Assessment of the Possibility of Imitating Experts' Aesthetic Judgments about the Impact of Knots on the Beauty of Furniture Fronts Made of Pine Wood

Krzysztof Gajowniczek[1], Marcin Bator[1,*],
Katarzyna Śmietańska[2], and Jarosław Górski[2]
[1]*Institute of Information Technology, Warsaw University of Life Sciences – SGGW, Warsaw, Poland*
[2]*Institute of Wood Sciences and Furniture, Warsaw University of Life Sciences – SGGW, Warsaw, Poland*
[*]*Corresponding author: M. Bator (marcin_bator@sggw.edu.pl)*

Abstract. Our research aims to reconstruct expert preferences regarding the visual attractiveness of furniture fronts made of pine wood using machine learning algorithms. A numerical experiment was performed using five machine learning algorithms of various paradigms. To find the answer to the question of what determines the expert's decision, we determined the importance of variables for some machine learning models. For random forest and classification trees, it involves the overall reduction in node impurities resulting from variable splitting, while for neural networks it uses the Garson algorithm. Based on the numerical experiments we can conclude that the best results of expert decision reconstruction are provided by a neural network model. The expert's decision is better reconstructed for more beautiful images. The decision for nice images is made based on the best 4 or 5 variables, while for ugly images many more features are important. Prettier images and those for which the expert's decision is better reconstructed have fewer knots.

Key words: image processing, knots on the fronts, machine learning, preference learning, solid wood furniture, quality control, importance of variables.

1. Introduction

Computer aided control of product quality in furniture manufacturing has been a relevant research problem for a long time. It turned out that image processing can be an effective way of automatically ensuring whether different types of wood products conform to set specifications [37]. However, without a doubt, the quality of wooden furniture has not only objective, but also subjective aspects – especially aesthetic value, in which case, current computer technology is not enough. Wood, being a product of nature, is not homogeneous. It contains natural features, often formally recognized as defects (knots, cracks, defects in shape, color, even mechanical damage), for many people can be an advantage, thanks to which this material gains its unique, individual character. Unlike objective characteristics that can be defined by standards (for example, dimension, strength properties), the only reliable measure of product quality in this case seems to be subjective human assessment. Therefore, a key issue in the case of wood furniture is to know the human preferences, which are additionally subject to constant change due to

product innovations and changing lifestyles [31]. Such knowledge should be a fundamental step towards improving the production process in furniture factories, consequently striving for automation in those features identification and classification [4,15,17]. Evaluation and quantification of furniture aesthetic value played an important role in the furniture design, and it is one of the difficult issues in this research field [5,34]. For example, nowadays you can easily use image processing algorithms to control the number of knots visible on a furniture front made of solid wood, but first you should know what the effect of this number on the relative, aesthetic attractiveness of this front is. Firstly, in order to know human preferences for wooden furniture, an adequate group of people should be examined [13].

In research where aesthetic values are concerned, it is common practice to use experts in a given field [1,6,14]. Unfortunately, we do not find many scientific studies presenting data on the real specific aesthetic preferences of furniture made of wood. Research conducted so far is mostly of marketing nature and does not focus on specific aesthetic features. However, they clearly emphasize that one of the most important factors influencing the purchase decision is the aesthetic value, including, among others, design or color [12,16,17,32]. The aesthetic functions (e.g. visual sensation, emotions), determinative forms and fashionable style play a very important role in furniture design and production [1,14].

The previous research [36] proves that there is a relationship between the occurrence of knots, their size and location, and aesthetic impressions. These studies used sets of questionnaires (based on standard 5 point Likert scale) for various groups of experts, and the analysis of the results included the proposed features describing the characteristics of each image. This study raised a question about the adequacy of these features for assessing aesthetics. Hence, numerous experiments in the field of machine learning were carried out to verify whether, using the features defined there and their values, it is possible to recreate the classification method (assessment method) performed by individual people or groups of people.

This research is part of the development of an automatic furniture front evaluation system. Its elements are image acquisition, image analysis (calculation of features), classification (evaluation). As in many tasks, a good approach is to complete the modules from the final module to the initial one, rather than from the initial one to the final one. In other words, we first answer the question of what features are important, then how to calculate them in the image, when we know what quality of the image is needed, only then the image acquisition begin.

Preference modeling can be done in three ways [19], treating each opinion independently as a separate class without maintaining order relations between ratings, as an ordinal regression/classification problem or using learning to rank approach [7,29]. Unfortunately, the second approach has produced only a few extensions to existing machine learning algorithms. The third approach assumes that we have all the images together

and rank them among themselves [8, 25, 30], but in our study we evaluate each image separately on an ordinal scale, so this approach is also not applicable to us. Due to that, in our paper, we will use the first approach because the results provided by the second approach are only slightly better. However, what is more important, the first approach allows us to use and compare machine learning algorithms of various paradigms and use methods deriving the importance of variables.

To name the subject of this paper, we will use the terms reconstruction, reproduction, mapping, and classification interchangeably. Therefore, our research aims to reconstruct expert preferences regarding the visual attractiveness of furniture fronts made of pine wood using machine learning algorithms. The numerical experiment will be performed using five machine learning algorithms, i.e. classification trees [3], artificial neural networks [22], k-nearest neighbors [28], random forest [2] and support vector machines [23]. From this point on, we will understand the label l as the true rating R (based on the Likert scale) given by an expert to a given image and its predicted value by a predictive model or benchmarking method. To achieve the intended research goal, we defined the following research questions:

1. Which machine learning-based model delivers the best results of reconstruction of the expert's preferences?
2. Is there a relationship between the quality of reconstruction of the expert's preferences and the beauty of the image?
3. Which group of experts is the best and worst reproducible?
4. Which features matter in the structure of a given machine-learning model?
5. What feature values characterize the best and the worst reproduced images?

The remainder of this paper is organized as follows. Section 2 presents the data used in this study. In Section 3, a detailed description of the methodology of the numerical experiment is presented. Section 4 provides the results of the numerical experiment. The paper ends with a discussion regarding the results and concluding remarks in Section 5.

2. Data characteristics

The furniture elements analyzed in the research was a furniture the front (600 mm × 600 mm) of the single-door cabinet made of solid wood. Each front was divided into 5 zones with surface area, but different location and shape (i.e. upper left, upper right, bottom left, bottom right and central). All of them were prepared by an expert based on visual analysis of knots presented on images. A subset of them was previously presented in [36]. A key question is "is it possible to reconstruct an expert preferences using those features based on knots position and size, or some other information from an image is needed?" In this study the condition question: "what kind of information could be also needed?" is not set. The focus was on arranging a selected set of knots in various, precisely defined configurations. This allowed us to eliminate the influence of unnecessary

features that are irrelevant to the research, such as the shape or color of knots. The aim of the experiment was to examine the impact of features of furniture fronts on their aesthetic quality in the opinion of the judges. Sample images are presented in Fig. 1 and Fig. 2. The list below presents the features with their description (with abbreviations used later on):

1. The number of knots (`qty1`).
2. Evenness of the number of knots (`qty2`): 0=odd; 1=even.
3. Four classes of the number of knots (`qty3`): 0=none, 1=small, i.e. from 1 to 2 knots; 2=medium, i.e. from 3 to 5 knots; 3=many, i.e. from 6 to 8 knots).
4. Dispersion (`disp1`): 0=concentrated in a specific zone; 1=dispersed, i.e. located in as many zones as possible.
5. Presence of at least one knot in the central zone (`pos1`): 0=not present; 1=present.
6. Presence of at least one knot in the upper left zone (`pos2`): 0=not present; 1=present.
7. Presence of at least one knot in the upper right zone (`pos3`): 0=not present; 1=present.
8. Presence of at least one knot in the bottom left zone (`pos4`): 0=not present; 1=present.
9. Presence of at least one knot in the bottom right zone (`pos5`): 0=not present; 1=present.
10. Presence of at least one knot in the bottom zone (`pos6`): 0=not present; 1=present.
11. Presence of at least one knot in the upper zone (`pos7`): 0=not present; 1=present.
12. Presence of at least one knot in the left zone (`pos8`): 0=not present; 1=present.
13. Presence of at least one knot in the right zone (`pos9`): 0=not present; 1=present.
14. Knot size overall (`siz2`): 0=no knots; 1=only small knots; 2=both small and large; 3=only large knots.
15. Size of knots in detail (`siz3`): 0=no knots; 1=only small ones; 2=both small and large but more small ones; 4=both small and large but more large ones; 5=only large ones.
16. Symmetry (`sym1`): 0=not present; 1=present.

The study employed the standard consensus-based assessment (CBA) technique. Four expert groups, totaling 50 participants, were invited to take part in the experiment. The selection of experts is extremely important. Based on the literature review (in chapter 1), it can be considered a sufficient or relatively large number. Below list presents group of experts with their description (with later used abbreviation):

1. The first group (Art), consisted of 12 interior design professionals, with ages ranging from 28 to 61 from the Academy of Fine Arts (Faculty of Interior Design) in Warsaw and the University of the Arts in Poznan (Department of Furniture Design and Department of Interior Design).
2. The second group of experts (WTD), comprised 12 individuals specializing in furniture design. They were members of the research and teaching faculty at the Institute of Wood Sciences and Furniture within the Warsaw University of Life Sciences (WULS). Their ages ranged from 30 to 56 years old.

3. Group 3 (Std) comprised 14 students from the Department of Furniture Manufacturing (WULS, Faculty of Wood Technology). These individuals were considered semi-professionals and fell within the age range of 20 to 21 years old.
4. Additionally, for comparative purposes, an entirely non-professional group of experts (WWa) was included. This group comprised 12 individuals, aged between 22 and 82, randomly selected from Warsaw residents. They were chosen based on their belief in their knowledge of furniture and their strong motivation to participate in scientific research.

The special designers method such as Questionnaire research was used in the study. All groups of judges (50 people in total) were asked to fill out an online questionnaire based on standard 5 point Likert scale (1-strongly disagree, 2-disagree, 3-neutral, 4-agree, 5-strongly agree) which contained 99 questions about each image. Each of them was the same: "Do you agree that furniture front shown in the photo above is more attractive than others?".

3. Research methodology

3.1. Machine learning based models

All simulations were prepared using R software [20] (version 4.3) and corresponding libraries implementing certain machine learning algorithms. The main infrastructure was the *caret* package [18] (short for *C*lassification *A*nd *RE*gression *T*raining) which is a set of functions aimed at improving the process of creating various predictive models. All algorithms were trained in the standard state-of-the-art cross-validation regime using a leave-one-out approach and checking various combinations of the hyper-parameters. To gain better numerical stability and have variables on comparable scales (which is required for some algorithms), the data were normalized using standardization. Each model was trained for the classification problem. The potential explanatory variables $\mathbf{x}$ were the variables described in the list 2 and the target variable was the expert's rating, $y \in \{1, 2, 3, 4, 5\}$.

The rpart package, implementing the CART (Classification and Regression Trees) algorithm [3], was utilized to train classification trees. Throughout the process of partitioning a multi-dimensional space, the criterion focused on minimizing the Gini impurity of the dependent variable for observations within the same leaf node. The node had a minimum requirement of 20 observations, and a leaf needed at least 6 observations to avoid further splitting. Rather than pruning the tree at the algorithm's conclusion, we employed a pruning technique during the tree's growth phase. This method halted the creation of new splits when prior splits only marginally improved predictive accuracy. The complexity parameter *cp* was tested using the following values 0, 0.001, 0.005, 0.01, 0.05, 0.1, and 0.2. The tree was constructed up to a depth of 30 levels.

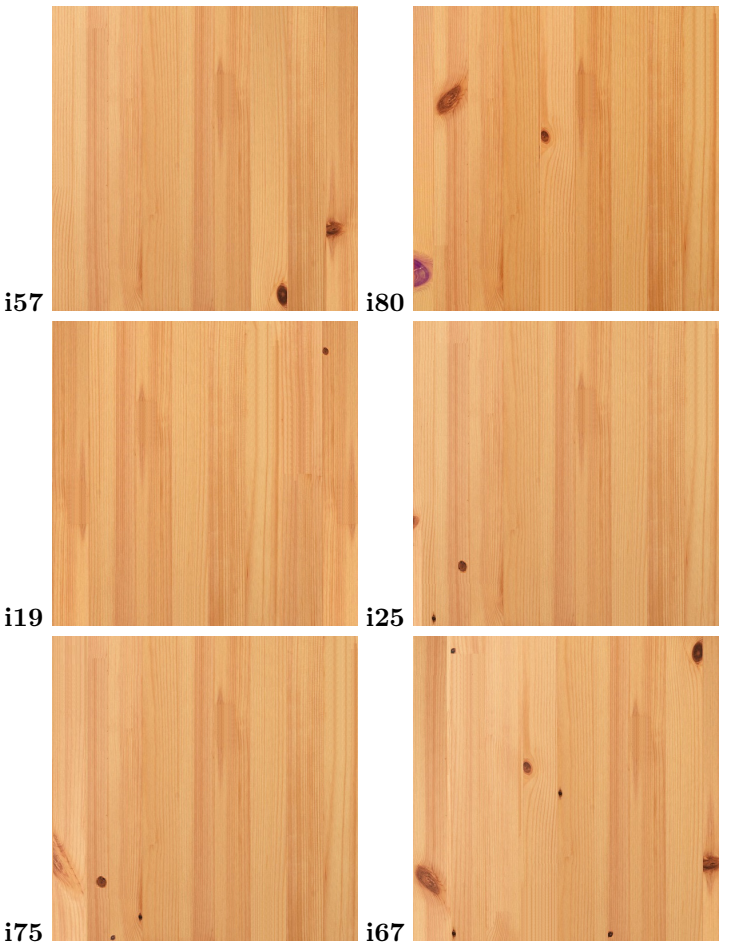

Fig. 1. Examples of images with the best (left-hand side) and the worst (right-hand side) classification results.

To train the neural networks, we used the BFGS (Broyden–Fletcher–Goldfarb–Shanno) algorithm [9,10,22], which belongs to the broad family of quasi-Newton optimization methods (available in the nnet library). Each neural network consisted of one input layer with 16 neurons (one for each feature), one hidden layer (a different number of neurons was tested, i.e. 1, 5, 10, 15, 20; *size* parameter) and one output layer with five neurons (one neuron for one class). The target feature was decoded using one-hot encoding, i.e. a matrix with five columns in which the number one indicates the true label, keeping

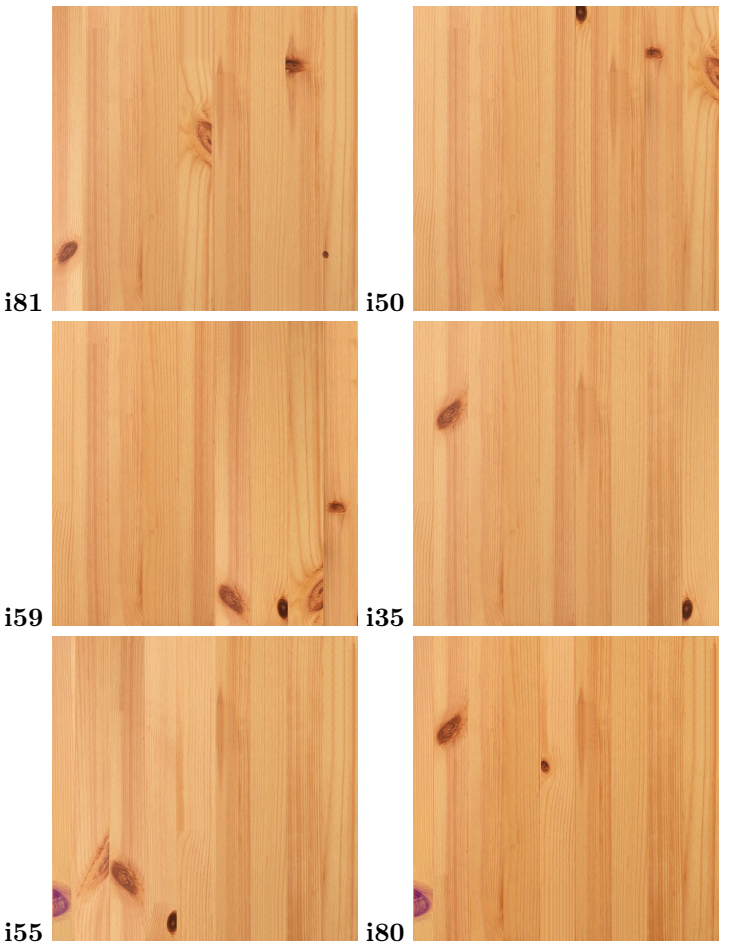

Fig. 2. Examples of the prettiest (left-hand side) and the ugliest (right-hand side) images.

zero for all other labels. A sigmoid function was used to activate all of the neurons in the network. To prevent overfitting the regularization term (weight decay) is employed which uses as the penalty the sum of squares of the weights was set at 0, 0.01, 0.5, and 0.1 (decay parameter). The maximum number of iterations was set at 200 with the stopping criterion set at $1.0e-4$.

Classification using k-nearest neighbors algorithm [28] was performed using knn3 function from the caret library. Different numbers of the k-values (k parameter) were proposed in the experiments including the following: 1, 5, 10, 15, 25, 50, 75.

The random forest [2] was trained using an algorithm sourced from the randomForest library. Preceding each training session, samples comprising n elements were drawn with replacement, representing around 63% of the population. These samples were employed to create $CART$ trees, each tree being constructed to its full size without any pruning, ensuring that no leaf contained 5 or fewer observations. The count of variables selected randomly as potential candidates at each split varied from 4 to 16 by 4 ($mtry$ parameter). The total number of trees in the forest was 500.

For building the support vector machine [23], the `ksvm` function from the kernlab library was employed, utilizing its Sequential Minimal Optimization algorithm. This algorithm was utilized to address the quadratic programming problem involved in the process. The radial basis was used as a kernel function, with `sigma` set at 0.01, 0.05, 0.1, 0.5 and 1. The regularization parameter `C` which controls the over-fitting, was arbitrarily set, and the simulations were run for the following values: 0.1, 0.5, 1, 2, and 5.

Importance of variables was calculated using `VarImp` which is a generic model-specific method. For the random forest and classification trees it involves the overall reduction in node impurities resulting from variable splitting (for rf averaged across all trees). The node impurity was assessed using the Gini index. For neural networks, it uses the Garson algorithm [11] which delineates the relative significance of features by dissecting the model weights. Assessing the relative importance (or strength of association) of a particular feature for the response variable involves identifying all weighted connections between relevant nodes. This encompasses pinpointing all weights that link the specific input node, traverse through the hidden layer, and culminate at the response variable.

3.2. Benchmarking methods

To examine the quality of the developed models, we proposed two benchmarking methods. Both indicate a baseline obtained through a kind of naive forecast. The first method uses a median value of the true labels for all 99 images for a given expert. Usually, it is a label of 3, but for some experts, it is a 2 or 4 because some experts have not given any image a label of 3 or other labels. The second approach employs random values from the empirical distribution for a given expert. In other words, let's assume that for the event space $\Omega = \{1, 2, 3, 4, 5\}$ the frequency (across all 99 images) of an expert selecting a given label (l) is $P(l = 1) = 0.10$, $P(l = 2) = 0.30$, $P(l = 3) = 0.20$, $P(l = 4) = 0.35$ and $P(l = 5) = 0.05$. Then as a predicted label for a given image, we take random value taking into account the aforementioned distribution.

3.3. Prediction quality measure

Instead of a standard accuracy measure where only exact matching (perfect prediction) increases accuracy, we have used Absolute Accuracy Error (AE). It provides robust information about how far the model predictions are from the original labels. This kind

of measure is crucial for our analysis because it provides deep insight into the distribution of errors. Our intuition was that importance of variables in the model depends on the error magnitude taking into account the true label. For example, we have assumed that different variables are important when: 1) the true label is 5 and the error is 2; 2) the true label is 5 and the error is 0; 3) the true label is 2 and the error is 1; etc. The measure is defined as follows:

$$\mathrm{AE} = \sum_{i=1}^{n} |t_i - p_i| \,, \tag{1}$$

where n is the number of images, t is the true label and p is the predicted label.

3.4. Additional configuration and notation

To be able to reproduce the simulation study, the initial value (`set.seed` function) of the pseudo-random number generator (Mersenne-Twister algorithm) was set for both machine learning models and benchmarking methods.

To make the message of the next part concise, we use the following notations and abbreviations: rpart – decision tree algorithm, nnet – neural network algorithm, knn – k-nearest neighbors algorithm, rf – random forest algorithm, svm – support vector machine algorithm, median – prediction based on the median value, and random – prediction based on the random value from the empirical distribution.

4. Results of the numerical experiment

4.1. Expert and image reconstruction

To perform the analysis presented in this section, it was first necessary to prepare a table with the true labels and the predicted labels for each model or benchmarking method. Results for each model present results for the best-tuned model (i.e. minimal error defined in 1) based on the aforementioned combinations of the hyper-parameters. Each table was of the size #`expert` × #`images` with the true or the predicted label for the expert for a particular image taken for a leave-on-out iteration.

On the Table 1 and Table 2 we present classification results for each image and each expert, respectively. Each value presents an accuracy measure defined in (1) and in the rounded brackets its standard error. Both tables are sorted (from the best to the worst result) based on the Avg. column (7th column) which is derived as the average accuracy measure for five machine learning-based models. We apply a sort of majority voting to be independent of the quality of the individual model and learning paradigm. In both tables first column presents the image number or expert number. As mentioned earlier, the best (e.g. i57 and i19) and the worst (e.g. i25 and i67) reproduced decisions for images are presented in the Figure 1.

Tab. 1. Classification results for each image. Each value presents distance defined in (1) and its standard error (the best result for each row is indicated in bold).

Image	rpart	nnet	knn	rf	svm	Avg.	Median	Random
i57	23(±0.76)	**18(±0.63)**	24(±0.68)	19(±0.60)	20(±0.64)	20.8(±0.66)	54(±0.80)	78(±1.05)
i19	20(±0.70)	**18(±0.69)**	31(±0.83)	25(±0.76)	20(±0.70)	22.8(±0.74)	53(±0.74)	63(±1.12)
i75	23(±0.73)	**20(±0.61)**	30(±0.86)	21(±0.64)	20(±0.67)	22.8(±0.70)	60(±0.76)	54(±1.16)
i54	28(±0.67)	18(±0.56)	35(±0.93)	18(±0.78)	**16(±0.62)**	23.0(±0.71)	97(±0.79)	81(±1.38)
i32	**19(±0.64)**	22(±0.73)	28(±0.81)	22(±0.70)	26(±0.71)	23.4(±0.72)	51(±0.80)	55(±1.13)
i90	**19(±0.64)**	29(±0.81)	19(±0.73)	29(±0.76)	22(±0.73)	23.6(±0.73)	48(±0.70)	55(±0.93)
i81	29(±0.76)	20(±0.73)	36(±1.01)	20(±0.81)	**19(±0.83)**	24.8(±0.83)	102(±0.7)	101(±1.25)
i63	28(±0.81)	24(±0.81)	25(±0.79)	**21(±0.73)**	28(±0.76)	25.2(±0.78)	52(±0.73)	51(±1.02)
i92	24(±0.81)	**20(±0.64)**	36(±0.76)	24(±0.71)	22(±0.50)	25.2(±0.68)	86(±0.67)	87(±1.43)
i89	31(±0.75)	23(±0.89)	31(±0.75)	26(±0.93)	**19(±0.67)**	26.0(±0.80)	84(±0.82)	87(±1.38)
i17	31(±0.97)	23(±0.79)	29(±0.88)	26(±0.91)	**22(±0.79)**	26.2(±0.87)	48(±0.70)	52(±0.99)
i6	**23(±0.84)**	28(±0.86)	25(±0.84)	29(±0.84)	27(±0.84)	26.4(±0.84)	58(±0.65)	67(±1.10)
i47	27(±0.86)	**24(±0.68)**	29(±0.76)	26(±0.68)	26(±0.71)	26.4(±0.74)	48(±0.78)	54(±1.14)
i33	29(±0.81)	27(±0.81)	**24(±0.79)**	27(±0.84)	26(±0.79)	26.6(±0.81)	55(±0.76)	48(±0.97)
i46	29(±0.84)	**23(±0.73)**	24(±0.76)	31(±0.78)	26(±0.76)	26.6(±0.77)	49(±0.77)	47(±1.02)
i66	30(±0.93)	27(±0.68)	27(±0.76)	**25(±0.76)**	29(±0.84)	27.6(±0.79)	57(±0.76)	58(±1.15)
i62	28(±0.91)	27(±0.79)	35(±0.71)	**20(±0.57)**	29(±0.70)	27.8(±0.74)	78(±0.76)	74(±1.33)
i61	**24(±0.79)**	26(±0.79)	34(±0.82)	27(±0.81)	29(±0.76)	28.0(±0.79)	46(±0.75)	48(±0.95)
i94	27(±0.81)	35(±0.89)	26(±0.71)	30(±0.76)	**25(±0.71)**	28.6(±0.78)	50(±0.76)	60(±1.01)
i21	28(±0.88)	30(±0.83)	**24(±0.76)**	37(±0.90)	25(±0.79)	28.8(±0.83)	50(±0.78)	56(±1.02)
i40	27(±0.86)	38(±1.06)	32(±0.90)	27(±0.89)	**20(±0.67)**	28.8(±0.88)	53(±0.82)	59(±1.12)
i26	33(±0.89)	**24(±0.71)**	30(±0.78)	28(±0.70)	30(±0.83)	29.0(±0.78)	56(±0.72)	68(±1.03)
i20	36(±0.83)	**22(±0.67)**	28(±0.79)	37(±0.88)	23(±0.71)	29.2(±0.78)	45(±0.65)	57(±0.99)
i36	**22(±0.67)**	29(±0.78)	39(±0.89)	31(±0.78)	27(±0.79)	29.6(±0.78)	48(±0.75)	71(±1.07)
i69	27(±0.76)	39(±0.97)	**21(±0.61)**	37(±0.78)	24(±0.65)	29.6(±0.75)	50(±0.78)	59(±1.12)
i59	33(±0.85)	25(±0.71)	39(±0.93)	30(±1.03)	**24(±0.74)**	30.2(±0.85)	103(±0.71)	91(±1.32)
i95	**26(±0.81)**	29(±0.95)	33(±1.00)	33(±0.85)	30(±0.81)	30.2(±0.88)	51(±0.82)	64(±1.20)
i88	36(±0.99)	24(±0.84)	42(±1.11)	**16(±0.65)**	34(±1.04)	30.4(±0.93)	60(±0.78)	73(±1.39)
i96	24(±0.74)	**22(±0.61)**	44(±1.00)	25(±0.71)	37(±0.90)	30.4(±0.79)	49(±0.77)	63(±1.17)
i31	39(±0.91)	30(±0.70)	**24(±0.71)**	30(±0.86)	30(±0.86)	30.6(±0.81)	50(±0.78)	45(±0.95)
i72	27(±0.73)	**24(±0.68)**	37(±0.80)	32(±0.80)	33(±0.80)	30.6(±0.76)	36(±0.70)	59(±1.02)
i77	29(±0.84)	**26(±0.84)**	37(±0.88)	28(±0.88)	34(±0.89)	30.8(±0.87)	60(±0.83)	69(±1.19)
i24	37(±0.96)	31(±1.05)	36(±1.03)	**25(±1.05)**	26(±1.01)	31.0(±1.02)	86(±0.81)	100(±1.26)
i30	36(±0.88)	**27(±0.73)**	34(±0.79)	28(±0.76)	30(±0.76)	31.0(±0.78)	70(±0.81)	57(±1.13)
i83	27(±0.76)	34(±0.96)	45(±0.95)	**19(±0.73)**	31(±0.92)	31.2(±0.86)	45(±0.76)	52(±1.07)
i43	33(±0.87)	34(±0.82)	36(±0.95)	**28(±0.79)**	30(±0.83)	32.2(±0.85)	59(±0.75)	54(±1.10)
i65	40(±0.95)	31(±0.81)	30(±0.76)	37(±0.92)	**24(±0.74)**	32.4(±0.84)	63(±0.75)	65(±1.22)
i52	26(±0.65)	**25(±0.76)**	46(±0.94)	31(±0.73)	39(±0.86)	33.4(±0.79)	54(±0.80)	63(±1.17)
i53	36(±1.01)	34(±0.96)	**31(±0.92)**	34(±0.96)	32(±0.90)	33.4(±0.95)	49(±0.77)	52(±1.12)
i18	30(±0.81)	34(±0.89)	**28(±0.84)**	41(±0.98)	35(±0.93)	33.6(±0.89)	58(±0.74)	60(±1.20)
i58	25(±0.76)	**25(±0.71)**	45(±0.91)	31(±0.78)	42(±0.98)	33.6(±0.83)	76(±0.68)	68(±1.10)
i76	32(±0.92)	**28(±0.79)**	38(±0.92)	38(±0.96)	32(±0.90)	33.6(±0.90)	58(±0.62)	70(±1.21)
i56	**29(±0.81)**	36(±0.93)	38(±0.87)	32(±0.75)	34(±0.91)	33.8(±0.85)	53(±0.68)	82(±1.27)
i14	41(±0.96)	36(±0.88)	**27(±0.84)**	38(±0.94)	28(±0.84)	34.0(±0.89)	54(±0.70)	51(±1.13)
i70	36(±0.86)	33(±0.92)	38(±0.87)	**26(±0.81)**	37(±0.88)	34.0(±0.87)	46(±0.75)	57(±0.99)
i71	34(±0.91)	**30(±0.81)**	38(±0.96)	35(±0.91)	34(±0.94)	34.2(±0.91)	47(±0.79)	66(±1.13)
i3	31(±0.90)	34(±0.94)	38(±0.92)	42(±0.96)	**27(±0.79)**	34.4(±0.90)	52(±0.67)	68(±1.12)
i86	33(±0.87)	37(±0.85)	39(±0.86)	**31(±0.78)**	32(±0.78)	34.4(±0.83)	52(±0.78)	65(±1.07)
i73	34(±0.94)	**32(±0.96)**	39(±0.97)	35(±0.95)	33(±0.92)	34.6(±0.95)	45(±0.76)	61(±1.17)
i49	34(±0.82)	**30(±0.88)**	39(±1.04)	35(±0.89)	36(±0.95)	34.8(±0.92)	56(±0.69)	60(±1.14)
i41	33(±0.92)	**31(±0.88)**	33(±0.98)	41(±1.04)	37(±0.90)	35.0(±0.94)	56(±0.82)	78(±1.16)
i51	**29(±0.81)**	35(±0.97)	43(±0.97)	33(±0.87)	35(±0.93)	35.0(±0.91)	54(±0.80)	56(±1.02)
i27	38(±0.89)	**29(±0.73)**	40(±0.83)	33(±0.89)	36(±0.73)	35.2(±0.81)	58(±0.91)	57(±0.99)
i37	35(±0.81)	40(±0.90)	41(±0.98)	**27(±0.73)**	33(±0.85)	35.2(±0.85)	55(±0.81)	70(±1.11)
i13	47(±1.00)	**26(±0.74)**	35(±0.89)	40(±0.95)	30(±0.81)	35.6(±0.88)	47(±0.71)	66(±0.89)
i99	40(±0.95)	37(±1.03)	39(±0.97)	**31(±0.85)**	32(±0.96)	35.8(±0.95)	56(±0.69)	56(±1.22)
i82	35(±0.95)	39(±1.02)	37(±0.96)	**34(±0.87)**	35(±0.93)	36.0(±0.95)	52(±0.81)	56(±1.06)
i85	**30(±0.81)**	39(±0.82)	38(±0.98)	41(±0.77)	32(±0.94)	36.0(±0.86)	63(±0.75)	63(±1.23)
i97	**29(±0.88)**	35(±0.81)	48(±1.09)	32(±0.78)	37(±0.88)	36.2(±0.89)	49(±0.71)	57(±1.23)
i74	32(±0.88)	39(±0.93)	40(±0.88)	39(±0.82)	**32(±0.83)**	36.4(±0.87)	51(±0.74)	65(±1.07)
i68	41(±1.00)	34(±0.89)	34(±0.84)	47(±1.00)	**29(±0.84)**	37.0(±0.91)	42(±0.71)	57(±1.11)
i87	39(±0.93)	40(±0.95)	34(±0.89)	42(±0.98)	**30(±0.88)**	37.0(±0.93)	67(±0.72)	69(±1.18)
i11	37(±0.80)	41(±0.87)	38(±0.72)	**34(±0.87)**	38(±0.94)	37.6(±0.84)	53(±0.87)	69(±1.18)
i4	**34(±0.96)**	38(±1.00)	47(±1.17)	34(±1.00)	36(±1.01)	37.8(±1.03)	42(±0.68)	74(±1.18)
i78	39(±0.93)	42(±0.93)	36(±0.83)	43(±0.90)	**31(±0.81)**	38.2(±0.88)	41(±0.69)	60(±1.03)
i44	42(±0.93)	**33(±0.94)**	40(±1.05)	38(±0.96)	40(±0.99)	38.6(±0.97)	69(±0.73)	62(±1.20)
i48	39(±0.82)	42(±0.79)	40(±0.83)	**36(±0.73)**	38(±0.82)	39.0(±0.80)	59(±0.94)	60(±1.16)
i79	**31(±0.83)**	46(±0.90)	43(±0.88)	36(±0.81)	41(±0.90)	39.4(±0.86)	49(±0.71)	72(±0.97)
i28	51(±0.98)	35(±0.84)	**34(±0.79)**	39(±0.89)	39(±0.89)	39.6(±0.88)	46(±0.90)	62(±1.08)
i45	44(±1.02)	39(±0.91)	39(±0.91)	42(±1.00)	**34(±0.96)**	39.6(±0.96)	64(±0.78)	63(±1.07)
i9	42(±1.13)	**33(±0.89)**	41(±1.02)	37(±0.99)	46(±1.03)	39.8(±1.01)	55(±0.74)	67(±1.21)
i1	39(±0.97)	36(±0.83)	40(±0.97)	51(±0.96)	**34(±0.84)**	40.0(±0.91)	62(±0.80)	73(±1.05)
i39	44(±1.08)	**37(±1.01)**	41(±1.02)	39(±0.93)	41(±1.02)	40.4(±1.01)	57(±0.88)	63(±1.27)
i38	**34(±0.87)**	45(±1.07)	45(±1.16)	42(±1.06)	38(±0.94)	40.8(±1.02)	60(±0.86)	70(±1.28)

i42	46(±0.97)	**34(±0.82)**	44(±1.02)	41(±0.87)	39(±0.97)	40.8(±0.93)	68(±0.83)	64(±1.09)
i10	43(±0.86)	**36(±0.93)**	39(±0.91)	45(±0.99)	43(±0.88)	41.2(±0.91)	66(±0.65)	68(±1.35)
i5	41(±0.87)	42(±1.13)	43(±1.01)	**38(±0.96)**	46(±0.99)	42.0(±0.99)	67(±0.82)	74(±1.11)
i55	44(±0.96)	**30(±0.70)**	54(±0.92)	43(±0.83)	40(±0.81)	42.2(±0.84)	81(±0.83)	74(±1.31)
i60	35(±0.79)	42(±1.06)	52(±0.97)	**34(±0.74)**	50(±0.88)	42.6(±0.89)	78(±0.84)	75(±1.16)
i34	45(±0.91)	44(±0.94)	43(±0.95)	43(±0.97)	**39(±0.95)**	42.8(±0.94)	52(±0.86)	70(±1.05)
i91	45(±0.99)	42(±1.04)	42(±0.96)	48(±1.18)	**37(±0.92)**	42.8(±1.02)	68(±0.72)	67(±1.29)
i84	49(±1.02)	39(±0.86)	49(±0.87)	44(±0.92)	**35(±0.79)**	43.2(±0.89)	61(±0.91)	73(±1.18)
i8	44(±0.98)	39(±0.86)	54(±0.92)	50(±1.11)	**37(±0.92)**	44.8(±0.96)	52(±0.81)	64(±1.03)
i93	50(±1.18)	43(±1.03)	**43(±0.97)**	44(±1.08)	44(±1.19)	44.8(±1.09)	58(±0.71)	83(±1.32)
i7	36(±0.97)	46(±1.03)	58(±1.02)	**33(±0.92)**	53(±0.96)	45.2(±0.98)	89(±0.76)	79(±1.18)
i98	48(±1.09)	44(±1.06)	46(±0.88)	48(±1.14)	**42(±1.02)**	45.6(±1.04)	70(±0.76)	81(±1.24)
i2	41(±1.02)	42(±0.91)	58(±1.08)	51(±1.00)	**39(±1.04)**	46.2(±1.01)	64(±0.76)	69(±1.21)
i35	49(±1.06)	40(±0.90)	64(±1.13)	**29(±0.78)**	51(±1.06)	46.6(±0.99)	62(±0.85)	64(±1.26)
i12	48(±1.09)	44(±1.02)	48(±1.11)	**43(±1.01)**	51(±1.08)	46.8(±1.06)	55(±0.89)	56(±1.06)
i22	44(±1.04)	**35(±0.89)**	50(±1.11)	56(±1.10)	49(±1.13)	46.8(±1.05)	56(±0.77)	66(±1.10)
i64	**38(±0.87)**	48(±1.23)	49(±1.06)	56(±1.24)	44(±1.10)	47.0(±1.10)	66(±0.79)	67(±1.17)
i15	51(±0.91)	40(±1.01)	59(±0.98)	**39(±0.89)**	55(±0.89)	48.8(±0.94)	62(±0.85)	70(±1.14)
i23	44(±1.04)	**43(±0.86)**	61(±1.04)	47(±0.98)	53(±0.98)	49.6(±0.98)	85(±0.76)	86(±1.20)
i16	**43(±0.93)**	52(±1.07)	51(±1.10)	56(±1.04)	51(±1.08)	50.6(±1.04)	47(±0.68)	63(±1.14)
i80	48(±0.95)	**41(±0.90)**	64(±1.13)	46(±1.08)	59(±1.06)	51.6(±1.02)	75(±0.71)	88(±1.36)
i25	65(±0.95)	49(±0.98)	**43(±0.95)**	59(±0.96)	46(±0.97)	52.4(±0.96)	62(±0.74)	50(±1.01)
i67	78(±1.13)	50(±1.05)	44(±1.04)	52(±1.09)	**38(±0.92)**	52.4(±1.05)	55(±0.68)	61(±1.09)
i29	**40(±0.95)**	58(±1.11)	65(±0.97)	52(±1.11)	56(±0.98)	54.2(±1.02)	76(±0.81)	68(±1.12)
i50	**54(±1.10)**	57(±1.11)	78(±1.09)	63(±1.17)	76(±1.18)	65.6(±1.13)	69(±0.75)	90(±1.23)

We can see that results provided by each model when aggregating predictions, on both, expert and image levels, are better than two baselines (Median and Random columns). This means that the models are able to reconstruct the expert's preferences (to some extent and with a certain degree of accuracy).

Spearman correlation coefficient between the beauty ranking of an image (average image beauty for all 50 experts) and the quality of reproduction of the expert's decision (order based on the Avg. column) about the image is 0.21. This value, of course, indicates a relationship between the rankings, but not as great as was initially assumed before the study began. This prompted us to look for hidden multidimensional relationships invisible at first glance, which depend on the prediction error and image beauty.

Let us now answer the 1-st research question. When creating a quality ranking (Table 3 is based on the Tables 1 and 2) of a given machine learning-based model (when aggregating results for an image), it can be noticed that the best classifier is the nnet. It appears most often in the first place (31.3%) and very often in the second place (26.3%). The second place belongs to svm, the third to rf, and the fourth to rpart. The worst model is knn, which takes last place 43%. For the expert level, the best model is rf getting the best result 40%, however, it also has a large share of last place, as much as 30%. The nnet is in second place, gaining as much as 76% on the podium. It seems that svm is in third place, rpart is in fourth place, and again knn closes the entire rank.

When it comes to the 2-nd research question it can be concluded that the answer is positive. The results are presented in Table 4. By aggregating the results for the 5 or 10 best and worst reproducible images, it can be seen that the average or median expert rating differs. For the entire population, the average expert rating for the top 5 is 3.67, while for the low 5, it is 2.72. When aggregating the results into individual expert groups, this trend is maintained. The smallest difference is for the Art group and the

Tab. 2. Classification results for each expert. Each value presents distance defined in (1) and its standard error (the best result for each row is indicated in bold).

Expert	rpart	nnet	knn	rf	svm	Avg.	Median	Random
WTD10	**38(±0.58)**	46(±0.66)	50(±0.75)	57(±0.72)	44(±0.69)	47.0(±0.68)	122(±0.70)	139(±1.06)
Art5	45(±0.56)	57(±0.67)	52(±0.56)	**42(±0.70)**	47(±0.54)	48.6(±0.61)	52(±0.56)	80(±0.72)
WWa1	57(±0.80)	**31(±0.62)**	57(±0.92)	68(±0.74)	33(±0.67)	49.2(±0.75)	87(±1.00)	141(±1.21)
WTD5	54(±0.67)	56(±0.61)	61(±0.70)	**27(±0.57)**	53(±0.72)	50.2(±0.65)	70(±0.67)	99(±0.81)
WTD4	**40(±0.60)**	47(±0.59)	55(±0.61)	63(±0.66)	47(±0.59)	50.4(±0.61)	96(±0.71)	88(±0.78)
WTD9	61(±0.68)	48(±0.63)	55(±0.56)	**39(±0.57)**	49(±0.54)	50.4(±0.60)	88(±0.64)	101(±0.93)
WTD12	61(±0.90)	60(±0.89)	54(±0.84)	**40(±0.64)**	53(±0.82)	53.6(±0.82)	136(±0.56)	142(±1.24)
Std2	40(±0.70)	**38(±0.68)**	79(±0.91)	75(±1.04)	50(±0.76)	56.4(±0.82)	148(±1.06)	145(±1.19)
WTD2	68(±0.99)	**48(±0.86)**	54(±0.94)	63(±0.79)	49(±0.92)	56.4(±0.90)	106(±0.29)	116(±1.09)
WTD6	**45(±0.73)**	46(±0.75)	83(±0.93)	48(±0.56)	64(±0.88)	57.2(±0.77)	127(±1.01)	135(±1.16)
WWa8	**35(±0.58)**	49(±0.68)	76(±0.83)	85(±1.11)	46(±0.67)	58.2(±0.77)	102(±0.54)	115(±1.01)
Std5	50(±0.69)	53(±0.70)	52(±0.66)	95(±1.10)	**47(±0.63)**	59.4(±0.76)	60(±0.75)	76(±0.81)
WWa9	78(±0.99)	**49(±0.87)**	56(±0.91)	60(±0.91)	54(±0.90)	59.4(±0.92)	83(±1.07)	98(±1.09)
Std1	55(±0.79)	**49(±0.75)**	49(±0.75)	95(±1.05)	52(±0.79)	60.0(±0.83)	104(±0.44)	123(±1.01)
Art11	74(±0.81)	**48(±0.60)**	59(±0.68)	70(±1.11)	51(±0.61)	60.4(±0.76)	93(±0.60)	107(±0.95)
Art6	58(±0.77)	61(±0.98)	**55(±0.94)**	80(±0.80)	58(±0.77)	62.4(±0.85)	220(±0.75)	77(±1.01)
WWa11	76(±0.89)	**46(±0.66)**	70(±0.80)	67(±0.90)	59(±0.71)	63.6(±0.79)	111(±0.61)	154(±1.07)
Art2	**51(±1.00)**	58(±0.99)	76(±1.09)	64(±0.88)	72(±1.11)	64.2(±1.01)	112(±1.01)	128(±1.11)
Std13	64(±0.88)	**56(±0.80)**	75(±0.94)	73(±0.94)	60(±0.83)	65.6(±0.88)	118(±0.42)	134(±1.17)
WWa4	67(±0.93)	63(±0.94)	60(±0.91)	83(±0.93)	**57(±0.88)**	66.0(±0.92)	72(±1.02)	102(±1.14)
Art3	**59(±0.79)**	68(±0.75)	77(±0.84)	67(±1.09)	64(±0.80)	67.0(±0.85)	131(±0.65)	133(±1.13)
Std4	76(±0.97)	73(±0.94)	73(±0.95)	**54(±0.64)**	60(±0.88)	67.2(±0.88)	114(±1.06)	104(±1.02)
WWa7	75(±0.94)	69(±0.92)	73(±0.93)	**59(±0.70)**	66(±0.88)	68.4(±0.87)	112(±0.42)	123(±1.03)
Std3	68(±0.86)	67(±0.89)	73(±0.91)	74(±0.95)	**63(±0.83)**	69.0(±0.89)	117(±0.41)	118(±1.11)
Std9	83(±0.87)	**51(±0.83)**	75(±0.92)	69(±0.85)	71(±0.88)	69.8(±0.87)	178(±1.15)	153(±1.37)
Art10	69(±0.83)	62(±0.78)	73(±0.86)	98(±1.01)	**52(±0.68)**	70.8(±0.83)	107(±0.55)	122(±1.04)
WTD7	**53(±0.75)**	58(±0.73)	91(±0.99)	88(±0.82)	66(±0.88)	71.2(±0.83)	136(±0.60)	140(±1.25)
Art4	74(±1.04)	60(±0.98)	94(±1.23)	**57(±0.67)**	75(±0.99)	72.0(±0.98)	157(±0.53)	148(±1.47)
Std11	77(±0.88)	**61(±0.79)**	70(±0.82)	93(±0.97)	63(±0.83)	72.8(±0.86)	107(±0.49)	149(±1.05)
WTD8	**55(±0.76)**	66(±0.83)	74(±0.87)	96(±1.22)	73(±0.86)	72.8(±0.91)	95(±0.60)	121(±1.03)
Art8	63(±0.97)	66(±0.98)	97(±1.20)	**60(±0.71)**	79(±1.07)	73.0(±0.99)	153(±0.58)	178(±1.46)
WTD11	82(±0.86)	72(±0.79)	90(±0.86)	**50(±0.92)**	75(±0.77)	73.8(±0.84)	103(±0.62)	138(±1.03)
Std14	84(±0.92)	73(±0.95)	93(±0.90)	**42(±0.73)**	79(±0.99)	74.2(±0.90)	110(±0.51)	139(±1.13)
WWa2	77(±1.06)	68(±0.95)	106(±1.05)	**54(±0.64)**	75(±1.03)	76.0(±0.95)	150(±1.14)	177(±1.51)
WWa12	88(±0.81)	**63(±0.72)**	88(±0.88)	81(±0.94)	72(±0.75)	78.4(±0.82)	111(±0.64)	143(±0.98)
Std8	70(±1.05)	86(±1.03)	95(±1.11)	**67(±0.82)**	81(±1.03)	79.8(±1.01)	102(±1.07)	152(±1.19)
Art1	**69(±0.99)**	77(±1.04)	79(±1.04)	98(±1.01)	79(±1.02)	80.4(±1.02)	119(±1.02)	148(±1.15)
Art9	85(±0.83)	78(±0.77)	89(±0.80)	77(±1.01)	**73(±0.79)**	80.4(±0.84)	101(±0.64)	119(±1.03)
WTD1	90(±0.99)	92(±0.91)	87(±0.97)	**59(±0.83)**	82(±0.98)	82.0(±0.94)	138(±0.53)	139(±1.21)
WWa6	91(±0.95)	86(±0.85)	97(±0.98)	**62(±0.93)**	91(±0.94)	85.4(±0.93)	118(±0.77)	149(±1.15)
Std6	81(±1.03)	95(±1.03)	95(±0.98)	**65(±0.85)**	95(±0.95)	86.2(±0.97)	139(±0.55)	147(±1.36)
WTD3	87(±1.11)	98(±1.22)	101(±1.14)	**55(±0.79)**	90(±1.20)	86.2(±1.09)	126(±0.51)	163(±1.15)
WWa3	96(±1.16)	83(±1.09)	**83(±1.03)**	84(±0.87)	89(±1.07)	87.0(±1.04)	150(±0.54)	131(±1.34)
Art7	102(±1.12)	94(±1.03)	108(±1.12)	**55(±0.63)**	85(±1.02)	88.8(±0.98)	139(±1.13)	166(±1.41)
WWa10	99(±0.94)	81(±0.95)	97(±0.98)	**77(±0.91)**	97(±0.97)	90.2(±0.95)	152(±1.12)	136(±1.23)
Art12	87(±1.03)	109(±1.01)	101(±0.91)	**56(±0.67)**	101(±0.91)	90.8(±0.91)	120(±0.54)	137(±1.10)
WWa5	110(±1.06)	**83(±0.97)**	92(±0.98)	86(±0.99)	86(±0.99)	92.7(±1.00)	93(±0.64)	129(±0.96)
Std12	85(±0.96)	**83(±1.00)**	93(±1.08)	120(±1.15)	92(±1.06)	94.6(±1.05)	133(±0.52)	125(±1.29)
Std7	98(±1.05)	117(±1.18)	105(±1.07)	**68(±0.89)**	116(±1.21)	100.8(±1.08)	122(±1.09)	150(±1.38)
Std10	96(±1.02)	108(±1.09)	**93(±1.00)**	145(±1.28)	96(±1.02)	107.6(±1.08)	121(±0.56)	143(±1.19)

largest for the WWa group. The difference also persists for the top and low 10. This means that nice pictures are reproduced better.

The answers to the question 3-rd are included in the Table 5. Taking into account the voting results of all models (Overall row), the best-reconstructed group is WTD in both the top 5 and 10, i.e. the mentioned sets contain 17% and 33% of all experts from this group. In the low 5 or 10, there are only 8% of experts from this group. The Std group is next in the ranking, while the Art group has the worst reproducibility (0% in the top 5). It should be noted that both the best groups come from the same environment, i.e.

Tab. 3. The frequency of occurrence of a given model on a given position in the classification quality ranking for a given image or expert.

Level	Rank	rpart	nnet	knn	rf	svm
Image	1	19.2%	31.3%	11.1%	21.2%	23.2%
	2	20.2%	26.3%	15.2%	21.2%	29.3%
	3	21.2%	19.2%	11.2%	22.2%	22.2%
	4	18.2%	16.1%	19.2%	18.2%	22.2%
	5	21.2%	7.1%	43.3%	17.2%	3.1%
Expert	1	18.0%	28.0%	8.0%	40.0%	11.0%
	2	24.0%	30.0%	2.0%	6.0%	42.0%
	3	14.0%	18.0%	28.0%	12.0%	36.0%
	4	20.0%	16.0%	28.0%	12.0%	12.0%
	5	24.0%	8.0%	34.0%	30.0%	0.0%

Tab. 4. Statistics (average and median) of true labels for the best (top 5 and 10) and for the worst (low 5 and 10) classified images. Presented results are for the entire population and for each group of the experts.

Statistic	Top 5	Low 5	Top 10	Low 10
Avg. Label	3.67(±0.51)	2.72(±1.09)	3.77(±0.61)	2.96(±0.95)
Med. Label	4.20(±0.45)	2.60(±1.52)	4.15(±0.88)	2.90(±1.29)
Avg. Label Art	3.48(±0.48)	2.77(±0.93)	3.53(±0.58)	2.94(±0.91)
Avg. Label Std	3.84(±0.59)	2.91(±1.02)	3.91(±0.67)	3.14(±0.87)
Avg. Label WTD	3.60(±0.60)	2.62(±1.28)	3.87(±0.72)	2.92(±1.07)
Avg. Label WWa	3.72(±0.50)	2.57(±1.19)	3.75(±0.68)	2.82(±1.04)
Med. Label Art	3.70(±0.57)	2.50(±1.41)	3.75(±0.82)	2.95(±1.40)
Med. Label Std	4.20(±0.45)	3.00(±1.41)	4.25(±0.59)	3.10(±1.20)
Med. Label WTD	3.90(±0.74)	2.50(±1.50)	4.10(±0.91)	2.90(±1.26)
Med. Label Wwa	4.10(±0.55)	2.60(±1.52)	4.05(±0.93)	2.85(±1.25)

students or teaching members at the Faculty of Wood Technology or Institute of Wood Sciences and Furniture, respectively.

4.2. Importance of variables

The answer to the question 4-th will be presented at two levels of data aggregation. First for all images in total and then broken down into the error made by a given model. In the Figure 3 we present the frequency of occurrence of a given variable in the model structure for each image. The results were obtained via the VarImp function (described in 3.1) only for nnet and rpart. Due to the fact that each tree in rf is built to the

Tab. 5. Frequencies showing the best and worst reproducible expert groups.

Model	Expert	Top 5	Low 5	Top 10	Low 10
rpart	Art	8%	8%	17%	17%
	Std	7%	7%	14%	14%
	WTD	17%	0%	42%	42%
	WWa	8%	25%	8%	8%
nnet	Art	0%	8%	8%	8%
	Std	7%	21%	14%	14%
	WTD	17%	8%	42%	42%
	WWa	17%	0%	17%	17%
knn	Art	8%	17%	17%	17%
	Std	14%	7%	14%	14%
	WTD	17%	8%	42%	42%
	WWa	0%	8%	8%	8%
rf	Art	0%	17%	8%	8%
	Std	14%	14%	21%	21%
	WTD	17%	8%	33%	33%
	WWa	8%	0%	17%	17%
svm	Art	8%	8%	17%	17%
	Std	7%	21%	14%	14%
	WTD	8%	0%	33%	33%
	WWa	17%	8%	17%	17%
Overall	Art	0%	17%	18%	17%
	Std	14%	14%	21%	29%
	WTD	17%	8%	33%	8%
	WWa	8%	0%	17%	25%

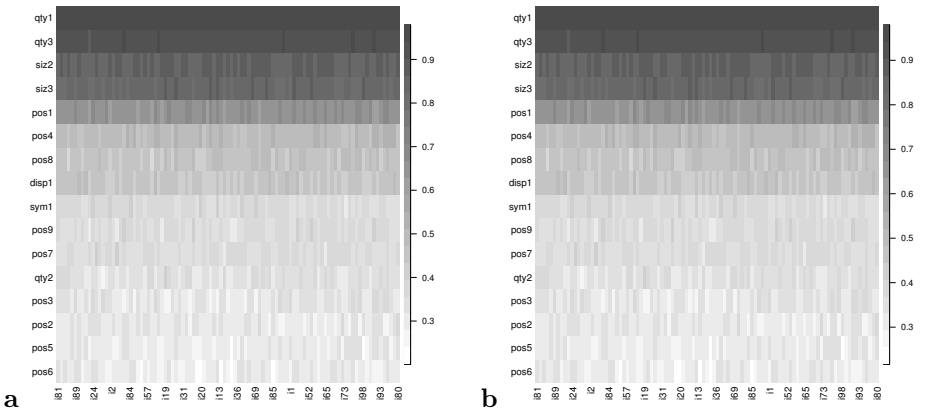

Fig. 3. Graphics showing the frequency of occurrence of a given variable in the model for each image: (**a**) rpart; (**b**) nnet. Each chart is sorted by columns (from the prettiest to the ugliest image) and by rows (from the most frequent to the least frequent variable). Due to the resolution and readability of the chart, labels of only some images are shown.

maximum level and the tree is very extensive, all variables are always included in the structure of a given tree. This results in the fact that each variable would receive the value 100% on this graph and there would be no cognitive value. For knn and svm there are no methods to derive importance of variables. The most common variables for nnet are `qty1`, `qty3`, `siz2`, and `siz3`, with a frequency of occurrence exceeding 80%. The `pos1` variable occurs in approximately 65% of images, the remaining variables appear less frequently than 45%. The results and conclusions for rpart are almost identical.

The next Figure 4 shows the average position in the importance ranking of a given variable for a given expert and a given image. This time results for rf are included as well. It is important to note that if a given variable did not appear in the model (i.e. the importance value was 0), when determining the ranking, the given variable received the lowest possible position, i.e. 16. In the case of rpart, one can see that 4 variables dominate (`qty1`, `qty3`, `siz2`, and `siz3`). It is the same set but with a slightly different order than in the case of frequency of occurrence. Additionally, unlike the previous analysis, all positional variables (`pos1-pos9`) occupy the last positions. In the case of nnet, the distribution of values is less polarized. The same group of variables as for the rpart leads the entire ranking. However, the remaining variables are no longer so far apart in the ranking and it can be seen that the values are similar. This means that the position in the ranking varied and depended on the expert and the image. In the case of rf, three more variables (`sym1`, `qty2`, and `pos1`) are added to the previously mentioned set. The variable `disp1` was ranked the lowest.

The next Figure 5 shows again the frequency of occurrence of a given variable in the

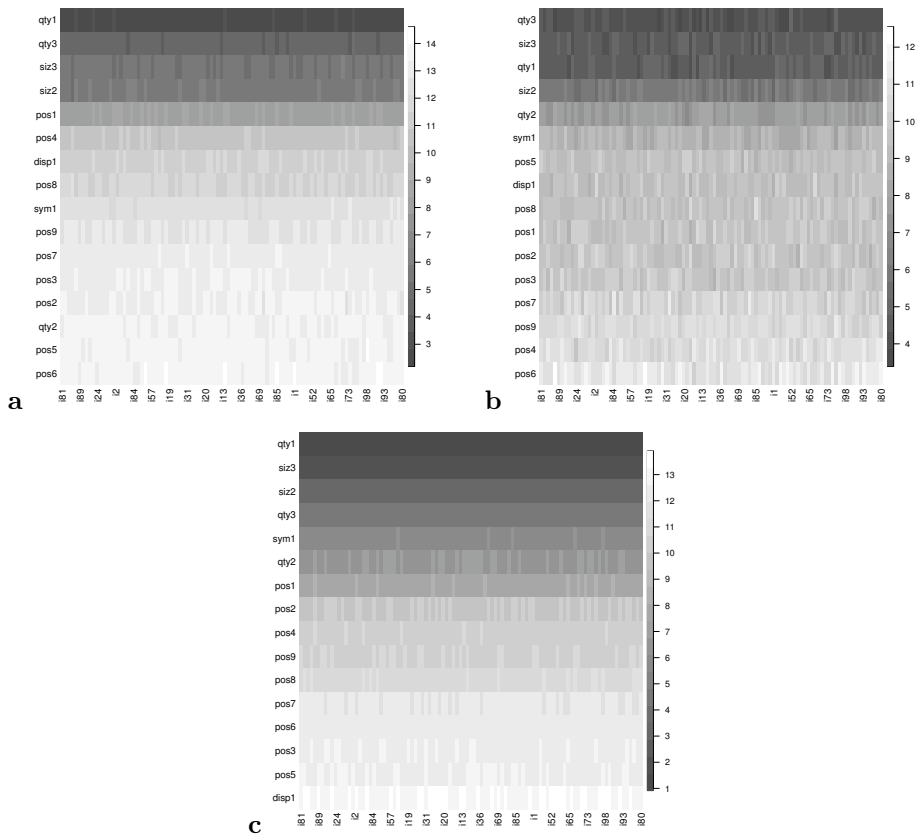

Fig. 4. Graphics showing the average importance ranking of a given variable in the model for each image: (**a**) rpart; (**b**) nnet; (**c**) rf. Each chart is sorted by columns (from the prettiest to the ugliest picture) and by rows (from the most relevant to the least relevant variable). Due to the resolution and readability of the chart, labels of only some images are shown.

model, but this time broken down into the error made by a given model for a given image. There are 9 possible errors from -4 (the predicted label is 5 while the true label is 1) to 4 (the predicted label is 1 while the true label is 5). Due to the length of the paper and the clarity of the message, we show the figure only for nnet, while the conclusions for the remaining models are quite similar.

Errors -4 and -3 usually occur with ugly images on the right side of the graph, this tendency, but to a lesser extent, is also visible for errors -2 and -1. This means that the rating was rather low and the model predicted value was rather higher. The better reproducible images (on the left-hand side) show a more clustered distribution of

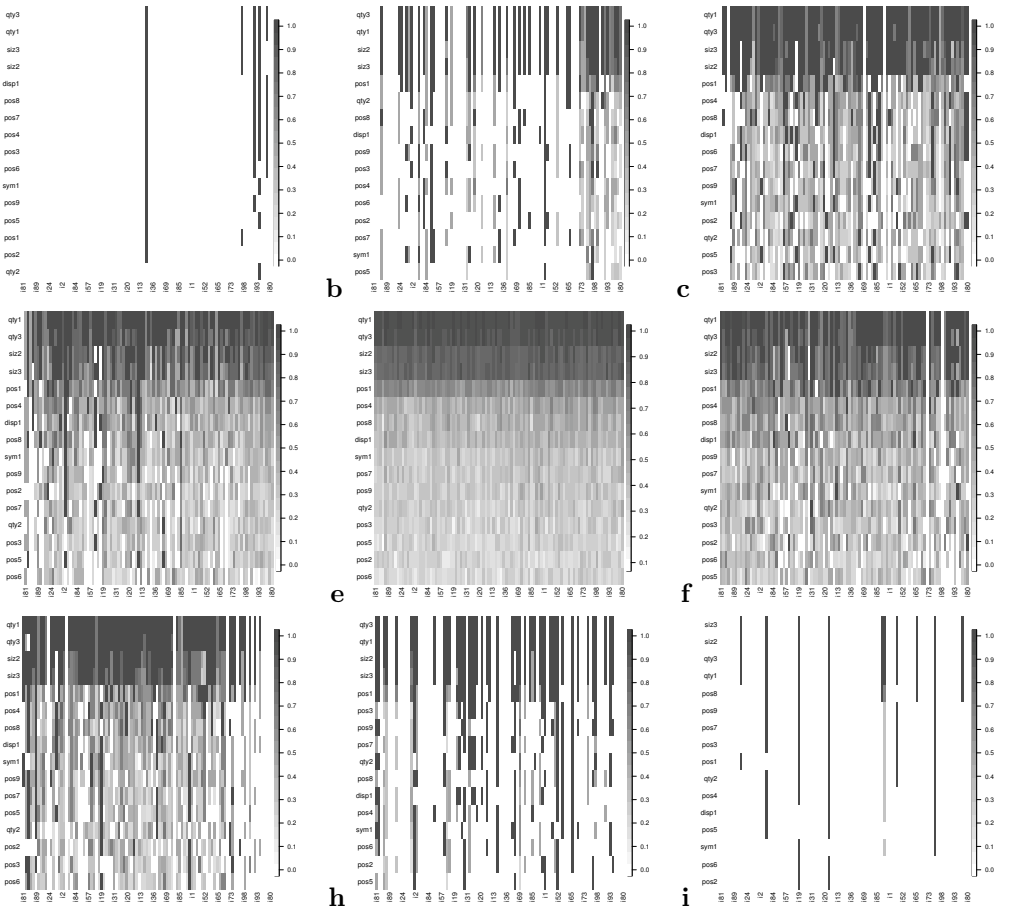

Fig. 5. Graphics showing the frequency of occurrence of a given variable in the nnet model for each image where prediction error equals (i.e. true label – prediction): (**a**) -4; (**b**) -3; (**c**) -2; (**d**) -1; (**e**) 0; (**f**) 1; (**g**) 2; (**h**) 3; (**i**) 4. Each chart is sorted by columns (from the prettiest to the ugliest image) and by rows (from the most frequent to the least frequent variable). Due to the resolution and readability of the chart, labels of only some images are shown.

frequencies to the best 4 or 5 variables (darker top right corner and lighter bottom left corner), while the images on the right-hand side have a more even distribution (longer vertical stripes of similar color). This proves that the decision for nice images is made based on the best 4 or 5 variables, while for images on the right, many more features are important. For the error equal to zero, the graph is very similar to the graph obtained for the entire population. In the case of errors directed in the opposite direction (positive

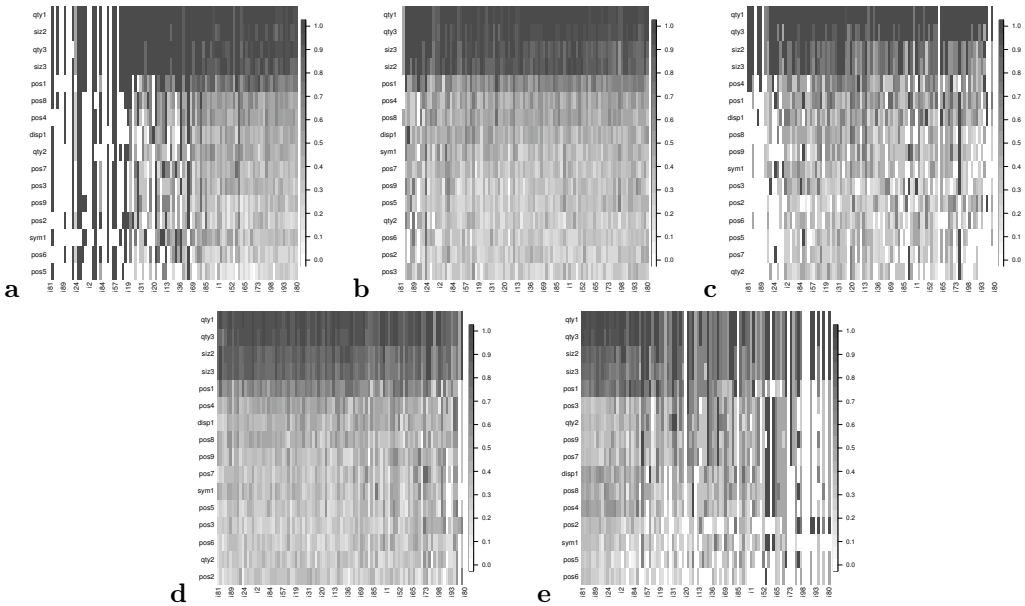

Fig. 6. Graphics showing the frequency of occurrence of a given variable in the neural network model for each true label of the image: (**a**) 1; (**b**) 2; (**c**) 3; (**d**) 4; (**e**) 5. Each chart is sorted by columns (from the prettiest to the ugliest image) and by rows (from the most frequent to the least frequent variable). Due to the resolution and readability of the chart, labels of only some images are shown.

errors 1 and 2), a slightly opposite tendency is visible, i.e. these errors occur less often for ugly images (white vertical stripes on the right). Finally, if an extreme error of 4 occurs, the models have more variables in their structure. Analyzing the order of occurrence of the variables, one can notice slightly opposite behavior of the variables `qty2` and `sym1`. For an error in the range of -2 to 2, the more important variable is `sym1` (however both are in the 3rd or 4th quadrant). For extreme negative and positive values, the order changes, and the `qty2` variable is more important.

The last set of charts (Figure 6), like the previous ones, shows the frequency of occurrence of a given variable in the nnet model, but this time divided into the true value of the label, from 1 to 5.

One can observe white stripes on the left-hand side for label 1 and on the right-hand side for label 5, this is a direct result of the fact that there are obviously no images from a given class here. The most important variables that always top the ranking regardless of the rating are `qty1`, `qt3`, `siz2`, and `siz3`. The `qty2` variable gains importance in extreme ratings (1 and 5), e.g. for 3 it is in last place. The `sym1` variable is more

Tab. 6. Average values (and their standard deviations) of the features (top 5 and 10; low 5 and 10) for the prettiest and ugliest images and the best and the worst reconstructive images.

	Beauty				Reconstruction			
Feature	Top 5	Low 5	Top 10	Low 10	Top 5	Low 5	Top 10	Low 10
`qty1`	2.6(±1.14)	4.0(±2.92)	2.3(±1.25)	3.0(±2.36)	3.6(±1.52)	5.4(±1.95)	2.9(±1.52)	3.9(±2.23)
`qty2`	0.4(±0.55)	0.6(±0.55)	0.3(±0.48)	0.6(±0.52)	0.6(±0.55)	0.8(±0.45)	0.5(±0.53)	0.5(±0.53)
`qty3`	1.6(±0.55)	2.0(±1.00)	1.5(±0.71)	1.6(±0.97)	2.0(±0.71)	2.6(±0.55)	1.7(±0.67)	2.1(±0.74)
`disp1`	0.2(±0.45)	0.6(±0.55)	0.3(±0.48)	0.6(±0.52)	0.4(±0.55)	0.8(±0.45)	0.3(±0.48)	0.4(±0.52)
`pos1`	0.2(±0.45)	0.6(±0.55)	0.2(±0.42)	0.4(±0.52)	0.6(±0.55)	0.8(±0.45)	0.4(±0.52)	0.4(±0.52)
`pos2`	0.4(±0.55)	0.4(±0.55)	0.3(±0.48)	0.5(±0.53)	0.4(±0.55)	0.6(±0.55)	0.4(±0.52)	0.5(±0.53)
`pos3`	0.0(±0.00)	0.6(±0.55)	0.1(±0.32)	0.4(±0.52)	0.4(±0.55)	1.0(±0.00)	0.2(±0.42)	0.6(±0.52)
`pos4`	0.4(±0.55)	0.6(±0.55)	0.3(±0.48)	0.5(±0.53)	0.6(±0.55)	0.8(±0.45)	0.5(±0.53)	0.6(±0.52)
`pos5`	0.2(±0.45)	0.6(±0.55)	0.3(±0.48)	0.3(±0.48)	0.4(±0.55)	0.8(±0.45)	0.4(±0.52)	0.4(±0.52)
`pos6`	0.4(±0.55)	0.4(±0.55)	0.4(±0.52)	0.5(±0.53)	0.4(±0.55)	0.2(±0.45)	0.3(±0.48)	0.4(±0.52)
`pos7`	0.6(±0.55)	0.4(±0.55)	0.6(±0.52)	0.4(±0.52)	0.4(±0.55)	0.0(±0.00)	0.5(±0.53)	0.2(±0.42)
`pos8`	0.2(±0.45)	0.4(±0.55)	0.5(±0.53)	0.3(±0.48)	0.2(±0.45)	0.2(±0.45)	0.3(±0.48)	0.2(±0.42)
`pos9`	0.8(±0.45)	0.2(±0.45)	0.6(±0.52)	0.5(±0.53)	0.6(±0.55)	0.0(±0.00)	0.6(±0.52)	0.4(±0.52)
`siz2`	2.0(±1.00)	1.8(±0.84)	1.8(±1.14)	1.7(±0.95)	2.0(±0.71)	2.0(±0.71)	2.3(±0.82)	2.1(±0.74)
`siz3`	2.8(±2.05)	3.0(±1.87)	2.6(±2.12)	2.9(±1.91)	2.8(±1.64)	3.2(±1.64)	3.5(±1.78)	3.3(±1.64)
`sym1`	0.2(±0.45)	0.4(±0.55)	0.2(±0.42)	0.4(±0.52)	0.2(±0.45)	0.0(±0.00)	0.1(±0.32)	0.0(±0.00)

important in the middle ratings (2-4) and less important for the extreme ones. The relation between `qty2` and `sym1` is quite similar as broken down into the errors. The `disp1` variable is always in the middle of the rank at the same level for all ratings.

4.3. Statistics of variables

To answer the question 5-th the Table 6 is prepared. It shows the actual average feature values divided into image beauty and reproducibility. As said before, the best variables are `qty1`, `qty3`, `siz2` and `siz3`. The mean values for the `qty1` variable for both divisions differ in the groups. For beauty in the top 5, one can see that prettier images have fewer knots (2.6) and uglier have 4. For reconstruction, these values are greater and the difference is bigger (3.6 vs. 5.4). A similar relationship exists for the `qty3` variable but not with the same explanatory power. For `siz2` and `siz3` variables, it cannot be concluded that there is a difference in the average values of these features.

Prettier images are less symmetrical (0.2 vs. 0.4), on the other hand, better reproducible images are more symmetrical (0.2 vs. 0). For both divisions, more beautiful and better reproducible images are less dispersed (`disp1`).

5. Conclusion

The findings from the questionnaire survey indicate that all five expert groups found the furniture featuring solid wood fronts sufficiently appealing to consider using them in both their personal residences and in a public building. Based on the research, it can be

concluded that specific features describing the image influence the expert's perception of attractiveness. The findings based on the numerical experiments can be summarized as follows:

1. The best results of expert decision reconstruction are provided by a neural network model.
2. The expert's decision is better reconstructed for more beautiful images.
3. The best reproducible groups of experts are groups with a similar background, i.e. WTD and Std.
4. The most significant features are `qty1`, `qty2`, `siz2`, `siz3` and `pos1`.
5. Adequate value of 4 or 5 features is enough to score as nice image, while to score as ugly more features are needed.
6. There is a slightly opposite behavior of the variables `qty2` and `sym1`. The `qty2` variable gains importance in extreme ratings, while `sym1` is more important for middle ratings.
7. Prettier images and those for which the expert's decision is better reconstructed have fewer knots. In other words, more knots, worst reconstruction what is also usual score as less attractive.

Our future research focuses on the development of an automatic scoring system for fronts made of pine wood. Since we roughly know what features are important for the expert, the next stage will be to calculate them automatically in the image. The calculation of these features can be done using pattern recognition and image analysis methods or using deep learning (DL) algorithms. DL algorithms also allow us to find other features not identified here that may improve the quality of reconstruction of the expert's decisions [21, 26, 33]. Future research is therefore part of the growing trend of explainable artificial intelligence [24, 27, 35].

References

[1] M. R. Antal, D. Domljan, and P. G. Horváth. Functionality and aesthetics of furniture – Numerical expression of subjective value. *Drvna industrija*, 67(4):323–332, 2017. doi:10.5552/drind.2016.1544.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi:10.1023/a:1010933404324.

[3] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification And Regression Trees*. Routledge, Oct 2017. doi:10.1201/9781315139470.

[4] I. Cetiner, A. Ali Var, and H. Cetiner. Classification of knot defect types using wavelets and KNN. *Elektronika ir Elektrotechnika*, 22(6), 2016. doi:10.5755/j01.eie.22.6.17227.

[5] M. Chen and J. H. Lyu. Aesthetic evaluation of furniture design based on anp method. *Applied Mechanics and Materials*, 574:318–323, 2014. doi:10.4028/www.scientific.net/amm.574.318.

[6] L. Deng and G. Wang. Quantitative evaluation of visual aesthetics of human-machine interaction interface layout. *Computational Intelligence and Neuroscience*, 2020:1–14, 2020. doi:10.1155/2020/9815937.

[7] J. Fürnkranz and E. Hüllermeier. Preference learning. In: C. Sammut and G. I. Webb, eds., *Encyclopedia of Machine Learning*, p. 789–795. Springer US, Boston, MA, 2011. doi:10.1007/978-0-387-30164-8_662.

[8] M. Gagolewski and J. Lasek. Learning experts' preferences from informetric data. In: *Advances in Intelligent Systems Research*, ifsa-eusflat-15. Atlantis Press, 2015. doi:10.2991/ifsa-eusflat-15.2015.70.

[9] K. Gajowniczek, Y. Liang, T. Friedman, T. Ząbkowski, and G. Van den Broeck. Semantic and generalized entropy loss functions for semi-supervised deep learning. *Entropy*, 22(3):334, 2020. doi:10.3390/e22030334.

[10] K. Gajowniczek, A. Orłowski, and T. Ząbkowski. Simulation study on the application of the generalized entropy concept in artificial neural networks. *Entropy*, 20(4):249, 2018. doi:10.3390/e20040249.

[11] G. D. Garson. Interpreting neural-network connection weights. *AI Expert*, 6(4):46–51, 1991. `https://dl.acm.org/doi/10.5555/129449.129452`.

[12] S. Gold and F. Rubik. Consumer attitudes towards timber as a construction material and towards timber frame houses – selected findings of a representative survey among the german population. *Journal of Cleaner Production*, 17(2):303–309, 2009. doi:10.1016/j.jclepro.2008.07.001.

[13] T. A. Guzel. Consumer attitudes toward preference and use of wood, woodenware, and furniture: A sample from kayseri, turkey. *BioResources*, 15(1):28–37, 2019. doi:10.15376/biores.15.1.28-37.

[14] J. Han, H. Forbes, and D. Schaefer. An exploration of how creativity, functionality, and aesthetics are related in design. *Research in Engineering Design*, 32(3):289–307, 2021. doi:10.1007/s00163-021-00366-9.

[15] U. R. Hashim, S. Z. Hashim, and A. K. Muda. Automated vision inspection of timber surface defect: A review. *Jurnal Teknologi*, 77(20), 2015. doi:10.11113/jt.v77.6562.

[16] S. Kizito, A. Y. Banana, M. Buyinza, J. R. S. Kabogozza, R. K. Kambugu, et al. Consumer satisfaction with wooden furniture: an empirical study of household products produced by small and medium scale enterprises in uganda. *Journal of the Indian Academy of Wood Science*, 9(1):1–13, 2012. doi:10.1007/s13196-012-0068-1.

[17] A. Krähenbühl, B. Kerautret, I. Debled-Rennesson, F. Longuetaud, and F. Mothe. *Knot Detection in X-Ray CT Images of Wood*, p. 209–218. Springer Berlin Heidelberg, 2012. doi:10.1007/978-3-642-33191-6_21.

[18] M. Kuhn. Building predictive models in R using the caret package. *Journal of Statistical Software*, 28(5), 2008. doi:10.18637/jss.v028.i05.

[19] J. Parmar, S. Chouhan, V. Raychoudhury, and S. Rathore. Open-world machine learning: Applications, challenges, and opportunities. *ACM Computing Surveys*, 55(10):1–37, 2023. doi:10.1145/3561381.

[20] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. `https://www.R-project.org/`.

[21] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In: *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, KDD '16, p. 1135–1144. ACM, New York, NY, USA, 13-17 Aug 2016. doi:10.1145/2939672.2939778.

[22] B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Jan 1996. doi:10.1017/cbo9780511812651.

[23] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *Neural Computation*, 12(5):1207–1245, 2000. doi:10.1162/089976600300015565.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, 2019. doi:10.1007/s11263-019-01228-7.

[25] J. Y. Shin, C. Kim, and H. J. Hwang. Prior preference learning from experts: Designing a reward with active inference. *Neurocomputing*, 492:508–515, 2022. doi:10.1016/j.neucom.2021.12.042.

[26] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *arXiv*, 2015. ArXiv.1412.6806. doi:10.48550/arXiv.1412.6806.

[27] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In: *Proc. 34th Int. Conf. Machine Learning*, vol. 70 of *ICML'17*, p. 3319–3328. JMLR.org, 6-11 Aug 2017. `https://dl.acm.org/doi/abs/10.5555/3305890.3306024`.

[28] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Springer New York, 2002. doi:10.1007/978-0-387-21706-2.

[29] M. N. Volkovs, H. Larochelle, and R. S. Zemel. Learning to rank by aggregating expert preferences. In: *Proc. 21st ACM Int. Conf. Information and Knowledge Management*, CIKM'12. ACM, Maui, Hawaii, USA, 29 Oct – 2 Nov 2012. doi:10.1145/2396761.2396868.

[30] C. van Winkelen and R. McDermott. Learning expert thinking processes: using KM to structure the development of expertise. *Journal of Knowledge Management*, 14(4):557–572, 2010. doi:10.1108/13673271011059527.

[31] C. Xiaolei, S. Jun, and L. Bing. Customer preferences for kitchen cabinets in China using conjoint analysis. *Journal of Chemical and Pharmaceutical Research*, 6(2):14–22, 2014. `https://www.jocpr.com/articles/customer-preferences-for-kitchen-cabinets-in-china-using-conjoint-analysis.pdf`.

[32] S. Yoon, H. Oh, and J. Y. Cho. Understanding furniture design choices using a 3D virtual showroom. *Journal of Interior Design*, 35(3):33–50, 2010. doi:10.1111/j.1939-1668.2010.01041.x.

[33] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In: D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, eds., *Computer Vision – Proc. ECCV 2014*, pp. 818–833. Springer International Publishing, Cham, Zurich, Switzerland, 6-12 Sep 2014. doi:10.1007/978-3-319-10590-1_53.

[34] L. Zeng and D. Liu. A study on the model of furniture aesthetic value based on fuzzy AHP comprehensive evaluation. In: *Proc. 2010 7th Int. Conf. Fuzzy Systems and Knowledge Discovery*. IEEE, Yantai, China, 10-12 Aug 2010. doi:10.1109/fskd.2010.5569152.

[35] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv*, 2017. ArXiv.1702.04595. doi:10.48550/arXiv.1702.04595.

[36] K. Śmietańska and J. Górski. Impact of visible knots on relative visual attractiveness of furniture fronts made of pine wood (pinus sylvestris l.). *Wood Material Science & Engineering*, 18(5):1749–1754, 2023. doi:10.1080/17480272.2023.2186263.

[37] K. Śmietańska, P. Podziewski, M. Bator, and J. Górski. Automated monitoring of delamination factor during up (conventional) and down (climb) milling of melamine-faced MDF using image processing methods. *European Journal of Wood and Wood Products*, 78(3):613–615, 2020. doi:10.1007/s00107-020-01518-9.