mapping. en. **Education and Information Technologies**, v. 26, n. 5, p. 5733–5778, Sept. 2021. ISSN 1573-7608.

GRESSE VON WANGENHEIM, Christiane; PORTO, João V. Araujo; HAUCK, Jean C. R.; BORGATTO, Adriano F. Do We Agree on User Interface Aesthetics of Android Apps? **arXiv**, p. 1–5, Dec. 2018. arXiv: 1812.09049.

GU, Zhenyu; JIN, Chenhao; CHANG, Danny; ZHANG, Liqun. Predicting Webpage Aesthetics with Heatmap Entropy. **Behaviour & Information Technology**, v. 40, n. 7, p. 676–690, Jan. 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2020.1717626. ISSN 0144-929X.

HALL, Richard H; HANNA, Patrick. The impact of web page text-background colour combinations on readability, retention, aesthetics and behavioural intention. **Behaviour & Information Technology**, v. 23, n. 3, p. 183–195, May 2004. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01449290410001669932. ISSN 0144-929X.

HAMBORG, Kai-Christoph; HÜLSMANN, Julia; KASPAR, Kai. The Interplay between Usability and Aesthetics: More Evidence for the “What Is Usable Is Beautiful” Notion. en. **Advances in Human-Computer Interaction**, v. 2014, e946239, Nov. 2014. Publisher: Hindawi. ISSN 1687-5893.

HARTMANN, Jan. Assessing the Attractiveness of Interactive Systems. In: CHI ’06 Extended Abstracts on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2006. (CHI EA ’06), p. 1755–1758.

HARTMANN, Jan; SUTCLIFFE, Alistair; DE ANGELI, Antonella. Investigating Attractiveness in Web User Interfaces. In: PROCEEDINGS of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2007. P. 387–396. ISBN 978-1-59593-593-9.

HARTMANN, Jan; SUTCLIFFE, Alistair; DE ANGELI, Antonella. Towards a Theory of User Judgment of Aesthetics and User Interface Quality. **ACM Transactions on Computer-Human Interaction**, v. 15, n. 4, 15:1–15:30, 2008. ISSN 1073-0516.

HASLER, David; SUESSTRUNK, Sabine E. Measuring colorfulness in natural images. In: HUMAN Vision and Electronic Imaging VIII. [S.l.]: SPIE, June 2003. v. 5007, p. 87–95.

HASSENZAHL, Marc. The Interplay of Beauty, Goodness, and Usability in Interactive Products. **Human–Computer Interaction**, v. 19, n. 4, p. 319–349, Dec. 2004. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1207/s15327051hci1904_2. ISSN 0737-0024.

HASSENZAHL, Marc; BURMESTER, Michael; KOLLER, Franz. AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität. In: SZWILLUS, Gerd; ZIEGLER, Jürgen (Eds.). **Mensch & Computer 2003: Interaktion in Bewegung**. Wiesbaden: Vieweg+Teubner Verlag, 2003. (Berichte des German Chapter of the ACM). P. 187–196. ISBN 978-3-322-80058-9.

HASSENZAHL, Marc; MONK, Andrew. The Inference of Perceived Usability From Beauty. **Human–Computer Interaction**, v. 25, n. 3, p. 235–260, Aug. 2010. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/07370024.2010.500139. ISSN 0737-0024.

HE, Kaiming; ZHANG, Xiangyu; REN, Shaoqing; SUN, Jian. Deep Residual Learning for Image Recognition. In: p. 770–778.

HEALE, Roberta; TWYCROSS, Alison. Validity and Reliability in Quantitative Studies. en. **Evidence-Based Nursing**, v. 18, n. 3, p. 66–67, July 2015. Publisher: Royal College of Nursing Section: Research made simple. ISSN 1367-6539, 1468-9618.

HOWARD, Jeremy; GUGGER, Sylvain. Fastai: A Layered API for Deep Learning. en. **Information**, v. 11, n. 2, p. 108, Feb. 2020. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

HU, Jie; SHEN, Li; ALBANIE, Samuel; SUN, Gang; WU, Enhua. **Squeeze-and-Excitation Networks**. [S.l.]: arXiv, May 2019. arXiv:1709.01507 [cs]. Available from: `http://arxiv.org/abs/1709.01507`. Visited on: 8 Nov. 2022.

HUANG, Shih-Miao. A Study of Affective Meanings Predicting Aesthetic Preferences of Interactive Skins. In: 2013 IEEE International Conference on Industrial Engineering and Engineering Management. [S.l.: s.n.], Dec. 2013. P. 781–785. ISSN: 2157-362X.

HUISMAN, Gijs; HOUT, Marco van; DIJK, Elisabeth van; GEEST, Thea van der; HEYLEN, Dirk. LEMtool: measuring emotions in visual interfaces. In: PROCEEDINGS of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2013. (CHI '13), p. 351–360.

HUYNH-THU, Quan; GARCIA, Marie-Neige; SPERANZA, Filippo; CORRIVEAU, Philip; RAAKE, Alexander. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. **IEEE Transactions on Broadcasting**, v. 57, n. 1, p. 1–14, Mar. 2011. Conference Name: IEEE Transactions on Broadcasting. ISSN 1557-9611.

IAKOVIDOU, Chryssanthi; ANAGNOSTOPOULOS, Nektarios; KAPOUTSIS, Athanasios Ch.; BOUTALIS, Yiannis; CHATZICHRISTOFIS, Savvas A. Searching images with MPEG-7 (& MPEG-7-like) Powered Localized dEscriptors: The SIMPLE answer to effective Content Based Image Retrieval. In: 2014 12th International Workshop on Content-Based Multimedia Indexing (CBMI). [S.l.: s.n.], June 2014. P. 1–6. ISSN: 1949-3991.

IOFFE, Sergey; SZEGEDY, Christian. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. en. In: PROCEEDINGS of the 32nd International Conference on Machine Learning. [S.l.]: PMLR, June 2015. P. 448–456. ISSN: 1938-7228.

ISMAIL, Nor Anita Fairos. **Classification of the Aesthetics Formula and Generating the Screen Designs for Balance Principle**. 2002. Doctoral Dissertation – Universiti Putra Malaysia.

ISO. **ISO 9241-11:2018(en), Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts**. [S.l.: s.n.], 2018. Available from: `https://www.iso.org/obp/ui/#iso:std:iso:9241:-11:ed-2:v1:en`. Visited on: 1 Aug. 2021.

ISO. **ISO/IEC 25010:2011, Systems and Software Engineering — Systems and Software Quality Requirements and Evaluation (SQuaRE) — System and Software Quality Models**. en. [S.l.: s.n.], 2011. Available from:

`https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/57/35733.html`. Visited on: 1 Aug. 2021.

ITEN, Glena H; TROENDLE, Antonin; OPWIS, Klaus. Aesthetics in Context—The Role of Aesthetics and Usage Mode for a Website's Success. **Interacting with Computers**, v. 30, n. 2, p. 133–149, Mar. 2018. ISSN 0953-5438.

IVANOV, Alex; SCHNEIDER, Christoph. The Effects of Perceived Visual Aesthetics on Process Satisfaction in GSS Use. In: 2010 43rd Hawaii International Conference on System Sciences. [S.l.: s.n.], Jan. 2010. P. 1–10. ISSN: 1530-1605.

JYLHÄ, Henrietta; HAMARI, Juho. Development of Measurement Instrument for Visual Qualities of Graphical User Interface Elements (VISQUAL): A Test in the Context of Mobile Game Icons. en. **User Modeling and User-Adapted Interaction**, v. 30, n. 5, p. 949–982, Nov. 2020. ISSN 1573-1391.

KARAYEV, Sergey; TRENTACOSTE, Matthew; HAN, Helen; AGARWALA, Aseem; DARRELL, Trevor; HERTZMANN, Aaron; WINNEMOELLER, Holger. Recognizing Image Style. **Proceedings of the British Machine Vision Conference 2014**, p. 122.1–122.11, 2014. arXiv: 1311.3715.

KHANI, Masoud Ganj; MAZINANI, Mohammad Reza; FAYYAZ, Mohsen; HOSEINI, Mojtaba. A Novel Approach for Website Aesthetic Evaluation Based on Convolutional Neural Networks. In: PROCEEDINGS of the 2016 Second International Conference on Web Research (ICWR). [S.l.: s.n.], Apr. 2016. P. 48–53.

KIMBERLIN, Carole L.; WINTERSTEIN, Almut G. Validity and Reliability of Measurement Instruments Used in Research. **American Journal of Health-System Pharmacy**, v. 65, n. 23, p. 2276–2284, Dec. 2008. ISSN 1079-2082.

KIRCHNER, Jens; HEBERLE, Andreas; LÖWE, Welf. Classification vs. Regression - Machine Learning Approaches for Service Recommendation Based on Measured Consumer Experiences. In: 2015 IEEE World Congress on Services. [S.l.: s.n.], June 2015. P. 278–285. ISSN: 2378-3818.

KLARE, George R. Understandability and Indefinite Answers to Public Opinion Questions. **International Journal of Opinion and Attitude Research**, v. 4, n. 1, p. 91–96, 1950.

KO, Chihhsiang; LIU, Yuchun. Old and Young Users' White Space Preferences for Online News Web Pages. **IEEE Access**, v. 7, p. 57284–57297, 2019. Conference Name: IEEE Access. ISSN 2169-3536.

KOLESNIKOV, Alexander; BEYER, Lucas; ZHAI, Xiaohua; PUIGCERVER, Joan; YUNG, Jessica; GELLY, Sylvain; HOULSBY, Neil. Big Transfer (BiT): General Visual Representation Learning. en. In: VEDALDI, Andrea; BISCHOF, Horst; BROX, Thomas; FRAHM, Jan-Michael (Eds.). **Computer Vision – ECCV 2020**. Cham: Springer International Publishing, 2020. (Lecture Notes in Computer Science), p. 491–507.

KONG, Qing; GUO, Qi. Comprehensive Evaluation Method of Interface Elements Layout Aesthetics Based on Improved AHP. en. In: REBELO, Francisco; SOARES, Marcelo M. (Eds.). **Advances in Ergonomics in Design**. Cham: Springer International Publishing, 2019. (Advances in Intelligent Systems and Computing), p. 509–520.

KONONENKO, Igor; KUKAR, Matjaž. Chapter 11 - Artificial Neural Networks. In: KONONENKO, Igor; KUKAR, Matjaž (Eds.). **Machine Learning and Data Mining**. [S.l.]: Woodhead Publishing, Jan. 2007. P. 275–320. ISBN 978-1-904275-21-3.

KOO, Terry K.; LI, Mae Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. en. **Journal of Chiropractic Medicine**, v. 15, n. 2, p. 155–163, June 2016. ISSN 1556-3707.

KOTTNER, Jan; AUDIGE, Laurent; BRORSON, Stig; DONNER, Allan; GAJEWSKI, Byron J.; HRÓBJARTSSON, Asbjørn; ROBERTS, Chris; SHOUKRI, Mohamed; STREINER, David L. Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. en. **International Journal of Nursing Studies**, v. 48, n. 6, p. 661–671, June 2011. ISSN 0020-7489.

KOUTSABASIS, Panayiotis; ISTIKOPOULOU, Theano G. Perceived Website Aesthetics by Users and Designers: Implications for Evaluation Practice. en. **International Journal of Technology and Human Interaction (IJTHI)**, v. 9, n. 2, p. 39–52, 2013. Publisher: IGI Global. ISSN 1548-3908.

KRIZHEVSKY, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. ImageNet classification with deep convolutional neural networks. **Communications of the ACM**, v. 60, n. 6, p. 84–90, May 2017. ISSN 0001-0782.

KÜCHLER, Louisa; HERTEL, Guido; THIELSCH, Meinald T. Are You Willing to Donate? Relationship Between Perceived Website Design, Trust and Donation Decisions Online. In: PROCEEDINGS of the Conference on Mensch und Computer. New York, NY, USA: Association for Computing Machinery, Sept. 2020. (MuC '20), p. 223–227.

KUMAR, Bimal Aklesh; MOHITE, Priya. Usability of Mobile Learning Applications: A Systematic Literature Review. en. **Journal of Computers in Education**, v. 5, n. 1, p. 1–17, Mar. 2018. ISSN 2197-9995.

LARCO, Andrés; MONTENEGRO, Carlos; DIAZ, Esteban; LUJÁN-MORA, Sergio. Underlying Quality Factors in Spanish Language Apps for People with Disabilities. In: 2018 International Conference on eDemocracy eGovernment (ICEDEG). [S.l.: s.n.], Apr. 2018. P. 110–116. ISSN: 2573-1998.

LAVIE, Talia; TRACTINSKY, Noam. Assessing Dimensions of Perceived Visual Aesthetics of Web Sites. en. **International Journal of Human-Computer Studies**, v. 60, n. 3, p. 269–298, Mar. 2004. ISSN 10715819.

LAZARD, Allison J.; KING, Andy J. Objective Design to Subjective Evaluations: Connecting Visual Complexity to Aesthetic and Usability Assessments of eHealth. **International Journal of Human–Computer Interaction**, v. 36, n. 1, p. 95–104, Jan. 2020. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2019.1606976. ISSN 1044-7318.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep Learning. en. **Nature**, v. 521, n. 7553, p. 436–444, May 2015. ISSN 1476-4687.

LEWIS, James R. Measuring Perceived Usability: The CSUQ, SUS, and UMUX. **International Journal of Human–Computer Interaction**, v. 34, n. 12, p. 1148–1156, Dec. 2018. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2017.1418805. ISSN 1044-7318.

LEWIS, James R. Measuring User Experience With 3, 5, 7, or 11 Points: Does It Matter? en. **Human Factors**, v. 63, n. 6, p. 999–1011, Sept. 2021. Publisher: SAGE Publications Inc. ISSN 0018-7208.

LEWIS, James R.; ERDINÇ, Oğuzhan. User Experience Rating Scales with 7, 11, or 101 Points: Does It Matter? **Journal of Usability Studies**, v. 12, n. 2, p. 73–91, Feb. 2017. ISSN 1931-3357.

LEYS, Christophe; LEY, Christophe; KLEIN, Olivier; BERNARD, Philippe; LICATA, Laurent. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. en. **Journal of Experimental Social Psychology**, v. 49, n. 4, p. 764–766, 2013. ISSN 0022-1031.

LI, X; ZHANG, G; LI, K; ZHENG, W. Deep Learning and Its Parallelization. In: BUYYA, Rajkumar; CALHEIROS, Rodrigo N.; DASTJERDI, Amir Vahid (Eds.). **Big Data: Principles and Paradigms**. [S.l.]: Morgan Kaufmann, June 2016. Google-Books-ID: MfOeCwAAQBAJ. ISBN 978-0-12-809346-7.

LIMA, Adriano Luiz de Souza; GRESSE VON WANGENHEIM, Christiane. A Deep Learning Model for the Distribution of Visual Aesthetics Degree of Mobile User Interfaces. **International Journal of Human-Computer Studies**.

LIMA, Adriano Luiz de Souza; GRESSE VON WANGENHEIM, Christiane. Assessing the Visual Esthetics of User Interfaces: A Ten-Year Systematic Mapping. **International Journal of Human–Computer Interaction**, p. 1–21, June 2021. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2021.1926118. ISSN 1044-7318.

LIMA, Adriano Luiz de Souza; GRESSE VON WANGENHEIM, Christiane; BORGATTO, Adriano Ferreti. Assessment of Visual Aesthetics through Human Judgments: a Systematic Mapping. In: PROCEEDINGS of the 21st Brazilian Symposium on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Oct. 2022. (IHC '22), p. 1–14.

LIMA, Adriano Luiz de Souza; GRESSE VON WANGENHEIM, Christiane; BORGATTO, Adriano Ferreti. Comparing Scales for the Assessment of Visual Aesthetics of Mobile GUIs Through Human Judgments. en. **International Journal of Mobile Human Computer Interaction (IJMHCI)**, v. 14, n. 1, p. 1–28, 2022. Publisher: IGI Global. ISSN 1942-390X.

LIMA, Adriano Luiz de Souza; GRESSE VON WANGENHEIM, Christiane; MARTINS, Osvaldo P. Heiderscheidt Roberge; WANGENHEIM, Aldo von; HAUCK, Jean Carlo Rossa; BORGATTO, Adriano Ferreti. A Deep Learning Model for the Assessment of the Visual Aesthetics of Mobile User Interfaces. **Journal of the Brazilian Computer Society**.

LIMA, Adriano Luiz de Souza; MARTINS, Osvaldo P. Heiderscheidt Roberge; GRESSE VON WANGENHEIM, Christiane; WANGENHEIM, Aldo von;

HAUCK, Jean Carlo Rossa; BORGATTO, Adriano Ferreti. Automated Assessment of Visual aesthetics of Android User Interfaces with Deep Learning. In: PROCEEDINGS of the 21st Brazilian Symposium on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Oct. 2022. (IHC '22), p. 1–11.

LIN, Rui. Augmenting Image Aesthetic Assessment with Diverse Deep Features. In: 2021 4th Artificial Intelligence and Cloud Computing Conference. New York, NY, USA: Association for Computing Machinery, Mar. 2022. (AICCC '21), p. 30–38.

LINDGAARD, Gitte. Aesthetics, Visual Appeal, Usability and User Satisfaction: What Do the User's Eyes Tell the User's Brain? **Australian Journal of Emerging Technologies & Society**, v. 5, n. 1, p. 1–14, May 2007. ISSN 14490706.

LINDGAARD, Gitte; DUDEK, Cathy. User Satisfaction, Aesthetics and Usability. In: HAMMOND, Judy; GROSS, Tom; WESSON, Janet (Eds.). **Usability: Gaining a Competitive Edge**. Boston, MA: Springer US, 2002. (IFIP — The International Federation for Information Processing). P. 231–246. ISBN 978-0-387-35610-5.

LINDGAARD, Gitte; DUDEK, Cathy. What Is This Evasive Beast We Call User Satisfaction? **Interacting with Computers**, v. 15, n. 3, p. 429–452, 2003. ISSN 0953-5438.

LINDGAARD, Gitte; FERNANDES, Gary; DUDEK, Cathy; BROWN, J. Attention Web Designers: You Have 50 Milliseconds to Make a Good First Impression! **Behaviour & Information Technology**, v. 25, n. 2, p. 115–126, Mar. 2006. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01449290500330448. ISSN 0144-929X.

LINDGAARD, Gitte; LITWINSKA, Justyna; DUDEK, Cathy. Judging Web Page Visual Appeal: Do East and West Really Differ? en. In: PROC. of Interfaces and Human Computer Interaction 2008. Amsterdam, The Netherlands: [s.n.], 2008. v. 8, p. 157–164.

LIU, Shuqing; LIANG, Tianyi; SHAO, Shuai; KONG, Jun. Evaluating Localized MOOCs: The Role of Culture on Interface Design and User Experience. **IEEE Access**, v. 8, p. 107927–107940, 2020. Conference Name: IEEE Access. ISSN 2169-3536.

LIU, Xin; JIANG, Yujia. Aesthetic assessment of website design based on multimodal fusion. en. **Future Generation Computer Systems**, v. 117, p. 433–438, Apr. 2021. ISSN 0167-739X.

LOZANO, Luis M.; GARCÍA-CUETO, Eduardo; MUÑIZ, José. Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. **Methodology: European Journal of Research Methods for the Behavioral and Social Sciences**, v. 4, n. 2, p. 73–79, 2008. Place: Germany Publisher: Hogrefe & Huber Publishers. ISSN 1614-2241.

LU, Xin; LIN, Zhe; JIN, Hailin; YANG, Jianchao; WANG, James Z. RAPID: Rating Pictorial Aesthetics Using Deep Learning. In: PROCEEDINGS of the 22nd ACM international conference on Multimedia. New York, NY, USA: Association for Computing Machinery, 2014. (MM '14), p. 457–466.

MAHMOOD, Ammar; BENNAMOUN, Mohammed; AN, Senjian; SOHEL, Ferdous; BOUSSAID, Farid; HOVEY, Renae; KENDRICK, Gary; FISHER, Robert B. Chapter 21 - Deep Learning for Coral Classification. In: SAMUI, Pijush; SEKHAR, Sanjiban; BALAS, Valentina E. (Eds.). **Handbook of Neural Computation**. [S.l.]: Academic Press, Jan. 2017. P. 383–401. ISBN 978-0-12-811318-9.

MAITY, Ranjan; BHATTACHARYA, Samit. A Model to Compute Webpage Aesthetics Quality Based on Wireframe Geometry. en. In: BERNHAUPT, Regina; DALVI, Girish; JOSHI, Anirudha; K. BALKRISHAN, Devanuj; O'NEILL, Jacki; WINCKLER, Marco (Eds.). **Human-Computer Interaction – INTERACT 2017**. Cham: Springer International Publishing, 2017. (Lecture Notes in Computer Science), p. 85–94.

MAITY, Ranjan; BHATTACHARYA, Samit. A Quantitative Approach to Measure Webpage Aesthetics. en. **International Journal of Technology and Human Interaction (IJTHI)**, v. 16, n. 2, p. 53–68, 2020. Publisher: IGI Global. ISSN 1548-3908.

MAITY, Ranjan; BHATTACHARYA, Samit. Is My Interface Beautiful?—A Computational Model-Based Approach. **IEEE Transactions on Computational Social Systems**, v. 6, n. 1, p. 149–161, Feb. 2019. Conference Name: IEEE Transactions on Computational Social Systems. ISSN 2329-924X.

MAITY, Ranjan; BHATTACHARYA, Samit. Relating Aesthetics of the GUI Text Elements with Readability using Font Family. In: PROCEEDINGS of the 2018 ACM Companion International Conference on Interactive Surfaces and Spaces. New York, NY, USA: Association for Computing Machinery, Nov. 2018. (ISS '18 Companion), p. 63–68.

MAITY, Ranjan; MADROSIYA, Akshay; BHATTACHARYA, Samit. A Computational Model to Predict Aesthetic Quality of Text Elements of GUI. en. **Procedia Computer Science**, v. 84, p. 152–159, Jan. 2016. ISSN 1877-0509.

MALU, Gautam; BAPI, Raju S.; INDURKHYA, Bipin. Learning Photography Aesthetics with Deep CNNs. **arXiv:1707.03981 [cs]**, July 2017. arXiv: 1707.03981.

MANJUNATH, B. S.; SALEMBIER, Philippe; SIKORA, Thomas. **Introduction to MPEG-7: Multimedia Content Description Interface**. [S.l.]: John Wiley & Sons, June 2002. Google-Books-ID: CmSPGXF1yB4C. ISBN 978-0-471-48678-7.

MAYDEU-OLIVARES, Alberto; KRAMP, Uwe; GARCÍA-FORERO, Carlos; GALLARDO-PUJOL, David; COFFMAN, Donna. The Effect of Varying the Number of Response Alternatives in Rating Scales: Experimental Evidence from Intra-individual Effects. en. **Behavior Research Methods**, v. 41, n. 2, p. 295–308, May 2009. ISSN 1554-3528.

MAZUMDAR, Suvodeep; PETRELLI, Daniela; ELBEDWEIHY, Khadija; LANFRANCHI, Vitaveska; CIRAVEGNA, Fabio. Affective graphs: The visual appeal of Linked Data. en. **Semantic Web**, v. 6, n. 3, p. 277–312, Jan. 2015. Publisher: IOS Press. ISSN 1570-0844.

MBENZA, Patrick; BURNY, Nicolas. Computing aesthetics of concrete user interfaces. In: COMPANION Proceedings of the 12th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. New York, NY, USA: Association for Computing Machinery, June 2020. (EICS '20 Companion), p. 1–8.

MCDONNELL, Marian; LEE, Alex. An Investigation of Visual Appeal and Trust in Websites. In: p. 53–60.

MCDONNELL, Marian; O'REILLY, Tara. The Effect of Interaction on Visual Appeal and Trust in Online Health Information. In: p. 100–108.

MCGRAW, Kenneth O.; WONG, S. P. Forming Inferences about Some Intraclass Correlation Coefficients. **Psychological Methods**, v. 1, n. 1, p. 30–46, 1996. Place: US Publisher: American Psychological Association. ISSN 1939-1463(Electronic),1082-989X(Print).

MEDEIROS, Giselle Araújo e Silva de; GRESSE VON WANGENHEIM, Christiane; HAUCK, Jean Carlo Rossa. O protagonismo de estudantes da Educação Básica a partir do desenvolvimento de aplicativos para smartphone. pt. **Perspectiva**, v. 39, n. 1, p. 1–18, Feb. 2021. Number: 1. ISSN 2175-795X.

MEHRA, Aashish; PAUL, Justin; KAURAV, Rahul Pratap Singh. Determinants of Mobile Apps Adoption Among Young Adults: Theoretical Extension and Analysis. **Journal of Marketing Communications**, v. 0, n. 0, p. 1–29, Feb. 2020. Publisher: Routledge _eprint: https://doi.org/10.1080/13527266.2020.1725780. ISSN 1352-7266.

MILLER, Matthew; CHOI, Gilok; CHELL, Lindsay. Comparison of Three Digital Library Interfaces: Open Library, Google Books, and Hathi Trust. In: PROCEEDINGS of the 12th ACM/IEEE-CS joint conference on Digital Libraries. New York, NY, USA: Association for Computing Machinery, 2012. (JCDL '12), p. 367–368.

MINIUKOVICH, Aliaksei; DE ANGELI, Antonella. Computation of Interface Aesthetics. In: PROCEEDINGS of the 33rd Annual ACM Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2015. P. 1163–1172. ISBN 978-1-4503-3145-6.

MINIUKOVICH, Aliaksei; DE ANGELI, Antonella. Quantification of Interface Visual Complexity. In: PROCEEDINGS of the 2014 International Working Conference on Advanced Visual Interfaces. New York, NY, USA: Association for Computing Machinery, May 2014. (AVI '14), p. 153–160.

MINIUKOVICH, Aliaksei; DE ANGELI, Antonella. Visual Diversity and User Interface Quality. In: PROCEEDINGS of the 2015 British HCI Conference. New York, NY, USA: Association for Computing Machinery, July 2015. (British HCI '15), p. 101–109.

MINIUKOVICH, Aliaksei; DE ANGELI, Antonella. Visual Impressions of Mobile App Interfaces. In: PROCEEDINGS of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational. New York, NY, USA: Association for Computing Machinery, Oct. 2014. (NordiCHI '14), p. 31–40.

MINIUKOVICH, Aliaksei; MARCHESE, Maurizio. Relationship Between Visual Complexity and Aesthetics of Webpages. In: PROCEEDINGS of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2020. (CHI '20), p. 1–13.

MINIUKOVICH, Aliaksei; SULPIZIO, Simone; DE ANGELI, Antonella. Visual Complexity of Graphical User Interfaces. In: PROCEEDINGS of the 2018 International Conference on Advanced Visual Interfaces. New York, NY, USA: Association for Computing Machinery, May 2018. (AVI '18), p. 1–9.

MIT APP INVENTOR. **MIT App Inventor | Explore MIT App Inventor**. [S.l.: s.n.], 2022. Available from: `http://appinventor.mit.edu/`. Visited on: 28 Apr. 2022.

MONTIEL, Hugo; GOMEZ-ZERMEÑO, Marcela Georgina. Educational Challenges for Computational Thinking in K–12 Education: A Systematic Literature Review of "Scratch" as an Innovative Programming Tool. en. **Computers**, v. 10, n. 6, p. 69, June 2021. Number: 6 Publisher: Multidisciplinary Digital Publishing Institute. ISSN 2073-431X.

MORAN, Kevin; LI, Boyang; BERNAL-CÁRDENAS, Carlos; JELF, Dan; POSHYVANYK, Denys. Automated Reporting of GUI Design Violations for Mobile Apps. In: PROCEEDINGS of the 40th International Conference on Software Engineering. New York, NY, USA: Association for Computing Machinery, May 2018. (ICSE '18), p. 165–175.

MOSHAGEN, Morten; THIELSCH, Meinald T. A Short Version of the Visual Aesthetics of Websites Inventory. en. **Behaviour & Information Technology**, v. 32, n. 12, p. 1305–1311, Dec. 2013. ISSN 0144-929X, 1362-3001.

MOSHAGEN, Morten; THIELSCH, Meinald T. Facets of Visual Aesthetics. en. **International Journal of Human-Computer Studies**, v. 68, n. 10, p. 689–709, Oct. 2010. ISSN 10715819.

MOSS, G. A.; GUNN, R. W. Gender Differences in Website Production and Preference Aesthetics: Preliminary Implications for ICT in Education and Beyond. **Behaviour & Information Technology**, v. 28, n. 5, p. 447–460, Sept. 2009. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01449290802332662. ISSN 0144-929X.

MÕTTUS, Mati; LAMAS, David. Aesthetics of Interaction Design: A Literature Review. In: PROCEEDINGS of the Mulitimedia, Interaction, Design and Innnovation. New York, NY, USA: Association for Computing Machinery, June 2015. (MIDI '15), p. 1–10.

MÕTTUS, Mati; LAMAS, David; PAJUSALU, Maarja; TORRES, Rui. The Evaluation of Interface Aesthetics. In: PROCEEDINGS of the International Conference on

Multimedia, Interaction, Design and Innovation. New York, NY, USA: Association for Computing Machinery, June 2013. (MIDI '13), p. 1–10.

MURPHY, Gardner; LIKERT, Rensis. **Public Opinion and the Individual. A Psychological Study of Student Attitudes on Public Questions, with a Retest Five Years Later**. Oxford, England: Harper, 1938. (Public opinion and the individual. A psychological study of student attitudes on public questions, with a retest five years later). Pages: x, 316.

NANNI, Loris; GHIDONI, Stefano; BRAHNAM, Sheryl. Handcrafted vs. Non-handcrafted Features for Computer Vision Classification. en. **Pattern Recognition**, v. 71, p. 158–172, Nov. 2017. ISSN 0031-3203.

NGO, David Chek Ling. Measuring the Aesthetic Elements of Screen Designs. en. **Displays**, v. 22, n. 3, p. 73–78, 2001. ISSN 0141-9382.

NGO, David Chek Ling; BYRNE, John G. Aesthetic Measures for Screen Design. In: PROCEEDINGS 1998 Australasian Computer Human Interaction Conference. OzCHI'98 (Cat. No.98EX234). [S.l.: s.n.], Nov. 1998. P. 64–71.

NGO, David Chek Ling; SAMSUDIN, Azman; ABDULLAH, Rosni. Aesthetic measures for assessing graphic screens. en. **Journal of Inforamtion Science and Engineering**, v. 16, n. 1, p. 97–116, Jan. 2000. Number: 1 Publisher: INST INFORMATION SCIENCE, ACADEMIA SINICA, TAIPEI 115, TAIWAN. ISSN 1016-2364.

NGO, David Chek Ling; TEO, Lian Seng; BYRNE, John G. Evaluating Interface Esthetics. en. **Knowledge and Information Systems**, v. 4, n. 1, p. 46–79, Jan. 2002. ISSN 0219-1377.

NGO, David Chek Ling; TEO, Lian Seng; BYRNE, John G. Formalising guidelines for the design of screen layouts. en. **Displays**, v. 21, n. 1, p. 3–15, Mar. 2000. ISSN 0141-9382.

NGO, David Chek Ling; TEO, Lian Seng; BYRNE, John G. Modelling interface aesthetics. en. **Information Sciences**, v. 152, p. 25–46, June 2003. ISSN 0020-0255.

NORMAN, Don. Emotion & Design: Attractive Things Work Better. **Interactions**, v. 9, n. 4, p. 36–42, July 2002. ISSN 1072-5520.

NUNES, Andreia; CASTRO, São Luís; LIMPO, Teresa. A Review of Mindfulness-Based Apps for Children. en. **Mindfulness**, v. 11, n. 9, p. 2089–2101, Sept. 2020. ISSN 1868-8535.

NUNNALLY, Jum C.; BERNSTEIN, Ira H. **Psychometric Theory**. 3rd. New York: McGraw-Hill, 1994. ISBN 978-0-07-047849-7.

OTTEN, Raphael; SCHREPP, Martin; THOMASCHEWSKI, Jörg. Visual clarity as mediator between usability and aesthetics. In: PROCEEDINGS of the Conference on Mensch und Computer. New York, NY, USA: Association for Computing Machinery, Sept. 2020. (MuC '20), p. 11–15.

OYIBO, Kiemute; ADAJI, Ifeoma; ORJI, Rita; VASSILEVA, Julita. What Drives the Perceived Credibility of Mobile Websites: Classical or Expressive Aesthetics? en. In: KUROSU, Masaaki (Ed.). **Human-Computer Interaction. Interaction in Context**. Cham: Springer International Publishing, 2018. (Lecture Notes in Computer Science), p. 576–594.

PALMER, Stephen E.; SCHLOSS, Karen B.; SAMMARTINO, Jonathan. Visual Aesthetics and Human Preference. **Annual Review of Psychology**, v. 64, n. 1, p. 77–107, Jan. 2013. Publisher: Annual Reviews. ISSN 0066-4308.

PANDIR, Muzeyyen; KNIGHT, John. Homepage Aesthetics: The Search for Preference Factors and the Challenges of Subjectivity. **Interacting with Computers**, v. 18, n. 6, p. 1351–1370, Dec. 2006. ISSN 0953-5438.

PAPACHRISTOS, Eleftherios; AVOURIS, Nikolaos. Are First Impressions about Websites Only Related to Visual Appeal? en. In: CAMPOS, Pedro; GRAHAM, Nicholas; JORGE, Joaquim; NUNES, Nuno; PALANQUE, Philippe; WINCKLER, Marco (Eds.). **Human-Computer Interaction – INTERACT 2011**. Berlin, Heidelberg: Springer, 2011. (Lecture Notes in Computer Science), p. 489–496.

PAPACHRISTOS, Eleftherios; AVOURIS, Nikolaos. The Influence of Website Category on Aesthetic Preferences. en. In: KOTZÉ, Paula; MARSDEN, Gary; LINDGAARD, Gitte; WESSON, Janet; WINCKLER, Marco (Eds.). **Human-Computer Interaction – INTERACT 2013**. Berlin, Heidelberg: Springer, 2013. (Lecture Notes in Computer Science), p. 445–452.

PAPACHRISTOS, Eleftherios; AVOURIS, Nikolaos. The Subjective and Objective Nature of Website Aesthetic Impressions. en. In: GROSS, Tom; GULLIKSEN, Jan; KOTZÉ, Paula; OESTREICHER, Lars; PALANQUE, Philippe; PRATES, Raquel Oliveira; WINCKLER, Marco (Eds.). **Human-Computer Interaction – INTERACT 2009**. Berlin, Heidelberg: Springer, 2009. (Lecture Notes in Computer Science), p. 119–122.

PAPPAS, Ilias O.; SHARMA, Kshitij; MIKALEF, Patrick; GIANNAKOS, Michail. A Comparison of Gaze Behavior of Experts and Novices to Explain Website Visual Appeal. **PACIS 2018 Proceedings**, June 2018.

PAPPAS, Ilias O.; SHARMA, Kshitij; MIKALEF, Patrick; GIANNAKOS, Michail N. How Quickly Can We Predict Users' Ratings on Aesthetic Evaluations of Websites? Employing Machine Learning on Eye-Tracking Data. en. In: HATTINGH, Marié; MATTHEE, Machdel; SMUTS, Hanlie; PAPPAS, Ilias; DWIVEDI, Yogesh K.; MÄNTYMÄKI, Matti (Eds.). **Responsible Design, Implementation and Use of Information and Communication Technology**. Cham: Springer International Publishing, 2020. (Lecture Notes in Computer Science), p. 429–440.

PATERNÒ, Fabio. End User Development: Survey of an Emerging Field for Empowering People. en. **ISRN Software Engineering**, v. 2013, e532659, June 2013. Publisher: Hindawi.

PATTON, Evan W.; TISSENBAUM, Michael; HARUNANI, Farzeen. MIT App Inventor: Objectives, Design, and Development. In: KONG, Siu-Cheung; ABELSON, Harold (Eds.). **Computational Thinking Education**. Singapore: Springer, 2019. P. 31–49. ISBN 9789811365287.

PAVLAS, Davin; LUM, Heather C.; SALAS, Eduardo. The influence of aesthetic and usability web design elements on viewing patterns and user response: 54th Human Factors and Ergonomics Society Annual Meeting 2010, HFES 2010. **54th Human Factors and Ergonomics Society Annual Meeting 2010, HFES 2010**, p. 1244–1248, Dec. 2010. ISSN 9781617820885.

PENGNATE, Supavich; SARATHY, Rathindra. An experimental investigation of the influence of website emotional design features on trust in unfamiliar online vendors. en. **Computers in Human Behavior**, v. 67, p. 49–60, Feb. 2017. ISSN 0747-5632.

PENGNATE, Supavich; SARATHY, Rathindra; ARNOLD, Todd J. The Influence of the Centrality of Visual Website Aesthetics on Online User Responses: Measure Development and Empirical Investigation. en. **Information Systems Frontiers**, v. 23, n. 2, p. 435–452, Apr. 2021. ISSN 1572-9419.

PENGNATE, Supavich; SARATHY, Rathindra; LEE, JinKyu. The Engagement of Website Initial Aesthetic Impressions: An Experimental Investigation. **International Journal of Human–Computer Interaction**, v. 35, n. 16, p. 1517–1531, Oct. 2019. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2018.1554319. ISSN 1044-7318.

PETERSEN, Kai; FELDT, Robert; MUJTABA, Shahid; MATTSSON, Michael. Systematic Mapping Studies in Software Engineering. In: PROCEEDINGS of the 12th International Conference on Evaluation and Assessment in Software Engineering. Bari, Italy: BCS Learning & Development Ltd, June 2008. P. 1–10.

POLYZOTIS, Neoklis; ROY, Sudip; WHANG, Steven Euijong; ZINKEVICH, Martin. Data Management Challenges in Production Machine Learning. In: PROCEEDINGS of the 2017 ACM International Conference on Management of Data. New York, NY, USA: Association for Computing Machinery, May 2017. (SIGMOD '17), p. 1723–1726.

PORTNEY, Leslie G. **Foundations of Clinical Research: Applications to Evidence-Based Practice**. Fourth edition. Philadelphia: F.A. Davis Company, Jan. 2020. ISBN 978-0-8036-6113-4.

PREECE, Jennifer; SHARP, Helen; ROGERS, Yvonne. **Interaction Design: Beyond Human-Computer Interaction**. 4th edition. Chichester: Wiley, May 2015. ISBN 978-1-119-02075-2.

PROPST, Dennis B.; SWIERENGA, Sarah J.; PIERCE, Graham L.; JEONG, Eunseong; COURSARIS, Constantinos K. From the Ground-Up: Role of Usability and Aesthetics Evaluation in Creating a Knowledge-Based Website for the U.S. Army Corps of Engineers. en. In: MARCUS, Aaron (Ed.). **Design, User Experience, and Usability. Web, Mobile, and Product Design**. Berlin, Heidelberg: Springer, 2013. (Lecture Notes in Computer Science), p. 274–283.

PUNCH, Keith F. **Introduction to Social Research: Quantitative and Qualitative Approaches**. 1st edition. London ; Thousand Oaks, Calif: SAGE Publications Ltd, Dec. 1998. ISBN 978-0-7619-5812-3.

PUNCHOOJIT, Lumpapun; HONGWARITTORRN, Nuttanont. Usability Studies on Mobile User Interface Design Patterns: A Systematic Literature Review. en. **Advances in Human-Computer Interaction**, v. 2017, e6787504, Nov. 2017. Publisher: Hindawi. ISSN 1687-5893.

PURCHASE, Helen C.; HAMER, John; JAMIESON, Adrian; RYAN, Oran. Investigating objective measures of web page aesthetics and usability. In: PROCEEDINGS of the Twelfth Australasian User Interface Conference - Volume 117. AUS: Australian Computer Society, Inc., Jan. 2011. (AUIC '11), p. 19–28.

RAHMAT, Hazwani; ZULZALIL, Hazura; GHANI, Abdul Azim Abd; KAMARUDDIN, Azrina. A Comprehensive Usability Model for Evaluating Smartphone Apps. **Advanced Science Letters**, v. 24, n. 3, p. 1633–1637, Mar. 2018.

RAMEZANI NIA, Masoud; SHOKOUHYAR, Sajjad. Analyzing the effects of visual aesthetic of Web pages on users' responses in online retailing using the VisAWI method. **Journal of Research in Interactive Marketing**, v. 14, n. 4, p. 357–389, Jan. 2020. Publisher: Emerald Publishing Limited. ISSN 2040-7122.

RANDOLPH, Justus J. **Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa**. en. [S.l.], Oct. 2005. Publication Title: Online Submission ERIC Number: ED490661. Available from: `https://eric.ed.gov/?id=ED490661`. Visited on: 8 Apr. 2023.

REINECKE, Katharina; BERNSTEIN, Abraham. Improving performance, perceived usability, and aesthetics with culturally adaptive user interfaces. **ACM Transactions on Computer-Human Interaction**, v. 18, n. 2, 8:1–8:29, 2011. ISSN 1073-0516.

REINECKE, Katharina; GAJOS, Krzysztof Z. Quantifying visual preferences around the world. In: PROCEEDINGS of the SIGCHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery, Apr. 2014. (CHI '14), p. 11–20.

REINECKE, Katharina; YEH, Tom; MIRATRIX, Luke; MARDIKO, Rahmatri; ZHAO, Yuechen; LIU, Jenny; GAJOS, Krzysztof Z. 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. New York, NY, USA: Association for Computing Machinery, 2013. P. 2049–2058. ISBN 978-1-4503-1899-0.

ROBERTS, Paula; PRIEST, Helena; TRAYNOR, Michael. Reliability and Validity in Research. **Nursing Standard**, v. 20, n. 44, p. 41–45, July 2006. ISSN 00296570.

ROBINS, David; HOLMES, Jason. Aesthetics and credibility in web site design. **Information Processing and Management: an International Journal**, v. 44, n. 1, p. 386–399, Jan. 2008. ISSN 0306-4573.

RUSSAKOVSKY, Olga et al. ImageNet Large Scale Visual Recognition Challenge. en. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, Dec. 2015. ISSN 1573-1405.

SADITA, Lia; SANTOSO, Harry Budi; WINDRAWAN, Luqman Iffan; KHOTIMAH, Purnomo Husnul. An Indonesian Adaption of Visual Aesthetics of Website Inventory (VisAWI) Questionnaire for Evaluating Video Game User Interface. In: PROCEEDINGS of the 2022 International Conference on Computer, Control, Informatics and Its Applications. New York, NY, USA: Association for Computing Machinery, 2022. (IC3INA '22), p. 382–386.

SAKAGUCHI, Daichi; TAKIMOTO, Hironori; KANAGAWA, Akihiro. Study on relationship between composition and prediction of photo aesthetics using CNN. Ed. by Moulay Akhloufi. **Cogent Engineering**, v. 9, n. 1, p. 2107472, Dec. 2022. Publisher: Cogent OA _eprint: https://doi.org/10.1080/23311916.2022.2107472. ISSN null.

SALIMUN, Carolyn; PURCHASE, Helen C.; SIMMONS, David R.; BREWSTER, Stephen. Preference Ranking of Screen Layout Principles, Sept. 2010. Publisher: BCS Learning & Development.

SALIMUN, Carolyn; PURCHASE, Helen C.; SIMMONS, David R.; BREWSTER, Stephen. The effect of aesthetically pleasing composition on visual search performance. In: PROCEEDINGS of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries. New York, NY, USA: Association for Computing Machinery, Oct. 2010. (NordiCHI '10), p. 422–431.

SALMAN, Hasanin Mohammed; WAN AHMAD, Wan Fatimah; SULAIMAN, Suziah. Usability Evaluation of the Smartphone User Interface in Supporting Elderly Users From Experts' Perspective. **IEEE Access**, v. 6, p. 22578–22591, 2018. Conference Name: IEEE Access. ISSN 2169-3536.

SANTAYANA, George. **The Sense of Beauty**. [S.l.]: Dover Publications, Aug. 2012.

SAREMI, Mahnaz; SADEGHI, Vahid; KHODAKARIM, Soheila; MALEKI-GHAHFAROKHI, Azam. Farsi Version of Visual Aesthetics of Website Inventory (FV-VisAWI): Translation and Psychometric Evaluation. **International Journal of Human–Computer Interaction**, v. 39, n. 4, p. 834–841, 2022. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2022.2049138. ISSN 1044-7318.

SARTWELL, Crispin. **Beauty**. Winter 2017. [S.l.: s.n.], 2017. Last Modified: 2016-10-05. Available from: `https://plato.stanford.edu/archives/win2017/entries/beauty/`. Visited on: 1 Aug. 2021.

SCHAIK, Paul van; HASSENZAHL, Marc; LING, Jonathan. User-Experience from an Inference Perspective. **ACM Transactions on Computer-Human Interaction**, v. 19, n. 2, 11:1–11:25, July 2012. ISSN 1073-0516.

SCHAIK, Paul van; LING, Jonathan. Five Psychometric Scales for Online Measurement of the Quality of Human-Computer Interaction in Web Sites. **International Journal of Human–Computer Interaction**, v. 18, n. 3, p. 309–322, 2005. Publisher: Taylor & Francis _eprint: https://doi.org/10.1207/s15327590ijhc1803_4. ISSN 1044-7318.

SCHAIK, Paul van; LING, Jonathan. Modelling user experience with web sites: Usability, hedonic value, beauty and goodness. **Interacting with Computers**, v. 20, n. 3, p. 419–432, May 2008. ISSN 0953-5438.

SCHAIK, Paul van; LING, Jonathan. The Role of Context in Perceptions of the Aesthetics of Web Pages over Time. en. **International Journal of Human-Computer Studies**, v. 67, n. 1, p. 79–89, Jan. 2009. ISSN 1071-5819.

SCHENKMAN, Bo N.; JÖNSSON, Fredrik U. Aesthetics and Preferences of Web Pages. **Behaviour & Information Technology**, v. 19, n. 5, p. 367–377, Jan. 2000. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/014492900750000063. ISSN 0144-929X.

SCHLATTER, Tania; LEVINSON, Deborah. **Visual Usability: Principles and Practices for Designing Digital Applications**. 1st. Amsterdam: Morgan Kaufmann, 2013. ISBN 978-0-12-398536-1.

SCHMUTZ, Sven; SONDEREGGER, Andreas; SAUER, Juergen. Implementing Recommendations From Web Accessibility Guidelines: A Comparative Study of

Nondisabled Users and Users With Visual Impairments. en. **Human Factors**, v. 59, n. 6, p. 956–972, Sept. 2017. Publisher: SAGE Publications Inc. ISSN 0018-7208.

SCHOBER, Patrick; BOER, Christa; SCHWARTE, Lothar A. Correlation Coefficients: Appropriate Use and Interpretation. en-US. **Anesthesia & Analgesia**, v. 126, n. 5, p. 1763–1768, May 2018. ISSN 0003-2999.

SCHREPP, Martin; HELD, Theo; LAUGWITZ, Bettina. The influence of hedonic quality on the attractiveness of user interfaces of business management software. **Interacting with Computers**, v. 18, n. 5, p. 1055–1069, Sept. 2006. ISSN 0953-5438.

SCHULTCHEN, Dana; TERHORST, Yannik; HOLDERIED, Tanja; STACH, Michael; MESSNER, Eva-Maria; BAUMEISTER, Harald; SANDER, Lasse B. Stay Present with Your Phone: A Systematic Review and Standardized Rating of Mindfulness Apps in European App Stores. en. **International Journal of Behavioral Medicine**, v. 28, n. 5, p. 552–560, Oct. 2021. ISSN 1532-7558.

SECKLER, Mirjam; OPWIS, Klaus; TUCH, Alexandre N. 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**, v. 49, p. 375–389, 2015. Place: Netherlands Publisher: Elsevier Science. ISSN 1873-7692(Electronic),0747-5632(Print).

SECKLER, Mirjam; TUCH, Alexandre N. Linking Objective Web-design Factors to Facets of Subjective Aesthetic Perception. In: PROCEEDINGS of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design. New York, NY, USA: Association for Computing Machinery, Oct. 2012. (NordiCHI '12), p. 809–810.

SEO, Kwang-Kyu; LEE, Sangwon; CHUNG, Byung Do; PARK, Changsoon. Users' Emotional Valence, Arousal, and Engagement Based on Perceived Usability and Aesthetics for Web Sites. **International Journal of Human–Computer Interaction**, v. 31, n. 1, p. 72–87, Jan. 2015. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2014.959103. ISSN 1044-7318.

AL-SHAMAILEH, Ons; SUTCLIFFE, Alistair. Investigating a Multi-faceted View of User Experience. In: PROCEEDINGS of the 24th Australian Computer-Human Interaction Conference. New York, NY, USA: Association for Computing Machinery, Nov. 2012. (OzCHI '12), p. 9–18.

AL-SHAMAILEH, Ons; SUTCLIFFE, Alistair. The Effect of Website Interactivity and Repeated Exposure on User Experience. In: PROCEEDINGS of the 4th Mexican Conference on Human-Computer Interaction. New York, NY, USA: Association for Computing Machinery, Oct. 2012. (MexIHC '12), p. 1–8.

SHANNON, C. E. A mathematical theory of communication. **The Bell System Technical Journal**, v. 27, n. 3, p. 379–423, July 1948. Conference Name: The Bell System Technical Journal. ISSN 0005-8580.

SHARMA, Tushar; MISHRA, Pratibha; TIWARI, Rohit. Designite: A Software Design Quality Assessment Tool. In: PROCEEDINGS of the 1st International Workshop on Bringing Architectural Design Thinking into Developers' Daily Activities. New York, NY, USA: Association for Computing Machinery, May 2016. (BRIDGE '16), p. 1–4.

SILVA, Edna Lucia da; MENEZES, Estera Muszkat. **Metodologia da pesquisa e elaboração de dissertação**. 3ª edição. Florianópolis: edUFSC, 2001.

SIMONYAN, Karen; ZISSERMAN, Andrew. **Very Deep Convolutional Networks for Large-Scale Image Recognition**. [S.l.]: arXiv, Apr. 2015. arXiv:1409.1556 [cs]. Available from: `http://arxiv.org/abs/1409.1556`. Visited on: 9 Nov. 2022.

SINGH, Nahar; BHATTACHARYA, Samit. A GA-based Approach to Improve Web Page Aesthetics. In: PROCEEDINGS of the First International Conference on Intelligent Interactive Technologies and Multimedia. New York, NY, USA: Association for Computing Machinery, 2011. (IITM '10), p. 29–32.

SKULMOWSKI, Alexander; AUGUSTIN, Yannik; PRADEL, Simon; NEBEL, Steve; SCHNEIDER, Sascha; REY, Günter Daniel. The negative impact of saturation on website trustworthiness and appeal: A temporal model of aesthetic website perception. en. **Computers in Human Behavior**, v. 61, p. 386–393, Aug. 2016. ISSN 0747-5632.

SMITH, Leslie N. A Disciplined Approach to Neural Network Hyper-parameters: Part 1 – Learning Rate, Batch Size, Momentum, and Weight Decay. **arXiv:1803.09820 [cs, stat]**, Apr. 2018. arXiv: 1803.09820.

SMITH, Leslie N.; TOPIN, Nicholay. Super-convergence: Very Fast Training of Neural Networks Using Large Learning Rates. In: PROCEEDINGS of the Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications. [S.l.]: International Society for Optics and Photonics, May 2019. v. 11006, p. 1100612.

SOKOLOVA, Marina; LAPALME, Guy. A Systematic Analysis of Performance Measures for Classification Tasks. en. **Information Processing & Management**, v. 45, n. 4, p. 427–437, 2009. ISSN 0306-4573.

SOLECKI, Igor da Silva; JUSTEN, Karla Aparecida; PORTO, João Vitor Araujo; WANGENHEIM, Christiane Anneliese Gresse von; HAUCK, Jean Carlo Rossa; BORGATTO, Adriano Ferreti. Estado da Prática do Design Visual de Aplicativos Móveis desenvolvidos com App Inventor. pt. **Revista Brasileira de Informática na Educação**, v. 28, n. 0, p. 30–47, Feb. 2020. Number: 0. ISSN 2317-6121.

SOLECKI, Igor da Silva; PORTO, João; ALVES, Nathalia da Cruz; GRESSE VON WANGENHEIM, Christiane; HAUCK, Jean; BORGATTO, Adriano Ferreti. Automated Assessment of the Visual Design of Android Apps Developed with App Inventor. In: PROCEEDINGS of the 51st ACM Technical Symposium on Computer Science Education. New York, NY, USA: Association for Computing Machinery, Feb. 2020. (SIGCSE '20), p. 51–57.

SONDEREGGER, Andreas; SAUER, Juergen; EICHENBERGER, Janine. Expressive and Classical Aesthetics: Two Distinct Concepts with Highly Similar Effect Patterns in User–artefact Interaction. **Behaviour & Information Technology**, v. 33, n. 11, p. 1180–1191, Nov. 2014. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2013.853835. ISSN 0144-929X.

SOUI, Makram; CHOUCHANE, Mabrouka; BESSGHAIER, Narjes; MKAOUER, Mohamed Wiem; KESSENTINI, Marouane. On the Impact of Aesthetic Defects on the Maintainability of Mobile Graphical User Interfaces: An Empirical Study. en. **Information Systems Frontiers**, Feb. 2021. ISSN 1572-9419.

SOUI, Makram; CHOUCHANE, Mabrouka; GASMI, Ines; MKAOUER, Mohamed Wiem. PLAIN: PLugin for predicting the usAbility of Mobile User INterface. In: p. 127–136.

SOUI, Makram; CHOUCHANE, Mabrouka; MKAOUER, Mohamed Wiem; KESSENTINI, Marouane; GHEDIRA, Khaled. Assessing the Quality of Mobile Graphical User Interfaces Using Multi-objective Optimization. en. **Soft Computing**, v. 24, n. 10, p. 7685–7714, May 2020. ISSN 1433-7479.

STATISTA. **Number of mobile devices worldwide 2020-2025**. en. [S.l.: s.n.], 2023. Available from: `https://www.statista.com/statistics/245501/multiple-mobile-device-ownership-worldwide/`. Visited on: 4 Apr. 2023.

STEVENS, S. S.; MARKS, Lawrence E. **Psychophysics: Introduction to Its Perceptual, Neural, and Social Prospects**. New York: Routledge, Oct. 2017. ISBN 978-1-315-12767-5.

STONE, Debbie; JARRETT, Caroline; WOODROFFE, Mark; MINOCHA, Shailey. **User Interface Design and Evaluation**. [S.l.]: Morgan Kaufmann, 2005.

STOYANOV, Stoyan R.; HIDES, Leanne; KAVANAGH, David J.; ZELENKO, Oksana; TJONDRONEGORO, Dian; MANI, Madhavan. Mobile App Rating Scale: A New Tool for Assessing the Quality of Health Mobile Apps. EN. **JMIR mHealth and uHealth**, v. 3, n. 1, e3422, Mar. 2015. Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

STREBE, Rita. Aesthetics on the web: effects on approach and avoidance behaviour. **Behaviour & Information Technology**, v. 35, n. 1, p. 4–20, Jan. 2016. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0144929X.2015.1070202. ISSN 0144-929X.

STREINER, David L. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. **Journal of Personality Assessment**, v. 80, n. 1, p. 99–103, Feb. 2003. Publisher: Routledge _eprint: https://doi.org/10.1207/S15327752JPA8001_18. ISSN 0022-3891.

SUCHECKI, Maciej; TRZCISKI, Tomasz. Understanding Aesthetics in Photography Using Deep Convolutional Neural Networks. In: 2017 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). [S.l.: s.n.], Sept. 2017. P. 149–153. ISSN: 2326-0319.

SUTCLIFFE, Alistair; DE ANGELI, Antonella. Assessing Interaction Styles in Web User Interfaces. en. In: COSTABILE, Maria Francesca; PATERNÒ, Fabio (Eds.). **Human-Computer Interaction - INTERACT 2005**. Berlin, Heidelberg: Springer, 2005. (Lecture Notes in Computer Science), p. 405–417.

SUTCLIFFE, Alistair; NAMOUNE, Abdallah. Getting the message across: visual attention, aesthetic design and what users remember. In: PROCEEDINGS of the 7th ACM conference on Designing interactive systems. New York, NY, USA: Association for Computing Machinery, Feb. 2008. (DIS '08), p. 11–20.

SZEGEDY, Christian; LIU, Wei; JIA, Yangqing; SERMANET, Pierre; REED, Scott; ANGUELOV, Dragomir; ERHAN, Dumitru; VANHOUCKE, Vincent; RABINOVICH, Andrew. Going Deeper With Convolutions. In: p. 1–9.

TABA, Seyyed Ehsan Salamati; KEIVANLOO, Iman; ZOU, Ying; NG, Joanna; NG, Tinny. An Exploratory Study on the Relation between User Interface Complexity and the Perceived Quality. en. In: CASTELEYN, Sven; ROSSI, Gustavo; WINCKLER, Marco (Eds.). **Web Engineering**. Cham: Springer International Publishing, 2014. (Lecture Notes in Computer Science), p. 370–379.

TAMURA, Hideyuki; MORI, Shunji; YAMAWAKI, Takashi. Textural Features Corresponding to Visual Perception. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 8, n. 6, p. 460–473, June 1978. Conference Name: IEEE Transactions on Systems, Man, and Cybernetics. ISSN 2168-2909.

TAN, Mingxing; LE, Quoc V. **EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks**. [S.l.]: arXiv, Sept. 2020. arXiv:1905.11946 [cs, stat]. Available from: http://arxiv.org/abs/1905.11946. Visited on: 10 Nov. 2022.

THANH VI, Chi; HORNBÆK, Kasper; SUBRAMANIAN, Sriram. Neuroanatomical Correlates of Perceived Usability. In: PROCEEDINGS of the 30th Annual ACM Symposium on User Interface Software and Technology. New York, NY, USA: Association for Computing Machinery, Oct. 2017. (UIST '17), p. 519–532.

THIELSCH, Meinald T.; BLOTENBERG, Iris; JARON, Rafael. User Evaluation of Websites: From First Impression to Recommendation. **Interacting with Computers**, v. 26, n. 1, p. 89–102, Jan. 2014. ISSN 0953-5438.

THIELSCH, Meinald T.; ENGEL, Ronja; HIRSCHFELD, Gerrit. Expected usability is not a valid indicator of experienced usability. en. **PeerJ Computer Science**, v. 1, e19, Sept. 2015. Publisher: PeerJ Inc. ISSN 2376-5992.

THORNDIKE, Robert M.; THORNDIKE-CHRIST, Tracy. **Measurement and Evaluation in Psychology and Education**. 8. ed. [S.l.]: Pearson Education Limited, 2014. ISBN 978-1-292-04111-7.

TRACTINSKY, Noam. **Visual Aesthetics**. en. 2. ed. [S.l.: s.n.], 2013. Available from: https://www.interaction-design.org/literature/book/the-encyclopedia-of-human-computer-interaction-2nd-ed/visual-aesthetics. Visited on: 1 Aug. 2021.

TRACTINSKY, Noam; COKHAVI, Avivit; KIRSCHENBAUM, Moti; SHARFI, Tal. Evaluating the Consistency of Immediate Aesthetic Perceptions of Web Pages. en. **International Journal of Human-Computer Studies**, v. 64, n. 11, p. 1071–1083, Nov. 2006. ISSN 1071-5819.

TUCH, Alexandre N.; PRESSLABER, Eva E.; STÖCKLIN, Markus; OPWIS, Klaus; BARGAS-AVILA, Javier A. The Role of Visual Complexity and Prototypicality Regarding First Impression of Websites: Working Towards Understanding Aesthetic Judgments. en. **International Journal of Human-Computer Studies**, v. 70, n. 11, p. 794–811, Nov. 2012. ISSN 1071-5819.

TUCH, Alexandre N.; ROTH, Sandra P.; HORNBÆK, Kasper; OPWIS, Klaus; BARGAS-AVILA, Javier A. Is Beautiful Really Usable? Toward Understanding the Relation Between Usability, Aesthetics, and Affect in HCI. en. **Computers in Human Behavior**, v. 28, n. 5, p. 1596–1607, Sept. 2012. ISSN 0747-5632.

ULRICH, Karl. **Design: Creation of Artifacts in Society**. [S.l.]: University of Philadelphia, 2006.

URIBE, Silvia; ÁLVAREZ, Federico; MENÉNDEZ, José Manuel. User's Web Page Aesthetics Opinion: A Matter of Low-Level Image Descriptors Based on MPEG-7. **ACM Transactions on the Web**, v. 11, n. 1, 5:1–5:25, Mar. 2017. ISSN 1559-1131.

VANDERDONCKT, Jean; GILLO, Xavier. Visual techniques for traditional and multimedia layouts. In: PROCEEDINGS of the workshop on Advanced visual interfaces. New York, NY, USA: Association for Computing Machinery, June 1994. (AVI '94), p. 95–104.

VARELA, Martin; SKORIN-KAPOV, Lea; MÄKI, Toni; HOSSFELD, Tobias. QoE in the Web: A dance of design and performance. In: 2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX). [S.l.: s.n.], May 2015. P. 1–7.

VARELA, Martín; MÄKI, Toni; SKORIN-KAPOV, Lea; HOSSFELD, Tobias. Towards an Understanding of Visual Appeal in Website Design. In: 2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX). [S.l.: s.n.], July 2013. P. 70–75.

VASSEUR, Aurélie; LÉGER, Pierre-Majorique; SÉNÉCAL, Sylvain. The Impact of Symmetric Web-Design: A Pilot Study. en. In: DAVIS, Fred D.; RIEDL, René;

BROCKE, Jan vom; LÉGER, Pierre-Majorique; RANDOLPH, Adriane; FISCHER, Thomas (Eds.). **Information Systems and Neuroscience**. Cham: Springer International Publishing, 2020. (Lecture Notes in Information Systems and Organisation), p. 173–180.

VELEZ, Pauline; ASHWORTH, Steven D. The Impact of Item Readability on the Endorsement of the Midpoint Response in Surveys. en. **Survey Research Methods**, v. 1, n. 2, p. 69–74, June 2007. Number: 2. ISSN 1864-3361.

VENNI, Julien; BÉTRANCOURT, Mireille. Aesthetics in Hypermedia: Impact of Colour Harmony on Implicit Memory and User Experience. In: COMPANION Publication of the 2020 International Conference on Multimodal Interaction. New York, NY, USA: Association for Computing Machinery, Oct. 2020. (ICMI '20 Companion), p. 215–219.

VET, Henrica C. W. de; TERWEE, Caroline B.; KNOL, Dirk L.; BOUTER, Lex M. When to Use Agreement Versus Reliability Measures. en. **Journal of Clinical Epidemiology**, v. 59, n. 10, p. 1033–1039, Oct. 2006. ISSN 0895-4356.

VOGEL, Marlene. Temporal Evaluation of Aesthetics of User Interfaces as one Component of User Experience. In: PROCEEDINGS of the Fourteenth Australasian User Interface Conference (AUIC2013). [S.l.: s.n.], 2013. P. 131–132.

WAN, Hongyan; JI, Wanting; WU, Guoqing; JIA, Xiaoyun; ZHAN, Xue; YUAN, Mengting; WANG, Ruili. A novel webpage layout aesthetic evaluation model for quantifying webpage layout design. en. **Information Sciences**, v. 576, p. 589–608, Oct. 2021. ISSN 0020-0255.

WANG, Chen; REN, Xiangshi. An Entropy-based Approach for Computing the Aesthetics of Interfaces. In: PROCEEDINGS of the 2018 ACM Companion International Conference on Interactive Surfaces and Spaces. New York, NY, USA: Association for Computing Machinery, Nov. 2018. (ISS '18 Companion), p. 57–61.

WEATHERS, Danny; SHARMA, Subhash; NIEDRICH, Ronald W. The Impact of the Number of Scale Points, Dispositional Factors, and the Status Quo Decision Heuristic on Scale Reliability and Response Accuracy. en. **Journal of Business Research**, v. 58, n. 11, p. 1516–1524, Nov. 2005. ISSN 0148-2963.

WEICHBROTH, Paweł. Usability of Mobile Applications: A Systematic Literature Study. **IEEE Access**, v. 8, p. 55563–55577, 2020. Conference Name: IEEE Access. ISSN 2169-3536.

WEN, Quan; WANG, Chen; SUN, Xiaoying; REN, Xiangshi. Exploration of the Relationship between UIDL and Interface Computational Aesthetics. In: PROCEEDINGS of the 2018 ACM Companion International Conference on Interactive Surfaces and Spaces. New York, NY, USA: Association for Computing Machinery, Nov. 2018. (ISS '18 Companion), p. 47–51.

WOHLIN, Claes; RUNESON, Per; HÖST, Martin; OHLSSON, Magnus C.; REGNELL, Björn; WESSLÉN, Anders. **Experimentation in Software Engineering**. [S.l.]: Springer Science & Business Media, June 2012. Google-Books-ID: QPVsM1_U8nkC. ISBN 978-3-642-29044-2.

WU, Chun Mao; LI, Pei. The Visual Aesthetics Measurement on Interface Design Education. en. **Journal of the Society for Information Display**, v. 27, n. 3, p. 138–146, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/jsid.751. ISSN 1938-3657.

WU, Ou; CHEN, Yunfei; LI, Bing; HU, Weiming. Evaluating the Visual Quality of Web Pages Using a Computational Aesthetic Approach. In: PROCEEDINGS of the fourth ACM international conference on Web search and data mining. New York, NY, USA: Association for Computing Machinery, 2011. (WSDM '11), p. 337–346.

WU, Ou; ZUO, Haiqiang; HU, Weiming; LI, Bing. Multimodal Web Aesthetics Assessment Based on Structural SVM and Multitask Fusion Learning. **IEEE Transactions on Multimedia**, v. 18, n. 6, p. 1062–1076, June 2016. Conference Name: IEEE Transactions on Multimedia. ISSN 1941-0077.

XING, Baixi; CAO, Hanfei; SHI, Lei; SI, Huahao; ZHAO, Lina. AI-driven user aesthetics preference prediction for UI layouts via deep convolutional neural networks. en. **Cognitive Computation and Systems**, v. 4, n. 3, p. 250–264, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1049/ccs2.12055. ISSN 2517-7567.

XING, Baixi; SI, Huahao; CHEN, Junbin; YE, Minchao; SHI, Lei. Computational model for predicting user aesthetic preference for GUI using DCNNs. en. **CCF Transactions on Pervasive Computing and Interaction**, v. 3, n. 2, p. 147–169, June 2021. ISSN 2524-5228.

YANG, Yaxin; WEN, Haiying; WU, Wenyu; ZHANG, Zhisheng; ZHU, Jianxiong. Evaluation of Digital Twin Interface Based on Aesthetics. In: 2022 28th International Conference on Mechatronics and Machine Vision in Practice (M2VIP). [S.l.: s.n.], Nov. 2022. P. 1–4.

YOSINSKI, Jason; CLUNE, Jeff; BENGIO, Yoshua; LIPSON, Hod. How Transferable Are Features in Deep Neural Networks? **arXiv:1411.1792 [cs]**, Nov. 2014. arXiv: 1411.1792.

ZEN, Mathieu. Metric-based evaluation of graphical user interfaces: model, method, and software support. In: PROCEEDINGS of the 5th ACM SIGCHI symposium on Engineering interactive computing systems. New York, NY, USA: Association for Computing Machinery, June 2013. (EICS '13), p. 183–186.

ZEN, Mathieu; VANDERDONCKT, Jean. Assessing User Interface Aesthetics based on the Inter-subjectivity of Judgment. In: PROCEEDINGS of the 30th International BCS Human Computer Interaction Conference. Poole, UK: BCS Learning & Development, July 2016.

ZEN, Mathieu; VANDERDONCKT, Jean. Towards an Evaluation of Graphical User Interfaces Aesthetics Based on Metrics. In: 2014 IEEE Eighth International Conference on Research Challenges in Information Science (RCIS). [S.l.: s.n.], May 2014. P. 1–12. ISSN: 2151-1357.

ZHANG, Jingfeng; HAN, Bo; WYNTER, Laura; LOW, Kian Hsiang; KANKANHALLI, Mohan. Towards Robust ResNet: A Small Step but A Giant Leap. **arXiv:1902.10887 [cs]**, July 2019. arXiv: 1902.10887.

ZORZO, Avelino Francisco; NUNES, Daltro; MATOS, Ecivaldo; STEINMACHER, Igor; ARAUJO, Renata Mendes de; CORREIA, Ronaldo; MARTINS, Simone. **Referenciais de Formação para os Cursos de Graduação em Computação**. [S.l.]: SBC, 2017. ISBN 85-7669-424-7.

GLOSSARY