Department: Head
Editor: Name, xxxx@email

How Much Data Does ML in HCI Need?: Re-Estimating the Dataset Size for CNN-Based Models of Visual Perception

Maxim Bakaev*, maxis81@gmail.com
Independent UX Consultant, Novosibirsk, Russia

Sebastian Heil
Technische Universität Chemnitz, Germany

Vladimir Khvorostov
Independent Software Developer, Novosibirsk, Russia

Martin Gaedke
Technische Universität Chemnitz, Germany

***Abstract*—AI-based user interface (UI) design and evaluation are currently constrained by the scarcity of human-generated training data. Correspondingly, choosing appropriate neural network architecture and carefully planning the sample size is essential for building accurate ML models. Previously, we have estimated that for a convolutional neural network (CNN) to produce better mean squared errors (MSE) than feature-based models, the required training dataset size should be about 3000. Our current validation study with about 4000 web UIs and 233 subjects suggests that the estimation should be closer to 17,000. We propose corrected regression models suggesting that the dataset size effect is better described with a logarithmic function. We also report significant differences in MSEs between the employed perception dimensions, with Aesthetics models having MSE 21.5% worse than Complexity and 12.1% worse than Orderliness.**

Index Terms: User Interfaces, Aesthetics, Visual Complexity, Convolutional Neural Networks.

Introduction

Data in HCI and Where to Find Them

Human-generated data tend to be scarce and/or expensive, unless it emerges as a by-product of

 Published by the IEEE Computer Society

some human activity. In HCI, interaction logs provide abundant data deposits for machine learning (ML) models related to mouse movements, scrolling, clicking, time on task, etc. (see e.g. in [15]). However, data on user-subjective dimensions of UX – satisfaction, aesthetic impression, emotional usability – are still largely collected via surveys. Correspondingly, a typical HCI dataset that is not logs-based has merely 1000s of records, which understandably prevents straightforward application of Deep Learning methods.

Moreover, the mainstream ML approaches for overcoming the limited dataset sizes have a hard time in some areas of HCI. Data Augmentation does not work easily e.g. for aesthetic impressions – rotating or brightening a UI design could result in a radically different feedback from a user. Resizing or distorting a graphical interface might bias visual complexity perception, and so on.

Transfer Learning can be feasible if established pre-trained models from e.g. Computer Vision can be brought in (like in [8]), though sometimes disadvantages of their direct usage are noted [6]. However, ML models within HCI itself are not so well developed or organized. Finding and successfully re-using an appropriate user behavior model is generally problematic. Many of them are task- and UI-dependent, so the accuracy suffers if the application context is modified – e.g., web UIs belong to a different domain [4].

Among the approaches that remain is the choice of the model architecture, the features (if they are involved), the hyperparameters, and of course careful planning of the dataset size.

Related Work

One substantial field that is similarly chocked with "expensive" data and limited dataset sizes is Medicine. In it, a particular research focus is no justifying the necessary and sufficient sample size for a study [14]. There are more or less universal methods and rules of thumb, but they often disagree considerably: e.g., recommending either 50 or 10 samples per the number of weights in NN [1].

Most researchers agree though that the general dependence between the sample size and the model accuracy, particularly for vision-related tasks, is logarithmic [18], [16], [20]. This is seemingly confirmed in practice: in [19], as a dataset became about 16 times larger, the increase in the models' accuracy was about 1.11 times. Moreover, it is believed that ML algorithms and models, depending of their complexity and some other factors, have "saturation points", after which they are not able to make good usage of extra training data [20]. Correspondingly, the choice of the proper model architecture becomes essential.

Research Question

Previously, we benchmarked convolutional neural network (CNN) models and feature-based ones in modelling subjective dimensions of user visual perception. The performance comparison was done on a training dataset of just over 2000 websites, which is a rather typical size in HCI. We found that the CNN models were inferior in all the considered subjective dimensions (Complexity, Aesthetics, Orderliness), but speculated that at dataset size of about 2900 the situation should reverse. The results were presented in HCII 2022 conference and published as [2].

In our current paper, the research question is whether this prediction comes true. For the validation, we extend the volume of the datasets that we use to about 4000 in total (with the corresponding increase in the number of human annotators to 233). We believe that replication and validation of the scientific results matter, particularly given the current "replication crisis" that has affected not just softer scientific fields like Psychology, but also HCI and Computer Science [7].

The rest of our paper is structured as follows. In Section 2 we describe our experimental study and detail the derived CNN models. In Section 3 we use statistical analysis to validate our previous prediction regarding the more effective NN architecture and further explore the effect of the dataset size. In the final section we summarize our findings, discuss the limitations and outline the direction for further research.

Experiment Description

Design and Material

Our goal in the current study was to validate some outcomes of the previous one (hereinafter Experiment 1) [2]. Experiment 1 was mainly

dedicated to the comparison of feature-based (ANN) and CNN architectures with respect to modelling three user-subjective visual perception dimensions, assessed in a dedicated survey on a Likert scale from 1 to 7 [4]:

1) Complexity: how visually complex the web UI appears in the screenshot;
2) Aesthetics: how aesthetically pleasant the web UI appears;
3) Orderliness: how orderly the web UI appears.

We found that on average the ANN models had better mean squared error (MSE) of 0.739, compared to 0.859 for the CNN models. We discovered that as the training dataset size (N) increased, MSE for the CNN models would improve, whereas for the ANN models it would not. Thus, we speculated that **at $N > 2912$ the convolutional NN architecture will start yielding superior results** over the feature-based one.

However, for the dataset of website homepage screenshots employed in Experiment 1 (hereinafter Dataset 1) we had N ranging from 263 to only 2154. Correspondingly, to validate our prediction, we initiated the new experiment (hereinafter Experiment 2), which added another dataset (hereinafter Dataset 2) to Dataset 1. Both datasets consisted of several sub-datasets identified by the websites' thematic domains (see in Table 1). Since we would capture the homepages, each screenshot is a different website.

Screenshots for the Dataset 1 were automatically collected using our dedicated script that followed the homepage URIs provided by student volunteers and captured the rendered webpages as 1280x960 or 1280x900 pixels images. We made sure that the websites were not dedicated to a famous brand or a company, to avoid the bias in the subjective assessments (see [4] for more detail on the Dataset 1 collection process). Besides the [2] that we are validating, parts of the dataset have been used in some other research projects ([4], [12]). Some representative screenshots can be found in [5, Fig. 2]. The Dataset 2 was merged from several very different sub-datasets collected by various researchers throughout the past decade, as detailed in Table 1.

The quality of the models in our study was operationalized as Mean Squared Error (MSE), which is arguably the most widely used loss function for neural network models that perform regression tasks:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad (1)$$

$\hat{y}_i$ is the predicted value and y_i the true value. The closer the MSE is to 0, the better is the forecast of a model.

The independent variables in our study were:

- The training dataset size in Experiment 1 (Dataset 1): N_1, varying in the range of $263 \leq N_1 \leq 2154$,
- The training dataset size in Experiment 2 (Dataset 1 plus Dataset 2): N_2, varying in the range of $263 \leq N_2 \leq 3379$,
- The subjective visual impression scale: *Scale* $\in \{$Complexity / Aesthetics / Orderliness$\}$.

The intermediate dependent variables were the models, while the derived dependent variables actually used in the study were the models' MSEs:

- MSE for CNN models in Experiment 1: MSE_1,
- MSE for CNN models in Experiment 2 that have $N_2 > 2912$: MSE_2.

Thus, we had the following null hypotheses in Experiment 2:

$H_0$1: MSE_2 for the models with $N_2 > 2912$ is not lower than the value of 0.739 obtained for the feature-based models in Experiment 1.

$H_0$2: There is no effect of dataset size on MSE.

$H_0$3: There is no effect of scales on MSE.

Subjects

Each of the three visual perception dimensions was represented as a Likert scale ranging from 1 (the lowest degree of the characteristic) to 7 (the highest degree). The subjective evaluations of the websites per the scales were assessed in two dedicated surveys (with about 2.5 years between them) by two groups of subjects (most of them were Bachelor's and Master's students, but also university staff and IT specialists):

1) For Dataset 1: 137 participants (67 female, 70 male), whose ages ranged from 17 to 46 (mean 21.18, SD = 2.68). The majority of

Table 1. Dataset: the screenshots collected from 6 domains and 8 sub-datasets, year of creation, resolution, and number of screenshots used from each dataset.

Domain/ sub-dataset	Description	Year	Resolution, px	Screens
Dataset 1				
Culture	Websites of museums, libraries, exhibition centers, other cultural institutions.	2018	W: 1280 H: 960/900	746
Food	Websites dedicated to food, cooking, healthy eating, etc.	2018	W: 1280 H: 960/900	369
Games	Websites dedicated to computer games.	2018	W: 1280 H: 960/900	362
Gov	E-government, non-governmental organizations' and foundations' websites.	2018	W: 1280 H: 960/900	346
Health	Websites dedicated to health, hospitals, pharmacies, medicaments.	2018	W: 1280 H: 960/900	541
News	Online and offline news editions' websites, news portals.	2018	W: 1280 H: 960/900	328
			Dataset 1 Total:	2692
Dataset 2				
AVI_14	From [10].	2014	W: 1278-1294 H: 799-800	124
Banks	Screenshots of banks' websites.*	2022	W: 1440 H: 960	287
CHI_15	From [11].	2015	W: 1280 H: 800	68
CHI_20	From [12].	2020	W: 720 H: 500-800	262
ECommerce	Screenshots of e-commerce websites.*	2022	W: 1440 H: 960	148
English	From [13].	2013	W: 1018-1024 H: 675-768	303
Foreign	From [13].	2013	W: 1024 H: 768	51
IJHCS	Part of the dataset from [17] via [12].	2012	W: 1000 H: 798-819	149
			Dataset 2 Total:	1371
			Dataset 1+2 Total:	4063

* Provided by A. Miniukovich within the framework of the project FWF M2827-N.

the participants were Russians (89.1%), the rest being from Bulgaria, Germany, South Africa, etc. In total, they provided 35,265 assessments for the 2,692 screenshots from the 6 domains.

2) For Dataset 2: 96 participants (27 female, 69 male), whose ages ranged from 19 to 25 (mean 21.02, SD = 1.30). The majority of the participants (93.8%) were Russian, with the others representing Uzbekistan. In total, they provided 24,114 assessments for the 1,371 screenshots from the 8 domains.

The subjects took part in the experiment voluntary and no random selection was performed. More details on the procedure and our specially developed online survey system can be found in [2]. The assessments were averaged per website and used as the output data for the models.

The Models

For greater statistical power of the comparisons and in order to explore the effect of the datasets sizes, we trained the models using combinations of the sub-datasets (e.g., Culture, Culture + News, Food + News + Gov, etc.). In Experiment 1, for each of the three subjective scales we trained 2^6-1 = 63 CNN models, covering all the possible combinations of the 6 sub-domains. In Experiment 2, we had $8 + 6 = 14$ sub-domains, but we were interested in the combinations that resulted in training dataset sizes over 2912. There were

118 such possible combinations for each of the three scales.

To construct and train the models, we used the `Colab` service freely offered by Google (TensorFlow 2.5 environment with Keras 2.4). The CNN models were built using a modified GoogLeNet architecture and Adam optimization algorithm. The architecture was modified to work with the input images size of 900x600 (to which all the screenshots were resized). Also, the output layer was replaced with a single neuron layer to accommodate to the regression task. The machine that we used to train the models had four i7-3930K CPUs @ 3.20GHz, 16 GB of memory and NVIDIA Quadro RTX 5000. The models were trained until the verification accuracy began to decrease for several epochs in a row, i.e., a stopping mechanism was employed. In accordance with the usual ML practices, 80% of the samples were used for training and 20% were used for testing.

Results

Descriptive Statistics

In addition to the 189 CNN models that we had from Experiment 1, in Experiment 2 we built and trained another 1624 models, which took another 250.1 hours, i.e. 554.5 s per model (SD=346.5). ANOVA suggests significant difference in the training time ($F_{2,1621} = 14.2, p < 0.001$) for the Aesthetics models (mean=617.9) compared to both Complexity (mean=531.7) and Orderliness (mean=513.9).

The mean training dataset size was 1094.3 ($SD = 448.1$) in Experiment 1, but already 1925.1 ($SD = 1127.9$) in Experiment 2. Pearson correlation between the training time and N_1 was significant ($r_{189} = 0.765, p < 0.001$), as well as between the time and N_2 ($r_{1624} = 0.761, p < 0.001$). These close positive correlations are in line with the ML theory and NN engineering practice.

Validation of the Previous Results

Of the models obtained in Experiment 2, 348 had $N_2 > 2912$ and could be used in the validation. The mean values and standard deviations for the MSE values obtained in Experiment 1 and Experiment 2 are presented in Table 2. The Shapiro-Wilk's tests suggest that the normality hypotheses had to be rejected for both MSE_1 ($W_{189} = 0.901, p < 0.001$) and MSE_2 ($W_{348} = 0.974, p < 0.001$). There were no significant Pearson correlations between the training times and either MSE_1 ($p = 0.143$) or MSE_2 for the 348 models ($p = 0.995$). This suggests that the training stopping mechanism worked properly and training the models longer would not yield better errors.

Table 2. Descriptive statistics for MSE in the two experiments, per the scales.

Scale	MSE_1 (n=189)	MSE_2 (n=348)
Complexity	0.750 (0.127)	0.729 (0.066)
Aesthetics	0.968 (0.182)	0.929 (0.083)
Orderliness	0.859 (0.122)	0.817 (0.083)
All	0.859 (0.170)	0.825 (0.112)

For the 348 models, the mean training dataset size amounted to 3031.9 ($SD = 102.1$). Thus, by increasing the mean dataset size by 177%, we managed to obtain the overall increase of MSE_2 over MSE_1 equal to 4.12%. ANOVA suggests that this difference in MSEs was significant ($F_{1,535} = 7.697, p = 0.006$).

However, MSE_2 was still inferior to the MSE of 0.739 that we previously obtained in Experiment 1 for the feature-based models [2, Table 2]. So, our prediction that at $N > 2912$ the models built with CNN architecture will gain superior MSEs **was not fulfilled and $H_0$1 could not be rejected**.

Improvement of the MSE Models from Experiment 1

Previously, in Experiment 1, we had proposed a linear regression model [2, Eq. (2)] for MSE in CNN models with the training dataset size as the factor. The model had rather low $R^2 = 0.03$, but was significant ($F_{1,187} = 5.85, p = 0.017$):

$$MSE_1 = 0.932 - 0.633 \cdot 10^{-4} N_1 \qquad (2)$$

We had also constructed the linear regression models per the scales, for the **difference between the MSE values** obtained when training CNN and feature-based (MSE_{ANN}) models on the same sub-datasets. The models were significant for Aesthetics ($F_{1,61} = 3.99, p = 0.050, R^2 =$

0.06) and Orderliness ($F_{1,61} = 5.34, p = 0.024, R^2 = 0.08$), but not for Complexity ($F_{1,61} = 0.42, p = 0.520$):

$$(MSE_1 - MSE_{ANN})_{Aesthetics} = \\ = 0.315 - 1.082 \cdot 10^{-4} N_1 \quad (3)$$

$$(MSE_1 - MSE_{ANN})_{Orderliness} = \\ = 0.188 - 0.900 \cdot 10^{-4} N_1 \quad (4)$$

Since in Experiment 2 we witnessed that the improvement in MSE for CNN models happens slower than we predicted, we re-considered the regression models from Experiment 1 using logarithmic function for N_1. The logarithmic regression models for MSE_1 ($F_{1,187} = 8.295, p = 0.004, R^2 = 0.042$), Aesthetics ($F_{1,61} = 7.639, p = 0.008, R^2 = 0.111$) and Orderliness ($F_{1,187} = 7.512, p = 0.008, R^2 = 0.110$) all had higher R^2s and p-values than their linear counterparts. The model for the Complexity scale remained non-significant ($F_{1,61} = 0.051, p = 0.823$).

$$MSE_1 = 1.351 - 0.071 \cdot ln(N_1) \quad (5)$$

$$(MSE_1 - MSE_{ANN})_{Aesthetics} = \\ = 1.110 - 0.133 \cdot ln(N_1) \quad (6)$$

$$(MSE_1 - MSE_{ANN})_{Orderliness} = \\ = 0.749 - 0.096 \cdot ln(N_1) \quad (7)$$

We should especially note that the linear regression model for MSE_{ANN} was not significant ($F_{1,187} = 0.200, p = 0.655$) in Experiment 1. We now tried the **logarithmic** function in regression model for MSE_{ANN}. The model was not significant ($F_{1,187} = 0.210, p = 0.648$), so the value of 0.739 can still be considered the valid threshold for the feature-based models' MSE.

Re-Consideration of the Dataset Size Effect

We built the corrected regression model for MSE due to the dataset size (*N*) with logarithmic function and using the CNN models from both Experiment 1 (n=189) and from Experiment 2 (n=1624). The model was highly significant ($F_{1,1811} = 86.404, p < 0.001, R^2 = 0.046$), **thus rejecting $\mathbf{H_0}$2**:

$$MSE = 1.179 - 0.045 \cdot ln(N) \quad (8)$$

The corrected model suggests that the threshold value of 0.739 will be reached at the dataset size of 16,714. This is over 5 times higher than our prediction of 2912, previously made with the linear model in Experiment 1.

Analysis of the Visual Perception Scales

We further tested the effect of the scales on MSE_2 in Experiment 2. ANOVA and post-hoc analysis revealed highly significant differences (at $\alpha = 0.001$) between all the three scales ($F_{2,1621} = 413.7, p < 0.001$) – hence, **$\mathbf{H_0}$3 had to be rejected**. This finding is completely in line with the effect of the scales we have previously discovered in Experiment 1.

Similarly, we constructed the corrected regression model (cf. [2, Eq. (5)]) for MSE in Experiment 2 with dummy variables (i.e., having the values 0/1): $Scale_A$ (has the value of 1 if the current model predicts Aesthetics) and $Scale_O$ (has the value of 1 if the current model predicts Orderliness). The rational scale factor in the regression was logarithm of the training dataset size for the current model. All the variables turned out to be significant (at $\alpha = 0.05$) in the resulting model, which had $R^2 = 0.374$ ($F_{3,1809} = 360.6, p < 0.001$):

$$MSE = 1.080 - 0.045 \cdot ln(N) + \\ + 0.227 Scale_A + 0.067 Scale_O \quad (9)$$

Thus, the model explained 9.68% more variance in MSE than the corresponding model in Experiment 1, even though that model ([2, Eq. (5)]) had more factors.

Discussion and Conclusion

In HCI, ML models of user behavior are often based on "expensive" data collected via dedicated surveys, so the field has to cope with datasets of limited size. These might be too scarce for the data-hungry Deep Learning models, so feature-based approaches still have their place in visual analysis of UIs.

Previously (Experiment 1), benchmarking these two approaches in predicting three dimensions of user visual perception, we speculated that once the number of webpages screenshots in the training dataset reaches about 3000, CNN models would start having better MSEs [2]. In our current paper (Experiment 2), dedicated to the validation of this prediction on a larger dataset of about 4000 websites, we found that it was too optimistic ($H_0 1$ not rejected). With the increase of the average training dataset size by 177%, the MSE of the CNN models had improved merely by 4.12%.

Now, the corrected prediction ($H_0 2$ rejected), based on logarithmic dependence and 1813 models (instead of 189 in [2]), suggests that **the dataset would need to include over 16,000 web page screenshots** for the Deep Learning to reveal its advantage. In our defence, we can note that the 5-fold disparity in recommended sample sizes is not something unprecedented [1], and in our case can be explained by the replacement of the linear function with the logarithmic one, which also better aligns with the theoretical considerations [16]. We have improved the regression model (9) for foretelling the models' errors in similar ML experiments, which now explains 9.68% more variance in MSE (cf. [2, Eq. (5)]).

The differences in the visual perception scales with regard to MSE that we previously found in Experiment 1 were confirmed in Experiment 2 ($H_0 3$ rejected). The models predicting Aesthetics had significantly higher errors and took more time to train, compared to the ones for Complexity (MSE -21.5%) and Orderliness (MSE -12.1%). This finding is in line with the related work in HCI, where aesthetic impression is generally harder to predict (typical R^2 values of about 0.5 [9, Table 15]), and complexity (typical R^2 values of about 0.65 [13], [3]) is sometimes used as the mediator for it [12].

As for the limitations of our study, we should note that the absolute MSE and R^2 values were rather high. One of the reasons might be that the standard GoogLeNet convolutional network architecture, which we relied on, is primarily intended for image classification, not predicting subjective impressions. Also, in one of our own previous experiments [4], we obtained comparable MSEs for ANN models (on average, 0.928 for Complexity, 1.09 for Aesthetics, 1.119 for Orderliness), which however had simpler architecture and did not optimize hyperparameters. In any case, our goal was not to develop ML models for production use, but to compare their parameters. In our statistical analyses we would employ samples of over 1600 models, so we assume reasonably high validity.

All in all, we hope that our study both reinforces the importance of validation and replication in science, and provides useful insights for researchers and practitioners who apply AI methods in HCI.

Acknowledgements

The reported study was funded by RFBR according to the research project No. 19-29-01017 and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 416228727 – SFB 1410.

REFERENCES

1. Alwosheel, A., van Cranenburgh, S., Chorus, C.G.: Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. Journal of choice modelling **28**, 167–182 (2018)
2. Bakaev, M., Heil, S., Chirkov, L., Gaedke, M.: Benchmarking neural networks-based approaches for predicting visual perception of user interfaces. In: International Conference on Human-Computer Interaction. pp. 217–231. Springer (2022)
3. Bakaev, M., Heil, S., Khvorostov, V., Gaedke, M.: Auto-Extraction and Integration of Metrics for Web User Interfaces. Journal of Web Engineering **17**(6), 561–590 (2019). https://doi.org/10.13052/jwe1540-9589.17676
4. Bakaev, M., Speicher, M., Heil, S., Gaedke, M.: I Don't Have That Much Data! Reusing User Behavior Models for Websites from Different Domains. In: Bielikova, M., Mikkonen, T., Pautasso, C. (eds.) Web Engineering, Lecture Notes in Computer Science, vol. 12128, pp. 146–162. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-50578-3_11
5. Boychuk, E., Bakaev, M.: Entropy and Compression Based Analysis of Web User Interfaces. In: Lecture Notes in Computer Science, pp. 253–261. Springer International Publishing (2019). https://doi.org/10.1007/978-3-030-19274-7_19
6. Chen, J., Xie, M., Xing, Z., Chen, C., Xu, X., Zhu, L., Li, G.: Object detection for graphical user interface: Old fashioned or deep learning or a combination? In:

proceedings of the 28th ACM joint meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering. pp. 1202–1214 (2020)
7. Cockburn, A., Dragicevic, P., Besançon, L., Gutwin, C.: Threats of a replication crisis in empirical computer science. Communications of the ACM **63**(8), 70–79 (2020)
8. Dou, Q., Zheng, X.S., Sun, T., Heng, P.A.: Webthetics: quantifying webpage aesthetics with deep learning. International Journal of Human-Computer Studies **124**, 56–66 (2019)
9. Lima, A.L.d.S., Gresse von Wangenheim, C.: Assessing the visual esthetics of user interfaces: a ten-year systematic mapping. International Journal of Human–Computer Interaction **38**(2), 144–164 (2022)
10. Miniukovich, A., De Angeli, A.: Quantification of interface visual complexity. In: Proceedings of the 2014 international working conference on advanced visual interfaces. pp. 153–160 (2014)
11. Miniukovich, A., De Angeli, A.: Computation of interface aesthetics. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 1163–1172 (2015)
12. Miniukovich, A., Marchese, M.: Relationship between visual complexity and aesthetics of webpages. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–13 (2020)
13. Reinecke, K., Yeh, T., Miratrix, L., Mardiko, R., Zhao, Y., Liu, J., Gajos, K.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. pp. 2049–2058 (2013)
14. Riley, R.D., Ensor, J., Snell, K.I., Harrell, F.E., Martin, G.P., Reitsma, J.B., Moons, K.G., Collins, G., Van Smeden, M.: Calculating the sample size required for developing a clinical prediction model. Bmj **368** (2020)
15. Speicher, M., Both, A., Gaedke, M.: TellMyRelevance! predicting the relevance of web search results from cursor interactions. In: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. pp. 1281–1290 (2013)
16. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017)
17. Tuch, A.N., Presslaber, E.E., Stöcklin, M., Opwis, K., Bargas-Avila, J.A.: The role of visual complexity and prototypicality regarding first impression of websites: Working towards understanding aesthetic judgments. International journal of human-computer studies **70**(11), 794–811 (2012)
18. Vabalas, A., Gowen, E., Poliakoff, E., Casson, A.J.: Machine learning algorithm validation with a limited sample size. PloS one **14**(11), e0224365 (2019)
19. Zhang, X., Gao, X., He, L., Lu, W.: Mscan: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks. Neurocomputing **430**, 14–23 (2021)
20. Zhu, X., Vondrick, C., Fowlkes, C.C., Ramanan, D.: Do we need more training data? International Journal of Computer Vision **119**(1), 76–92 (2016)