Hindawi
Complexity
Volume 2019, Article ID 4659809, 12 pages
https://doi.org/10.1155/2019/4659809

Research Article

Avoiding the Inherent Limitations in Datasets Used for Measuring Aesthetics When Using a Machine Learning Approach

Adrian Carballal ,[1] Carlos Fernandez-Lozano ,[1,2] Nereida Rodriguez-Fernandez,[3] Luz Castro,[3] and Antonino Santos[1]

[1] *Computer Science Department, Faculty of Computer Science, University of A Coruña, A Coruña 15071, Spain*
[2] *Investigación Biomédica de A Coruña (INIBIC), Complexo Hospitalario Universitario de A Coruña (CHUAC), A Coruña 15006, Spain*
[3] *Computer Science Department, Faculty of Communication Science, University of A Coruña, A Coruña 15071, Spain*

Correspondence should be addressed to Adrian Carballal; adrian.carballal@udc.es

Received 17 August 2018; Revised 14 November 2018; Accepted 20 December 2018; Published 8 January 2019

Guest Editor: Jon McCormack

An important topic in evolutionary art is the development of systems that can mimic the aesthetics decisions made by human begins, e.g., fitness evaluations made by humans using interactive evolution in generative art. This paper focuses on the analysis of several datasets used for aesthetic prediction based on ratings from photography websites and psychological experiments. Since these datasets present problems, we proposed a new dataset that is a subset of DPChallenge.com. Subsequently, three different evaluation methods were considered, one derived from the ratings available at DPChallenge.com and two obtained under experimental conditions related to the aesthetics and quality of images. We observed different criteria in the DPChallenge.com ratings, which had more to do with the photographic quality than with the aesthetic value. Finally, we explored learning systems other than state-of-the-art ones, in order to predict these three values. The obtained results were similar to those using state-of-the-art procedures.

1. Introduction

Estimating aesthetic value and the complexity of an image is a technological challenge that has recently been addressed by numerous fields, including psychology and artificial intelligence. Several research groups have attempted to create computer systems that are able to learn the aesthetics perception of a group of human beings as a part of a generative system (such as evolutionary art systems) or that can be used for automatic image selection or ordering. Given the subjective nature of the aesthetic problem, the selection of the dataset for the training is vital. This paper explores a new way to build a dataset and provide initial results by using machine learning techniques.

Previous research studies [1, 2] have concluded that the degree of generalisation of some existing sample sets was not enough to take them as reference in the training of automated prediction and classification of images. Other functional limitations were identified in these datasets, which are also mentioned in this paper.

In order to solve the problems identified in these datasets, this paper describes the creation of a new set of images from the website DPChallenge.com, with greater statistical consistency. Besides, this new dataset was evaluated in terms of aesthetics and quality by a group of individuals under controlled experimental conditions. This makes it the first dataset evaluated by two different populations (the one evaluating at the DPChallenge.com portal and the one evaluating it in person).

With the new dataset created, several Machine Learning-based models were trained for the automated prediction of the aesthetic and quality value and that of DPChallenge.com.

This paper starts with a state-of-the-art section on the datasets created for the automated prediction and classification of images. In Section 3, the limitations found in such sample sets are provided. Section 4 describes the method for the creation of a new dataset with greater statistical coherence and the results of the evaluation procedure obtained under experimental conditions. Section 5 presents the Machine Learning models based on the prediction that were used as well as the results obtained in the training based on the three available criteria for the images of the proposed set. There is a section discussing the results and another one with the final conclusions.

2. State of the Art

Some authors, such as Datta et al. [3], Wang et al. [4], Ke et al. [5], and Luo et al. [6], conducted studies aimed at automated aesthetic classification using a number of technical characteristics such as lightness, saturation, Rule of Thirds, etc. For these experiments, sets of large-format photographs from websites and the evaluations made by the users of such sites were used. On the other hand, other authors, including Cela-Conde et al. [7], Forsythe et al. [8], and Nadal et al. [9], carried out aesthetic perception and image complexity experiments using a sample set with a more limited number of images, but evaluated by a specific set of people under controlled experimental conditions. A brief analysis of these sample sets is presented below.

2.1. Photo.net (2006). Datta et al. [3] created a dataset based on the photography website Photo.net, which has over a million images and 400,000 users. In this dataset, each image is rated on a range from 1 to 7 (1 being the worst possible score and 7 the best) based on aesthetics and originality. Statistical information on the rating can be found on the website. It does not provide information on the image evaluators, though. The full dataset comprises 3,581 images rated by at least 2 persons and has an average score between 3.55 and 7 and an overall total average of 5.06, with a standard deviation of 0.83. The high correlation found between the criterion of originality and aesthetics (Pearson's r = 0.891) might indicate that users most assuredly are not making such distinctions.

Datta et al. [3] and other researchers such as Wong et al. [8], who used this sample group, have established a division to obtain two different groups: (i) the images with an average score equal to or higher than 5.8 were branded high quality and (ii) those with scores equal to or lower than 4.2 were branded low quality. In the case of the study conducted by Datta et al. [3] a success rate of 70.12% was achieved in the global classification using Support Vector Machines (SVM): 68.08% for high quality images and 72.31% for low quality images.

2.2. Photo.net (2008). In 2008, a new study was published by Datta et al. [10], which introduced a second set of data from the website consisting of 20,278 images rated by an average of 16.81 persons with a standard deviation of 16.19. It should be noted that there were images evaluated by a minimum of four people and others by a maximum of 395. When comparing this study with the previous one, it becomes apparent that this statistical analysis is more complete, as it provides specific data for each image. The total set of images had at least four ratings per image, with scores ranging between 2.33 and 6.99, and a global average of 5.15, with a mean standard deviation of 0.58. From the same set, Wong et al. [8] displayed 44 metrics grouped into three categories with global characteristics, for which they used a reduced set of images from the original experiment down to a total of 3,161. After performing a classification using SVM with linear kernel and resorting to a crossed validation with 5 independent runs, 78.2% of the images were successfully classified (82.9% high quality and 75.6% low quality).

2.3. DPChallenge.com. Ke et al. [5] created a different sample set, which became one of the most commonly used in aesthetic classification experiments. For the construction of this set, the photography portal DPChallenge.com was used, with a total of 60,000 images rated by at least 100 persons being selected.

For the aesthetic classification experiments, two sets of 6,000 photographs were created by selecting the top and bottom 10% after arranging them according to their mean score. Subsequently, Ke et al. [5] carried out a subdivision into two new random subsets, thus obtaining 4 sets of 3,000 images (two high quality and two low quality sets). A set of each type was used to train the proposed systems, while the other was used to validate their capacity and efficacy.

2.4. Dataset Created by Psychologists. Cela-Conde et al. [7] created a dataset consisting of a final standardized set of 800 images divided into 5 categories: artistic abstract (AA), non-artistic abstract (AN), artistic representational (RA), non-artistic representational (RN), and photographs of natural scenes and human constructions (NHS).

The images were shown to a group of 240 participants (112 men and 128 women, with a mean age of 22.03 years and a standard deviation of 3.75), randomly divided into subgroups of 30 persons in a controlled experimental environment. The images were shown for five seconds and participants were asked to rate the visual complexity of a subset of stimuli on a Likert scale from 1 to 5 (1 being the worst possible score and 5 the best). Consequently, each image had a total of 30 ratings. The mean value obtained by each subgroup for each stimulus was the value considered to represent the complexity of this stimulus in the final set. The stimuli in this set were used by Cela-Conde et al. [7], Forsythe et al. [8], Nadal et al. [9], and Machado et al. [11].

3. Limitations Found in the Dataset Available

The study of the generalisation capacity of the analysed datasets led to the conclusion that they did not provide a satisfactory degree of generalisation: the correlation is greater when the validation set belongs to the same source of data as the training set. However, in experiments where the test was performed with a set from a source different from the training set, the correlation results decreased notably. A clear example in this regard can be seen in experiments conducted in previous research studies [1, 2]: when training a subset of 6,000 images from DPChallenge.com carried out by Ke et al. [5], the result of the correlation was 91.38%. If validated with another subset from the same source, however, the resulting percentage decreased down to 56.21%, when using, for example, the dataset from Photo.net created by Datta et al. in 2006 [3], and down to 55.39% with the dataset from Photo.net created by Datta et al. in 2008 [10].

Besides, the sample sets trained with ratings from the photography portals had some defects: the evaluation system did not have the same control as a psychological test because it was not possible to obtain all the information about the evaluating users or about the device used to see the image (smartphone, computer), or distance or lighting conditions; the amount of images might be insufficient as there was no justified reason to choose a sample size and there was a huge difference in the number of people rating each image; user evaluations could be easily conditioned by personal relationships with the creator of the work or a momentary surge in popularity of certain styles. Lastly, in one of the cases [3] it was shown that the users of these portals did not have enough basis to differentiate between aesthetic and originality criteria, with Pearson's correlation coefficient of 0.891. Furthermore, as these datasets were designed for binary classification, only the images rated with extreme scores (those obtaining the highest and lowest ratings) were used, leaving out of the set the images with intermediate ratings.

In the set created by Ke et al. [5] there was another limitation in the collected evaluations, as the DPChallenge.com portal operated as if it were a photography competition and there was no specification of any criteria to assess the images. Consequently, any user can evaluate the image on their own criteria, which may have nothing to do with those of other people.

On the other hand, in the dataset created by Cela-Conde et al. [7] the number of images presented by category was not balanced. Therefore, the obtained results cannot be considered as representative of the whole. Besides, the set was built on the basis of a considerable number of subsets of images, which resulted in the dataset eventually becoming a number of datasets of independent themes of smaller size, with less internal coherence.

Once the limitations of the studied datasets were identified, a new dataset was built for the aesthetic prediction of images. This dataset was evaluated by humans under controlled experimental conditions using a coherent set of images.

4. Building a New Dataset

After identifying the limitations discussed above in the existing sets of images, we created a new dataset for the prediction and classification of images, in which the process of human evaluation of the images was carried out under controlled experimental conditions. This new method is generally put forward in [1] and includes the advantages of the sets of images studied in this article. This new method of creation makes it possible to build a set of images with greater statistical coherence from the rating results on the photography website DPChallenge.com and is subsequently evaluated in a manner similar to the procedure used by Forsythe et al. [8]. Thus, we shall be able to analyse the correlation between the results obtained with subjects under controlled circumstances and those obtained through the photography portal.

4.1. Source Data. We began by collecting a set of images from the DPChallenge.com photography portal. The images on the DPChallenge.com portal are rated by users within the range $[1, 10]$, where 1 is the lowest possible score and 10 the highest. The only information about the score in DPChallenge.com is that a score of 1 is a "bad" photo, and a score of 10 is a "good" photo. So the score is not clearly related to aesthetics, photographic quality, or originality. Nevertheless, this portal has been used in the past to obtain data for aesthetic classification experiments [5, 12, 13]. The original idea behind this site was for it to be a place where friends could teach themselves to be better photographers by giving each other a "challenge" for the week. Methodologically, DPChallenge organises weekly competitions into "themes" represented by a word of phrase (e.g., "Alfred Hitchcock", "Abstract: Black and White II", "Color Portrait IV"). For the current study, this aspect of the evaluation is not taken into account.

Images were collected using a brute force process whereby all data from all images whose identifiers were between 10,000 and 172,000 in May 2012. All statistical information of the ratings was available for only 40,047 images. The images in this initial set were rated by an average of 233 subjects and the mean rating was 5.23 ± 0.78. All descriptive data are shown in Figure 1(a). The file with the evaluation data and the links to the images used (for copyright reasons) are publicly available at https://doi.org/10.6084/m9.figshare.6127295.v1. Figure 1(c) shows the arrangement of votes based on each range and Figure 1(b) displays the distribution of the mean evaluations of the images within the range of scores, showing in both cases that they apparently follow a Gaussian model.

4.2. Dataset Proposed. As noted above, only the images in which all the evaluation data were available were used. Then, only the images with at least 100 ratings were selected. The objective was that the mean value subsequently attributed to each image was the least biased possible.

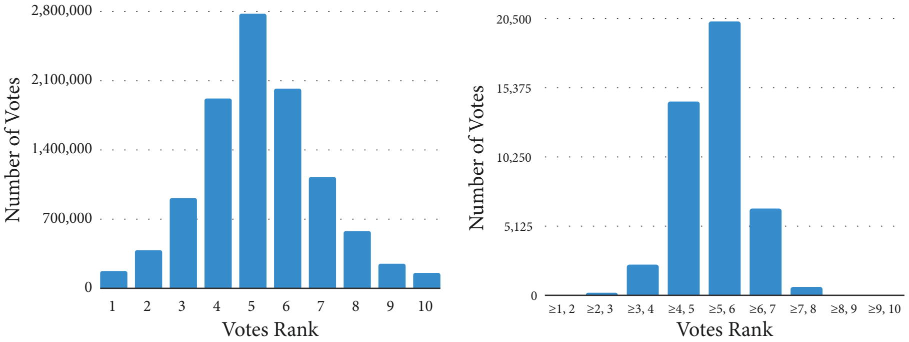
Once this selection was made, images were arranged in groups according to the mean ratings given on DPChallenge.com. The images in our selection were classified according to 9 scoring ranges, one for each integer value of valid evaluation. Then, a minimum number of images were set for all groups. In our case, the minimum number was 200 (see Figure 2(b)). There were no sets of images numerous enough with mean scores below 3 or higher than 8. Consequently, the used groups were collected from the $[3, 8]$ range. From these groups, 200 images with the lowest standard deviation were selected. In other words, these were images with the most internally consistent scores. We used the more consistent image set in order to build a dataset that can be used as ground truth dataset. The descriptive data for each of the ranges are detailed in Table 1. Figure 2 shows (a) the distribution of the number of votes within the range of valid

Table 1: Descriptive data for each of the five sets of 200 images that make up the proposed dataset.

Range	[3, 4)	[4, 5)	[5, 6)	[6, 7)	[7, 8)
Average	3.5943	4.4695	5.4975	6.4715	7.3112
Deviation	0.2613	0.2868	0.2894	0.2845	0.2335
Variance	0.0683	0.0822	0.0837	0.0809	0.0545
Kurtosis	-0.8224	-1.2370	-1.1765	-1.1595	-0.4500
Bias	-0.3998	0.1611	-0.0005	0.1099	0.6879
Minimum	3.0070	4.0130	5.0060	6.0030	7.0000
Maximum	3.9970	4.9970	5.9970	6.9940	7.9530

Descriptive data	
Images	44047
Average	5.2405
Deviation	0.7821
Variance	0.6117
Kurtosis	0.2480
Bias	-0.0182
Minimum	1.9951
Maximum	8.3900

(a)



(b) (c)

Figure 1: Characterisation of all 44,047 images initially obtained from DPChallenge. (a) Descriptive data, (b) arrangement of the number of votes within the range of valid ratings, and (c) distribution of mean image evaluations within the range of valid ratings.

scores and (b) the distribution of the mean ratings within the range of valid scores for the 1000-image dataset.

This process provides a set of images with equal number of elements for each range, with high scoring consistence, and which could eventually be the most representative.

4.3. Human Evaluation. The dataset proposed above was evaluated by a number of humans under controlled experimental conditions. According to Infinite Population Sampling [14] with a minimum sample size of 8 individuals and 95% of confidence level, the true population rating of an image can be obtained, with a margin of error of 3%.

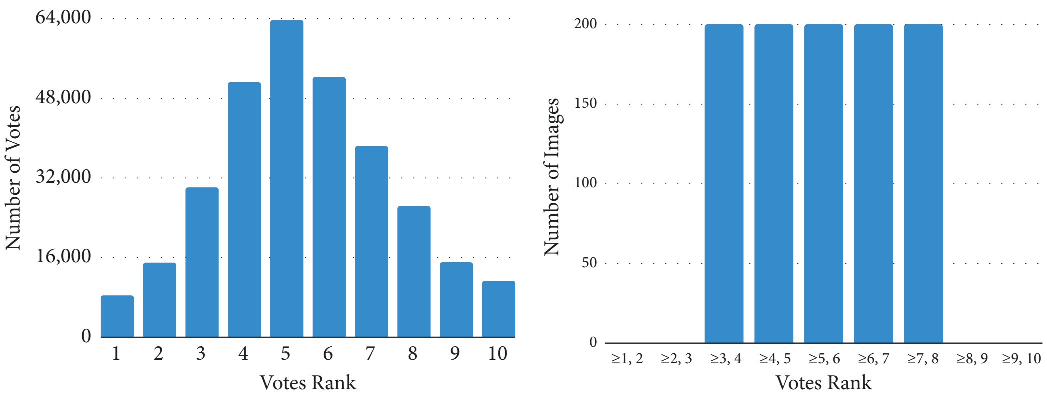
To this end, 5 subsets were created with randomly selected images out of a total of 1,000 available. Each person could rate the images in one or several of these subsets with a score between 1 and 5, where 1 is the lowest possible score and 5 the highest. Each set was evaluated by at least 10 persons (a total of 10,000 ratings).

Evaluations were carried out on February 1st and March 5th, 2018, by student volunteers of the University of A Coruña, Spain (mainly, students at the School of Communication Sciences). Ninety (33 male and 66 female) participants (18.7 years, age range 18-30) took part in this study. Each participant evaluated at least 200 images before the research study and under the same viewing conditions: screens with the same specifications, same lighting conditions, and same distance between evaluators and the screens.

For every image, users independently rated its aesthetic value and quality. The English translation of the text of the survey questions verbatim is: "In this task we want you to evaluate the quality and aesthetic value of each of the images that we propose. To score the "quality" you should look at the framing, focus, colors, etc. In general, professional photographs have higher quality than photographs taken by amateurs. The editing of images (use of Photoshop, filters, etc.) does not have to affect its quality. It may be that you do not like an image, but if it is well made, your quality score should be high. For the aesthetic score value we look for your personal opinion about the image, whether you like it or not. The semantic value should not influence. That is, a nice picture of a crying baby can have a high aesthetic value score."

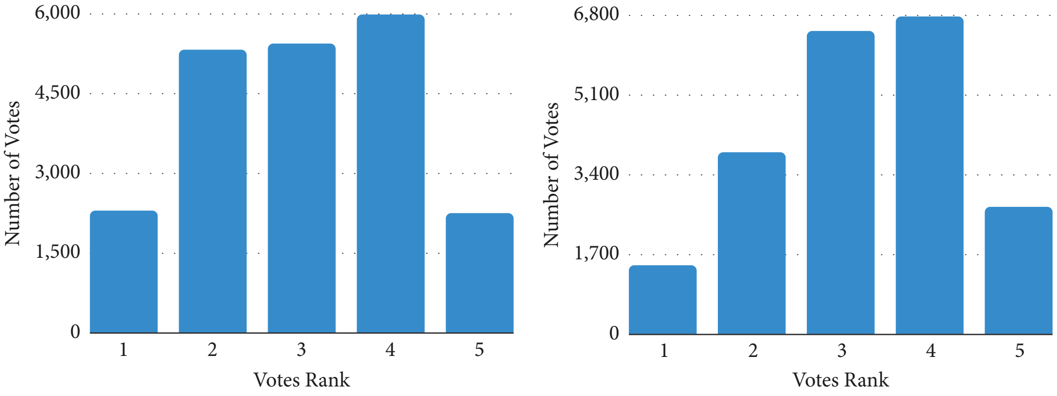
The data shown in Figure 3 correspond to the mean obtained for each image from the different evaluations made for both aesthetic and quality criteria.

The correlation between the scores given in person and those registered on the Dpchallenge.com platform was calculated (see Figure 4). Pearson's correlation between the mean score on Dpchallenge.com and the mean score was 0.692 according to the aesthetic criterion and 0.690 according to Spearman's. The mean correlation between DPChallenge.com and the mean according to the quality value was 0.748 according to Pearson's and 0.756 according to Spearman's. Lastly, the correlation between the two measures obtained in the in-person experiment (aesthetics/quality) was 0.787 according to Pearson's and 0.786 according to Spearman's, higher than in the other two correlations. Figure 4 shows the Scatterplots between ranks for the three possible combinations given the three criteria that are evaluated for the entire study.



(a) (b)

Figure 2: Characterisation of the 1000 images in the proposed set. (a) Distribution of the number of votes within the scoring range and (b) distribution of mean ratings in the images within the valid range of scores.



(a) (b)

Figure 3: Distribution of the mean aesthetic (a) and quality (b) ratings obtained in the control group.

5. Machine Learning Approach

In this study, some state-of-the-art models based on Machine Learning applied to the proposed input were proposed. The aim of these experiments was to study whether the existing correlation values between both human populations seen in the previous sections (DPChallenge and control group) can be *replicated* by a computer system for the proposed dataset.

5.1. Materials and Methods. To characterise the images that make up the study set, a feature extractor available in WND-CHARM [15] was used, which is a multipurpose image classifier that can be applied to a wide variety of image tasks. According to its developers, the system extracts a large set of image features, including polynomial decompositions, high contrast features, pixel statistics, and textures, among others. These features are computed on the raw image, transforms of the image, and transforms of transforms of the image. The final feature vector comprises 2905 variables, each of which reporting on a different aspect of image content. All features are based on greyscale images, so colour information is not currently used.

The authors tested the different computational models using a 10-fold cross-validation to split the data and 50 runs per model in order to evaluate the performance across different experiments. The performance of the models was evaluated using Spearman's correlation coefficient (rho) and Pearson's correlation coefficient (Pearson's r).

5.2. Computational Models. The authors performed several experiments in order to select the best model using the R package and MATLAB©. Some of the used computational models looked for the smallest subset of variables of the original set which provided a better performance [16], or at least equal to that obtained when using all the possible variables, considering this was a Feature Selection (FS) approach [17–19].

More specifically, the used methods were the following: the well-known Support Vector Machines-Recursive Feature Elimination (SVM-RFE) [20, 21] and the Generalized Linear Model with Stepwise Feature Selection (GLM) [22] which selects features that minimise the AIC score and the most basic standard Multiple Linear Regression (LM) without FS. The abilities of the RRegrs Package [23] were enhanced in

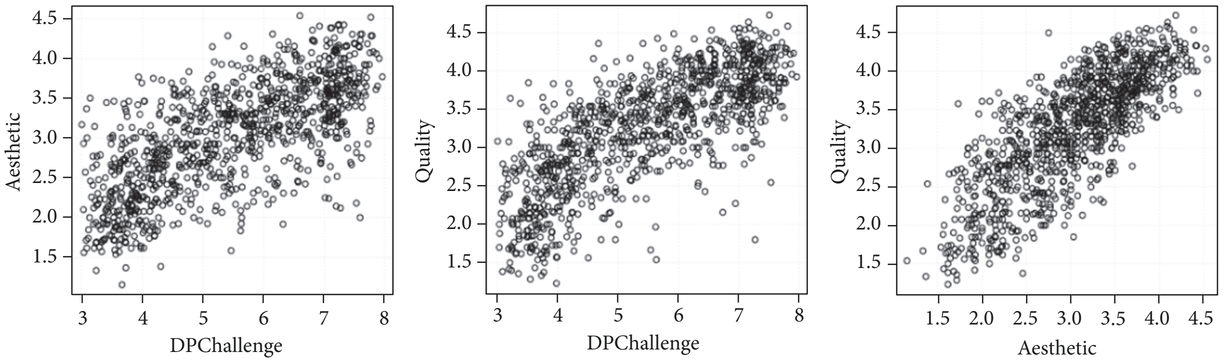


Figure 4: Scatterplots between ranks for the three possible combinations given the criteria evaluated for the entire study.

order to implement the SVM-RFE and GLM to avoid finding the best model according to the proposed methodology as, according to [24], it should be performed based on a null hypothesis test. This package was also enhanced in order to avoid the initial splitting process, and an external cross-validation process was performed to avoid selection bias as suggested by [25]. The last step was modified in order to easily extract the results for all the models.

The K-nearest neighbour (k-NN) algorithm is a technique based on the cluster theory. In this case, a variant called weighted k-NN [26] was used. It is based on the fact that a new observation particularly close to an observation within the learning set should have great importance in the decision-making process and, conversely, an observation that is at a further distance should have much less importance [27]. For this algorithm, only the hyperparameter k was tuned, which represented the number of neighbour data points that were considered closest. The range of values was from 1 to 5.

The generalized boosted models (GBM) applied the approach described in [28], to establish the foundation of boosting algorithms. GBM estimation involves an iterative process with multiple regression trees to capture complex and nonlinear relationships without overfitting the data [29, 30]. It works with continuous and discrete variables and is invariant to their monotonic transformations [31]. For this algorithm, the interactive depth was represented by the number of splits it had to perform on a tree (starting from a single node) and the number of trees that were tuned. The range of values used was from 1 to 4 and 100, 250, and 500 for the number of trees.

The design of our experiments was based on a novel methodology for the development of experimental designs in regression problems with multiple machine learning regression algorithms [32]. For each model described above, the optimal set of parameters was sought using hyperparameter optimisation.

5.3. Results. Figure 5 and Table 2 show the results obtained for each of the four methods studied according to Pearson's and Spearman's correlation value, using as reference the average ratings from the DPChallenge.com portal. Firstly, examining Spearman's correlation values, the maximum value of the SVM-based model was 0.574, using 1024 variables. The input set could be decreased down to 256 with no significant loss of performance (0.570), as the correlation values remained statistically constant between both figures. On the other hand, if we look at the values for Pearson's r, the same pattern remained, since, with 1024 input variables, 0.581 was obtained, whereas, with 256, 0.574 was obtained (with no significant difference in performance). In any case, both Spearman's and Pearson's values show a moderate uphill (positive) relationship, with the exception of k-NN.

The authors checked the significance of the difference between GLM, SVM, GBM, and k-NN with 256 input variables (see Figure 6) using a Kruskal-Wallis test, and our results showed that, with a very high level of confidence, SVM (cost = 2^{-6} y gamma=2^{-9}) was significantly better than the others with a p-value < 2.2 x 10^{-16}. Consequently, it could be stated that the minimum input set with the best results was the one with 256 input variables in combination with an SVM prediction model with specified parameters.

Once the method with the best results was identified using the average ratings obtained by the users of DPChallenge.com, the best SVM hyperparameters were calculated (cost = 2^{-4} and gamma=2^{-12} in both cases) training the scores for "aesthetics" and "quality" obtained in the above-mentioned experiment with humans.

As shown in Figure 7, the values for any of the 3 cases are below 0.60 on average. Specifically, it was 0.578 for DPChallenge, 0.456 for aesthetics, and 0.539 for quality, using as mean of performance Spearman's rho and 0.574, 0.451, and 0.562, respectively, using Pearson's r. On the negative side, it is particularly relevant that in the case of "aesthetics" there is a weak uphill (positive) relationship given the average value obtained with both measures.

6. Discussion

A correlation of 0.78 was obtained between the ratings based on aesthetics and those based on quality. This indicated that the evaluation teams distinguished between both criteria when compared with the measurements made by Datta et al. [3], where Pearson's correlation between aesthetics and originality was 0.891.

Table 2: Average results presented in Figure 5, identifying hyperparameters and input size for each model.

Size	Model	Pearson	SD	Hyperparameters
16	GLMNET	0.5320	0.0717	Alpha=0
16	GBM	0.5234	0.0738	Interaction.depth=4, n.trees=500
16	k-NN	0.4831	0.0744	k=12; distance=2
16	SVM	0.5389	0.0713	Cost = 16 Gamma=0.00984
32	GLMNET	0.5451	0.0709	Alpha=0
32	GBM	0.5266	0.0733	Interaction.depth=4, n.trees=500
32	k-NN	0.4851	0.0750	k=12; distance=2
32	SVM	0.5581	0.0732	Cost=0.397 Gamma=0.00984
64	GLMNET	0.5406	0.0669	Alpha=0
64	GBM	0.5474	0.0723	Interaction.depth=4, n.trees=500
64	k-NN	0.4898	0.0752	k=12; distance=2
64	SVM	0.5503	0.0691	Cost=2.52 Gamma=0.000244
128	GLMNET	0.5473	0.0745	Alpha=0
128	GBM	0.5425	0.0679	Interaction.depth=4, n.trees=500
128	k-NN	0.4926	0.0720	k=12; distance=2
128	SVM	0.5687	0.0676	Cost=0.397 Gamma=0.00155
256	GLMNET	0.5555	0.0719	Alpha=0,15
256	GBM	0.5479	0.0704	Interaction.depth=4, n.trees=500
256	k-NN	0.4776	0.0774	k=12; distance=2
256	SVM	0.5778	0.0671	Cost=2.52 Gamma=0.000244
512	GLMNET	0.5748	0.0701	Alpha=0,15
512	GBM	0.5482	0.0765	Interaction.depth=4, n.trees=500
512	k-NN	0.4845	0.0758	k=12; distance=2
512	SVM	0.5747	0.0683	Cost=2.52 Gamma=0.000244
1024	GLMNET	0.5644	0.0708	Alpha=0,15
1024	GBM	0.5473	0.0685	Interaction.depth=4, n.trees=500
1024	k-NN	0.4908	0.0777	k=12; distance=2
1024	SVM	0.5782	0.0670	Cost=0.397 Gamma=0.000244
2048	GLMNET	0.5602	0.0733	Alpha=0,15
2048	GBM	0.5465	0.0692	Interaction.depth=4, n.trees=500
2048	k-NN	0.4482	0.0815	k=12; distance=2
2048	SVM	0.5723	0.0685	Cost = 2.52 Gamma=0.000244
fulldataset	GLMNET	0.5590	0.0719	Alpha=0,15
fulldataset	GBM	0.5476	0.0690	Interaction.depth=4, n.trees=500
fulldataset	k-NN	0.4299	0.0825	k=12; distance=2
fulldataset	SVM	0.5554	0.0721	Cost=2.52 Gamma=0.000244

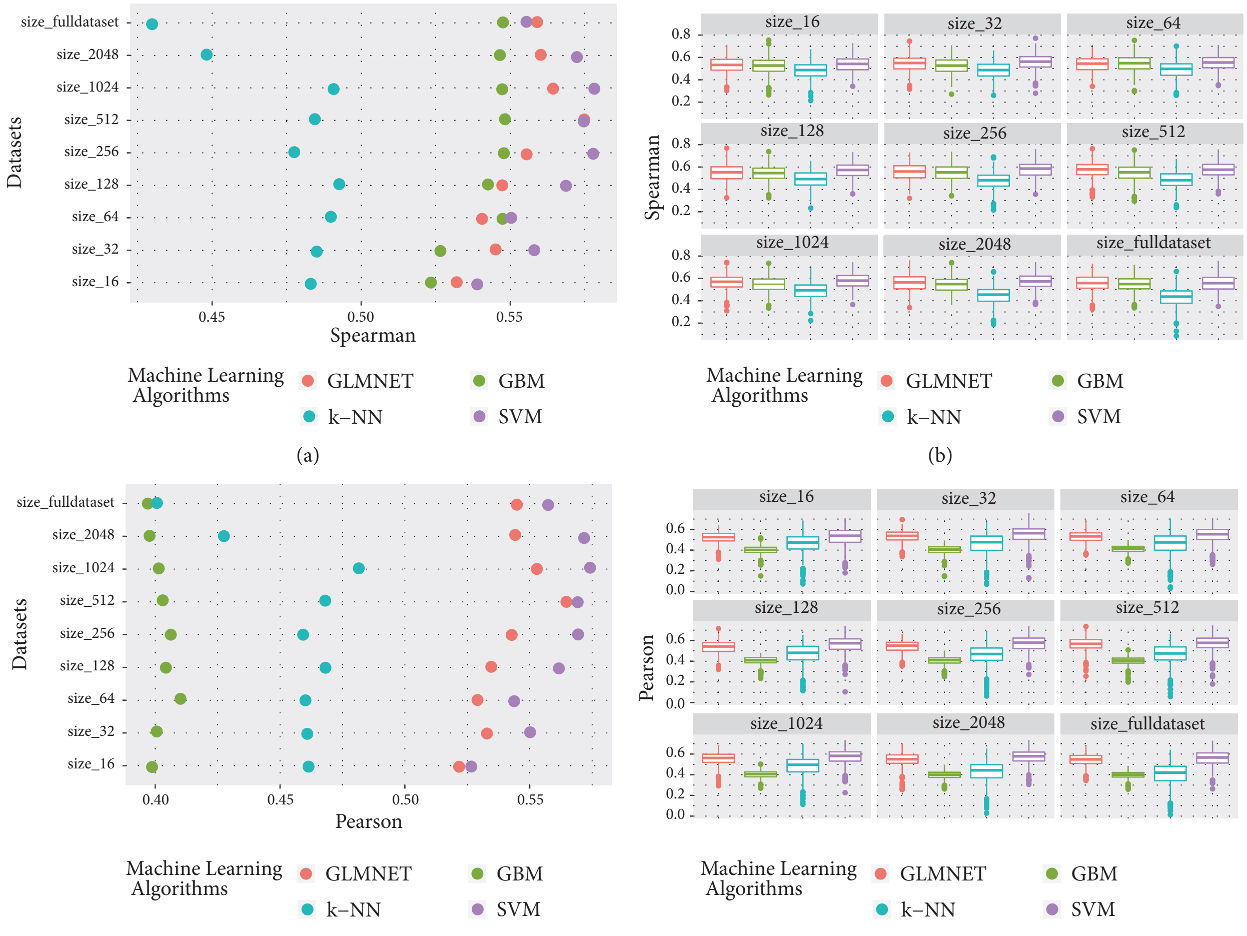
Regarding the correlation between DPChallenge and quality and aesthetics individually, we should begin by underscoring that the highest correlation was between DPChallenge and quality, which suggests that, at DPChallenge, the photographic quality is valued over the aesthetic value of the image.

In our opinion, there was no single reason that explained the difference between the correlations regarding DPChallenge, as far as the aesthetic and quality values were concerned:

(i) In the case of DPChallenge, the users' rating may be conditioned by affinity with the author of the photograph as we were dealing with a competition whereas in the case of the control group, the experimental conditions were controlled (for instance, everyone used the same screen model, at the same distance, with the same ambient light, etc.).

(ii) At DPChallenge, numerous devices can be used (smartphones, tablets, and high resolution screens) and conditions such as viewing distance and ambient light are heterogeneous.

(iii) In the case of the in-person group, the evaluation criteria were established: aesthetics and quality. At DPChallenge, as mentioned above, we were dealing with a photography competition and many different things may be evaluated such as quality, aesthetics, originality, etc.



(a)

(b)

(c)

(d)

Figure 5: Results obtained for the four models proposed and optimised by hyperparameterisation. On the right, the mean values for Spearman's (top) and Pearson's r (bottom) are shown. On the left, the distributions of all 50 independent runs for each optimum model (Spearman top and Pearson bottom) are shown with different input sizes tested using FS.

(iv) On the other hand, in the case of in-person ratings, the minimum per image was 10 whereas, for the evaluations from DPChallenge, the minimum was 100 for each image. It should be borne in mind that the used images had the lowest standard deviation at DPChallenge, which means that the mean rating at DPChallenge had a standard deviation (0.27) lower than that of in-person ratings (1.18 for aesthetics and 1.10 for quality).

If we pay attention to the visual characteristics of some of the images of the set (Figure 8), some noteworthy cases were found:

(i) Figure 8(a) was wrongly rated by the users of the photography portal as having some overexposed areas. It showed a palm tree on the foreground which was slightly incorrectly exposed. However, it obtained a high score in aesthetics because it had some aesthetic value for the evaluators (these motifs tend to have certain aesthetic value). The value of quality was closer to that of DPChallenge in this case.

(ii) Figure 8(b) in DPChallenge obtained a low rating, whereas as far as quality and aesthetics were concerned, it was clearly over average. This difference could be due to the experimental conditions in which the in-person evaluation took place (good quality of image on a big-enough screen, well exposed sky). Under these conditions, evaluators might have paid more attention to the drop and the sky to the detriment of darker area.

(iii) Figure 8(c) at DPChallenge had a high rating, which may be due to the fact that its originality and editing were taken into account.

(iv) Lastly, in Figure 8(d) quality was again closer to DPChallenge. However, a lower score was given in aesthetics.

All this shows that, at DPChallenge, in specific cases, different parameters might be evaluated: originality, quality, aesthetics, photo editing, etc.

As to the use of machine learning techniques to predict each of the three criteria studied, the highest correlation

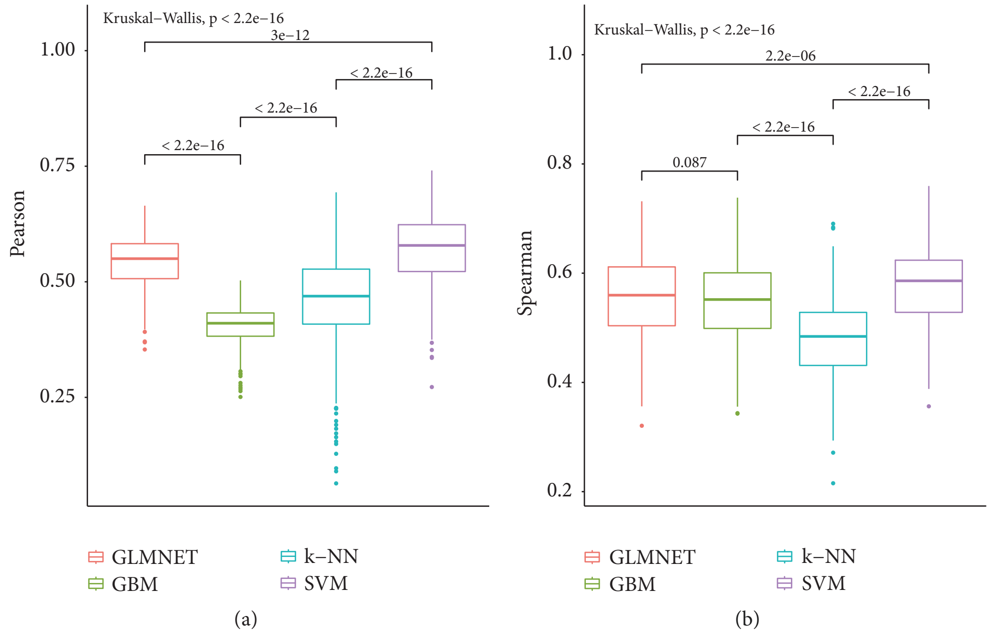


Figure 6: Distribution of the correlations obtained for each optimised model (Pearson's on the right and Spearman's on the left). For each pair, the p-value obtained using a Kruskal-Wallis test is shown.

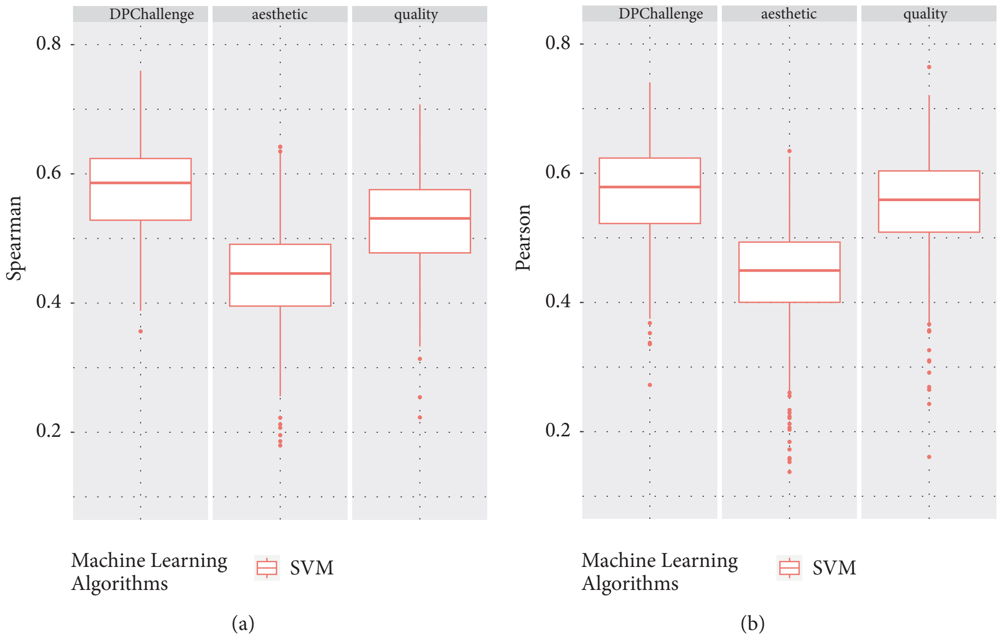


Figure 7: Distribution of correlations (Spearman's on the right and Pearson's on the left) obtained for each of the three criteria (DPChallenge, Aesthetic, and Quality) using 256 input variables and an SVM model optimised using hyperparameterisation.

obtained was 0.578 using SVM. This value is similar to those obtained by Marin and Leder [33] using as criteria "arousal" (Spearman's rho=0.44) and "pleasantness" (Spearman's rho=0.64) or by Nadal [9] with "beauty" (Spearman's rho=0.648) under similar experimental conditions with humans. These values were obtained using numerous state-of-the-art methods in predicting and determining the best configuration for each of them through hyperparameterisation.

As to the correlation between the SVM model with quality and aesthetics individually (Figure 8), it follows that for the system it was simpler to learn the quality values than

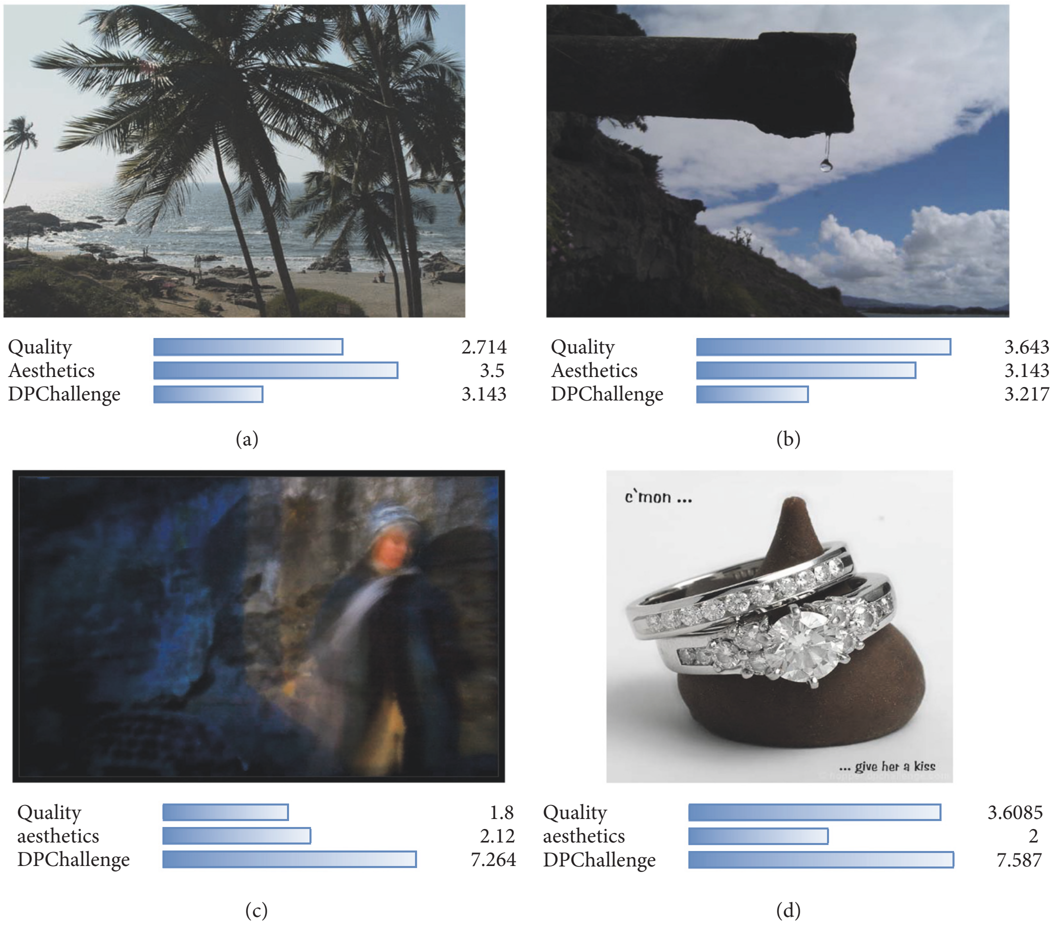


Figure 8: Examples of images with different scores based on the three evaluation criteria. For each image, a number value is given according to each criterion, while bars show the normalised weight of such value within each assessment range (DPChallenge in the [1, 10] range and aesthetics and quality in the [1, 5] range).

the aesthetic ones, which makes sense considering that the former is a less subjective component and more related to the characteristics of the image.

7. Conclusions

Taking into account a number of problems found regarding the state-of-the-art datasets, a dataset was developed following a new methodology. This dataset consists of 1000 images from the DPChallenge portal, which were evaluated in 3 different ways: (1) evaluation from the DPChallenge portal with at least 100 scores per image; (2) an aesthetic evaluation conducted under controlled experimental conditions and a minimum of 10 votes per image; (3) a quality assessment made under the same conditions as (2). As far as the authors are aware, this is the first time a dataset is evaluated based on three different criteria by two different populations.

The results of the correlation suggest that the evaluation of DPChallenge is closer to a quality criterion than to an aesthetic one. The DPChallenge users and in-person evaluators rate images differently and it is apparent that at DPChallenge each user may be following different criteria for the evaluation of images, such as originality, image editing, quality, aesthetics, etc.

Numerous state-of-the-art computational techniques were used and their optimal configurations were identified and applied to all three criteria (DPChallenge, aesthetic, and quality) and correlations of 0.578, 0.456, and 0.539, respectively, were achieved. These results are similar to those obtained in the state-of-the-art experiments. They show that machine learning techniques are more able to learn human assessment of technical quality than aesthetic value, despite the fact that the gap between them is very narrow.

It should be emphasized that machine learning approaches are better at predicting quality than aesthetics, perhaps because of their lower subjective component and their greater association with the intrinsic characteristics of the images.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work is supported by the General Directorate of Culture, Education and University Management of Xunta de Galicia (Ref. GRC2014/049) and the European Fund for Regional Development (FEDER) allocated by the European Union, the Portuguese Foundation for Science and Technology for the development of project SBIRC (Ref. PTDC/EIA–EIA/115667/2009), Xunta de Galicia (Ref. XUGA-PGIDIT-10TIC105008-PR) and the Spanish Ministry for Science and Technology (Ref. TIN2008-06562/TIN), and the Juan de la Cierva fellowship programme by the Spanish Ministry of Economy and Competitiveness (Carlos Fernandez-Lozano, Ref. FJCI-2015-26071). We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

[1] A. Carballal, L. Castro, N. Rodríguez-Fernández, I. Santos, A. Santos, and J. Romero, "Approach to minimize bias on aesthetic image datasets," in *Interface Support for Creativity, Productivity, and Expression in Computer Graphics*, p. 131, IGI Global, 2019.

[2] A. Carballal, L. Castro, R. Perez, and J. Correia, "Detecting bias on aesthetic image datasets," *International Journal of Creative Interfaces and Computer Graphics*, vol. 5, pp. 62–74, 2014.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *Computer Vision – ECCV 2006*, A. Leonardis, H. Bischof, and A. Pinz, Eds., pp. 288–301, Springer, Berlin, Germany, 2006.

[4] W. Wang, D. Cai, L. Wang, Q. Huang, X. Xu, and X. Li, "Synthesized computational aesthetic evaluation of photos," *Neurocomputing*, vol. 172, pp. 244–252, 2016.

[5] Y. Ke, X. Tang, and F. Jing, "The design of high-level features for photo quality assessment," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, vol. 1, pp. 419–426, IEEE, New York, NY, USA, 2006.

[6] L.-K. Wong and K.-L. Low, "Saliency-enhanced image aesthetics class prediction," in *Proceedings of the 2009 16th IEEE International Conference on Image Processing ICIP 2009*, pp. 997–1000, Cairo, Egypt, November 2009.

[7] C. J. Cela-Conde, F. J. Ayala, E. Munar et al., "Sex-related similarities and differences in the neural correlates of beauty," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 106, no. 10, pp. 3847–3852, 2009.

[8] A. Forsythe, M. Nadal, N. Sheehy, C. J. Cela-Conde, and M. Sawey, "Predicting beauty: fractal dimension and visual complexity in art," *British Journal of Psychology*, vol. 102, no. 1, pp. 49–70, 2011.

[9] M. Nadal, E. Munar, G. Marty, and C. Cela-Conde, "Visual complexity and beauty appreciation: Explaining the divergence of results," *Empirical Studies of the Arts*, vol. 28, no. 2, pp. 173–191, 2010.

[10] R. Datta, J. Li, and J. Z. Wang, "Algorithmic inferencing of aesthetics and emotion in natural images: an exposition," in *Proceedings of the IEEE International Conference on Image Processing (ICIP '08)*, pp. 105–108, IEEE Press, San Diego, Calif, USA, October 2008.

[11] P. Machado, J. Romero, M. Nadal, A. Santos, J. Correia, and A. Carballal, "Computerized measures of visual complexity," *Acta Psychologica*, vol. 160, pp. 43–57, 2015.

[12] X. Tang, W. Luo, and X. Wang, "Content-based photo quality assessment," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1930–1943, 2013.

[13] Y. Luo and X. Tang, "Photo and video quality evaluation: Focusing on the subjec," in *Computer Vision – ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., pp. 386–399, Springer, Berlin, Germany, 2008.

[14] J. Neyman, "Basic ideas and some recent results of the theory of testing statistical hypotheses," *Journal of the Royal Statistical Society*, vol. 105, pp. 292–327, 1942.

[15] N. Orlov, L. Shamir, T. Macura, J. Johnston, D. M. Eckley, and I. G. Goldberg, "WND-CHARM: multi-purpose image classification using compound image transforms," *Pattern Recognition Letters*, vol. 29, no. 11, pp. 1684–1693, 2008.

[16] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial Intelligence*, vol. 97, no. 1-2, pp. 245–271, 1997.

[17] C. Fernandez-Lozano, J. A. Seoane, M. Gestal, T. R. Gaunt, J. Dorado, and C. Campbell, "Texture classification using feature selection and kernel-based techniques," *Soft Computing*, vol. 19, no. 9, pp. 2469–2480, 2015.

[18] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[19] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, "A review of feature selection methods on synthetic data," *Knowledge and Information Systems*, vol. 34, no. 3, pp. 483–519, 2013.

[20] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*, Springer-Verlag, New Jersy, NJ, USA, 2006.

[21] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," *Information Sciences*, vol. 181, no. 1, pp. 115–128, 2011.

[22] R. R. Hocking, "The analysis and selection of variables in linear regression," *Biometrics*, vol. 32, no. 1, pp. 1–49, 1976.

[23] G. Tsiliki, C. R. Munteanu, J. A. Seoane, C. Fernandez-Lozano, H. Sarimveis, and E. L. Willighagen, "RRegrs: An R package for computer-aided model selection with multiple regression models," *Journal of Cheminformatics*, vol. 7, no. 1, pp. 1–16, 2015.

[24] S. Garcia, A. Fernandez, J. Luengo, and F. Herrera, "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power," *Information Sciences*, vol. 180, no. 10, pp. 2044–2064, 2010.

[25] C. Ambroise and G. J. McLachlan, "Selection bias in gene extraction on the basis of microarray gene-expression data," *Proceedings of the National Acadamy of Sciences of the United States of America*, vol. 99, no. 10, pp. 6562–6566, 2002.

[26] K. Hechenbichler and K. Schliep, "Weighted k-nearest-neighbor techniques and ordinal classification," *Sonderforschungsbereich*, vol. 386, 2004, Discussion Paper 399.

[27] W. Liu and S. Chawla, "Class Confidence Weighted kNN Algorithms for Imbalanced Data Sets," *Advances in Knowledge Discovery and Data Mining*, vol. 6635, pp. 345–356, 2011.

[28] https://CRAN.R-project.org/package=gbm.

[29] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics and Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

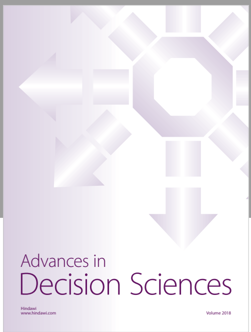
[30] https://projecteuclid.org/euclid.aos/1016218223.

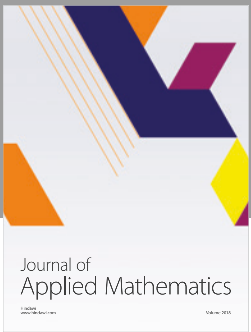
[31] D. F. McCaffrey, B. A. Griffin, D. Almirall, M. E. Slaughter, R. Ramchand, and L. F. Burgette, "A tutorial on propensity score estimation for multiple treatments using generalized boosted models," *Statistics in Medicine*, vol. 32, no. 19, pp. 3388–3414, 2013.

[32] C. Fernandez-Lozano, M. Gestal, C. R. Munteanu, J. Dorado, and A. Pazos, "A methodology for the design of experiments in computational intelligence with multiple regression models," *PeerJ*, vol. 2016, no. 12, 2016.

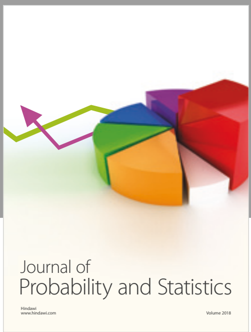
[33] M. M. Marin and H. Leder, "Examining complexity across domains: relating subjective and objective measures of affective environmental scenes, paintings and music," *PLoS ONE*, vol. 8, no. 8, Article ID e72412, 2013.

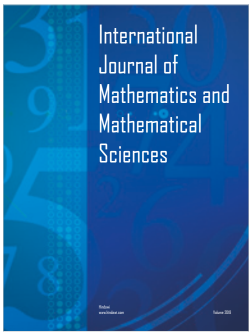
Advances in
Operations Research


Advances in
Decision Sciences

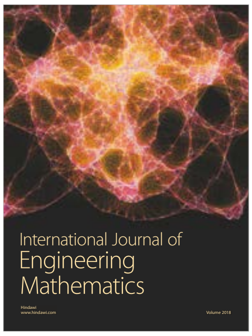

Journal of
Applied Mathematics

The Scientific
World Journal

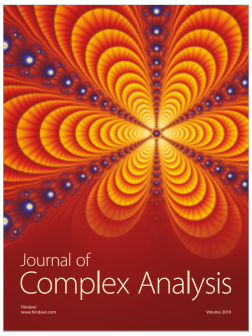

Journal of
Probability and Statistics

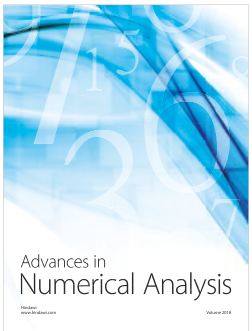

International
Journal of
Mathematics and
Mathematical
Sciences

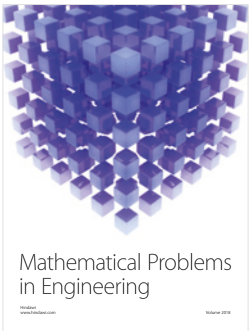
Journal of
Optimization

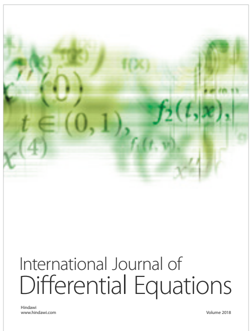

International Journal of
Engineering
Mathematics

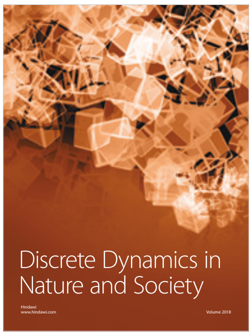
International Journal of
Analysis

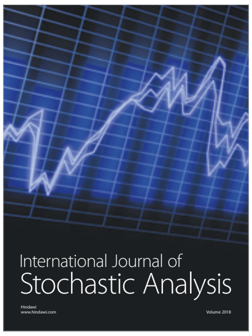

Journal of
Complex Analysis


Advances in
Numerical Analysis


Mathematical Problems
in Engineering

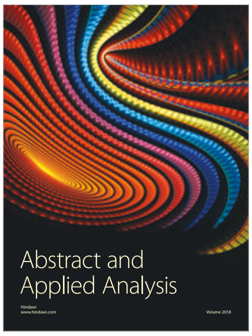

International Journal of
Differential Equations

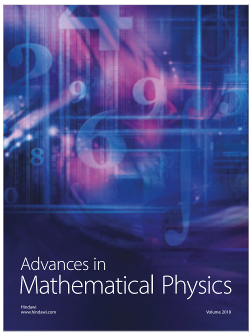

Discrete Dynamics in
Nature and Society


International Journal of
Stochastic Analysis

Journal of
Mathematics

Journal of
Function Spaces


Abstract and
Applied Analysis


Advances in
Mathematical Physics