IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 23 May 2025, accepted 20 June 2025, date of publication 25 June 2025, date of current version 2 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3583029

RESEARCH ARTICLE

Comparing Blind Image Quality Metrics for Reliable Image Assessment

CRISTIAN GEORGE FIERARU[1], MARIA BISERICĂ[1], IOANA CRISTINA PLAJER[1], AND MIHAI IVANOVICI[2], (Senior Member, IEEE)

[1]Department of Mathematics and Computer Science, Faculty of Mathematics and Computer Science, Transilvania University of Braşov, 500036 Braşov, Romania
[2]Department of Electronics and Computers, Faculty of Electrical Engineering and Computer Science, Transilvania University of Braşov, 500036 Braşov, Romania

Corresponding author: Ioana Cristina Plajer (ioana.plajer@unitbv.ro)

Funded by the European Union. The AI4AGRI Project titled "Romanian Excellence Center on Artificial Intelligence on Earth Observation Data for Agriculture" received funding through the European Union's Horizon Europe Research and Innovation Program under Grant Agreement no. 101079136.

ABSTRACT Reliable image quality assessment is essential not only in digital photography but also as a key metric for evaluating the performance of algorithms and models designed for image quality enhancement or generation. In recent years, a wide range of image quality assessment metrics, both traditional and learning-based, have been proposed, making it a challenge to select the appropriate method for a given task. This study presents a comparative analysis between five widely used traditional no-reference image quality assessment techniques and five machine learning-based approaches, evaluating their effectiveness in computing image quality scores. The evaluation is carried out comprehensively using a set of standard and advanced performance metrics. Furthermore, we analyze how characteristics of the training datasets, such as score distribution, influence model performance. The machine learning models considered vary significantly in architectural complexity, in terms of both the number of layers and parameters, and we investigate whether this variability has a considerable impact on prediction accuracy. The analysis also extends to non-photographic imagery, with a comparative evaluation of the methods on hyperspectral satellite image visualizations. For full transparency and reproducibility of the current study, all training parameters and hardware specifications are reported.

INDEX TERMS Blind image quality assessment, machine learning, Pearson correlation, hyperspectral image visualization quality assessment.

I. INTRODUCTION

Image quality assessment (IQA) focuses on evaluating various attributes of an image to determine its overall quality. This evaluation may target specific features such as noise levels, sharpness, brightness, contrast, color fidelity, or the presence of distortions. Frequently, a single aggregate score is used to represent the overall quality of the image. Accurate and consistent IQA is significant not only for applications reliant on high-quality visual data but also as a key benchmark for measuring the effectiveness of algorithms and techniques in image restoration [1], [2], [3]. IQA is of significant importance in various applications concerning the generation of digital images, such as image acquisition [4], image fusion [5], or synthesizing [6], [7]. The evaluation of image acquisition and enhancement methods is essential in various fields, including everyday photography, medical imaging [4], [8], spectral satellite visualization [9], and recently even for security tasks [10]. This evaluation enables the comparison of existing methods with new developments and allows for both quantitative and qualitative validation of different approaches.

IQA methods can be categorized as subjective and objective [11]. Subjective approaches are based on human assessment and are considered the most precise. However, they are resource-intensive, lack real-time capability, and

The associate editor coordinating the review of this manuscript and approving it for publication was Yun Zhang.

appear to be challenging in practical applications. Objective image quality analysis, on the other hand, uses computational algorithms to evaluate image quality without human intervention, based on different quality metrics [12], [13]. These methods extract image features such as noise, contrast, sharpness, colorfulness [14], and then predict perceived quality by simulating aspects of human visual perception or by advanced mathematical and statistical techniques.

A fundamental distinction exists between IQA methods, notably between those that require a reference image for comparison and those that operate without such a reference. Objective image quality evaluation can be categorized into three groups based on the availability and completeness of the reference image, namely Full-Reference (FR) [15], Reduced-Reference (RR) [16], [17] and No-Reference (NR) image quality assessment [18], [19].

Another distinction can be made between traditional algorithms and the more recent ones, based on neural networks and deep learning (DL) [20], [21]. In order to evaluate picture quality, traditional IQA techniques mostly rely on manually created features and mathematical models. To estimate a quality score, these methods frequently require extracting low-level features like texture, sharpness, and contrast, combining them according to predetermined guidelines or algorithms. The Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) [22] are a few examples of the most used traditional FR-quality measures. For example, SSIM uses three factors to calculate the similarity of two images: luminance, contrast, and structure. The result is a measure that closely matches human perception. A significant problem with these methods is that they require a reference image in order to calculate the quality score. They can be effectively used to evaluate, for example, the outcomes of machine learning (ML) models where the neural network's output is compared to a label. However, they are not suitable for assessing the quality of an image when no reference is available.

Therefore, a series of no-reference quality metrics have been developed in the past years, many of them based on learned statistical features. Feature-based IQA methods use a series of image features, which can either be hand-crafted or learned, in order to estimate the quality of an image. These features may capture various properties such as sharpness, blur, noise content, compression with the Joint Photographic Experts Group (JPEG) format, contrast, or other characteristics [23], [24]. Alternatively, they can be derived from statistical properties of undistorted natural images. The most significant metrics in this context are the blind image quality index (BIQI) [25], the distortion-identification-based image verity and integrity evaluation (DIIVINE) [26], the blind/reference-less image spatial quality evaluator (BRISQUE) [27] or the natural image quality evaluator (NIQE) [28], which are shortly described in Section II-B.

In recent years, ML techniques have demonstrated their ability to automatically identify complex patterns and relationships directly from data, contributing to their growing popularity in this area. ML models predict quality scores and extract hierarchical information from images in an end-to-end manner. Especially suited for such a task, convolutional neural networks (CNNs) show the ability to adaptively learn from large-scale image datasets, capturing nuances and fine details that may elude manually crafted feature-based techniques.

In this context, several research papers exploit this ability for IQA prediction, ranging from very basic CNNs, like in [29], where 32×32 image patches are input to a model consisting of one convolutional layer followed by two fully connected layers, to deeper architectures, like the 31-layer network in [30], which processes color images by first sampling 224×224 patches and subtracting the ImageNet mean image. These patch-based models typically estimate the overall image quality by averaging the scores predicted for individual patches.

Input image patches can be obtained by sampling the original image with a specific stride [30], or by selecting them by using an algorithm such as Grey-Level Co-occurrence Matrix (GLCM) [31], which measures texture complexity and thus aims to identify the most informative regions of the image. Furthermore, in [31], the authors combine a CNN-based feature extractor with two multi-layer perceptrons (MLPs), one for classification and one for a quality score prediction.

Several research papers propose multi-stage approaches for image quality prediction. For example, in [32] the authors propose a two-stage CNN with 8 convolutional layers, which first learns distortion-related features via objective error mapping and then refines predictions to align with human visual perception.

Multiple networks can be combined in more complex architectures, like the two stream model in [33] with separate image and gradient image subcomponents, which enable the extraction of different levels of information and simplify feature extraction. The scores of the input patches are then averaged to obtain the final score. In the context of complex networks, the Very Deep Convolutional Networks for Large-Scale Image Recognition (VGG) [34] and the Residual Network (ResNet) [35] serve as key components of the model, owing to their strong feature extraction capabilities. For example, in [36] and [37], the authors employ a pretrained ResNet50, which is subsequently fine-tuned for IQA and used as a feature extractor in combination with another network for quality score prediction. Conversely, studies such as [38] and [39] choose VGG16 for feature extraction.

In the last few years transformers have gained popularity, as they are powerful tools in all tasks traditionally solved by ML architectures. In the context of IQA we can mention the transformer based approaches of [40]. Moreover, [41] provides an excellent overview of recent transformer-based IQA architectures.

When evaluating the quality of images generated by a specific algorithm, researchers face the challenge of selecting the most suitable method among the many available to ensure a reliable validation of their algorithm or a consistent quality-based comparison of different outputs.

Given the wide range of approaches, particularly in the fields of ML and DL, factors beyond accuracy also come into play. The complexity of the solution, along with the hardware and software resources required to implement and deploy the model, are equally important considerations. Not every user or researcher has access to sufficient computational resources to effectively utilize complex models, such as transformer-based architectures.

In this study, our objective is to compare several blind image quality assessment (BIQA) methods to help select the most appropriate approach for various scenarios. We evaluated the performance of neural network-based methods compared to traditional ones, analyzing the impact of various datasets on the accuracy of quality estimation. Furthermore, we trained five different CNN-based models, under various conditions and with three different image datasets, and we present a discussion of the results obtained. Additionally, we provide comparative results of these methods when applied to non-photographic images, specifically the visualization of hyperspectral satellite imagery.

The selection of the compared ML models was done in the light of the previously mentioned trade-off between accuracy, complexity, and model size. As shown in Table 1, 5GG models have a comparatively large number of parameters, while ResNet and Inception models are significantly lighter and MobileNet is one of the smallest networks considered. The number of parameters reported in Table 1 is expressed in millions (M) and reflects the sizes of the standard Keras classification models. The accuracy results were obtained on the validation set of ImageNet [42] and illustrate that a larger model does not inherently guarantee better performance. The inference time values originate from [42] and are the average of 30 batches, calculated on a central processing unit (CPU) AMD EPYC Processor with indirect branch prediction barrier (IBPB) (92 core), respectively, a graphic processing unit (GPU) Tesla A100, with a batch size of 32 images.

From the models presented in Table 1, which comprises two VGG, two ResNet, one Inception [43], two Neural Architecture Search Network (NASNet) [44] and three EfficientNet architectures [45], all existing pretrained in the Python Keras library, we selected for training and testing one of the large models, VGG16, the smaller models ResNet50 and InceptionV3, the NASNetMobile, which is one of the smallest CNN pretrained models, and the EfficientNet version 2 small (EfficientNetV2S). Their efficiency and usability in IQA tasks were highlighted by different studies [36], [37], [38], [39]. In the second part of Table 1, several transformer models, also explored for IQA in [41], are presented, along with their number of parameters and accuracy as reported in [46] and [47]. The inference time per image was measured using an NVIDIA V100 GPU with 32GB of memory [47]. Transformer models were not included in our experiments, as our study focused on less complex, smaller-scale systems that are already available as pretrained models in common Python ML libraries, making them more accessible to a broader range of users.

TABLE 1. Machine learning models - size, top-1 accuracy and inference time per step. The data for the CNNs was obtained from [42], while the data for the transformers from [46] and [47].

Model	Params	Accuracy Top-1	Inference Time (CPU) (ms)	Inference Time (GPU) (ms)
VGG16	138.4M	71.3%	69.5	4.2
VGG19	143.7M	71.3%	84.8	4.4
ResNet50	25.6M	74.9%	58.2	4.6
ResNet101	44.7M	76.4%	89.6	5.2
InceptionV3	23.9M	77.9%	42.2	6.9
NASNetMobile	5.3M	74.4%	27.0	6.7
NASNetLarge	88.9M	82.5%	344.5	20.0
EfficientNetV2S	21.6M	83.9%	-	-
EfficientNetV2M	54.4M	85.3%	-	-
EfficientNetV2L	119.0M	85.7%	-	-
ViT-B/16	86M	77.9%	-	36.6
ViT-L/16	307M	76.5%	-	11.6
DeiT-S	22M	79.8%	-	1.06
DeiT-B	86M	81.8%	-	3.42

The main contributions of this work are:

- the comparison of different BIQA classical methods with five of the existing CNNs, adapted for the aim and selected, as to exhibit various characteristics of size, accuracy and inference time;
- an exploration of the connection between the performance of the different models and the statistical properties of the datasets, in the attempt to explain the results obtained by the different models;
- the reliable usage of these BIQA methods on assessing the quality of other images than photographic ones, specifically visualizations of hyperspectral images, to investigate both their robustness and effectiveness regardless of the nature of the images.

The paper is organized as follows. In Section II we present the training and testing datasets together with the selected BIQA methods, which are subsequently compared. The performance of the methods and ML models is then evaluated on the considered datasets, using a series of assessment metrics. The results are presented and discussed in Section III. In this section, we also discuss the impact of the datasets, as well as the usability of these BIQA methods on different kinds of image, specifically visual representation of hyperspectral images. Finally, in Section IV we draw some conclusions and indicate potential future work in the area of general and reliable IQA.

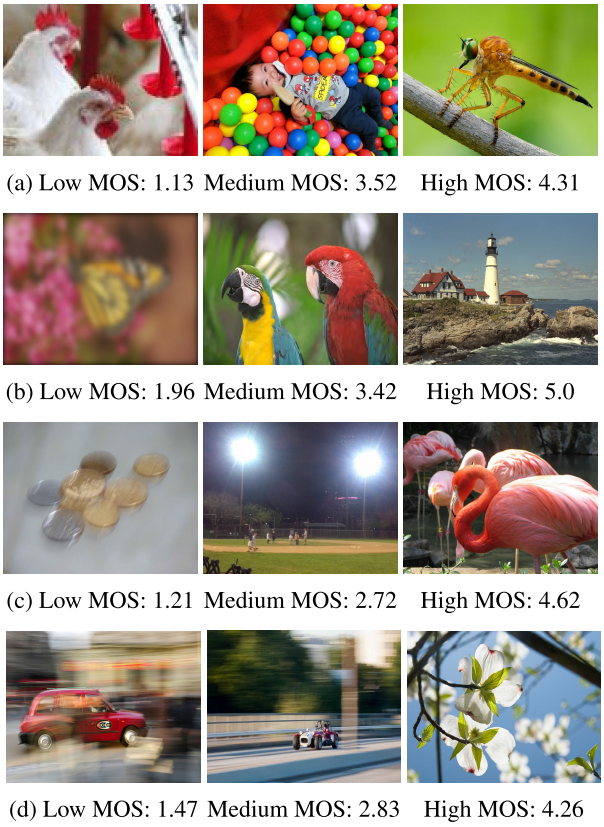
II. MATERIALS AND METHODS

The assessment of the different considered methods for image quality assessment was performed using several popular publicly available datasets. These datasets consist of images, each accompanied by a corresponding quality score reflecting subjective human judgments of quality. The selected methods were applied to estimate the quality of the images in these

datasets, and their performance was evaluated based on their ability to predict the assigned quality scores. The Pearson linear correlation coefficient (PLCC), the Spearman's rank correlation coefficient (SRCC), the Kendall rank correlation coefficient (KRCC), and the root mean square error (RMSE) were used as performance metrics. Additionally, the receiver operating characteristic (ROC) curve and the corresponding area under the curve (AUC) were employed to assess the effectiveness of the ML models.

The results of this assessment are discussed in detail in Section III, where we also analyze the influence of dataset construction on the accuracy of the results. Specifically, we explore how the subjective nature of human quality assessments and the variability of the dataset affect the overall reliability and accuracy of the models' predictions. Additionally, we investigate the generalization capability of the methods, which rely on learned features, across different datasets or in the context of images other than those obtained from natural photography.

A. DATASETS



(a) Low MOS: 1.13 Medium MOS: 3.52 High MOS: 4.31

(b) Low MOS: 1.96 Medium MOS: 3.42 High MOS: 5.0

(c) Low MOS: 1.21 Medium MOS: 2.72 High MOS: 4.62

(d) Low MOS: 1.47 Medium MOS: 2.83 High MOS: 4.26

FIGURE 1. Example images with low, medium, and high quality scores from each dataset: (a) KonIQ-10K, (b) LIVE2, (c) LIVE-itW, and (d) FLIVE.

IQA aims to automatically predict the perceptual quality of images as perceived by human observers. To train and evaluate IQA methods, large and diverse publicly available labeled datasets were used, such as KonIQ-10K, developed at the University of Konstanz [48]; two databases from the Laboratory for Image and Video Engineering (LIVE), namely, the LIVE Image Quality Assessment Database - Version 2 (LIVE2) [49] and the LIVE In the Wild Image Quality Challenge Database (LIVE-itW) [50], [51]; as well as the Flickr Labeled Image Verification for Esthetics (FLIVE) dataset [52]. These datasets offer a diverse collection of images with varying distortions, content, and quality levels, accompanied by Mean Opinion Scores (MOS) obtained from human ratings, which serve as labels. Building such datasets based on the subjective quality estimation based on human ratings is time-consuming and should meet different requirements as recommended by the International Telecommunication Union (ITU) [53], [54]. They enable researchers to validate, benchmark, and refine IQA models, driving progress in fields such as image enhancement, compression, and restoration. A selection of samples from these datasets, together with the associated MOS values, is presented in Figure 1.

KonIQ-10K is a large-scale IQA dataset comprising 10,073 images sourced from a wide variety of origins, showcasing diverse content and quality levels. Image quality scores were obtained through a crowdsourcing process in which multiple human raters assessed each image based on its perceptual quality. The MOS for each image was calculated as the average of the human ratings.

LIVE2 dataset contains 982 images derived from 29 original images. The original images were distorted using five different types of distortions [49]: compression with the JPEG standard, as well as the next generation JPEG compression JPEG2000, white noise, Gaussian blur, and fast fading. The quality scores were obtained through subjective experiments in which human participants rated the images, and the MOS was thus computed.

LIVE-itW is a dataset of natural images with real-world distortions. It contains 1,169 images captured using various mobile devices under different conditions. The images were rated by a large number of human subjects using an online crowd-sourcing platform. The MOS was calculated as the average of MOS values specified by the online users for each image to represent its perceptual quality.

FLIVE dataset contains a large number of images collected from Flickr. The images are labeled with aesthetic quality scores. The scores were obtained through a crowdsourcing platform where human raters provided their assessments. The MOS was calculated to reflect the aesthetic quality of each image.

The MOS distribution over these datasets is illustrated in Figure 2. The normalized histograms show the frequency of quality scores within each dataset, providing insights into their statistical characteristics. As can be observed in Figure 2 most of these datasets are unbalanced, the number of images with good scores being significantly larger than those with small scores. LIVE2 and FLIVE contain significantly more images with high scores, while the low scores are almost nonexistent, as can be observed in the normalized score histogram from Figure 2b and 2d. KonIQ-10K has most of the images with a medium score, while images with large

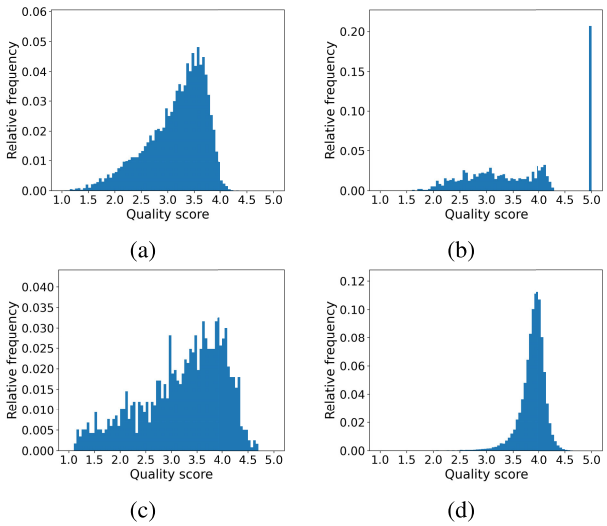


FIGURE 2. Histogram of score distributions for (a) KonIQ-10K, (b) LIVE2, (c) LIVE-itW, and (d) FLIVE datasets.

or small scores are underrepresented (Figure 2a). Although the best score distribution is present in the LIVE-itW dataset (Figure 2c), the number of samples present in this dataset is too small to perform a good generalization of the model. As a consequence, this dataset is not used in the training of our model but only for testing purposes.

These datasets were used to compare some of the most commonly used blind feature-based IQA methods alongside neural network-based approaches. Below we briefly describe the measures considered, which were tested and compared using the aforementioned image datasets.

B. FEATURE BASED QUALITY ASSESSMENT

In our study, we compared some of the most frequently used IQA methods as described below.

The BRISQUE evaluator [27] estimates image quality using natural scene statistics (NSS). It first computes mean-subtracted contrast-normalized (MSCN) coefficients and fits a generalized Gaussian distribution (GGD) to extract distortion-related parameters. In the second stage, spatial correlations among neighboring pixels are modeled, and a support vector machine (SVM) maps these features to a quality score.

After preprocessing through mean subtraction and divisive normalization, the NIQE estimator [28] extracts NSS features from 96×96 patches with sufficient sharpness, which are modeled using a multivariate Gaussian (MVG). The NIQE score is then computed as the distance between the MVG of the test image and the MVG of high-quality natural images.

BIQI evaluator [25] assesses image quality in two stages. It first uses NSS and wavelet decomposition to extract a GGD-based 18-D feature vector. Then, a SVM maps this vector to a quality score.

The BLIINDS-2 metric [55] assesses image quality using NSS features from blockwise discrete cosine transform (DCT) coefficients across scales. A Bayesian model trained on labeled data maps GGD-based features to a quality score.

Unlike the other methods, the DIIVINE [26] first identifies the type of distortion, then performs distortion-specific quality assessment. It uses wavelet-based statistical features, with SVM for classification and support vector regression (SVR) for predicting the final quality score.

C. ML-BASED QUALITY ASSESSMENT

Estimating a quality score within the range of $[1, 5]$ is a relatively straightforward regression problem. Therefore, we believe that classical ML models with a moderate number of parameters should suffice for this task. Although recent literature has proposed more advanced models, such as transformers, the results obtained with these models are comparable to those obtained by simpler networks [19]. In this study, we trained and compared five CNN architectures, inspired by the approach in [48]. The models share a similar structure, based respectively on **VGG16**, **ResNet50**, **InceptionV3**, **NASNetMobile**, and **EfficientNetV2S**. The selection of these models was carried out as described in Section I, taking into account various properties such as architecture, size, and inference time. Each model was pre-trained on the ImageNet dataset and subsequently modified by removing its classification head. We refer to these modified structures as backbones.

1) ARCHITECTURE DETAILS

For all models, a Global Average Pooling (GAP) layer is appended to the backbone, followed by three Fully Connected (FC) layers, each with Rectified Linear Unit (ReLU) activation, Batch Normalization, and Dropout (Figure 3). To maintain architectural consistency, normalization layers are omitted for the VGG16 backbone, which does not include them in its original design. A final linear FC layer outputs the predicted MOS. The weights of the top FC layers were initialized using the He method [56], which is a technique designed to maintain the variance of activations across layers. It sets weights using a random normal distribution scaled by $\sqrt{2/n}$, where n is the number of inputs to the neuron. This method is particularly effective for activation functions like the ReLU, helping to prevent vanishing or exploding gradients.

2) DATASET AUGMENTATION

To facilitate training while preserving the integrity of image distortions, as an essential factor for quality assessment, we applied minimal data augmentation techniques. Specifically, we used random cropping as the primary strategy to introduce variability in the training data, along with random horizontal flipping. This careful selection of augmentation methods was driven by the need to maintain the original distortions in the images. Geometric transformations such as rotations, scaling, or color modifications were avoided,

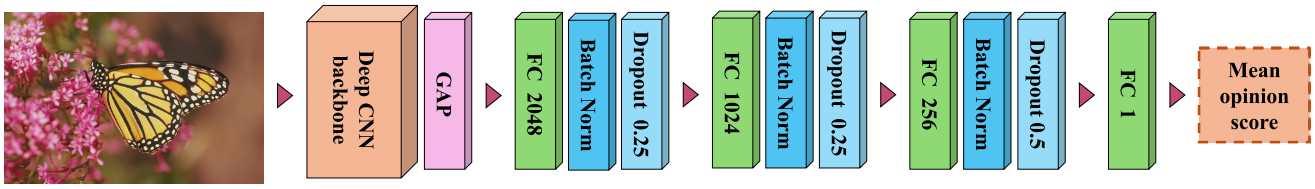


FIGURE 3. Network architecture for quality score prediction. It employs a deep ImageNet-pretrained CNN backbone with top layers removed. A regression head consisting of Fully Connected (FC), Batch Normalization and Dropout layers is appended to the backbone.

as they could introduce unintended changes that could interfere with the model's ability to accurately assess image quality.

3) TRAINING PROCESS

Prior to training, each dataset was shuffled and then divided into three subsets: 75% for training, 10% for validation, and 15% for testing. This partitioning was chosen to support effective model learning while enabling reliable validation and unbiased performance evaluation.

All our models were trained in the same manner, by using the MSE loss function, which significantly penalizes larger deviations from the ground truth. The optimization was handled by the adaptive moment estimator (Adam) optimizer [57]. Additionally, we employed a learning rate scheduler, combining both exponential decay and step decay mechanisms to fine-tune the learning process. This hybrid approach allowed for a gradual reduction in the learning rate while adjusting it based on specific epochs.

The training process itself was divided into two main stages, each characterized by different learning rate schedules and levels of model flexibility. In the first stage, we froze the backbone in order to focus on training the additional layers added on top. For this phase, the initial learning rate was set to 10^{-4}, which decayed gradually to 5×10^{-5} across 40 epochs. During this stage, we monitored the model's performance on the validation set using the PLCC as the primary evaluation metric. The model corresponding to the best PLCC score on the validation set was saved for further use. Additionally, early stopping was employed to prevent overfitting, with the patience set to 5 and the minimum delta set to 5×10^{-4}.

In the second stage of training, we unfroze the backbone network to allow fine-tuning of the entire model, including the backbone layers. For this phase, we loaded the best-performing model from the first stage and continued training for an additional 30 epochs, starting with a reduced learning rate of 10^{-5} and decaying it to 5×10^{-6} over time. Once again, we relied on the PLCC metric to monitor validation performance, saving the best-performing model based on this metric.

To better evaluate the impact of the datasets on model performance, we constructed a combined dataset consisting of 743 images from LIVE2 and 650 images from KonIQ-10K. A slightly modified training procedure was employed for this dataset, involving fine-tuning of models pre-trained on KonIQ-10K. The initial learning rate was set to 10^{-6} and progressively reduced to 5×10^{-7} across 30 epochs.

Figure 4 illustrates the loss decay curves from the best training runs on the KonIQ-10K dataset for each of the ML models evaluated. The corresponding performance metrics are discussed in detail in Section III. As shown, the training error converges around epoch 40 for VGG16 and ResNet50, around epoch 50 for InceptionV3, and around epoch 60 for both NASNetMobile and EfficientNetV2S. Among these, VGG16 and EfficientNetV2S exhibit the lowest training MSE (approximately 0.05), whereas ResNet50 records the highest MSE (approximately 0.1).

4) HARDWARE AND SOFTWARE CONFIGURATION

The training experiments were conducted on a workstation equipped with an Intel Core i7-8700 CPU with 32 GB of random access memory (RAM), and an NVIDIA GTX 1080 Ti GPU with 11 GB of video random access memory (VRAM). Due to memory limitations of the GPU, the batch size during training was fixed at 8. The models were developed and trained using TensorFlow 2.15.

5) REPRODUCIBILITY AND CODE AVAILABILITY

To support reproducibility, all relevant source code and scripts used for model training and evaluation are available on Zenodo: DOI 10.5281/zenodo.15491771. While no novel mathematical model was proposed, several modifications were made to existing estimators from the MATLAB Blind Image Quality Assessment Toolbox [58] to ensure compatibility with MATLAB R2022b version.

III. RESULTS AND DISCUSSIONS

In evaluating image quality assessment models, it is essential to measure how well model predictions align with subjective human judgments. We use as key performance metrics the PLCC [59], SRCC [60], KRCC [61], RMSE, and advanced statistical methods such as ROC analysis and AUC [62], [63], [64].

A. CONSIDERED ASSESSMENT METRICS

The PLCC for the set of pairs $\{(y_i, \hat{y}_i)|i = \overline{1,n}\}$, as given by:

$$\text{PLCC}(y, \hat{y}) = \frac{\sum_{i=1}^{n}(y_i - E[y])(\hat{y}_i - E[\hat{y}])}{\sqrt{\sum_{i=1}^{n}(y_i - E[y])^2}\sqrt{\sum_{i=1}^{n}(\hat{y}_i - E[\hat{y}])^2}} \quad (1)$$

measures the linear relationship the sets represented by the true labels $y = \{y_i|i = \overline{1,n}\}$ and the predicted labels

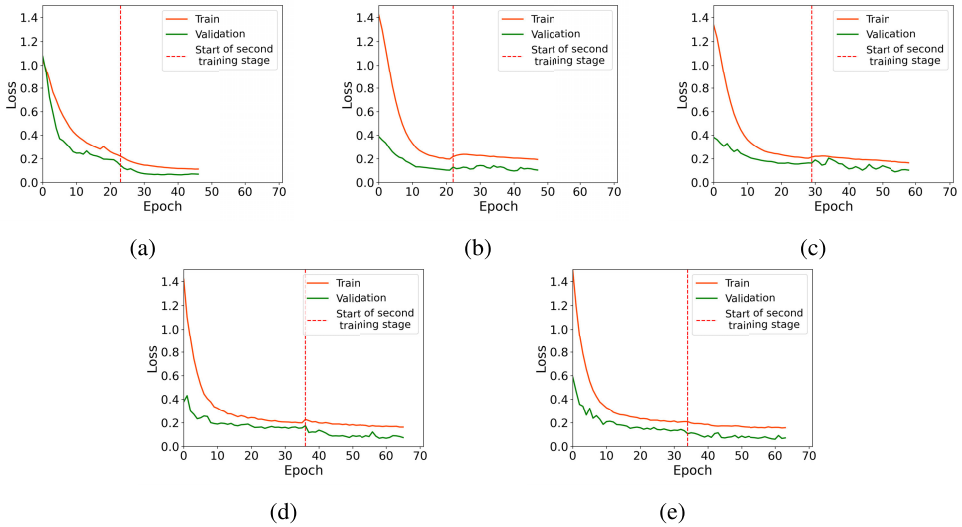


FIGURE 4. Training loss curves for all models on the KonIQ-10K training set: (a) VGG16, (b) ResNet50, (c) InceptionV3, (d) NASNetMobile, and (e) EfficientNetV2S.

$\hat{y} = \{\hat{y}_i | i = \overline{1, n}\}$, where $E[y]$ and $E[\hat{y}]$ represent the averages of these respective sets.

The SRCC is defined as:

$$\text{SRCC}(y, \hat{y}) = 1 - \frac{6 \sum_{i=1}^{n} diff_i^2}{n(n^2 - 1)} \tag{2}$$

where $diff_i$ is the rank difference between y_i and $\hat{y}_i$. The SRCC measures the strength and direction of the association between the ranks of y and $\hat{y}$.

The KRCC is a non-parametric statistic that measures the ordinal association between two variables. It is defined as:

$$\text{KRCC}(y, \hat{y}) = \frac{n_c - n_d}{\frac{1}{2}n(n-1)} \tag{3}$$

where n_c and n_d represent the number of concordant and discordant $(y_i, \hat{y}_i)$ pairs, respectively, and n is the total number of observations. Compared to SRCC, KRCC is generally more robust to small sample sizes and tied ranks.

All three correlation metrics, PLCC, SRCC, and KRCC, produce values in the range $[-1, 1]$, where -1 indicates a perfect negative correlation, 0 denotes no correlation, and 1 represents a perfect positive correlation.

The RMSE is a common regression metric used to evaluate the magnitude of prediction errors. It is defined as:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{4}$$

where y_i is the ground truth and $\hat{y}_i$ is the predicted value. Lower RMSE values indicate more accurate predictions, with 0 representing a perfect prediction.

The ROC analysis is used to evaluate the performance of a binary classifier. In our case, this method serves to analyze how well the considered ML models preserve the similarity and dissimilarity between image pairs in terms of their perceived overall quality, and whether they can correctly identify which image has higher quality for the dissimilar pairs. Following the approach in [65], we consider all possible pairs (i, j) of images from a dataset and classify them as either similar or dissimilar, based on a computed z-score:

$$z(i, j) = \frac{|y_i - y_j|}{\sqrt{\frac{var(i)}{N(i)} + \frac{var(j)}{N(j)}}} \tag{5}$$

where y_i and y_j are the ground-truth quality scores of images i and j, $var(i)$, $var(j)$ represent the variance of the subjective votes, and $N(i)$, $N(j)$ are the number of votes for each respective image.

The likelihood that two images differ is estimated using the cumulative distribution function (CDF) of the normal distribution. Pairs for which the resulting probability CDF(z) exceeds the chosen significance threshold $\alpha = 0.95$ are considered significantly different.

The image pairs from the two classes are then evaluated using the ML model, and the difference in their predicted scores is computed as:

$$\Delta_{model}(i, j) = \hat{y}_i - \hat{y}_j \tag{6}$$

Perceptual similarity between two images is assumed when $|\Delta_{\text{model}}(i, j)| < T$, with T being a predefined threshold.

Using Equations 5 and 6, we computed the ROC curve and the corresponding AUC for each ML model, and analyzed the results in Section III-C. The AUC values range from 0 to 1, with higher values indicating better model performance in preserving perceptual similarity relationships.

B. ASSESSMENT OF THE IQA METHODS

In order to perform the comparison of different methods for blind image quality assessment, we applied all the

TABLE 2. Training and inference time for the different ML models, expressed in hours (h), minutes (min) and seconds (s).

Method	LIVE2		KonIQ-10K		Combined set	
	Training	Inference	Training	Inference	Training	Inference
VGG16	16 min 58 s	0.215 s	4 h 24 min	0.255 s	32 min 28 s	0.251 s
ResNet50	15 min 22 s	0.255 s	3 h 34 min	0.281 s	30 min 56 s	0.294 s
InceptionV3	22 min 11 s	0.290 s	3 h 25 min	0.343 s	26 min 30 s	0.344 s
NASNetMobile	28 min 30 s	0.497 s	4 h 46 min	0.510 s	37 min 36 s	0.512 s
EfficientNetV2S	27 min 03 s	0.421 s	5 h 28 min	0.446 s	41 min 26 s	0.436 s

selected measures to estimate the quality of images in the chosen datasets and compared the results to the provided labels, by calculating the considered metrics, as discussed in Subsection III-A.

To compare the performance of the ML models, we conducted experiments using three specific input resolutions: 224×224, 320×320, and 384×512. Preprocessing was applied to most datasets to ensure compatibility with these resolutions, as follows.

For the LIVE2, LIVE-itW and FLIVE datasets, we adopted the approach described in [66], applying a white-fill transformation to ensure that each image had a height of at least 384 pixels and a width of at least 512 pixels, while maintaining the original aspect ratio. No pre-processing was applied to the KonIQ-10K dataset, as its images already matched the resolution of 384×512.

For training the ML models, the images were randomly cropped in order to fit the input in each case. For validation and testing, we adopted a more comprehensive sampling approach. To mitigate any potential bias due to random cropping and to account for the variability across different regions of an image, we extracted five sub-images from each test image: one from each corner and one from the center. The final predicted score for the entire image was obtained by averaging the predicted scores from these five regions.

For each of the models, we finally considered the input image size, for which the best results in most of the experiments were obtained, namely 224×224 when training on the LIVE2 dataset and 384×512 when training on KonIQ-10K and the combined dataset, respectively.

Table 2 reflects the analysis of computational efficiency and predictive performance of the ML models, and Table 3 provides a comparison of these models based on their architectural complexity, quantified by the number of parameters and layers.

TABLE 3. Model size of the different ML architectures.

Method	Number of parameters	Number of layers
VGG16	18.1M	29
ResNet50	30.1M	189
InceptionV3	28.3M	325
NASNetMobile	8.8M	783
EfficientNetV2S	25.3M	527

The results for the ML models are compared with other feature-based BIQA methods in Table 4. We evaluated the performance of the ML models when trained separately on LIVE2, KonIQ-10K, and the combined dataset comprising images from both LIVE2 and KonIQ-10K.

For some of the other BIQA methods, we also compared their performance across different training datasets, as detailed in the second column of Table 4. These values were calculated using a specific toolbox offered by [58]. Performance was evaluated by computing the SRCC, PLCC, KRCC, and RMSE between the predicted labels and the original labels for the images in each dataset.

As observed, the ML models achieved their best performance on the datasets they were specifically trained on. In Table 4, we highlighted all SRCC and PLCC values greater than 0.8 for LIVE2 and KonIQ-10K, indicating a high correlation. These measures are considered reliable for evaluating the images in the respective datasets. For LIVE-itW, we highlighted the top two results, which are at least 0.7, representing the best-performing measures for this dataset. The KRCC measure did not exceed 0.8 in any of the experiments; therefore, we highlighted values greater than 0.65. For RMSE, values smaller than 0.35 were highlighted. This enables us to easily identify, in Table 4, the models with the highest number of highlighted values, indicating their relatively superior performance.

For FLIVE, all methods failed to achieve satisfactory results and are not included in Table 4. Very low correlation values indicate that the measure is not adequate for the respective dataset.

C. IMPACT OF THE DATASET ON IQA

As can be observed from the results presented in the result tables, one of the main problems with the learning-based IQA methods is the dependency of the dataset. All the CNNs trained on the LIVE2 dataset perform poorly on the naturally distorted images from KonIQ-10K or LIVE-itW. When trained on KonIQ-10K, VGG16 and ResNet50 achieve moderate results on LIVE2, while InceptionV3 and EfficientNetV2S achieve relatively good results across all datasets. Moreover, we observe that training on the combined dataset, comprising images from both LIVE2 and KonIQ-10K, increases the generalization potential of all models. The BRISQUE descriptor from the well-known and frequently used OpenCV library, which is trained on LIVE2, performs very poorly in the natural context, as can be observed in Table 4. Some recent works like [21] and [66] train models on combined datasets. This certainly enhances the overall results on those datasets, as it can be observed in Table 4, which still does not guarantee the generalization on unknown data. Some of the methods, like NIQE, poorly estimate an overall score on naturally distorted images and are designed to recognize specific distortions, as indicated in [28].

TABLE 4. Results of the estimators.

Method	Train set	LIVE2				KonIQ-10K				LIVE-itW			
		SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE	SRCC	PLCC	KRCC	RMSE
VGG16	LIVE2	**0.914**	**0.893**	**0.744**	0.706	0.405	0.397	0.277	0.906	0.437	0.466	0.305	0.748
	KonIQ-10K	0.715	0.662	0.51	0.884	**0.884**	**0.909**	**0.705**	**0.241**	**0.787**	**0.808**	0.59	0.664
	Combined Set	**0.873**	**0.839**	**0.683**	0.735	**0.849**	**0.86**	**0.664**	**0.295**	**0.768**	**0.775**	0.569	0.677
ResNet50	LIVE2	**0.914**	**0.904**	**0.738**	0.522	0.487	0.503	0.334	0.744	0.325	0.36	0.218	0.812
	KonIQ-10K	0.662	0.627	0.511	1.136	**0.871**	**0.885**	**0.686**	**0.339**	**0.752**	**0.765**	0.555	0.921
	Combined Set	**0.876**	**0.899**	**0.719**	0.739	**0.853**	**0.864**	**0.662**	**0.32**	0.678	0.696	0.485	0.723
InceptionV3	LIVE2	**0.897**	**0.874**	**0.713**	0.923	0.388	0.417	0.262	0.623	0.263	0.331	0.177	0.804
	KonIQ-10K	**0.806**	**0.764**	0.594	1.092	**0.894**	**0.921**	**0.719**	**0.315**	**0.781**	**0.803**	0.585	0.903
	Combined Set	**0.935**	**0.912**	**0.779**	0.72	**0.862**	**0.879**	**0.673**	**0.288**	**0.727**	**0.754**	0.532	0.65
NASNet Mobile	LIVE2	**0.91**	**0.885**	**0.743**	1.065	0.45	0.484	0.308	0.568	0.401	0.432	0.271	0.829
	KonIQ-10K	0.787	0.754	0.583	0.98	**0.868**	**0.892**	**0.68**	**0.292**	**0.782**	**0.793**	0.583	0.766
	Combined Set	**0.901**	**0.887**	**0.72**	0.777	**0.859**	**0.885**	**0.67**	**0.286**	**0.75**	**0.769**	0.554	0.655
EfficientNet V2S	LIVE2	**0.852**	**0.822**	**0.665**	1.011	0.323	0.3	0.202	0.567	0.283	0.3	0.189	0.849
	KonIQ-10K	**0.818**	**0.793**	0.61	1.053	**0.892**	**0.919**	**0.715**	**0.26**	**0.826**	**0.836**	0.632	0.806
	Combined Set	**0.907**	**0.897**	**0.733**	0.786	**0.883**	**0.904**	**0.703**	**0.319**	**0.8**	**0.811**	0.602	0.757
BRISQUE	LIVE2	**0.903**	**0.89**	**0.73**	0.449	0.32	0.329	0.218	0.831	0.302	0.327	0.206	0.958
	KonIQ-10K	0.472	0.414	0.32	0.927	0.686	**0.711**	0.49	0.391	0.499	0.51	0.341	0.74
BRISQUE (OpenCV)	LIVE2	**0.937**	**0.908**	0.782	0.498	0.267	0.269	0.18	1.239	0.385	0.416	0.264	1.184
NIQE	LIVE	**0.829**	0.422	0.62	1.41	0.077	0.037	0.052	1.766	0.319	0.325	0.215	1.455
	LIVE2	0.542	0.53	0.377	0.89	0.016	0.011	0.01	1.68	0.054	0.06	0.037	1.507
	KonIQ-10K	**0.733**	0.452	0.528	1.316	-0.033	0.011	-0.021	1.646	0.124	0.108	0.083	1.52
BIQI	LIVE	**0.896**	**0.824**	**0.718**	0.631	0.336	0.33	0.227	0.871	0.288	0.328	0.195	0.944
BLIINDS-II	LIVE	**0.912**	**0.888**	**0.741**	0.52	0.023	0.045	0.015	1.457	0.15	0.196	0.101	1.147
DIIVINE	LIVE	**0.851**	**0.827**	**0.663**	0.567	0.471	0.439	0.326	1.107	0.471	0.459	0.32	1.013

Another problem of the considered datasets is the fact that most of them are unbalanced, as indicated in Section II-A, and the number of images with good scores is considerably larger than those with small scores. The best results in our experiments were obtained by training with KonIQ-10K and subsequently fine-tuning using a mixed dataset that combines LIVE2 and KonIQ-10K. This can be observed in the results of Table 4. The methods were also tested on the FLIVE dataset, but the results for the correlation were so small, under 0.5, that we did not include them in Table 4.

To further compare the performance of the models, we calculated the AUC, based on the ROC analysis, as indicated in Section III-A. The results for the ML models trained on KonIQ-10K and the combined set, are presented in Table 5 for the *Different versus Similar* analysis and in Table 6 for the *Better versus Worse* analysis. We did not consider the models trained on LIVE2, because they exhibit a low generalization ability, as can be observed from Table 4. The results confirm the observations from Table 4, with InceptionV3 and EfficientNetV2S achieving the best overall performance. The AUC values obtained on KonIQ-10K, which was also used for training, are generally slightly higher than those on LIVE-itW, with EfficientNetV2S demonstrating the best generalization performance on LIVE-itW.

In the *Better versus Worse* analysis, the AUC values are higher overall, indicating that the models generally succeed in selecting the images with better perceptual quality from the pairs.

In Figure 5 are presented the ROC curves for the models which obtained the best AUC values, respectively InceptionV3 for KonIQ-10K, and EfficientNetV2S for LIVE-itW. The horizontal axis indicates the False Positive Rate (FPR), while the vertical axis shows the True Positive Rate (TPR). In the *Different versus Similar* analysis, FPR denotes the probability of classifying a similar pair as different, while TPR indicates the probability of correctly identifying a different pair. In the *Better versus Worse* analysis, FPR corresponds to the rate at which the model selects the lower-quality image, and TPR reflects the rate of correctly selecting the higher-quality one.

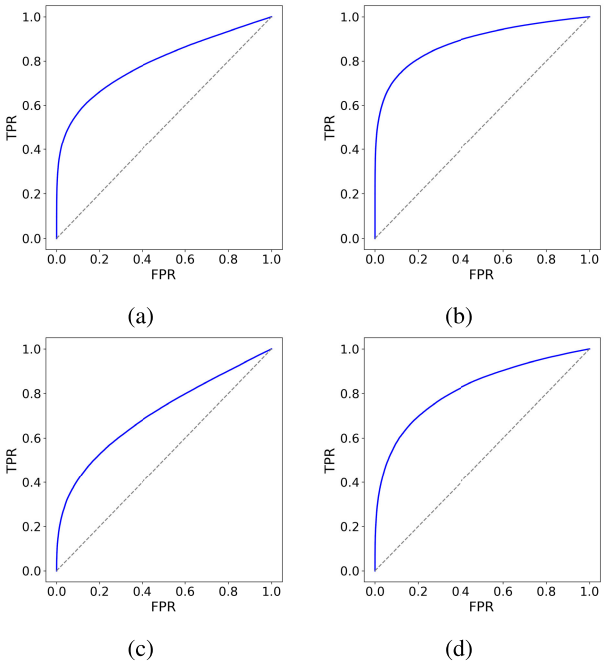


FIGURE 5. Examples of ROC curves: (a) Different versus similar and (b) Better versus worse for InceptionV3, and (c) Different versus similar and (d) Better versus worse for EfficientNetV2S.

TABLE 5. Different versus similar AUC.

Method	Train Set	KonIQ-10K	LIVE-itW
VGG16	KonIQ-10K	0.7774	0.6909
	Combined Set	0.7368	0.6633
ResNet50	KonIQ-10K	0.7594	0.6622
	Combined Set	0.7409	0.6165
InceptionV3	KonIQ-10K	**0.7913**	0.6844
	Combined Set	0.7454	0.6466
NASNetMobile	KonIQ-10K	0.7562	0.6773
	Combined Set	0.7474	0.652
EfficientNetV2S	KonIQ-10K	0.7878	**0.7081**
	Combined Set	0.7735	0.6843

TABLE 6. Better versus Worse AUC.

Method	Train Set	KonIQ-10K	LIVE-itW
VGG16	KonIQ-10K	0.8741	0.7982
	Combined Set	0.8401	0.7797
ResNet50	KonIQ-10K	0.8547	0.7759
	Combined Set	0.8422	0.7356
InceptionV3	KonIQ-10K	**0.8854**	0.7919
	Combined Set	0.8534	0.7572
NASNetMobile	KonIQ-10K	0.8626	0.7941
	Combined Set	0.8596	0.7711
EfficientNetV2S	KonIQ-10K	0.8837	**0.8188**
	Combined Set	0.8758	0.8017

To gain deeper insight into the impact of imbalanced training datasets on score prediction, we further analyze the differences between the true labels and the predicted scores across the various datasets used to evaluate the models trained on KonIQ-10K. The score difference for an image i is given by:

$$\Delta(i) = y_i - \hat{y}_i \tag{7}$$

where y_i is the true score and $\hat{y}_i$ is the predicted score for image i. These differences are illustrated in Figure 6 for InceptionV3 and EfficientNetV2S, which achieved the overall best results. For the other ML models tested, the results are very similar. When testing on KonIQ-10K, the negative values, indicating overestimation of the quality scores, and the positive values, indicating underestimation, are distributed relatively evenly throughout the MOS range. This is an expected outcome, as both the training and test samples come from the same distribution. However, when evaluating the KonIQ-10K trained models on the LIVE2 and LIVE-itW datasets, the distribution of the score differences changes significantly. We observe that, influenced by the distinct nature of distortions in LIVE2 and the imbalanced score distribution, the predictions exhibit a consistent pattern. All ML models tend to overestimate the scores for lower-quality images, an effect more pronounced for LIVE-itW, and to underestimate the scores for higher-quality images. This behavior is remarkably similar across all ML models. A likely explanation is that, in the KonIQ-10K dataset, the number of images with MOS scores below 2.0 is very limited, while images with scores close to 5 are virtually absent. Consequently, models have limited exposure to such cases during training and tend to overrate images with true scores below 2.0 and significantly underrate those with true scores above 4.0.

All labels in the datasets are derived from human quality assessments of photographic images depicting natural scenes. Consequently, quality evaluators trained on these datasets are designed to model human perception of quality for this specific type of imagery. This poses a challenge when seeking an effective quality measure for other types of images. For instance, in fields such as medical imaging, denoising, or image visualization - particularly with spectral images [9] - researchers often need quantitative tools to evaluate the performance of their enhancement methods. In such cases, IQA estimators trained on existing datasets fail to provide reliable instruments for assessment.

Such an example are medical images, which need specialized IQA tools, taking into account the specific features of such images [4]. Using classical tools does not provide reliable results, as can be observed in [67], where the results of the BRISQUE and the NIQUE estimators available contradict in various cases the enhancement results for angiogram quality enhancement. Another example is the quality estimation of the spectral image visualization algorithms of satellite images (Figure 7[1]) [9], where, again, the results as presented in Tables 7 and 8 do only partially correspond to the visually best results in Figure 7. In both tables, the arrows adjacent to the names of the IQA methods indicate whether the resulting value increases or decreases with the quality of the image. An upward arrow signifies that a higher score corresponds to better image quality, while a downward arrow denotes that a lower score indicates better image quality. The satellite images were acquired by the PRecursore IperSpettrale della Missione Applicativa (PRISMA) mission of the Italian Space Agency over Brasov county.

TABLE 7. Scores for the October 2022 PRISMA visualization results.

Method	Band Sel.	Decolorization	MPCNN	FCNN
BRISQUE ↓	15.463	17.004	**14.182**	14.982
NIQE ↓	2.577	**2.56**	2.884	2.659
BIQI ↓	50.297	55.981	**24.921**	24.981
BLIINDS ↓	29.5	**26.5**	27.5	30.5
DIIVINE ↓	24.445	15.757	14.88	**14.369**
VGG16 ↑	2.939	3.378	3.232	**3.439**
ResNet50 ↑	2.620	3.116	3.123	**3.17**
InceptionV3 ↑	2.727	**3.338**	3.268	3.337
NASNetMobile ↑	2.556	3.19	3.234	**3.549**
EfficientNetV2S ↑	2.796	**3.375**	3.106	3.233

TABLE 8. Scores for the March 2023 PRISMA visualization results.

Method	Band Sel.	Decolorization	MPCNN	FCNN
BRISQUE ↓	11.544	13.334	**11.213**	11.9303
NIQE ↓	2.521	2.973	2.617	**2.385**
BIQI ↓	**17.244**	29.239	28.727	26.786
BLIINDS ↓	24.5	24.5	**23.5**	24.5
DIIVINE ↓	15.286	20.302	14.955	**14.368**
VGG16 ↑	2.938	2.808	2.924	**3.201**
ResNet50 ↑	2.364	2.399	**2.932**	2.915
InceptionV3 ↑	2.709	2.899	3.013	**3.179**
NASNetMobile ↑	2.664	2.802	3.112	**3.167**
EfficientNetV2S ↑	2.9	2.909	2.912	**3.209**

[1]These images were published with the permission of the authors of [9].

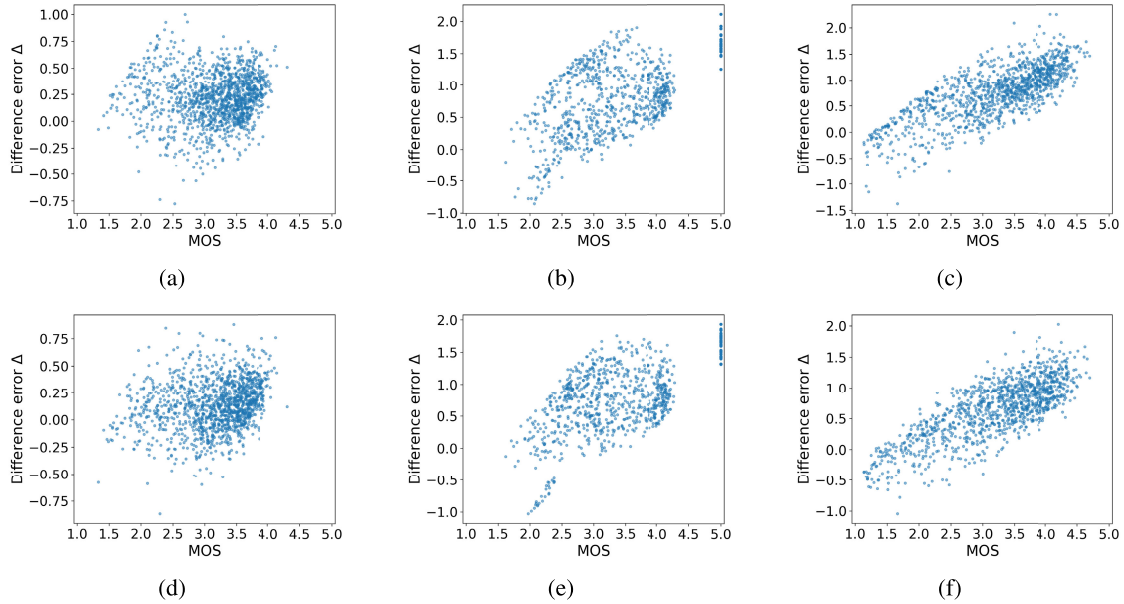


FIGURE 6. Score differences Δ between true and predicted MOS values. In the first row for InceptionV3 when tested on: (a) KonIQ-10K, (b) LIVE2, (c) LIVE-itW, and in the second row for EfficientNetV2S when tested on: (d) KonIQ-10K, (e) LIVE2, (f) LIVE-itW. Both models were trained on KonIQ-10K.

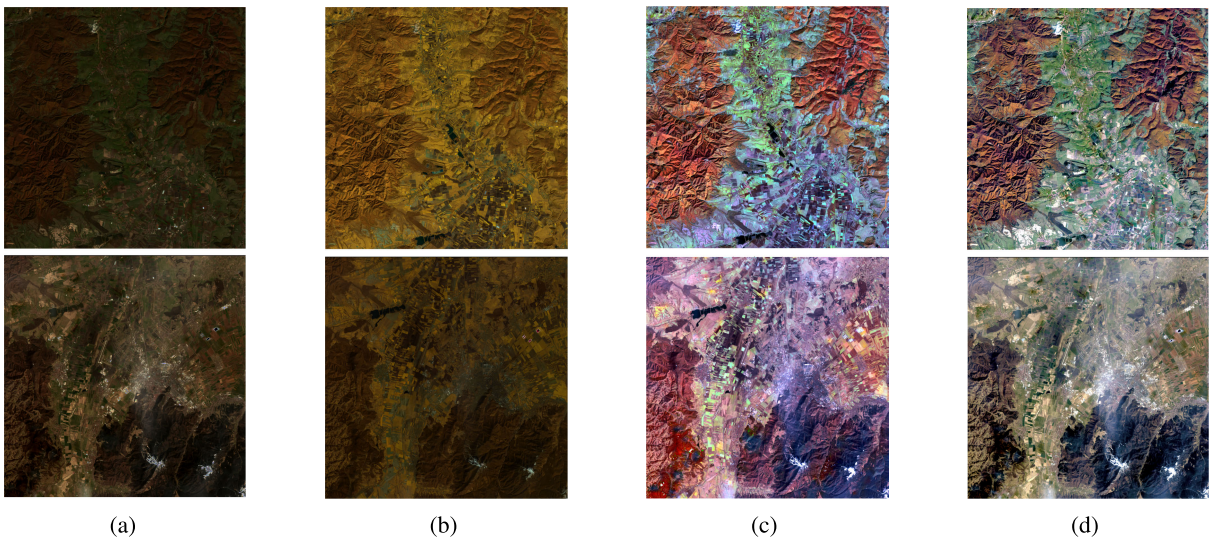


FIGURE 7. Visualization of two hyperspectral images acquired by the PRISMA satellite over Brasov county in October 2022 (first row), and in March 2023 (second row) using different methods: a) Band selection without further enhancement, b) decolorization-based HSI visualization [68], c) Multichannel Pulse-Coupled Neural Network-Based Hyperspectral Image Visualization (MPCNN) [69], d) Fully connected neural network [9].

In Table 7 we can observe that most of the ML-models and DIIVINE obtained the best scores for the images in Figure 7d top, which is the best visualization of the October PRISMA image. InceptionV3 obtains almost identical scores for the decolorization-based HSI and the neural-network visualizations, while EfficientNetV2S obtains a slightly higher score for the decolorization-based HSI visualization, than for the visually best image. BRISQUE and BIQI obtain the best scores for Figure 7c top, which exhibits larger contrast and more vivid colors, but has an unnatural look. On the other hand, NIQE and BLIINDS do not obtain reliable results for the October PRISMA image. The results in Table 8 show that NIQE, DIIVINE, and most of the ML models yield the best scores for the fully connected neural network (FCNN) visualization of the March PRISMA image. In contrast, BRISQUE, BLIINDS, and ResNet50 rate the MPCNN visualization as superior, with ResNet50 showing only minimal differences between the MPCNN and FCNN methods. Furthermore, as observed in both Table 7 and Table 8, the ML models tend to rank the PRISMA images in a manner consistent with human perceptual judgments.

In order to obtain a deeper analysis of the reliability and consistency of the score estimation of the ML models, we performed some statistical calculations. As explained in

TABLE 9. Mean and standard deviation for the patch scores of the PRISMA images.

Image	Viz. Meth.	VGG16		ResNet50		InceptionV3		NASNetMobile		EfficientNetV2S	
		mean	std	mean	std	mean	std	mean	std	mean	std
PRISMA Oct	Band selection	2.887	0.081	2.574	0.085	2.74	0.064	2.524	0.098	2.782	0.127
	Decolorization	3.335	0.092	3.122	0.052	**3.398**	0.101	3.283	0.122	**3.346**	0.073
	MPCNN	3.204	0.073	3.113	0.082	3.269	0.085	3.258	0.136	3.152	0.054
	FCNN	**3.488**	0.071	**3.199**	0.138	3.348	0.066	**3.49**	0.137	3.281	0.055
PRISMA Mar	Band selection	2.971	0.079	2.366	0.103	2.695	0.118	2.591	0.155	2.92	0.074
	Decolorization	2.778	0.09	2.326	0.13	2.946	0.178	2.793	0.152	2.915	0.061
	MPCNN	2.927	0.068	**2.872**	0.067	3.024	0.061	3.002	0.157	2.907	0.089
	FCNN	**3.219**	0.084	2.871	0.099	**3.258**	0.096	**3.201**	0.148	**3.24**	0.082

Section III-B, for larger inputs than the model input, as not to alter the quality of the image, no resizing was performed. The overall score of the input image was obtained, by averaging the scores of 5 patches extracted from this target image. For images in the considered datasets, this is not of much consequence for the overall score, as these images may have only slightly different sizes than the input of the model. In the case of the PRISMA visualization, the situation changes, as each image is of size 1000×1000 pixels, significantly larger than the input of the networks, in each of the training scenarios.

To analyze if an overall score can be reliably calculated for such images considering only 5 patches, we performed the following calculations. For each image, we considered all patches of the model input size, starting from the top left corner of the image, with a stride of 16 pixels, resulting in a number of 1209 patches. All patches were passed through the considered model and the average score, as well as the standard deviation (std) for these scores, were calculated. The results are presented in Table 9.

Comparing the scores of Table 9 with those of Tables 7 and 8 we can observe that the differences are not significant. Moreover, the ranking of the images is the same. We also can observe that the standard deviation values are all very small, indicating an insignificant variability of score values for the patches of the same image. These results confirm the consistency of overall score calculation, as well as the reliability of ML models for BIQA.

IV. CONCLUSION

Blind Image Quality Assessment is a critical task, not only for the rapid, automatic estimation of image quality but also as a tool for quantitatively evaluating the performance of various algorithms for image restoration, generation, or enhancement. To address the need for quantitative assessment, we reviewed and compared five of the most prominent blind quality estimators from recent literature. Additionally, we trained five CNN models, VGG16, ResNet50, InceptionV3, NASNetMobile, and EfficientNetV2S, using publicly available BIQA datasets.

Our experiments and analysis indicate that neural network-based estimators achieve the best performance, particularly on natural images and the datasets they were trained on. Despite variations in complexity, models such as NASNetMobile and EfficientNetV2S demonstrated competitive performance even on previously unseen data. InceptionV3 and EfficientNetV2S slightly outperformed the other models, while larger architectures like VGG16 did not yield significant accuracy improvements over more compact networks.

However, the effectiveness of these models is highly influenced by dataset characteristics, particularly label quality and class imbalance. Datasets with a limited number of extremely low- or high-quality images can lead to inaccurate score predictions due to the limited exposure of the models to such cases. On the other hand, classical estimators such as BRISQUE and NIQE struggle when applied beyond their original calibration domains, highlighting the limited generalization of many BIQA techniques.

In conclusion, while ML-based BIQA methods hold strong potential, no single estimator proves universally reliable. Accurate quality assessment still requires careful alignment between the estimator and the specific image domain, making the objective evaluation of image generation or enhancement algorithms an ongoing and complex challenge.

As a future direction, we propose further exploration of both model architectures and the construction of appropriate datasets. Transformer-based models may offer promising potential, along with the development of new BIQA metrics that reflect the specific properties of the target images. These metrics could also incorporate existing quality measures in an effective way.

ACKNOWLEDGMENT

Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them. The hyperspectral image from the PRISMA satellite presented in this article was kindly provided by the Italian Space Agency (ASI).

REFERENCES

[1] B. Hu, L. Li, J. Wu, and J. Qian, "Subjective and objective quality assessment for image restoration: A critical survey," *Signal Process., Image Commun.*, vol. 85, Jul. 2020, Art. no. 115839.

[2] Z. Wang, J. Zhuang, S. Ye, N. Xu, J. Xiao, and C. Peng, "Image restoration quality assessment based on regional differential information entropy," *Entropy*, vol. 25, no. 1, p. 144, Jan. 2023.

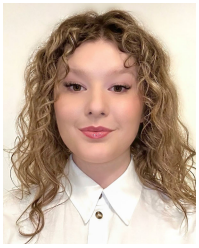
[3] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Comparison of full-reference image quality models for optimization of image processing systems," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1258–1281, Apr. 2021.

[4] S. Kastryulin, J. Zakirov, N. Pezzotti, and D. V. Dylov, "Image quality assessment for magnetic resonance imaging," *IEEE Access*, vol. 11, pp. 14154–14168, 2023.

[5] P. Jagalingam and A. V. Hegde, "A review of quality metrics for fused image," *Aquatic Proc.*, vol. 4, pp. 133–142, Jan. 2015.
[6] W. Yu, X. Zhang, Y. Zhang, Z. Zhang, and J. Zhou, "Blind image quality assessment for a single image from text-to-image synthesis," *IEEE Access*, vol. 9, pp. 94656–94667, 2021.
[7] X. Sui, M. Ding, J. Yan, Y. Fang, Y. Zuo, and Z. Tan, "Objective quality assessment of synthesized images by local variation measurement," *Signal Process., Image Commun.*, vol. 92, Mar. 2021, Art. no. 116096.
[8] X.-L. Zhang, Z.-F. Liu, Y. Kou, J.-B. Dai, and Z.-M. Cheng, "Quality assessment of image fusion based on image content and structural similarity," in *Proc. 2nd Int. Conf. Inf. Eng. Comput. Sci.*, Dec. 2010, pp. 1–4.
[9] I. C. Plajer, A. Băicoianu, L. Majercsik, and M. Ivanovici, "Multisource remote sensing data visualization using machine learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, vol. 62, 2024, Art. no. 5510912.
[10] X. Liu, R. Wang, W. Liu, L. Zhang, and X. Wang, "Quality matters: Boosting face presentation attack detection with image quality metrics," *IEEE Access*, vol. 12, pp. 94654–94672, 2024.
[11] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 1–52, Nov. 2020.
[12] S. Winkler, *Digital Video Quality: Vision Models and Metrics*. Hoboken, NJ, USA: Wiley, 2005.
[13] M. Ivanovici, N. Richard, and C. Fernandez-Maloigne, "Towards video quality metrics based on colour fractal geometry," *EURASIP J. Image Video Process.*, vol. 2010, pp. 1–18, Dec. 2010.
[14] S. Winkler, "Visual fidelity and perceived quality: Toward comprehensive metrics," in *Human Vision and Electronic Imaging VI*, vol. 4299. Bellingham, WA, USA: SPIE, 2001, pp. 114–125.
[15] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "A comprehensive evaluation of full reference image quality assessment algorithms," in *Proc. 19th IEEE Int. Conf. Image Process.*, Sep. 2012, pp. 1477–1480.
[16] A. Rehman and Z. Wang, "Reduced-reference image quality assessment by structural similarity estimation," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3378–3389, Aug. 2012.
[17] L. Ma, S. Li, and K. N. Ngan, "Reduced-reference image quality assessment in reorganized DCT domain," *Signal Process., Image Commun.*, vol. 28, no. 8, pp. 884–902, Sep. 2013.
[18] V. Kamble and K. M. Bhurchandi, "No-reference image quality assessment algorithms: A survey," *Optik*, vol. 126, nos. 11–12, pp. 1090–1097, Jun. 2015.
[19] P. Yang, J. Sturtz, and L. Qingge, "Progress in blind image quality assessment: A brief review," *Mathematics*, vol. 11, no. 12, p. 2766, Jun. 2023.
[20] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.
[21] P. Zhang, X. Shao, and Z. Li, "Cycleiqa: Blind image quality assessment via cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.
[22] U. Sara, M. Akter, and M. S. Uddin, "Image quality assessment through FSIM, SSIM, MSE and PSNR—A comparative study," *J. Comput. Commun.*, vol. 7, no. 3, pp. 8–18, 2019.
[23] D. Liu, F. Li, and H. Song, "Image quality assessment using regularity of color distribution," *IEEE Access*, vol. 4, pp. 4478–4483, 2016.
[24] C. Liu, C. Yang, M. Wei, and J. Wang, "Texture smoothing quality assessment via information entropy," *IEEE Access*, vol. 8, pp. 88410–88421, 2020.
[25] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.
[26] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3350–3364, Dec. 2011.
[27] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[28] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[29] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.
[30] Y. Li, L.-M. Po, L. Feng, and F. Yuan, "No-reference image quality assessment with deep convolutional neural networks," in *Proc. IEEE Int. Conf. Digit. Signal Process. (DSP)*, Oct. 2016, pp. 685–689.
[31] W. Lu, W. Sun, X. Min, W. Zhu, Q. Zhou, J. He, Q. Wang, Z. Zhang, T. Wang, and G. Zhai, "Deep neural network for blind visual quality assessment of 4K content," *IEEE Trans. Broadcast.*, vol. 69, no. 2, pp. 406–421, Jun. 2023.
[32] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.
[33] Q. Yan, D. Gong, and Y. Zhang, "Two-stream convolutional networks for blind image quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2200–2211, May 2019.
[34] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[36] X. Wang, Y. Pang, and X. Ma, "Real distorted images quality assessment based on multi-layer visual perception mechanism and high-level semantics," *Multimedia Tools Appl.*, vol. 79, nos. 35–36, pp. 25905–25920, Sep. 2020.
[37] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3667–3676.
[38] Y. Gao, X. Min, Y. Zhu, J. Li, X.-P. Zhang, and G. Zhai, "Image quality assessment: From mean opinion score to opinion score distribution," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 997–1005.
[39] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2020.
[40] Y. Zhu, Y. Li, W. Sun, X. Min, G. Zhai, and X. Yang, "Blind image quality assessment via cross-view consistency," *IEEE Trans. Multimedia*, vol. 25, pp. 7607–7620, 2022.
[41] M. U. Rehman, I. F. Nizami, F. Ullah, and I. Hussain, "IQA vision transformed: A survey of transformer architectures in perceptual image quality assessment," *IEEE Access*, vol. 12, pp. 183369–183393, 2024.
[42] F. Chollet and K. Team. (2024). *Keras Applications*. Accessed: Mar. 31, 2024. [Online]. Available: https://keras.io/api/applications/
[43] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[44] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.
[45] M. Tan and Q. Le, "EfficientNetV2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
[46] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, and S. Gelly, "An image is worth 16×16 words: Transformers for image recognition at scale," 2021, *arXiv:2010.11929*.
[47] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.
[48] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.
[49] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.
[50] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.
[51] D. Ghadiyaram and A. C. Bovik. (2015). *Live in the Wild Image Quality Challenge Database*. Accessed: Mar. 2017. [Online]. Available: http://live.ece.utexas.edu/research/ChallengeDB/index.html
[52] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3575–3585.

[53] T. Installations and L. Line, "Subjective video quality assessment methods for multimedia applications," *Networks*, vol. 910, no. 37, p. 5, 1999.
[54] International Telecommunication Union, "User requirements for objective perceptual video quality measurements in digital cable television," ITU-T Recommendation J.143, ITU, Geneva, Switzerland, 2000. [Online]. Available: https://handle.itu.int/11.1002/1000/5071
[55] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.
[56] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1026–1034.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–13.
[58] D. Söllinger. (2020). *Blind Image Quality Toolbox*. [Online]. Available: https://github.com/dsoellinger/blind_image_quality_toolbox
[59] D. Freedman, R. Pisani, and R. Purves, *Statistics (International Student Edition)*, 4th ed., R. Purves, Ed., New York, NY, USA: WW Norton & Company, 2007.
[60] J. H. Zar, "Spearman rank correlation," in *Encyclopedia of Biostatistics*, vol. 7, P. Armitage and T. Colton, Eds., 2nd ed. New York, NY, USA: Wiley, 2005, pp. 4191–4196.
[61] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–93, 1938.
[62] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognit.*, vol. 30, no. 7, pp. 1145–1159, Jul. 1997.
[63] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[64] M. Simka, L. Polak, M. Novotny, J. Kufa, and K. Fliegel, "Performance evaluation of objective quality assessment methods for omnidirectional images under emerging compressions," *IEEE Access*, vol. 12, pp. 150419–150429, 2024.
[65] L. Krasula, K. Fliegel, P. Le Callet, and M. Klíma, "On the accuracy of objective image and video quality models: New methodology for performance evaluation," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Jun. 2016, pp. 1–6.
[66] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-Wild images via hierarchical feature fusion and iterative mixed database training," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 6, pp. 1178–1192, 2023.
[67] A. D. Dinescu, R. Miron, L. M. Itu, I. C. Plajer, and A. Turcea, "XCAE: Deep neural network for X-ray coronary angiograms quality enhancement," in *Proc. IEEE 28th Int. Conf. Emerg. Technol. Factory Automat. (ETFA)*, Sep. 2023, pp. 1–6.
[68] X. Kang, P. Duan, S. Li, and J. A. Benediktsson, "Decolorization-based hyperspectral image visualization," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4346–4360, Aug. 2018.
[69] P. Duan, X. Kang, S. Li, and P. Ghamisi, "Multichannel pulse-coupled neural network-based hyperspectral image visualization," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2444–2456, Apr. 2020.

CRISTIAN GEORGE FIERARU received the bachelor's degree in computer science from the Transilvania University of Braşov, in 2023, where he is currently pursuing the master's degree in modern technologies in software systems engineering. Notable work includes a full-stack application for grayscale image colorization and a library management system using enterprise technologies. His research interests include artificial intelligence, software architecture, and leveraging modern technologies to address real-world challenges.



MARIA BISERICĂ received the bachelor's degree in computer science from the Faculty of Mathematics and Computer Science, Transilvania University of Braşov, in 2024. Notable achievements include developing a full-stack application for analyzing and classifying digital images using convolutional neural networks and NR-IQA algorithms.

IOANA CRISTINA PLAJER received the Ph.D. degree in computer science from the Transilvania University of Braşov, Romania, in 2011. She is currently a Lecturer with the Faculty of Mathematics and Computer Sciences, Transilvania University of Braşov. She is also a member of the Department's Machine Learning Research Group and the Multispectral Imaging and Vision Research Laboratory. Her research interests include machine learning, image processing, spectral imaging and remote sensing, and formal languages.

MIHAI IVANOVICI (Senior Member, IEEE) received the Ph.D. degree in electronics and telecommunications from Politehnica University, Bucharest, Romania. He is currently a Full Professor with the Electronics and Computers Department, Transilvania University of Braşov, Romania. He is the Head of the Multispectral Imaging and Vision Research Laboratory. He was an Invited Researcher, in 2008, 2010, and 2014, and a Visiting Professor with the University of Poitiers, in 2018, University Toulouse 3 Paul Sabatier, France, in 2019, and the Technical University of Moldova, Chişinău, Moldova, in 2023. His research interests and expertise are in the field of algorithms and electronic system design for signal and data acquisition, processing and analysis-including color, multi- and hyper-spectral images, remote sensing and earth observation data, and data from particle detectors in the ATLAS Experiment at CERN, Geneva. He has been a member of the IEEE Signal Processing Society, since 2008, and the IEEE Geoscience and Remote Sensing Society, since 2018.