Decomposition of Perceptual Scores in Multi-Component IQA: Pathways to Enhanced Model Interpretability and Efficiency

A.M. Trykin [1], V.E. Turlapov [1]
[1] Artificial Intelligence Research Center, Lobachevsky State University, 603022, Nizhny Novgorod, Russia, Gagarin Ave., 23

Abstract

Traditional Image Quality Assessment (IQA) often aggregates human perception into a single score, overlooking its multidimensional nature, which is crucial for interpretable IQA modeling. This study presents a systematic cross-dataset analysis of how technical distortions or aesthetic components jointly shape overall quality perception. By evaluating seven large-scale datasets (SPAQ, KonIQ-10k, GISET, ICAA20K, AADB, EVA, and PARA), we investigate the relationships between human-annotated attribute-level features and Mean Opinion Scores (MOS) using Pairwise Correlation (PC), Principal Component Analysis (PCA), and cross-validated regression models. Our analysis uncovers four distinct empirical patterns in perceptual score modeling: (1) high attribute redundancy (SPAQ, GISET, EVA, PARA), where inter-attribute correlations exceed 0.7–0.9, achieving near-optimal predictive power ($\mathrm{PLCC/SRCC} > 0.96$) and indicating that existing multi-attribute annotations often collapse into a single dimension, rendering complex architectural partitioning redundant for these specific tasks; (2) effective orthogonality in AADB, where aesthetic attributes show weak inter-correlations ($r < 0.3$) yet enable accurate regression ($\mathrm{PLCC/SRCC} \approx 0.90$), demonstrating that orthogonal features contribute unique variance; (3) an extraction and complexity gap in KonIQ-10k, where structurally simple, automated attributes fail to align with actual subjective perception ($\mathrm{PLCC/SRCC} \approx 0.74$); and (4) insufficient specificity in ICAA20K, where isolated perceptual color attributes (temperature, harmony) show weak correlation with MOS ($\mathrm{PLCC} \approx 0.34$), proving inadequate without semantic context. Based on these findings, we propose actionable design principles for future IQA benchmarks, prioritizing semantically distinct attributes alongside annotation protocols designed to mitigate cognitive biases (e.g., the halo effect), thereby ensuring interpretable and robust IQA models.

Keywords: image quality assessment, image quality aesthetic assessment, multidimensional image quality assessment, correlation analysis, principal component analysis, machine learning, regression.

Citation: Trykin AM, Turlapov VE. Decomposition of Perceptual Scores in Multi-Component IQA: Pathways to Enhanced Model Interpretability and Efficiency. Computer Optics 2026; 50(4): 2082. DOI: 10.18287/COJ2082.

Introduction

Image Quality Assessment (IQA) plays a critical role in various computer vision applications, including image compression, enhancement, restoration, and generation [1 – 3]. The ultimate goal of IQA is to predict the perceptual quality of images as accurately as possible, aligning with human subjective opinions. Traditionally, this has been achieved by predicting a single Mean Opinion Score (MOS) or aesthetic rating, which serves as a ground truth for training and evaluating objective models. While significant progress has been made in developing blind IQA (BIQA) algorithms that achieve high correlation with human ratings, most existing approaches treat quality as a monolithic scalar value.

However, human perception of image quality is multifaceted and context-dependent. An observer's judgment is influenced by a combination of technical attributes (such as noise, blur, and compression artifacts) and aesthetic components (such as color harmony, composition, and content). Understanding how these individual quality components contribute to the overall assessment is essential for building interpretable models and guiding targeted image improvements.

Despite the availability of large-scale multi-attribute datasets, there is limited comprehensive analysis regarding the interdependencies between quality attributes and their collective impact on overall perception across different content domains [4]. Some attributes may be highly correlated, suggesting redundancy, while others may be orthogonal, providing unique information. Furthermore, the structure of these relationships may vary significantly depending on the type of content (e.g., authentic distortions in mobile images vs. artistic aesthetics in photography).

Our contributions are threefold:

- We present the first systematic cross-dataset comparison of attribute predictability across technical and aesthetic IQA benchmarks;
- We identify four empirically distinct scenarios of attribute-MOS relationships;
- We provide reproducible baselines and hyperparameter configurations for multi-component IQA modeling.

1. Datasets

To comprehensively evaluate the impact of quality components on overall Image Quality Assessment (IQA), we utilize several large-scale benchmark datasets. These datasets encompass a wide range of authentic distortions, capture conditions, and annotation granularities, ensuring the robustness of our analysis.

SPAQ [5]: The Smartphone Photography Attribute and Quality dataset contains 11 125 images captured by diverse smartphone cameras under naturalistic conditions. Human observers provided both MOS values and attribute scores across five dimensions: *Brightness*, *Colorfulness*, *Contrast*, *Noisiness*, and *Sharpness*. The dataset is designed to reflect authentic, co-occurring distortions typical of mobile photography.

KonIQ-10k [6]: An ecologically valid database of 10 073 web-collected images with MOS values gathered via crowdsourcing from 120 raters. Critically, attribute scores (*Brightness*, *Contrast*, *Colorfulness*, *Sharpness*) were not obtained from human judgments but computed algorithmically. This methodological distinction is central to interpreting the attribute–MOS correlation structure.

GISET [7]: A small-scale dataset of 164 game content images annotated for *Fragmentation* and *Unclearness* in addition to an overall quality score. It enables fine-grained analysis of how perceptually distinct degradations (compression artifacts and blur) manifest in synthetic visual content.

ICAA20K [8]: A color-centric benchmark comprising 20 355 images annotated with three perceptual color dimensions: *Color Temperature*, *Colorfulness*, and *Color Harmony*. The dataset supports both holistic and attribute-level Image Color and Aesthetics Assessment (ICAA).

AADB [9]: The Photo Aesthetics Ranking Network with Attributes and Content Adaptation dataset contains 9 958 images with human-annotated aesthetic scores across eleven attributes, including *Content*, *Light*, *Color Harmony*, *Rule of Thirds*, and *Symmetry*. Its broad attribute vocabulary makes it well-suited for studying compositional and semantic factors in aesthetic quality.

EVA [10]: The Explainable Visual Aesthetics dataset comprises 4 070 images rated by 30 – 40 observers on overall aesthetic quality and four sub-components: *Light & Color*, *Composition & Depth*, *Quality*, and *Semantic Content*. The multi-rater design provides stable estimates of inter-attribute agreement.

PARA [11]: A large-scale personalized aesthetic benchmark with 31 220 images annotated for seven attributes: *Quality*, *Composition*, *Color*, *Depth of Field*, *Light*, *Content*, and *Object Emphasis*. The inclusion of *Object Emphasis* as an attribute with potentially idiosyncratic perceptual salience makes PARA particularly informative for studying annotation artifacts.

Tab. 1. summarizes the statistical properties of the datasets used in this study.

Tab. 1.Summary of multi-attribute IQA datasets used in the study

Dataset	Images	#Attr	Attributes	Content	Annotation	Raters	MOS
SPAQ	11 125	5	Brightness, Colorfulness, Contrast, Noisiness, Sharpness	Smartphone photos	Human	~1 750	[0,100]
KonIQ-10k	10 073	4	Brightness, Contrast, Colorfulness, Sharpness	Web images	Algorithmic	120	[1,5]
GISET	164	2	Fragmentation, Unclearness	Game content	Human	–	[1,5]
ICAA20K	20 355	3	Color Temperature, Colorfulness, Color Harmony	Web images	Algorithmic	–	[0,10]
AADB	9 958	11	Balancing, Color Harmony, Content, DoF, Light, Motion Blur, Object, Repetition, Rule of Thirds, Symmetry, Vivid Color	Art-photo	Human	~5	[1,5]
EVA	4 070	4	Light & Color, Composition & Depth, Quality, Semantic Content	Art-photo	Human	30–40	[0,10]
PARA	31 220	7	Quality, Composition, Color, DoF, Light, Content, Object Emphasis	Art-photo	Human	~3	[1,5]

2. Methods

2.1. Evaluation Metrics

We employ four complementary metrics to quantify the agreement between model predictions and subjective MOS values.

All three correlation coefficients lie in $[-1,1]$: a value of $+1$ indicates perfect positive agreement, -1 indicates perfect inverse agreement, and 0 indicates no monotonic or linear relationship.

Pearson Linear Correlation Coefficient (PLCC) measures the strength of the *linear* relationship between predicted and ground-truth scores:

$$\mathrm{PLCC} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(\hat{y}_i-\bar{\hat{y}})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}}, \quad (1)$$

where y_i and $\hat{y}_i$ are the ground-truth and predicted scores for image i, $\bar{y}$ and $\bar{\hat{y}}$ are the respective means, and n is the number of images.

Spearman Rank Correlation Coefficient (SRCC) assesses *monotonic* agreement by operating on rank-transformed scores, making it robust to nonlinear scale transformations and outliers:

$$\mathrm{SRCC} = \frac{\sum_{i=1}^{n}(R_i-\bar{R})(S_i-\bar{S})}{\sqrt{\sum_{i=1}^{n}(R_i-\bar{R})^2}\sqrt{\sum_{i=1}^{n}(S_i-\bar{S})^2}}, \quad (2)$$

where $R_i = rank(y_i)$ and $S_i = rank(\hat{y}_i)$.

Kendall Rank Correlation Coefficient (KRCC) quantifies rank agreement via the proportion of concordant and discordant pairs. The τ_b variant, which corrects for tied ranks common in discrete MOS scales, is used:

$$\mathrm{KRCC} = \tau_b = \frac{N_c-N_d}{\sqrt{(N_0-N_1)(N_0-N_2)}}, \quad (3)$$

where N_c and N_d denote concordant and discordant pairs, $N_0 = n(n-1)/2$, and N_1, N_2 are tie-correction terms for y and $\hat{y}$, respectively. KRCC yields more conservative concordance estimates than SRCC.

Root Mean Square Error (RMSE) quantifies absolute prediction error on the original MOS scale:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}. \quad (4)$$

Unlike the rank-based metrics, RMSE is sensitive to the magnitude of individual errors and to outliers, making it a useful complement for assessing absolute predictive accuracy.

Jointly, PLCC and SRCC reveal whether observed associations are stable under both linear and rank-order assumptions: agreement between the two suggests a predominantly linear relationship, while a substantially higher SRCC points to a monotonic but nonlinear one.

2.2. Regression ML Models

Our regression task differs fundamentally from pixel-level IQA: the input space consists of a small number of highly aggregated scalar attribute scores rather than raw image data. Employing deep neural networks in this setting would introduce severe overfitting and undermine interpretability. Instead, we use classical machine learning regressors that are well suited to low-dimensional tabular inputs and that support transparent attribution of model decisions.

We evaluate five algorithms: Linear Regression (LR) [12] as a transparent baseline; Support Vector Regression (SVR) [13, 14] with an RBF kernel; Decision Tree regression (DT) [15]; k-Nearest Neighbors (KNN) [16]; and a shallow Multi-Layer Perceptron (MLP) [17] with a single hidden layer. All models are implemented in scikit-learn [18].

Prior to training, all input features were scaled to the range $[0,1]$ using min-max normalization to ensure uniform contribution of attributes with different scales. Hyperparameter optimization was performed using Optuna [19] with the Tree-structured Parzen Estimator (TPE) sampler [20], maximizing the SRCC over 10-fold cross-validation. The optimization was constrained to 100 trials per model. Among the hyperparameters tested were: for KNN (n_neighbors, weights, p), for SVR (C, gamma, epsilon, kernel), for MLP (hidden_layer_sizes, alpha, learning_rate_init, activation), for DT (max_depth, min_samples_split, min_samples_leaf, criterion). LR served as a baseline without hyperparameter tuning.

3. Results

3.1. Pairwise Correlation Analysis

We report both Pearson (r_P) and Spearman (r_S) correlation matrices for each dataset in Tables 2 – 8. Color intensity reflects the magnitude of the correlation: darker red indicates stronger positive association, blue indicates negative values. The two measures are complementary: PLCC captures linear association, while SRCC captures monotonic association. Convergence between the two indicates a stable, approximately linear relationship; a substantially higher SRCC suggests nonlinearity; a discrepancy in the opposite direction may indicate outlier influence.

*Tab. 2. Pairwise correlation matrix for the **SPAQ** dataset. Each cell: r_P / r_S*

	MOS	Brightness	Colorfulness	Contrast	Noisiness	Sharpness
MOS	–					
Brightness	0.77 / 0.78	–				
Colorfulness	0.83 / 0.84	0.88 / 0.87	–			
Contrast	0.86 / 0.87	0.90 / 0.90	0.92 / 0.92	–		
Noisiness	0.89 / 0.89	0.73 / 0.75	0.80 / 0.82	0.82 / 0.83	–	
Sharpness	0.96 / 0.96	0.67 / 0.70	0.75 / 0.78	0.79 / 0.81	0.87 / 0.88	–

SPAQ: All five technical attributes correlate strongly with MOS (both r_P and $r_S > 0.75$), reaching 0.96 for *Sharpness*. Inter-attribute correlations are equally high: *Contrast* and *Noisiness* co-vary at 0.92, reflecting the physical co-occurrence of distortions under adverse illumination. As a consequence, the attributes carry largely redundant information; improvement in any one attribute is accompanied by proportional improvement in the others.

*Tab. 3. Pairwise correlation matrix for the **KonIQ-10k** dataset. Each cell: r_P / r_S*

	MOS	Brightness	Contrast	Colorfulness	Sharpness
MOS	–				
Brightness	0.35 / 0.33	–			
Contrast	0.24 / 0.23	0.32 / 0.32	–		
Colorfulness	−0.05 / −0.05	−0.04 / −0.04	−0.23 / −0.22	–	
Sharpness	0.64 / 0.64	0.13 / 0.13	0.37 / 0.37	−0.21 / −0.21	–

*Tab. 4. Pairwise correlation matrix for the **GISET** dataset. Each cell: r_P / r_S*

	Overall	Fragmentation	Unclearness
Overall	–		
Fragmentation	0.91 / 0.91	–	
Unclearness	0.87 / 0.82	0.62 / 0.61	–

*Tab. 5. Pairwise correlation matrix for the **ICAA20K** dataset. Each cell: r_P / r_S*

	MOS	Color Temp.	Colorfulness	Color Harmony
MOS	–			
Color Temp.	0.29 / 0.27	–		
Colorfulness	0.12 / 0.22	0.11 / 0.11	–	
Color Harmony	−0.32 / −0.33	0.02 / −0.00	0.14 / −0.03	–

*Tab. 6. Pairwise correlation matrix for the **AADB** dataset. Each cell: r_P / r_S*

	Score	Balanc.	Color H.	Content	DoF	Light	Blur	Object	Repet.	RoT	Symm.	Vivid
Score	–											
Balanc.	0.34 / 0.35	–										
Color H.	0.57 / 0.57	0.23 / 0.23	–									
Content	0.69 / 0.70	0.28 / 0.27	0.35 / 0.35	–								
DoF	0.45 / 0.43	0.07 / 0.09	0.15 / 0.17	0.24 / 0.24	–							
Light	0.58 / 0.58	0.19 / 0.20	0.40 / 0.39	0.39 / 0.38	0.20 / 0.21	–						
Blur	0.22 / 0.21	0.06 / 0.06	0.10 / 0.11	0.12 / 0.12	0.15 / 0.15	0.11 / 0.11	–					
Object	0.55 / 0.56	0.13 / 0.13	0.19 / 0.19	0.42 / 0.41	0.40 / 0.38	0.24 / 0.23	0.09 / 0.09	–				
Repet.	0.10 / 0.08	0.04 / 0.05	0.05 / 0.06	0.05 / 0.03	0.05 / – 0.05	0.06 / 0.05	0.02 / 0.02	– 0.07 / −0.09	–			
RoT	0.36 / 0.36	0.18 / 0.18	0.20 / 0.20	0.23 / 0.23	0.13 / 0.14	0.15 / 0.15	0.08 / 0.08	0.17 / 0.16	0.02 / 0.01	–		
Symm.	0.13 / 0.12	0.07 / 0.07	0.07 / 0.07	0.09 / 0.08	– 0.03 / – 0.02	0.06 / 0.05	0.01 / 0.01	0.06 / 0.04	0.23 / 0.20	0.00 / 0.01	–	
Vivid	0.53 / 0.53	0.07 / 0.07	0.32 / 0.33	0.32 / 0.32	0.21 / 0.21	0.38 / 0.38	0.08 / 0.08	0.24 / 0.24	0.03 / 0.03	0.11 / 0.11	0.01 / 0.02	–

*Tab. 7. Pairwise correlation matrix for the **EVA** dataset. Each cell: r_P / r_S*

	Score	Light & Color	Comp. & Depth	Quality	Semantic
Score	–				
Light & Color	0.85 / 0.85	–			
Comp. & Depth	0.90 / 0.89	0.73 / 0.72	–		
Quality	0.77 / 0.76	0.76 / 0.76	0.73 / 0.71	–	
Semantic	0.88 / 0.87	0.67 / 0.67	0.82 / 0.81	0.59 / 0.59	–

KonIQ-10k: Attribute–MOS correlations are substantially weaker and more heterogeneous. *Colorfulness* is virtually uncorrelated with MOS (≈ -0.05), while *Sharpness* retains a moderate positive association (0.64). *Colorfulness* shows negative inter-attribute correlations with *Contrast* (-0.23) and *Sharpness* (-0.21), indicating genuine orthogonality. This pattern is directly attributable to the algorithmic origin of the attribute scores.

GISET: *Fragmentation* is the dominant attribute, correlating with the overall score at 0.91, higher than *Unclearness* (0.87). The two degradation attributes exhibit only moderate inter-correlation (0.62 / 0.61), confirming that compression artifacts and blur are perceived as perceptually distinct phenomena in game-rendered content.

*Tab. 8. Pairwise correlation matrix for the **PARA** dataset. Each cell: r_P / r_S*

	Score	Quality	Comp.	Color	DoF	Light	Content	Obj. Emp.
Score	–							
Quality	0.99 / 0.98	–						
Comp.	0.95 / 0.93	0.93 / 0.91	–					
Color	0.94 / 0.93	0.92 / 0.92	0.89 / 0.86	–				
DoF	0.95 / 0.93	0.94 / 0.92	0.93 / 0.91	0.89 / 0.86	–			
Light	0.95 / 0.93	0.95 / 0.93	0.89 / 0.85	0.92 / 0.91	0.91 / 0.88	–		
Content	0.96 / 0.94	0.94 / 0.92	0.93 / 0.90	0.91 / 0.89	0.93 / 0.90	0.91 / 0.87	–	
Obj. Emp.	– 0.69 / −0.5	– 0.69 / −0.58	– 0.70 / −0.60	– 0.63 / −0.54	– 0.73 / −0.67	– 0.64 / −0.54	– 0.68 / −0.58	–

ICAA20K: All pairwise correlations among the three color attributes are low (< 0.3), and *Color Harmony* shows a weakly negative association with *Colorfulness* (-0.32 / -0.33). While orthogonality is an asset for multi-attribute modeling, the correlations between individual color attributes and MOS are also very low, indicating that color descriptors alone are insufficient to account for holistic aesthetic quality.

AADB: The eleven aesthetic attributes exhibit marked heterogeneity. Semantically rich attributes—*Content* (0.69) and *Light* (0.58), are the strongest predictors of the aesthetic score, whereas *Motion Blur*, *Symmetry*, and *Repetition* show negligible correlation (< 0.15). This supports the view that in fine-art photography, perceived quality is governed primarily by compositional and semantic coherence rather than technical correctness.

EVA: All four components correlate strongly with the overall aesthetic score (0.77–0.90), with *Composition & Depth* being the highest (0.90). Strong inter-attribute correlations indicate holistic evaluation: a high rating on any one dimension is associated with high ratings on the others. The close agreement between r_P and r_S confirms predominantly linear relationships.

PARA: Six of the seven attributes correlate with the aesthetic score at 0.93–0.99, the highest inter-attribute cohesion across all datasets. The sole exception is *Object Emphasis*, which shows a consistent negative correlation (< -0.59), suggesting that within personalized aesthetic evaluation, excessive subject foregrounding is perceived as a compositional violation.

3.2. Principal Component Analysis

To characterize the intrinsic dimensionality of each attribute space, we applied PCA exclusively to the predictor variables, excluding MOS from the analysis. This isolates the internal covariance structure of the attributes from their relationship to subjective quality. All features were Z-score standardized prior to PCA. To also assess how well each principal component captures subjective quality, we additionally compute the Pearson (r_P) and Spearman (r_S) correlations between each component's projection scores and MOS. The results are summarized in Table 0.

Since PCA operates without supervision, the resulting components have no inherent mathematical obligation to correlate with MOS. The algorithm targets directions of maximum attribute variance, not directions of maximum perceptual relevance. Observing that PC_1 aligns closely with MOS in some datasets is therefore not a property of the method but a property of the data: it reflects either the halo effect in human annotations, where ratings on all attributes move together with the global quality impression, or a genuine physical co-occurrence of distortions in the image content, or both. Conversely, datasets where principal components show weak or no correlation with MOS represent the mathematically standard case, in which the axes of attribute variance are approximately orthogonal to the axis of subjective quality. For components beyond PC_1, the interpretation also differs by dataset type. In datasets with human-annotated attributes, higher-order components tend to capture real physical variation in the images that is simply ignored by raters when forming a quality judgment, so their near-zero MOS correlations confirm the effective one-dimensionality of human perception in those settings. In the algorithmically annotated KonIQ-10k, by contrast, the absence of a halo effect means that perceptual information is distributed across the component space rather than concentrated in PC_1, making the MOS correlations of higher-order components informative rather than negligible.

A useful additional lens is the joint reading of PC_1 explained variance and its MOS correlation, as these two quantities together constrain what can have happened during annotation. When PC_1 variance is high and its MOS correlation is also high, as in SPAQ (85 %, $r_P = 0.933$), EVA (79 %, $r_P = 0.957$), and PARA (88 %, $r_P = 0.984$), the attribute space has collapsed into a single axis that also tracks quality well. This is consistent with the halo effect, but also with genuine physical co-variance of distortions; the two explanations cannot be separated from the PCA output alone. When PC_1 variance is low but its MOS correlation is high, as in AADB (25 %, $r_P = 0.893$), the picture is different: attribute dimensions vary independently of each other, yet their linear combination defined by the PC_1 loadings correlates strongly with MOS. This means that all eleven attributes contribute to the overall aesthetic judgment in a weighted manner, and it is this specific combination, rather than any single attribute, that raters respond to. This pattern is harder to attribute to the halo effect alone, since that would typically produce a high-variance PC_1; it suggests instead that raters integrate multiple genuinely independent attribute dimensions into a single quality judgment through a weighting that PCA recovers as PC_1. When PC_1 variance is moderate and its MOS correlation is near zero or negative, as in ICAA20K (44 %, $r_P = -0.157$),

the dominant axis of attribute variation is simply unrelated to perceived quality, which argues against both the halo effect and physical co-variance as explanations for the dataset structure.

Tab. 9.PCA results per dataset and principal component. Var: individual explained variance (%). ΣVar: cumulative explained variance (%). r_P and r_S: Pearson and Spearman correlations of the component's projection scores with MOS. Color coding for r_P/r_S: $|r| > 0.7$; $|r| \in [0.3,0.7]$. The bar in the Var % column shows the proportion of total attribute variance explained by each component

Dataset	PC	Var, %	ΣVar, %	r_P	r_S
SPAQ	1	85.14	85.14	+0.933	+0.938
	2	8.88	94.03	+0.245	+0.287
	3	2.49	96.51	−0.149	−0.144
	4	2.24	98.75	−0.015	−0.027
	5	1.25	100.00	−0.029	−0.041
KonIQ-10k	1	42.02	42.02	+ 0.504	+ 0.502
	2	24.47	66.49	+ 0.105	+ 0.078
	3	19.40	85.89	+ 0.334	+ 0.336
	4	14.11	100.00	– 0.354	– 0.347
GISET	1	81.17	81.17	+ 0.988	+ 0.984
	2	18.83	100.00	– 0.054	– 0.065
ICAA20K	1	44.16	44.16	– 0.157	– 0.141
	2	35.99	80.15	+ 0.193	+ 0.186
	3	19.85	100.00	+ 0.192	+ 0.161
AADB	1	25.21	25.21	+ 0.893	+ 0.913
	2	11.81	37.02	+ 0.025	– 0.033
	3	9.44	46.46	+ 0.034	– 0.004
	4	9.25	55.71	+ 0.014	+ 0.024
	5	8.64	64.36	– 0.027	– 0.155
	6	7.55	71.90	+ 0.026	+ 0.021
	7	6.90	78.80	+ 0.012	+ 0.035
	8	5.96	84.76	– 0.018	– 0.039
	9	5.58	90.34	– 0.032	– 0.033
	10	5.27	95.61	– 0.029	– 0.045
	11	4.39	100.00	+ 0.026	+ 0.027
EVA	1	78.84	78.84	+ 0.957	+ 0.957
	2	11.39	90.24	+ 0.104	+ 0.109
	3	6.10	96.34	+ 0.057	+ 0.100
	4	3.66	100.00	+ 0.001	+ 0.028
PARA	1	87.64	87.64	+ 0.984	+ 0.978
	2	6.89	94.53	+ 0.086	+ 0.137
	3	1.91	96.44	+ 0.013	– 0.006
	4	1.19	97.64	– 0.031	+ 0.019
	5	0.93	98.57	+ 0.017	– 0.049
	6	0.85	99.42	– 0.008	+ 0.034
	7	0.58	100.00	+ 0.091	– 0.044

SPAQ: PC_1 accounts for 85 % of attribute variance and its projection scores correlate with MOS at $r_P = 0.933$, $r_S = 0.938$. Because PC_1 is a weighted linear combination of all five attributes, this means that a specific composite of brightness, noise, sharpness, contrast, and colorfulness tracks perceived quality closely. The remaining four components, which are by construction orthogonal to PC_1 and capture the residual variation in attribute space, show no meaningful association with MOS (all $|r| < 0.25$). In other words, whatever variation exists in the attributes beyond that single composite direction is perceptually irrelevant. The physical reason is that under adverse illumination these distortions co-occur and co-vary, so PCA finds a single axis along which all attributes move together, and that axis happens to align with the perceptual quality axis as well.

KonIQ-10k: Variance is distributed across all four components (42, 24, 19, 14 %), reflecting the greater statistical independence of the algorithmically derived attributes. The MOS correlation structure is also qualitatively different from the human-annotated datasets: PC_1 achieves $r_P = 0.504$, while PC_3 ($r_P = 0.334$) and PC_4 ($r_P = -0.354$) carry signals of comparable strength but opposite sign. This means that the weighted attribute combination defining PC_1 (the one capturing most dataset variance) moderately predicts quality, but two other orthogonal combinations, PC_3 and PC_4, also carry independent perceptual signals with opposite directions. PC_1 likely reflects the type of image variation most prevalent in the dataset, for instance global contrast or brightness level, rather than the variation humans find most salient. The combinations captured by PC_3 and PC_4 may correspond to distortions with qualitatively different perceptual effects: one group of attributes combines in a direction observers find tolerable or even pleasing, while another combines in a direction they find objectionable. This

separation of opposing perceptual signals across multiple components is characteristic of an attribute space without annotation-induced bias.

GISET: PC_1 explains $81\ \%$ of variance and its scores correlate with MOS at $r_P = 0.988$, $r_S = 0.984$, the highest in the table. PC_1 here is a weighted combination of both *Fragmentation* and *Unclearness* that reflects their common contribution to perceived degradation. PC_2, which captures the orthogonal contrast between the two attributes (images where one is high and the other low), accounts for $19\ \%$ of attribute variance but is essentially uncorrelated with MOS ($r_P = -0.054$). This means that the relative balance between fragmentation and unclearness in a given image does not affect the quality rating: observers integrate both degradations into a single judgment rather than weighting them differently depending on which is dominant.

ICAA20K: Variance is distributed across three components ($44,\ 36,\ 20\ \%$), consistent with the near-orthogonality of the color dimensions. Each component represents a different weighted combination of color temperature, colorfulness, and harmony, and none of these combinations correlates meaningfully with MOS. The dominant combination PC_1 (the one that captures most color variation in the dataset) even shows a slightly negative association ($r_P = -0.157$, $r_S = -0.141$), suggesting that the direction along which color attributes vary most in this dataset corresponds to a dimension that raters mildly penalise aesthetically, possibly excess saturation or warmth. The remaining components, capturing orthogonal color combinations, also fail to predict MOS. Taken together, no linear combination of the three color attributes recovers a strong quality signal, which points to a fundamental mismatch between the color attribute space and the perceptual space of aesthetic quality.

AADB: PC_1 explains only $25\ \%$ of attribute variance, the lowest among all datasets, yet its scores correlate strongly with MOS ($r_P = 0.893$, $r_S = 0.913$). PC_1 here is a weighted combination of all eleven attributes that reflects the shared direction along which the attributes tend to co-vary, however weakly. The fact that this combination predicts MOS well, despite each attribute varying largely independently, suggests that raters do respond to a composite of all eleven dimensions rather than to any single one; the weights of that composite are encoded in the PC_1 loadings. The remaining ten components, which are orthogonal to PC_1 and together explain $75\ \%$ of attribute variance, show $|r| < 0.16$ with MOS. These components capture directions along which attributes vary in ways that are orthogonal to the perceptually salient combination; for linear models they carry no predictive signal and receive near-zero regression weights. Their potential utility lies in nonlinear settings where interactions between attributes may matter, or in diagnostic tasks where the goal is characterising specific image properties rather than predicting the overall score.

EVA and PARA: Both datasets show PC_1 dominating ($79\ \%$ and $88\ \%$ of variance respectively) with strong MOS correlations ($r_P = 0.957$ for EVA; $r_P = 0.984$ for PARA). A high PC_1 variance means that all attributes in each dataset move nearly in unison: their loadings on the first component are all large and of the same sign, so PC_1 is essentially a nearly equal-weighted average of all attributes. The most likely cause is the halo effect: when raters score multiple semantically diverse attributes in a single session, their global impression of image quality drives all sub-scores simultaneously, producing near-uniform loadings on PC_1 and leaving little independent variation for the higher-order components. Those higher-order components capture the residual physical differences between images that raters did not consciously register, and their near-zero MOS correlations confirm this. The same annotation-induced structure would also arise if the images genuinely varied along a single quality axis, so the two explanations remain confounded. For PARA, PC_7, with the smallest explained variance ($0.58\ \%$) and a slightly negative MOS correlation ($r_S = -0.044$), is the most plausible candidate for *Object Emphasis*: its loading on the dominant component was reported as negative, meaning this attribute diverges from the shared quality axis, and PCA isolates that divergence into a residual component.

3.3. Regression Performance

All regression models were evaluated using stratified 10-fold cross-validation. Reporting Mean $\pm$ Std across splits allows assessment of both predictive accuracy and robustness to training-set variation. Because the input features are low-dimensional scalar attributes rather than raw pixels, the risk of semantic leakage across splits is negligible; the cross-validated scores therefore represent a reliable upper bound on attribute-to-MOS predictability. Results are presented in Table 10.

One important caveat concerns the comparability of RMSE values across datasets. Because MOS is measured on different scales in each dataset (e.g., $[0{,}100]$ in SPAQ versus $[1{,}5]$ in KonIQ-10k and AADB), absolute RMSE values are not directly comparable between datasets and should be interpreted only within each dataset separately. The correlation metrics (PLCC, SRCC, KRCC) are scale-invariant and do not carry this limitation.

The goal of Table 0 is not to identify a definitively superior regression algorithm but to establish the *ceiling* of attribute-to-MOS predictability for each dataset. Across methods, performance differences are minor and not evaluated for statistical significance. The dominant signal is the magnitude of correlation achieved, which varies dramatically across datasets and reflects their underlying attribute structure.

Based on the cross-dataset analysis, we categorize the results into four distinct scenarios:

• **High Redundancy** (SPAQ, GISET, EVA, PARA): These datasets exhibit strong inter-attribute correlation, where attributes essentially measure a single latent factor. In such cases, the true necessity of multi-component model architectures may be overstated, as overall quality remains highly predictable (PLCC > 0.96). While this pattern heavily implies a halo effect [21], a cognitive bias where an observer's overall perception artificially aligns sub-attribute scores,

it could also stem from natural co-variances in mobile sensor captures (e.g., poor illumination natively generating high noise and low contrast simultaneously).

Ta. 10. Regression performance (Mean ± Std) over 10-fold cross-validation. Best results per dataset and metric are in ***bold***

Dataset	Method	PLCC ↑	SRCC ↑	KRCC ↑	RMSE ↓
SPAQ	LR	0.9769 ± 0.0011	0.9755 ± 0.0017	0.8644 ± 0.0041	4.4686 ± 0.0935
	SVR	0.9797 ± 0.0009	0.9784 ± 0.0016	0.8729 ± 0.0038	4.1886 ± 0.0667
	DT	0.9747 ± 0.0013	0.9733 ± 0.0019	0.8611 ± 0.0042	4.6754 ± 0.0858
	KNN	0.9787 ± 0.0009	0.9774 ± 0.0016	0.8698 ± 0.0038	4.3091 ± 0.0803
	MLP	0.9796 ± 0.0009	0.9783 ± 0.0016	0.8724 ± 0.0040	4.2266 ± 0.0807
KonIQ-10k	LR	0.7078 ± 0.0091	0.6949 ± 0.0131	0.5000 ± 0.0104	0.3902 ± 0.0113
	SVR	0.7670 ± 0.0049	0.7434 ± 0.0110	0.5462 ± 0.0094	0.3559 ± 0.0095
	DT	0.7427 ± 0.0061	0.7152 ± 0.0109	0.5321 ± 0.0093	0.3725 ± 0.0107
	KNN	0.7577 ± 0.0065	0.7353 ± 0.0109	0.5376 ± 0.0090	0.3612 ± 0.0101
	MLP	0.7639 ± 0.0052	0.7357 ± 0.0105	0.5385 ± 0.0092	0.3581 ± 0.0092
GISET	LR	0.9901 ± 0.0048	0.9790 ± 0.0087	0.9214 ± 0.0222	0.1799 ± 0.0200
	SVR	0.9919 ± 0.0045	0.9838 ± 0.0101	0.9358 ± 0.0280	0.1649 ± 0.0241
	DT	0.9821 ± 0.0101	0.9721 ± 0.0151	0.9096 ± 0.0356	0.2339 ± 0.0667
	KNN	0.9904 ± 0.0045	0.9786 ± 0.0130	0.9193 ± 0.0399	0.1717 ± 0.0237
	MLP	0.9880 ± 0.0084	0.9807 ± 0.0153	0.9278 ± 0.0381	0.3899 ± 0.0789
ICAA20K	LR	0.3140 ± 0.0170	0.2944 ± 0.0263	0.1989 ± 0.0180	0.6330 ± 0.0095
	SVR	0.3399 ± 0.0123	0.3128 ± 0.0235	0.2119 ± 0.0163	0.6271 ± 0.0089
	DT	0.3182 ± 0.0184	0.3007 ± 0.0251	0.2163 ± 0.0175	0.6324 ± 0.0088
	KNN	0.3265 ± 0.0135	0.2973 ± 0.0242	0.2011 ± 0.0167	0.6302 ± 0.0082
	MLP	0.3213 ± 0.0150	0.3022 ± 0.0240	0.2042 ± 0.0164	0.6318 ± 0.0089
AADB	LR	0.8968 ± 0.0079	0.9155 ± 0.0063	0.7705 ± 0.0095	0.0823 ± 0.0032
	SVR	0.9022 ± 0.0072	0.9173 ± 0.0056	0.7722 ± 0.0084	0.0803 ± 0.0031
	DT	0.8535 ± 0.0077	0.8622 ± 0.0084	0.6956 ± 0.0093	0.0972 ± 0.0021
	KNN	0.9058 ± 0.0069	0.9105 ± 0.0064	0.7626 ± 0.0093	0.0792 ± 0.0029
	MLP	0.8973 ± 0.0076	0.9140 ± 0.0059	0.7671 ± 0.0083	0.0824 ± 0.0032
EVA	LR	0.9645 ± 0.0038	0.9643 ± 0.0043	0.8383 ± 0.0093	0.2693 ± 0.0102
	SVR	0.9666 ± 0.0037	0.9656 ± 0.0046	0.8415 ± 0.0102	0.2613 ± 0.0104
	DT	0.9527 ± 0.0042	0.9533 ± 0.0048	0.8161 ± 0.0091	0.3103 ± 0.0124
	KNN	0.9639 ± 0.0037	0.9646 ± 0.0047	0.8396 ± 0.0101	0.2786 ± 0.0130
	MLP	0.9647 ± 0.0038	0.9642 ± 0.0043	0.8383 ± 0.0094	0.2692 ± 0.0098
5*PARA	LR	0.9930 ± 0.0002	0.9891 ± 0.0005	0.9153 ± 0.0016	0.0642 ± 0.0006
	SVR	0.9932 ± 0.0002	0.9891 ± 0.0004	0.9156 ± 0.0015	0.0632 ± 0.0007
	DT	0.9914 ± 0.0003	0.9864 ± 0.0006	0.9065 ± 0.0019	0.0712 ± 0.0008
	KNN	0.9918 ± 0.0002	0.9874 ± 0.0005	0.9085 ± 0.0016	0.0696 ± 0.0010
	MLP	0.9930 ± 0.0002	0.9891 ± 0.0005	0.9153 ± 0.0016	0.0643 ± 0.0009

• **Effective Orthogonality** (AADB): Despite weak correlations between aesthetic attributes (composition, lighting, color), models achieve high accuracy (PLCC ≈ 0.90). This validates the hypothesis that aesthetic quality is a composite of independent dimensions. Here preserving the full attribute set is critical for model interpretability and performance.

• **Extraction and Complexity Gap** (KonIQ-10k): Predictive performance drops significantly (PLCC ≈ 0.76) when sub-attributes are derived algorithmically rather than human-annotated. Because features like FFT-sharpness, RMS contrast, or Hasler-Suesstrunk colors capture low-level mathematical variances rather than human cognitive thresholds, regression models cannot perfectly align them with actual crowdsourced MOS. This highlights a clear gap, proving that basic statistical descriptors cannot fully simulate biological feedback loops in multi-component modeling.

• **Insufficient Specificity** (ICAA20K): The low performance on color attributes (PLCC ≈ 0.34) indicates that standard color metrics (temperature, harmony) fail to capture the nuances of human aesthetic judgment. While this could imply complex perceptual mechanisms, it more likely highlights a gap between color descriptors and subjective perception.

4. Discussion

Attribute quantity versus attribute quality. The results show that adding more annotated attributes to a dataset does not necessarily produce a richer representation of image quality. The high-redundancy pattern found in SPAQ, EVA, and PARA is consistent with two distinct explanations that cannot be separated on the basis of the current data alone.

The first is the halo effect [21]: raters who score multiple attributes in a single session tend to anchor their sub-scores on their overall impression of the image, inflating inter-attribute correlations beyond what the underlying perceptual dimensions would produce. A rater who judges an image as globally poor may assign a lower Depth of Field score even when the focal sharpness is not itself deficient. The second explanation is genuine perceptual co-variance: certain distortions naturally co-occur (poor illumination simultaneously degrades sharpness, increases noise, and alters color rendition), and compositional

strength tends to correlate with lighting quality in well-crafted photographs. Under this account, high inter-attribute correlations are a property of the image content rather than an artifact of the annotation procedure.

Both mechanisms predict the same correlation structure, so the available data do not allow us to distinguish between them. A controlled experiment would be needed: one group of raters scores all attributes together in the standard crowdsourcing format, while a second group scores each attribute in isolation, with no access to their ratings on the other dimensions. Substantially different inter-attribute correlations across the two conditions would implicate annotation bias; convergence would point to genuine perceptual co-variance. To our knowledge, no large-scale IQA benchmark has been designed with this kind of within-dataset comparison, and constructing one would be a useful contribution to the field.

Same attribute names, different structures. A related observation concerns the inconsistent behavior of semantically similar attributes across datasets. AADB, EVA, and PARA all contain dimensions related to color, composition, and lighting, yet their correlation structures differ substantially. In EVA, *Light & Color* and *Composition & Depth* are strongly inter-correlated with each other and with the overall score (0.73–0.90), suggesting that raters evaluated images holistically. In AADB, the corresponding attributes (*Color Harmony*, *Light*, *Balancing Elements*) show weak mutual correlations and contribute more independently to the aesthetic score. In PARA, *Color*, *Light*, and *Composition* are nearly collinear, with inter-attribute correlations exceeding 0.89. Content differences across the three corpora are unlikely to fully account for this variation; a more plausible factor is the annotation protocol itself: rater training, attribute presentation order, and whether raters are explicitly encouraged to evaluate dimensions independently all affect how well sub-scores reflect distinct perceptual judgments. This means that attributes with the same name cannot be treated as equivalent across datasets, and cross-dataset comparisons of model performance must account for structural differences in the underlying annotation.

Annotation protocol as a possible source of attribute independence in AADB. One plausible explanation for AADB's low PC1 variance (25%, Table 9) has to do with how the dataset was annotated. According to the dataset's original authors [9], AADB attributes were not collected as single continuous scores. Instead, workers tagged each of the eleven attributes independently as "positive," "negative," or "null," and the final attribute value is simply the average of five such tags. This scheme also isn't applied the same way across all attributes: for *Repetition* and *Symmetry*, negative values aren't allowed, since a degrading effect doesn't really make sense for these two, so workers could only mark "positive" or "null." This probably explains why repetition and symmetry have the lowest MOS correlations of the eleven attributes in Table 6 (0.10 and 0.13, respectively): with only two possible states instead of three, they simply carry less discriminative information than the other attributes. More broadly, this kind of tagging forces raters to make a separate call for each attribute in turn, rather than letting one holistic quality impression bleed across the whole form, as could more easily happen with a shared continuous rating scale. We think this feature of the protocol may limit the halo effect at the annotation stage, which would help explain both the low inter-attribute correlations in AADB (Table 0) and the broad spread of variance across its principal components (Table 9).

This stands in contrast to EVA and PARA, both of which show high PC1 variance and strong inter-attribute correlations. In EVA, the same participant rates all four attributes and the overall aesthetic score in one sitting, on comparable numerical scales: the overall score on an 11-point scale (0–10) and the four sub-attributes on a 4-level Likert scale (very bad/bad/good/very good). PARA follows a similar pattern: image-oriented attributes are rated on a shared 1–5 scale, with all attributes for a given image collected through a single annotation interface. In both cases, one rater judges every attribute for an image in a single continuous pass, which plausibly makes it easier for a general impression of quality to spread across the different attribute fields than under AADB's separate, attribute-by-attribute tagging. The same idea has been raised independently for PARA by [22], who suggest that PARA's single-interface, multi-annotator averaging protocol may encourage consistent scoring across attributes and account for the elevated inter-attribute correlations they observe; for this reason, they use AADB rather than PARA as their main dataset for attribute-diversity probing. As with our own account of AADB, this remains a plausible, protocol-level explanation rather than a proven one: no controlled study has yet directly compared annotation outcomes across different interface designs on the same set of images.

Consequences for model interpretability. High collinearity among predictors also has a direct consequence for model interpretability: when attributes are strongly correlated, feature importance estimates become numerically unstable, and a model that appears to decompose quality into multiple components may in practice rely on a single composite signal. The AADB case illustrates the alternative; with attributes that vary independently, importance scores are stable and can be meaningfully linked to specific perceptual dimensions.

Semantic gap in algorithmic and partial attribute sets. The results for KonIQ-10k and ICAA20K reflect a different kind of problem. In KonIQ-10k, algorithmically extracted attributes are statistically orthogonal but poorly aligned with subjective ratings, because mathematical proxies such as FFT-based sharpness or RMS contrast measure signal properties rather than the perceptual thresholds that shape human quality judgments. In ICAA20K, color descriptors are both orthogonal and accurately computed, yet remain nearly uninformative about holistic aesthetic quality: color is only one factor among several, and how it is perceived depends heavily on the semantic and compositional context that the attribute set does not capture.

A limitation of the KonIQ-10k analysis. The explanation for the low predictive performance on KonIQ-10k rests on the assumption that algorithmically computed attributes diverge from how human raters would score the same dimensions.

This assumption is plausible but has not been verified directly. KonIQ-10k does not include human annotations of the individual attributes, so it is not possible to measure how closely the algorithmic scores track human perceptual judgments on the same images. Without this comparison, the claim that algorithmic extraction is responsible for the performance drop remains an inference rather than an established finding.

Two approaches could test this more directly. The first is to collect human attribute annotations for a subset of KonIQ-10k images and compare the resulting attribute–MOS correlations with those reported here; if the gap narrows substantially, the extraction methodology is confirmed as the limiting factor. The second does not require new annotation at all: datasets such as SPAQ share attribute dimensions with KonIQ-10k (brightness, sharpness, colorfulness, contrast), and both human-annotated scores and the corresponding algorithmic estimates could be computed for the same images. Comparing how well each type of score predicts MOS within the same dataset would provide a direct measure of how much the annotation source matters, independently of any differences in image content or MOS scale.

Conclusions

Our analysis highlights that developing datasets with orthogonal quality attributes is critical for meaningful multi-component IQA. When attributes are highly collinear, predicting them separately offers little advantage over a single holistic score, as they essentially measure the same latent factor of overall quality. In contrast, orthogonal attributes provide unique, non-redundant information about distinct perceptual dimensions (e.g., color harmony vs. composition vs. technical fidelity). This enables: [topsep=0px,itemsep= – 1px]

1. Interpretable modeling, where the contribution of each component can be assessed independently;

2. Targeted image improvement, by identifying which specific attribute requires adjustment (e.g., «enhance color balance without altering composition»);

3. Efficient feature selection, by avoiding redundancy and reducing dimensionality without information loss.

Therefore, we recommend that future IQA benchmarks prioritize the annotation of semantically distinct, statistically independent attributes that are validated for perceptual relevance. Simply ensuring orthogonality is insufficient if the attributes do not align with human judgment.

References

[1] Zhai G, Min X. Perceptual image quality assessment: A survey. Science China Information Sciences 2020; 63(11): 211301. DOI: 10.1007/s11432-019-2757-1.

[2] Kastryulin S, Zakirov J, Pezzotti N, Dylov DV. Image quality assessment for magnetic resonance imaging. IEEE Access 2023; 11: 14154-14168. DOI: 10.1109/ACCESS.2023.3243466.

[3] Jamil S. Review of image quality assessment methods for compressed images. Journal of Imaging 2024; 10(5): 113. DOI: 10.3390/jimaging10050113.

[4] Soydaner D, Wagemans J. Unveiling the factors of aesthetic preferences with explainable AI. British Journal of Psychology 2026; 117(2): 444-478. DOI: 10.1111/bjop.12707.

[5] Fang Y, Zhu H, Zeng Y, Ma K, Wang Z. Perceptual quality assessment of smartphone photography. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2020: 3674-3683. DOI: 10.1109/CVPR42600.2020.00373.

[6] Hosu V, Lin H, Sziranyi T, Saupe D. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. IEEE Transactions on Image Processing 2020; 29: 4041-4056. DOI: 10.1109/TIP.2020.2967829.

[7] Utke M, Zadtootaghaj S, Schmidt S, Bosse S, Möller S. NDNetGaming: Development of a no-reference deep CNN for gaming video quality prediction. Multimedia Tools and Applications 2022; 81: 3181-3203. DOI: 10.1007/s11042-020-09144-6.

[8] He S, Xiao Y, Ming A, Ma H. Prompt-guided image color aesthetics assessment: Models, datasets and benchmarks. Information Fusion 2025; 114: 102706. DOI: 10.1016/j.inffus.2024.102706.

[9] Kong S, Shen X, Lin Z, Mech R, Fowlkes C. Photo aesthetics ranking network with attributes and content adaptation. In: Leibe B, Matas J, Sebe N, Welling M, eds. Computer Vision – ECCV 2016. Lecture Notes in Computer Science. Vol 9912. Springer; 2016: 662-679. DOI: 10.1007/978-3-319-46448-0_40.

[10] Kang C, Valenzise G, Dufaux F. EVA: An explainable visual aesthetics dataset. In: Proceedings of the Joint Workshop on Aesthetic and Technical Quality Assessment of Multimedia and Media Analytics for Societal Trends. ACM; 2020: 5-13. DOI: 10.1145/3423268.3423590.

[11] Yang Y, Xu L, Li L, et al. Personalized image aesthetics assessment with rich attributes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022: 19829-19837. DOI: 10.1109/CVPR52688.2022.01924.

[12] Freedman DA. Statistical Models: Theory and Practice. 2nd ed. Cambridge: Cambridge University Press; 2009. DOI: 10.1017/CBO9780511815867.

[13] Cortes C, Vapnik V. Support-vector networks. Machine Learning 1995; 20(3): 273-297. DOI: 10.1007/BF00994018.

[14] Awad M, Khanna R. Support vector regression. In: Efficient Learning Machines. New York: Apress; 2015: 67-80. DOI: 10.1007/978-1-4302-5990-9_4.

[15] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. New York: Routledge; 2017. DOI: 10.1201/9781315139470.

[16] Goldberger J, Hinton GE, Roweis S, Salakhutdinov RR. Neighbourhood components analysis. In: Saul LK, Weiss Y, Bottou L, eds. Advances in Neural Information Processing Systems 17. Cambridge, MA: MIT Press; 2005: 513-520.

[17] Rosenblatt F. The perceptron: A probabilistic model for information storage and organization in the brain. Psychological Review 1958; 65(6): 386-408. DOI: 10.1037/H0042519.

[18] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 2011; 12: 2825-2830.
[19] Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2019: 2623-2631. DOI: 10.1145/3292500.3330701.
[20] Watanabe S. Tree-structured Parzen estimator: Understanding its algorithm components and their roles for better empirical performance. arXiv preprint arXiv:2304.11127; 2023.
[21] Thorndike EL. A constant error in psychological ratings. Journal of Applied Psychology 1920; 4(1): 25-29. DOI: 10.1037/H0071663.
[22] Ryu K, Yanaka H. What do vision-language models encode for personalized image aesthetics assessment? In: Findings of the Association for Computational Linguistics: ACL 2026. Association for Computational Linguistics; 2026: 34146-34167. DOI: 10.18653/v1/2026.findings-acl.1706.

Author's information

Alexander Mikhailovich Trykin, (b. 1998), graduated from Institute of Information Technology, Mathematics and Mechanics of Lobachevsky State University of Nizhny Novgorod. Works as a junior researcher in Lobachevsky State University of Nizhny Novgorod. Research interests: medical data segmentation, image processing, deep learning.
E-mail: *alexander.trykin@gmail.com* ORCID: 0000-0002-5731-9624

Vadim Evgenievich Turlapov, (b. 1949), graduated from Radio Physics faculty of N.I. Lobachevsky Gorky State University, specialty "Radiophysics". Works as a professor at the Lobachevsky State University of Nizhny Novgorod. Research interests: computational mathematics, computer graphics, image processing, geometric modeling, GIS, computer vision. E-mail: *vadim.turlapov@itmm.unn.ru* ORCID: 0000-0001-8484-0565

Received July 01, 2026. The final version – August 08, 2026.