# Improved Performance Measures for Video Quality Assessment Algorithms Using Training and Validation Sets

*Original*
Improved Performance Measures for Video Quality Assessment Algorithms Using Training and Validation Sets / Aldahdooh, A.; Masala, E.; Janssens, O.; Van Wallendael, G.; Barkowsky, M.; Le Callet, P.. - In: IEEE TRANSACTIONS ON MULTIMEDIA. - ISSN 1520-9210. - STAMPA. - 21:8(2019), pp. 2026-2041. [10.1109/TMM.2018.2882091]

*Publisher:*
IEEE (Institute of Electrical and Electronics Engineers Inc.)

*Published*
DOI:10.1109/TMM.2018.2882091

# Improved Performance Measures for Video Quality Assessment Algorithms Using Training and Validation Sets

Ahmed Aldahdooh, Enrico Masala, Olivier Janssens, Glenn Van Wallendael, Marcus Barkowsky, and Patrick Le Callet

***Abstract*—The training and performance analysis of objective video quality assessment algorithms is complex due to the huge variety of possible content classes and transmission distortions. Several secondary issues such as free parameters in machine learning algorithms and alignment of subjective datasets put additional burden on the developer. In this paper, three subsequent steps are presented to address such issues. First, the content and coding parameter space of a large-scale database is used to select dedicated subsets for training objective algorithms. This aims at providing a method for selecting the most significant contents and coding parameters from all imaginable combinations. In the practical case where only a limited set is available, it also helps to avoid redundancy in the training subset selection. The second step is a discussion on performance measures for algorithms that employ machine learning methods. The particularity of the performance measures is that the quality of the training and verification datasets is taken into consideration. Common issues with often used existing measures are presented and improved or complementary methods are proposed. The measures are applied to two examples of No-reference objective assessment algorithms using the aforementioned subsets of the large-scale database. While limited in terms of practical application, this sandbox approach of objectively predicting objectively evaluated video sequences allows for eliminating additional influence factors from subjective studies. In the third step, the proposed performance measures are applied to the practical case of training and analyzing assessment algorithms on readily available subjectively annotated image datasets. The presentation method in this part of the paper can also be used as an exemplified recommendation for reporting in-depth information on the performance. Using this presentation method, future publications presenting newly developed quality assessment algorithms may be significantly improved.**

***Index Terms*—Video quality, No-Reference VQA, HRC selection, Datasets evaluation, Content features.**

## I. INTRODUCTION

MEASURING the perceived video quality has an important role in enhancing the quality of experience (QoE). Moreover, measuring video quality is important in different stages of video delivery systems, e.g., processing, compression, transmission, post-processing [1] and displaying [2]. Video quality can be measured either subjectively or objectively. Subjective video quality is human dependent and it is generally measured with the mean opinion score (MOS) of a number of observers. Since this method is highly expensive, time consuming, and cannot be integrated in automated systems, a lot of efforts have been devoted to developing objective measures [3], [4]. Objective video quality approaches, instead, rely on algorithms that can be classified into three different types depending on how much data of the original video is available to them. Full reference (FR) objective measures can access the full original video, whereas reduced reference (RR) ones can only access some representative characteristics and features of the original video [5]. No reference (NR) objective measures, instead, rely on the capability to extract bitstream or content features from the distorted video [6]–[8]. A tremendous number of video quality algorithms have been proposed in the past. In recent years, Narwaria et al. [9] published a low-complexity FR algorithm relying on machine learning based pooling, Naccari et al. [10] developed a NR measure for video quality monitoring, Anegekuh et al. [11] took the content characteristics into account in their NR-bitstream measures, and also Hameed et al. [12] and Li et al. [13] proposed a NR-bitstream model enhanced with content features for controlling the amount of forward error correction and playout and packet scheduling respectively. *Stability and Consistency* of these measures over a wide range of conditions is important for those and many other common scenarios such as rate-distortion based mode decisions (Sung et al. [14]) and client- or server-side controlled HTTP Adaptive Streaming (Bouten et al. [15]). However, evaluating the *stability and consistency* of objective video quality measures is a difficult task due to the scarce availability of suitable datasets and the lack of good indicators to compare the measures. This paper investigates two methodologies to address such issues.

Three approaches can be used to select a suitable dataset for evaluating a video quality assessment (VQA) measures. The first one depends on the researchers' expertise. In video coding community, for instance, different quality levels can be obtained by using different bitrate budgets or different quantization parameters (QPs). This approach would not guarantee the same performance when another dataset is selected by another expert since the contents and coding conditions are varied. The second approach is to provide a large-scale database that covers wide range of contents and coding conditions. Although a large-scale database can be created and can help in providing insight into the performance of VQA measures, as we showed in our previous work [16] by using 1,984 hypothetical reference circuits (HRCs) for each content type and resolution, using such a large-scale database to iteratively develop a new VQA measure is not practical. Hence, a third approach is recommended, that we name hybrid approach. In this approach, a large scale database is created, then a set of proccessed video sequences (PVSs) are selected to create a database. In our previous work [17], the problem of finding good datasets is addressed and Algorithm 1 was proposed to generate suitable and small representative databases from a large-scale database. The performance of these algorithms were evaluated using Pearson Linear Correlation Coefficient (PCC) which might not report the *best* evaluation of the VQA measure. In this paper, the performance of the algorithms in [17] is emphasized using the proposed performance measures.

The second important factor in evaluating and comparing VQA measures is to use indicators that can effectively distinguish and compare the performance obtained using different measures and datasets. Typically, in the VQA community such comparisons are carried out by means of the Pearson Linear Correlation Coefficient (PCC), Spearman's Rank Order Correlation Coefficient (SROCC), and the Root Mean Squared Error (RMSE). However, in many cases such indicators are not able to capture the peculiarities of the VQA measures. Therefore, in this work we aim to find **which performance evaluation measures should be used if PCC and RMSE are not able to report the goodness of a subset of HRCs.** The main focus and contribution of this paper is to discuss the advantages and the shortcomings of these evaluation measures and propose further measures based on the analysis of the results obtained with the performed subset analysis. We took advantages of analyzing the

Mr. Aldahdooh, University College of Applied Sciences, Gaza, Palestine. *Email*: adahdooh@ucas.edu.ps

Mr. Le Callet are with L'UNIVERSITÉ BRETAGNE LOIRE, Université de Nantes, Nantes, France. *Email*: patrick.lecallet@univ-nantes.fr

Mr. Masala is with Control and Computer Engineering Department, Politecnico di Torino, Torino, Italy. *Email*: enrico.masala@polito.it

Mr. Van Wallendael and Janssens are with Ghent University - imec - IDLab, Ghent, Belgium. *Email*: firstname.lastname@ugent.be

Mr. Barkowsky is with Deggendorf Institute of Technology, University of Applied Sciences, Deggendorf, Germany. *Email*: Marcus.Barkowsky@th-deg.de

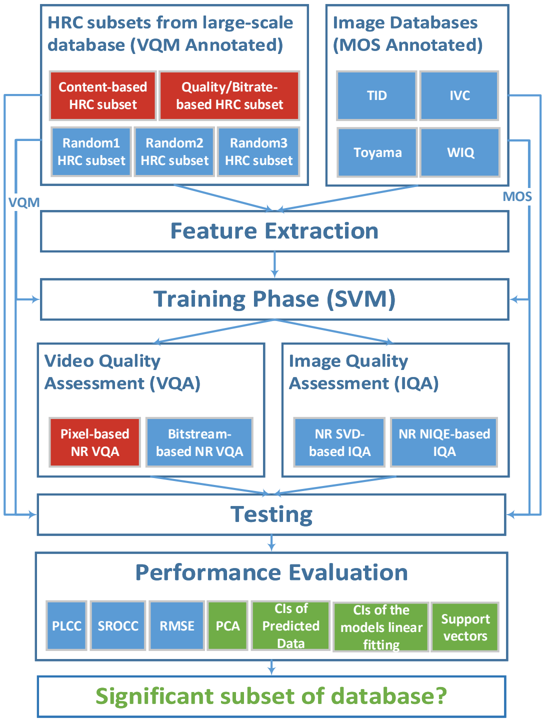


Fig. 1: Structural overview of the paper's experimental setup and data flow. Green boxes highlight this paper's contribution, while red boxes reviewed and extended from the author's previous publications.

residuals, the difference between the annotated and the predicted scores of the degraded videos, and of analysis the confidence intervals (CIs) of the fitting of the different models. The proposed performance measures are also highlighted by green-filled boxes in Figure 1.
In order to practically demonstrate how the proposed performance measures can measure the performance of the VQA measures and the datasets, many experiments are conducted and discussed in details in Sections II and III. Firstly, datasets are created from an existing database. Then, VQA/IQA measures are used to predict the quality of degraded videos/images. Finally, the VQA measures are assessed using the proposed performance indicators.

The rest of the paper is organized as follows. In Section II, an overview of the experiment is presented. Section III presents and explains the NR VQA models that are used in this paper. Section IV focuses on how to improve current performance evaluation measures and introduces the new VQA performance measures. The advantages achieved by the proposed methodologies are shown in Section V by applying them to two IQA algorithms from the literature. Conclusions are drawn in Section VI.

## II. EXPERIMENT OVERVIEW

Figure 1 visually summarizes our contributions and the conducted experiments by showing the overall structure used for evaluation. Our contributions are highlighted by green-filled boxes. The red-filled boxes show the main elements that emphasize the concept of the whole paper. They will be briefly summarized in the following.

### *A. Generation of small representative datasets from a large-scale database*

Differently from many other works that propose new VQA measures, we rely on a *non-subjectively* annotated large-scale database. However, the database contains several different objective VQA measures that have been computed for each HRC. In our previous work [17] we showed how, on the basis of such values, HRC subsets can be selected with limited loss of generality for the purpose of testing the performance of VQA methods. For completeness' sake we briefly report this approach here in Algorithm 1. Here we show, through practical examples using NR measures, the effectiveness of such a method. Note also that different optimization targets are possible. The second part (Lines 10-15) of Algorithm 1 is optimized to find HRCs that cover different ranges of (PSNR, Bitrate) values, whereas the first one (Lines 5-9) searches HRCs that have widely different impact depending on the type of content. Later in the paper, for comparison purposes, three randomly-generated subsets are also used. Note that the need for different comparison methods stems from the typical development cycle of an objective measure. First, a training dataset (i.e., one of a number of possible subsets) is selected; second, the VQA measures is iteratively developed; third, the VQA model is tested using validation dataset.

### *B. NR VQA measures*

As a further exemplification, we apply the presented methodologies to the specific case of NR VQA measures. A classification of no-reference quality estimation models for images and video has been presented in [18], which also discusses the most recent approaches, mainly focusing on coding artifacts due to JPEG or H.264/AVC. Although NR measures optimized for H.264/AVC could be adapted to the High Efficiency Video Coding (HEVC) case, some publications have specifically addressed HEVC [19]–[23]. Despite the fact that the state-of-the-art NR VQA measures provide promising results, it is not yet clear how well they can perform in real application contexts since the measures have almost all been implemented and tested using a small dataset that might not be sufficiently representative of real world conditions. This is currently an important limitation of such proposals. Hence, One approach to tackle the problem of general applicability is to use the large-scale database. In [24], the large-scale database is used to predict the behavior of the objective measures with full-reference (FR) video quality metrics for loss-impaired sequences using encoding, channel, and content features. Following the conclusions from [24], we extend this work in this paper by including pixel-based features to predict the video quality. These pixel-based features are used effectively in other works such as [25], [26]. Moreover, bitstream-based features [21] are used to build VQA models for HEVC coded videos. More details are discussed

**Algorithm 1** HRCs Selection process [17]

```
Input: data(PSNR,bitrate), K, N {for each HRC for each video sequence, K
    is number of clusters, and N is the number of sub-ranges}
Output: $HRC_1, HRC_2$ {set of selected HRCs for content and quality/bitrate driven subsets}
 1: $HRC_1, HRC_2 \Leftarrow null$
 2: $psnrRank \Leftarrow rank(PSNR)$
 3: $rateRank \Leftarrow rank(log(bitrate))$ ascending.
 4: $kmean++(psnrRank, rateRank, K)$

    subset: content-driven
 5: Find HRCs that distribute sources to the same clusters and assign them
    to a group #Groups=$G$
 6: for $g = 1 : G$ do
 7:     Compute the magnitude of the rank for each src in each hrc
 8:     $HRC_1 \leftarrow HRC_1 +$ Select the HRC that has the highest standard
        deviation
 9: end for

    subset: quality/bitrate-driven
10: for $k = 1 : K$ do
11:     Find the common HRCs (group(s)) between different sources
12:     For each group Get the COST = PSNR/log(bitrate)
13:     Divide the COST range into (N) sub-ranges
14:     $HRC_2 \leftarrow HRC_2 +$ For each subrange find the common HRC that
        is close to the mid-range point
15: end for
```

in Section III. In the final part of this paper we also investigate the specific case of two existing NR image quality assessment (IQA) algorithms [27], [28], showing the usefulness of our proposed methodologies to compare the performance of the two algorithms when trained using one out of four different datasets.

### C. Performance evaluation measures

Measuring the suitability of the different subsets for the training and verification stages is not straightforward. We will assess the PCC, SROCC, and RMSE indicators, discussing their advantages and shortcomings, and propose new comparison methodologies inspired by the previous development procedure. In particular, we propose to analyze residual errors and confidence intervals in the training and evaluation phases. For the specific case of support vector regression (SVR), we also found that a useful indicator could be the support vectors (SV) density. However, it should be noted that subset selection for the training stage inevitably introduces a bias in the quality measurement model which may be chosen such that particular degradations or content characteristics are better predicted at the expense of worse performance for others. In fact, in this paper we show this effect by using two subsets, one aimed at maximizing content variety and the other to maximize the coverage of rate-distortion tradeoff points.

## III. NO-REFERENCE VIDEO QUALITY MEASURES

### A. The pixel-based content features

The pixel-based content features used in this paper have been presented in our previous work [24] and used in [24], [29]. The features cover spatial and temporal characteristics that are extracted from the luminance frame (Y), and the chrominance frames (Cb and Cr), in the spatial or frequency domain. The features have been extracted on both block and frame levels. For the features that have been extracted at the block level, the Minkowski sum with different power has been applied to obtain a scalar value of each frame, then several statistical measures (e.g., mean, maximum, standard deviation, etc.) have been applied to get a scalar value that represents the video sequence. In addition to those features the standard deviation, the variance, the skewness, and the kurtosis of the motion intensity histogram that is computed using a pixel change ratio map (PCRM) [30] have been calculated. In total, 284 features have been extracted from a subset of the encoded sequences in the large-scale database [16].

### B. Bitstream features

In [21], Shahid et. al. use 52 bitstream features in order to perform perceptual quality estimation of HEVC coded videos. Ratios of various used CU sizes and of various prediction modes of intra and inter frames, and statistics of different levels of quantization parameters and motion vectors have been considered in these features. The features have been extracted as follows:

- The bitstream information extractor (HMIX) [16] has been used to generate the '.xml' file from the encoded stream file.
- HMIXParser, developed for this work, has been used to extract the bitstream features. Firstly, the frame-level features have been extracted and then sequence-level features have been calculated using a pooling strategy based on the average value.

### C. Subset description

Five HRC subsets have been used in this work. Two HRC subsets have been selected using HRC generation algorithms [17], see Algorithm 1. The first one shows the selection that is optimized for the HRCs that cover different ranges of (PSNR, Bitrate) values. The second shows the selection that is optimized for the HRCs in terms of contents, i.e. those that behave differently depending on the sources. The other three datasets use a random selection. Figure 3 shows the histograms of the quality scores (PSNR) for the five subsets. This histogram will be useful, for instance, when testing HRCs that are under-represented in the subset (i.e. PSNR $> 50$). These subsets will be named as $HRC_1$, $HRC_2$, $HRC_3$, $HRC_4$, and $HRC_5$ and correspond to the Content-driven subset, the quality/bitrate-driven subset, and the three random-based subsets respectively. Note that the number of HRCs in each subset are 97, 83, 100, 100, and 100, respectively. The number of HRCs is not identical due to the selection algorithms but sufficiently close for comparison.

### D. Feature selection process

Figure 2 shows the model that has been used in the experiments. First, all HRCs are encoded and then an objective full-reference measure, i.e., the VQM [31], has been used to estimate the quality. Then, the pixel-based features are extracted from the decoded output and finally the support vector regression (SVR) has been used to train the model. The feature selection algorithm in [24] has been used to get the features that are required for the support vector regression (SVR) process. Epsilon-SVR (LIBSVM tool [32]) with radial basis function has been used to train the model with 10-fold cross validation. Before the training is started, the parameters of the SVR (C, G, and epsilon) are optimized by selecting one combination of different C, G, and epsilon values. Five feature selection processes (SP) have been carried out: the first one $SP1$ for the content-driven subset $HRC_1$, the second one $SP2$ for the quality/bitrate-driven subset $HRC_2$, and the $SP3$, $SP4$, and $SP5$ for the three random subsets $HRC_3$, $HRC_4$, and $HRC_5$, respectively. These processes have been carried out for the pixel-based NR VQA and another five selection processes have been carried out for bitstream-based NR VQA. In the training phase, an exhaustive process of adding each feature one by one has been applied. In the training process of $SP1$, 16 features have been selected to be used for the SVR training. LIBSVM reports the squared correlation coefficient (SCC) as performance criterion. The SCC when 16 features are used is 0.9728. On the other hand, 14 features are selected in $SP2$ with an SCC equal to 0.9735. Figure 4 and Figure 5 show two features for the selected HRCs. The first feature (DCTHis13) is the histogram dissimilarity of DCT based feature maps using low and high frequency maps. The second feature (entrB_p4_mean) is the mean of entropy of 64x64 gray level co-occurrence matrix using Minkowski pooling (p=4). It can be observed that the features cover different ranges of values which make them useful for the training model. In $SP3/4/5$, 45, 11, and 8 features are selected, respectively, with SCC of 0.9883, 0.9830, and 0.9828. It can be observed that the number of selected features largely depend on the training data. Due to the over-fitting problem, the model that has the highest correlation is not necessarily the best one. This can be tested when the trained model is further validated with other datasets.

### E. Training and testing results: the impact of content features

After the features have been selected for the training model for the five HRC subsets, each training model has been trained and tested using all other HRC subsets including the subset that has been used in the training phase. The experiments have been divided into three categories: the first category will show the overall impact of the pixel-based content features in the different datasets. The second category will study the impact of pixel-based features per content. The third category will show which HRC group behaves differently when using the features. The performance of all experiments are measured using Pearson Linear Correlation Coefficient (PCC) and Root Mean Squared Error (RMSE).

In the first category, 25 experiments (referenced to as $X(row, column)$) have been conducted as shown in Figure 6. The rows of the figure represent the different training

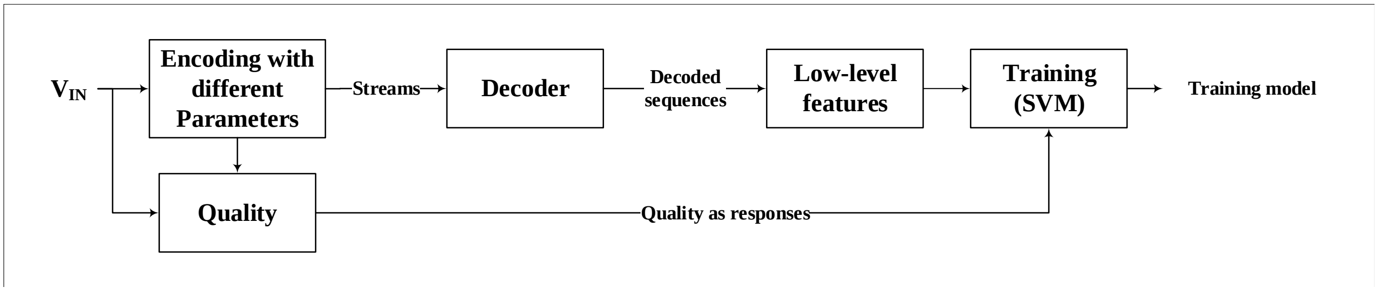


Fig. 2: NR video quality assessment (VQA) model

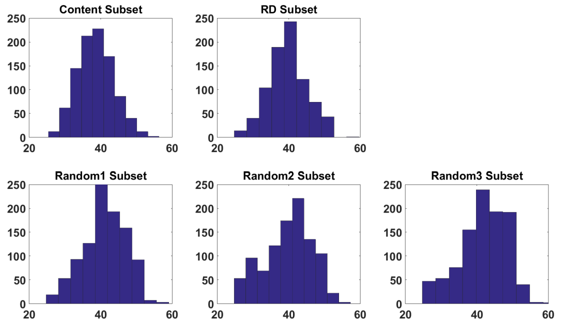


Fig. 3: Histograms of the quality scores for the five subsets

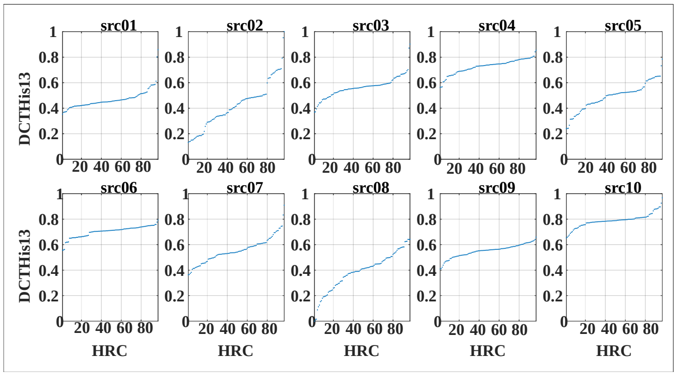


Fig. 4: DCT-based histogram dissimilarity feature of low and high frequency maps for content-driven subset.

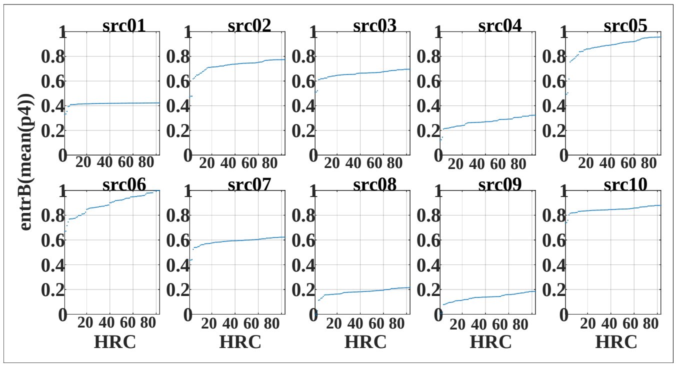


Fig. 5: The mean of entropy feature of 64x64 gray level co-occurrence matrix using Minkowski pooling (p=4) of all sources for quality/bitrate-driven subset.

models that have been trained, from top to bottom, using $HRC_1, HRC_2, HRC_3, HRC_4$, and $HRC_5$, while the columns represents the test data for each model and they are, from left to right, $HRC_1, HRC_2, HRC_3, HRC_4$, and $HRC_5$. Hence, the diagonal represents the evaluation of the model using the training data as the input. The green line represents $y = x$, while the red line represents the fitting line of each experiment. The first observation concerns the stability of content-driven, quality/bitrate-driven, and random 2 prediction models. It can be noticed that, when looking at the performance row by row, there is a stable and high performance (PCC higher than 0.95) for rows $HRC_1$, $HRC_2$ and $HRC_4$. Although the prediction model using $HRC_3$ is stable, it is still a random process and, as it can be noticed in the other random-based prediction models, the correlation scores are not stable when using $HRC_5$ and the fitting line deviates by a notable offset. Further discussion on the PCC performance measure will be presented in Section IV. Moreover, the predicted VQM in $X(2,1)$ is better correlated than $X(1,2)$. This can be explained as follows: both experiments try to predict VQM but the HRCs in $X(2,1)$ cover a wide range of quality/bitrate values while this is not the case for $X(1,2)$. Therefore, the training model has a better ability to predict the VQM value. Hence, this is an indication that the selection algorithm for quality/bitrate-driven yields good performance.

Another observation can be made by looking column-wise at the correlation of the experiments. This can suggest which HRCs are challenging to a certain model. The two challenging sets are, in order, quality/bitrate-based HRCs and content-based HRCs. Table I presents the analysis of PCC by calculating the absolute mean difference of the correlation coefficients, showing that the quality/bitrate-based and the $HRC_3$ subsets perform better in the pixel-based models, whereas the quality/bitrate-based and the content-based subsets perform, in general, better than others.

In the second category of experiments the influence of the content is presented. One content has been left out during the training and then the model has been tested on the content that has been left out. Table II shows a typical example where source 5 is left out of the training set and used for evaluation. Comparing this figure with the results of models that includes all contents in the training, it can be observed that the correlation is reduced and also the residual error is increased. That is an indication of content importance and how the absence of some content HRCs would affect the training model. Finally, in order to show that the HRCs subset selection algorithms work well, we compare the results on the diagonal. In general, on average, the Random 1 HRCs set has a lower correlation: this suggests that each content in the subset is valuable. The content and the quality/bitrate HRC subsets come next. When leaving one sample out from a subset that has many samples, a negligible drop in correlation means that this sample is redundant, whereas a huge drop in correlation means that the subset is not good enough. On the basis of this assumption, the content and the quality/bitrate HRC subsets are the ones that shows good performance.

In the third category of the experiments, the influence of individual HRCs cannot simply be seen by removing one HRC from the training phase and then testing with this HRC since, in this experiment, only 10 sources are used and there are HRCs that share the same encoder conditions. Therefore, one HRC group, i.e., coding condition, is removed from the training phase and the model is tested with this group. It is observed, as shown in Table III, that the main HRC groups that have the highest impact are the quality groups whereas other groups such as ‘Open/Closed GOP’, ‘Intraperiod’, and ‘Slice Arg.’ present stable results and higher PCC compared to quality groups. In general, removing one of these HRCs groups will highly impact the training model. For instance, including HRCs of low quality (QP=46) and high quality (8 Mbps, 16 Mbps, and QP=26) will help the model in better predicting the quality of new sample videos.

### *F. Training and testing results for bit-stream based no-reference model*

Figure 7 shows the same training/testing experiments done for the pixel-based features, but this time for bit-stream-based features NR VQA. Here, the samples (HRCs) are common between the pixel-based model and the bitstream-based model. It can be observed that all HRC subsets have a high correlation which makes it quite difficult to distinguish between them. Since, samples and features are important inputs for the training, the following conclusions can be drawn: first, the bit-stream features are optimal for the prediction, so all subsets have a high correlation. Second, the performance measures (PCC and RMSE) are not indicative of the significance of the HRCs. Further discussion will be elaborated in Section IV.

### *G. Results from different machine learning algorithms*

The two NR VQA models are trained using Stochastic Gradient Boosted Regression Trees algorithm, which recently has been shown objectively to be the state-of-the-art approach on structured data [33]. Furthermore, XGBoost is used, which is the state-of-the-art variation of the Stochastic Gradient Boosted Regression Trees algorithm [34]. In recent years, the popularity of this algorithm has risen dramatically due to its performance results in many machine learning competitions. For example, on the Kaggle platform, 17 out of 29 challenge winning solutions in 2015 used XGBoost. These XGBoost-based approaches outperformed both neural network and support vector machine-based solutions [35]. Apart from its success in machine learning competitions, XGBoost has also been proven to work well for practical applications such as train occupancy prediction [36], offshore wind turbine power prediction [37] and ads click-through prediction [38]. The success of XGBoost is often attributed to several aspects such as the fact that it is an ensemble model, requires little hyper parameter tuning, can deal with sparse data, requires no feature scaling and is very scalable due to its out-of-core learning ability [35]. The aim of this step is to observe some similarities and some dissimilarities when using different machine learning algorithms. The five models are trained using the same selected features for the pixel-based and bitstream-based VQA models. In comparisons with SVR technique and using the PCC performance measure, the XGBoost results confirm that $HRC_5$ is the worst subset and it disagrees in the performance order of other subsets. The absolute mean difference of PCC is considered as the stability measure between two different machine learning algorithms. Therefore, because they almost agree when using the absolute mean difference of the PCC performance measure, see Tables I and IV, we will complete the rest of this paper using the SVR technique.

After comparing the two machine learning algorithms, the two trained NR VQA models (using XGBoost) are now compared. From Table IV, it can be concluded that content-based and quality/bitrate-based HRC subsets provide the best results with respect to the correlation analysis; the absolute mean difference performance measure, regarding the pixel-based model, $HRC_3$ shows better performance on average, then $HRC_1$ $HRC_2$ $HRC_4$ $HRC_5$ come next in order. It should be noted that it is difficult to distinguish the difference between the $HRC_{1,2}$. On the other hand, in bitstream-based models, the average correlation shows that the content-based subset is the worst and it is very hard to distinguish the difference among $HRC_{2,3,4}$. But when analyzing the absolute mean difference of the PCC, the content-based and the quality/bitrate-based subsets perform better than others for both NR VQA models.

## IV. PERFORMANCE MEASURES FOR MODELS AND (SUB)SETS

As explained in the introduction, one of the main goals of the paper is to have an HRCs subset that can represent the large scale database. Hence, a set of analyses for the predicted values should be identified in order to judge the datasets. Since, as discussed in the previous section, the usual PCC and RMSE are not enough to judge a dataset, in this section other analyses are proposed for performance evaluation.

Please note that a prerequisite of all the following measures is that the input data is restricted to the unit interval, zero to one. This can be achieved by linear rescaling in most cases.

### *A. Analysis of the residuals using PCA*

#### *1) Redundancy in the training data $P_{RPCA\_T}$:*

***Purpose:*** measure the goodness of the training data in the training process by Analysis of the residuals using PCA.

***Idea:*** find the systematic redundancies in the training data that should be avoided such as redundant HRCs or redundant contents. By identifying similar behavior of the RMSE for two contents over all HRCs, redundancies can be identified. Optimality is reached if the HRCs behave differently for any two contents of the subset. The same applies to the HRC analysis: Optimality is reached if the contents behave differently for any two HRCs of the subset.

***Process:*** train the model and evaluate it on the training data. Calculate the residual errors of the prediction by the model. For the content analysis (dimension=SRC), first create a vector per content that contains the residual errors for each HRC. Perform a PCA on these $m$ vectors. Calculate the sum of the Eigenvalues of the first $n$ components of the total $m$ components. The default value should be $n = 0.2m$. Perform the same operations by creating a vector per HRC (dimension=HRC).

***Reporting:*** Use $P^{\text{dimension}}_{\text{RPCA\_T}}(\frac{n}{m}, m) = x$, i.e. $P^{\text{SRC}}_{\text{RPCA\_T}}(0.2, 10) = 0.9$.

***Interpretation:*** the lower the value, the better because the explained variance is low in the first $n$ components, i.e. the remaining components have significant information.

***Example and further explanations:*** The titles of the subplots in Figure 8 show the sum of the first two principal components, i.e $P^{\text{SRC}}_{\text{RPCA\_T}}(0.2, 10)$. The higher the value, the higher the possibility of existing systematic redundancy. As shown in the diagonal of Figure 8, it is observed that the quality/bitrate-based subset has the lowest explained variance in the first two components. That is an indication that the HRCs are valuable in the subset. Using the RMSE would not provide the same information, as can be seen from Fig. 6 because the best subset with respect to different HRCs is not easy to identify.

#### *2) Redundancy in the validation data $P_{RPCA\_V}$:*

***Purpose:*** measure the goodness of the validation data in the subset and model comparison process by analyzing the residuals using PCA. In other words, characterize which subset is challenging for the trained models.

***Idea:*** similar to $P_{\text{RPCA\_T}}$, find the systematic redundancies in the validation data. Redundancy should be avoided, both as redundant HRC or redundant content.

***Process:*** Evaluate the already trained model on the validation data without retraining. Then, follow the process of Section IV-A1 in order to obtain $P_{\text{RPCA\_V}}$.

***Reporting:*** Use $P^{\text{dimension}}_{\text{RPCA\_V}}(\frac{n}{m}, m) = x$, i.e. $P^{\text{SRC}}_{\text{RPCA\_V}}(0.2, 10) = 0.4$.

***Interpretation:*** The lower the value of $P_{\text{RPCA\_V}}$, the better the performance as less redundancy is found in the considered dimension in the validation. There are two sources of redundancy in the validation: The first one is the same as with training, i.e. the used set contains systematic redundancies; the second source is that the model may

TABLE I: Correlation analysis, expressed as a percentage, for the NR VQA models using SVR

| | | PCC | | | | | | Difference to train PCC | | | | | Absolute |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $HRC_1$ | $HRC_2$ | $HRC_3$ | $HRC_4$ | $HRC_5$ | Average | $HRC_1$ | $HRC_2$ | $HRC_3$ | $HRC_4$ | $HRC_5$ | mean difference |
| Pixel-based | $HRC_1$ | 98.8 | 97.1 | 96.3 | 96.6 | 96.5 | 96.6 | 0 | 1.7 | 2.53 | 2.21 | 2.29 | 2.19 |
| | $HRC_2$ | 98.3 | 98.8 | 98.8 | 99.1 | 98.9 | **98.8** | 0.53 | 0 | 0.01 | -0.32 | -0.13 | **0.02** |
| | $HRC_3$ | 97.8 | 96.1 | 99.5 | 97.2 | 90.5 | 95.4 | 1.71 | 3.44 | 0 | 2.30 | 9.03 | 4.12 |
| | $HRC_4$ | 95.1 | 98.2 | 98.7 | 99.3 | 99.0 | 97.7 | 4.13 | 1.12 | 0.60 | 0 | 0.27 | 1.53 |
| | $HRC_5$ | 62.7 | 60.0 | 68.6 | 91.9 | 99.2 | 70.8 | 36.52 | 39.22 | 30.66 | 7.33 | 0 | 28.43 |
| Bitstream-based | $HRC_1$ | 98.0 | 97.3 | 97.4 | 97.5 | 97.8 | 97.5 | 0 | 0.73 | 0.59 | 0.49 | 0.26 | 0.52 |
| | $HRC_2$ | 97.2 | 98.2 | 97.9 | 98.1 | 98.2 | **97.8** | 1.07 | 0 | 0.35 | 0.10 | 0.02 | **0.38** |
| | $HRC_3$ | 96.5 | 97.4 | 98.5 | 98.4 | 98.0 | 97.6 | 1.96 | 1.04 | 0 | 0.04 | 0.44 | 0.87 |
| | $HRC_4$ | 95.7 | 97.1 | 98.1 | 99.0 | 98.4 | 97.3 | 3.34 | 1.94 | 0.99 | 0 | 0.69 | 1.74 |
| | $HRC_5$ | 96.8 | 97.6 | 97.7 | 98.4 | 98.9 | 97.6 | 2.09 | 1.25 | 1.16 | 0.46 | 0 | 1.24 |

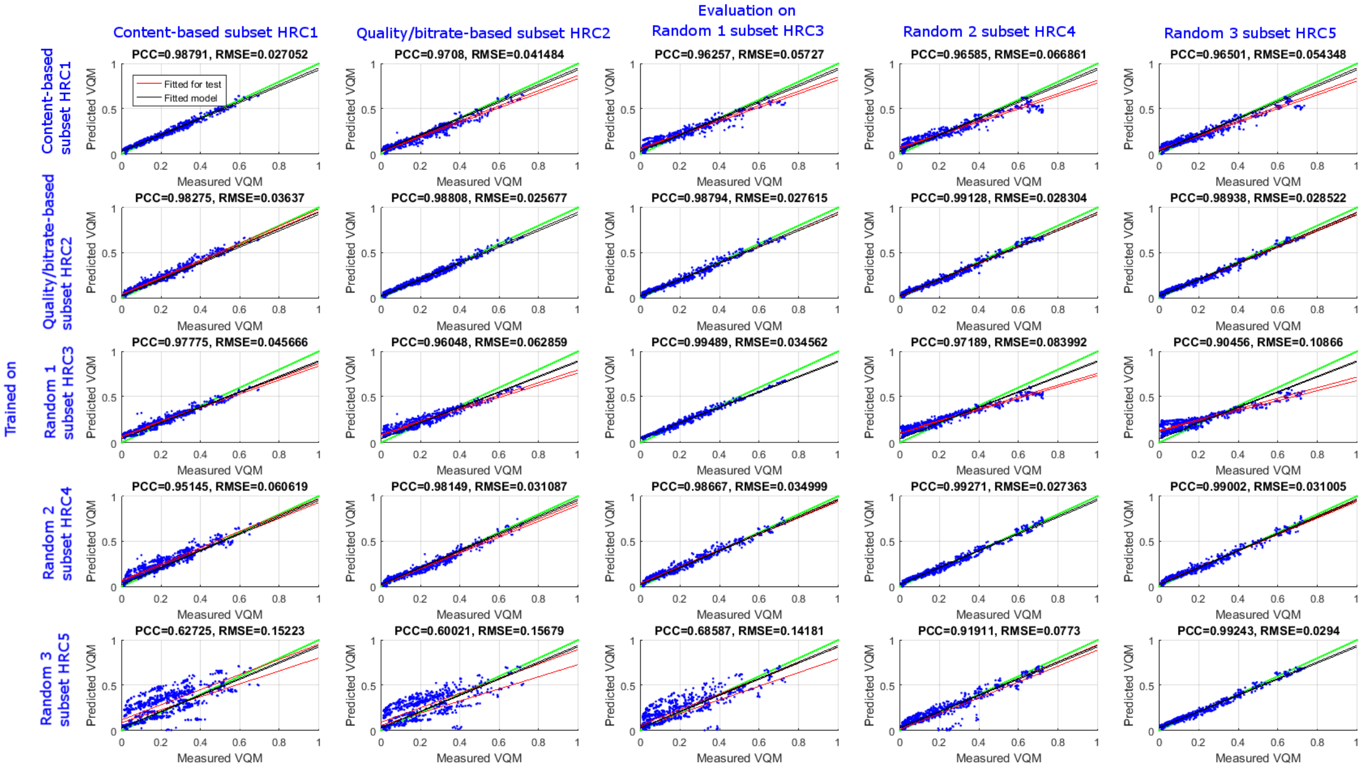


Fig. 6: The PCC and the RMSE for the 25 experiments of the pixel-based model. Rows: the different training models that are trained using $HRC_1, HRC_2, HRC_3, HRC_4$, and $HRC_5$,. Columns: the test data for each model, from the left, $HRC_1, HRC_2, HRC_3, HRC_4$, and $HRC_5$. The green line is the reference $(y = x)$.

behave similarly for different conditions, such as not considering a certain degradation at all. When using several subsets for training and validation, further analysis on these two can be obtained by comparing the graphs cross-wisely, i.e. $X(n, :)$ and $X(:, n)$. Good models should provide low values of $P_{\text{RPCA\_V}}$ row-wise in $X(n, :)$ and good subsets for verification are characterized by high values of $P_{\text{RPCA\_V}}$ column-wise, i.e. $X(:, m)$ because a high value indicates that a model is challenged by the subset $m$, i.e. the model can not reliably predict this subset.

***Example:*** following the above interpretation in Figure 8, models that were trained on the specific subsets are performing in the following rank order: quality/bitrate-based, random 2, random 1, content-based, random 3. The subsets in decreasing order of goodness are quality/bitrate-based, content-based, random 2, random 1 and finally random 3.

### *B. Analysis of confidence intervals (CIs) of the different models fittings*

In this section, further performance measures for the models are explained. These analyses are based on two different notions of confidence intervals. When fitting a model, the parameters of the model are determined based on training data. The more training data is available and the better the model fits, the smaller the confidence intervals for each calculated model parameter. In this text, this is called the model confidence, *model-C*. When the model is used for prediction, a certain percentage (usually 95%) of the predicted data lies in a corridor bounded by the upper and lower confidence intervals. This is called the data confidence, *data-C*.

#### *1) Model's prediction performance on particular validation dataset:*

***Purpose:*** measure the goodness of the trained model with respect to its reliability of predicting validation data.

***Idea:*** Determine the confidence interval corridor for the model predicting its own training data. Then, count the number of validation data points that fall into this corridor.

TABLE II: The PCC and the RMSE for the 25 experiments that are trained without source 5 and tested with source 5 HRCs.

| | | Tested on | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PCC | | | | | RMSE | | | | |
| | | Content | RD | Rand 1 | Rand 2 | Rand 3 | Content | RD | Rand 1 | Rand 2 | Rand 3 |
| Trained on | Content | 0.31 | 0.22 | 0.41 | 0.69 | 0.6 | 0.25 | 0.23 | 0.17 | 0.17 | 0.15 |
| | RD | 0.92 | 0.91 | 0.92 | 0.93 | 0.92 | 0.16 | 0.17 | 0.14 | 0.14 | 0.13 |
| | Rand 1 | -0.3 | 0.21 | 0.19 | 0.62 | 0.62 | 0.16 | 0.18 | 0.22 | 0.22 | 0.23 |
| | Rand 2 | 0.96 | 0.98 | 0.98 | 0.98 | 0.98 | 0.06 | 0.06 | 0.07 | 0.08 | 0.08 |
| | Rand 3 | 0.78 | 0.76 | 0.81 | 0.89 | 0.87 | 0.39 | 0.39 | 0.36 | 0.35 | 0.29 |

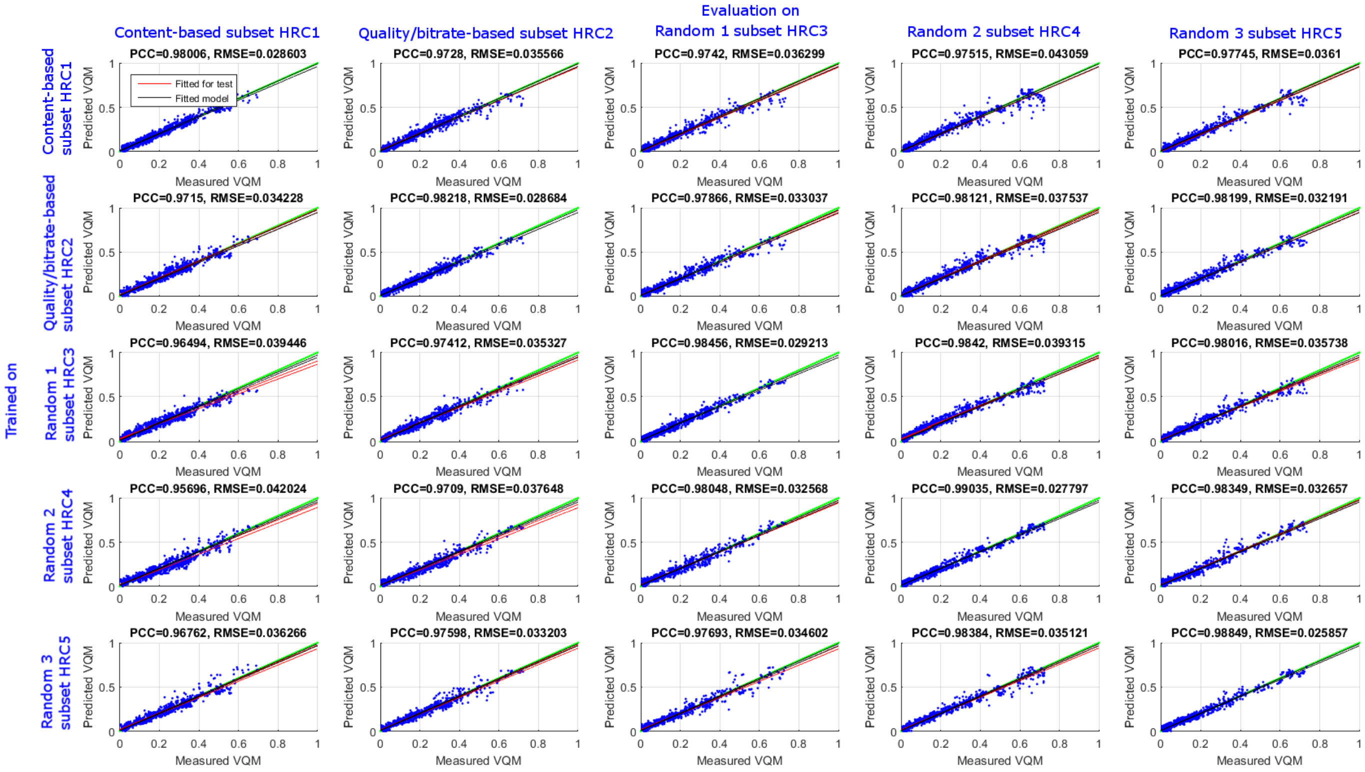


Fig. 7: The PCC and the RMSE for the 25 experiments of bitstream-based model. Rows: the different training models that are trained using $HRC_1, HRC_2, HRC_3, HRC_4$, and $HRC_5$,. Columns: the test data for each model, from the left, $HRC_1, HRC_2, HRC_3, HRC_4$, and $HRC_5$. The green line is the reference $(y = x)$.

***Process:*** Train the model on the training data. Evaluate it on the validation dataset. The 95% confidence interval boundaries for *data-C*are obtained by using a function such as MATLAB's[1] *polycon* function. This function is applied on the training data in order to get the two boundary lines that are parallel to the fitting line, i.e. $y \pm \delta$. The validation data is then predicted by the same model and for each datapoint it is determined whether it is inside the previously determined confidence interval boundary. The ratio of inliers $i$ and outliers $o$ of the total number of datapoints in the validation set $n$ is reported. This is similar to the well-known outlier ratio with respect to the standard error but takes into consideration training and validation. This analysis is further studied in Section IV-C.

***Reporting:*** Use $P_{\text{DCI\_V}}(\delta, n)=\frac{i}{o}$, i.e. $P_{\text{DCI\_V}}(0.12,\ 100)=0.3$.

***Interpretation:*** The higher the ratio, the better the model predicts the validation data with respect to its own training data.

***Example:*** as shown in Figure 9, the black lines are the boundaries of *data-C*of the trained model and the black points are the predicted data points of the trained data. The red points are the predicted data points of the validation data. In addition, the red lines show the boundaries of *data-C*using the validation subset (further exploited in IV-C). Each sub figure reports the $P_{\text{DCI\_V}}$. For instance, the fifth row $X(5, :)$, which refers to the validation of the model trained on Random3, shows that the spread of content and quality/bitrate based subset is the largest compared to the other subsets. This is reflected in the value of $P_{\text{DCI\_V}}$. Since the content-based model is not designed to have a wide-range of quality and bitrate, the predicted VQM values of content-based HRCs mostly lie outside the area of the *data-C*CIs for other models. Therefore, its HRCs are challenging for other models, especially random-based models.

### 2) Model determined by its training data:

***Purpose:*** measure the goodness of the model by analyzing the area of the confidence interval spread by the *model-C*.

***Idea:*** the size of the area of the model parameter's confidence spread provides information about the exactness with which the model parameters can be determined by the training data.

***Process:*** determine the confidence interval values for each of the trained model parameter on the training data of size $n$. For a linear model, gradient and offset have a confidence interval that is provided by the fitting function, e.g. the MATLAB[1] function $fitlm$ in conjunction with $coefCI$. Determine the maximum confidence boundaries that are spread by the uncertainty, e.g. for a linear model

---

[1] MATLAB functions are given here for exact reproducibility, other software such as Octave or R have similar functionality.

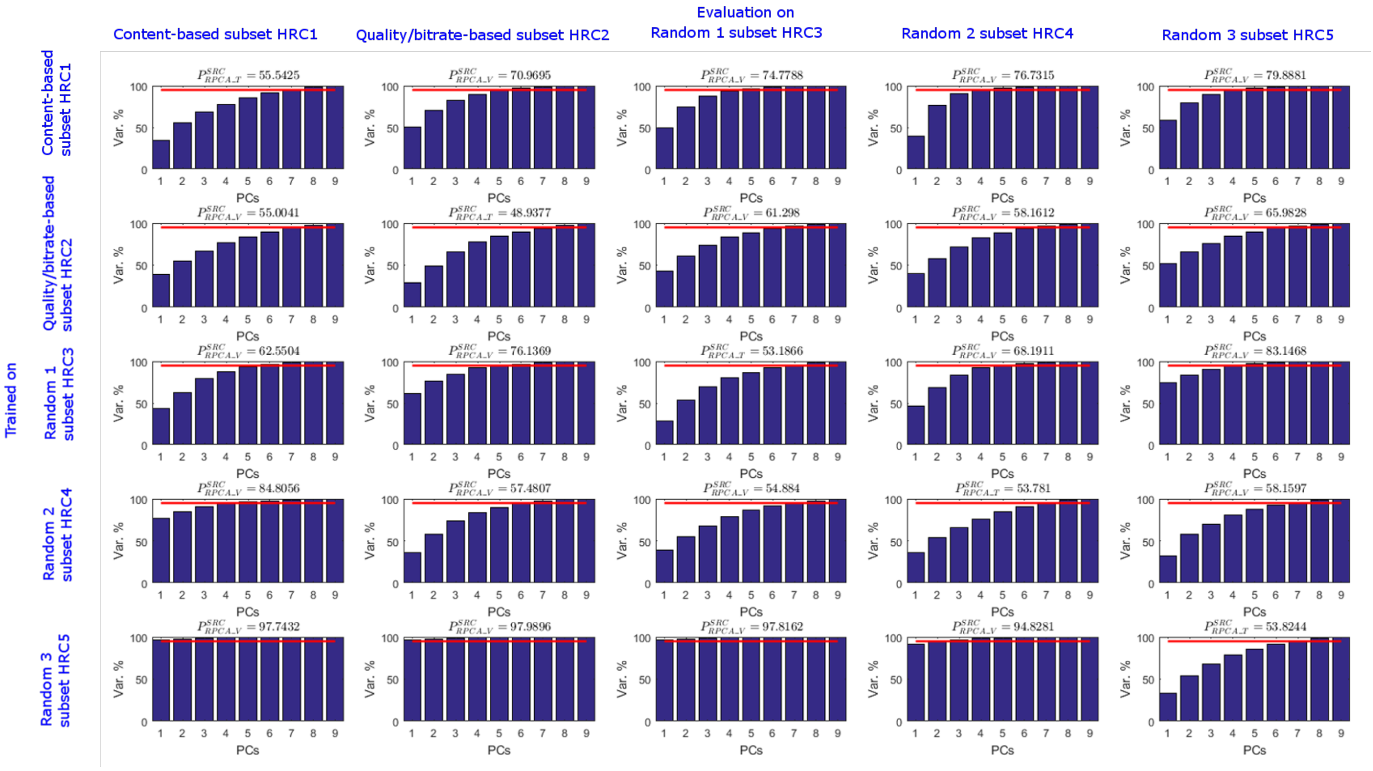


Fig. 8: The cumulative sum of explained variances of the principal components. The red lines indicate when the model reach a 95% of cumulative variances.

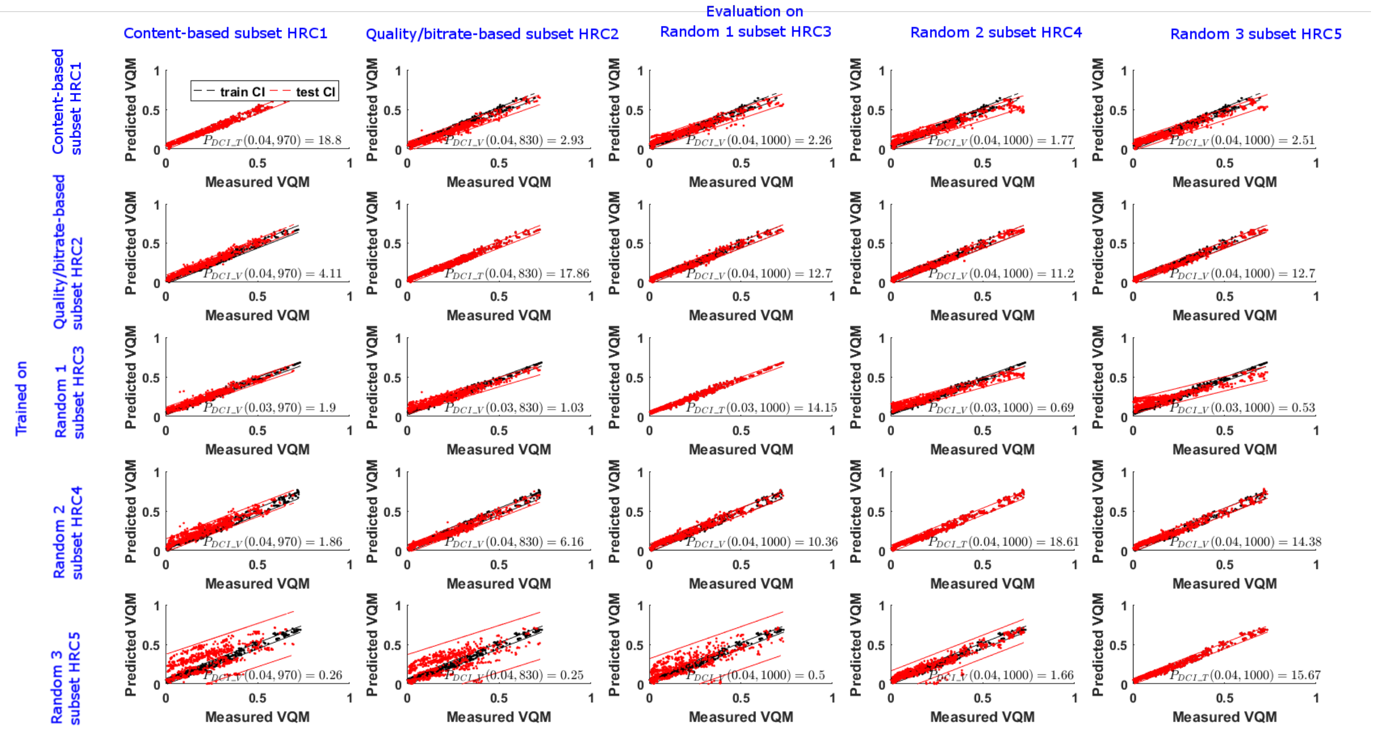


Fig. 9: How much the predicted VQM values lie in area of confidence interval of the fitted data.

the lower bound is determined by the line $y = (a - CI_a)x + (b - CI)$. Calculate the area of uncertainty $x$, e.g. for a linear model between the lower and upper bound $y = (a + CI_a)x + (b + CI)$. This analysis is studied further in Section IV-C.

***Reporting:*** Use $P_{\text{MCI\_T}}(n)$=$x$, i.e. $P_{\text{MCI\_T}}(120)$=0.4.

***Interpretation:*** the lower the value, the better the model is able to predict its training data. Please note that a low value may also indicate overtraining or irrelevant training data. Hence, we will consider this value when considering the interaction between the training data and the validation data in Section IV-C.

***Example:*** Figure 6 shows in each subplot 5 lines. The green line represents $y = x$. The two black ones are the CIs of a fitted model, therefore each row in the figures shows the same two black lines. The value of $P_{\mathrm{MCI\_T}}$(n) for the five models are, respectively, 0.014, 0.014, 0.008, 0.011, and 0.010. The sign of the over-fitting is obvious for random-based subsets. Going further than $P_{\mathrm{MCI\_T}}$, the two red lines represent the *model-C*CIs when trained on the validation set. This leads to the idea of observing the amount of overlap between training a model on one or the other subset. A good model is characterized by red lines located between black lines. As shown in Figure 6, the amount of overlap in the quality/bitrate-based model is the largest one.

### *C. Interaction between the model training, the training data, and the validation data*

In this subsection, the goodness of a the trained model on its own training data and on a specific validation subset is studied. The analysis can be applied to the model fit (*model-C*CI) or to the data fitting ability (*data-C*CI). The following three attributes of the CI analysis are used. The area between the CI boundaries of the training (black lines, denoted as $b$), the area between the CI boundaries of the validation (red lines, denoted as $r$), and finally the area of the intersection between the two areas (denoted as $i$). The main conditions with respect to line intersections are explained in Table V.

#### *1) Goodness of data prediction using a trained model on validation data:*

***Purpose:*** provide an absolute number for the prediction performance of a trained model on a validation dataset taking into consideration the training dataset.

TABLE III: PCC of the prediction using leave-one-out strategy, i.e. leave one HRC group out.

| HRC groups | | | Data sets | | | | |
|---|---|---|---|---|---|---|---|
| | | | $HRC_1$ | $HRC_2$ | $HRC_3$ | $HRC_4$ | $HRC_5$ |
| GOP | | GOP2 | 0.99 | 0.98 | 0.99 | 0.99 | 0.99 |
| | | GOP4 | 0.98 | 0.99 | 1.00 | 0.99 | 0.99 |
| | | GOP8 | 0.96 | 0.98 | 0.99 | 0.99 | 0.99 |
| | | LDGOP4 | 0.98 | 0.99 | 0.99 | 0.98 | 0.99 |
| Quality Control | Bitrate | 500000 | 0.96 | 0.97 | 0.96 | 0.97 | 0.97 |
| | | 500001 | 0.95 | 0.96 | 0.94 | 0.94 | 0.95 |
| | | 1000000 | 0.98 | 0.98 | 0.98 | 0.96 | 0.97 |
| | | 1000001 | 0.98 | 0.98 | 0.97 | 0.97 | 0.97 |
| | | 2000000 | 0.97 | 0.97 | 0.98 | 0.95 | 0.94 |
| | | 2000001 | 0.99 | 0.93 | 0.98 | 0.96 | 0.95 |
| | | 4000000 | 0.83 | 0.80 | 0.91 | 0.86 | 0.88 |
| | | 4000001 | 0.88 | 0.87 | 0.93 | 0.94 | 0.91 |
| | | 8000000 | 0.58 | 0.34 | 0.72 | 0.67 | 0.63 |
| | | 8000001 | 0.55 | 0.25 | 0.77 | 0.75 | 0.62 |
| | | 16000000 | - | 0.13 | 0.50 | 0.22 | 0.04 |
| | | 16000001 | 0.23 | 0.13 | 0.54 | 0.20 | -0.02 |
| | QP | 26 | 0.92 | 0.93 | 0.93 | 0.95 | 0.91 |
| | | 32 | 0.92 | 0.97 | 0.94 | 0.92 | 0.95 |
| | | 38 | 0.94 | 0.96 | 0.90 | 0.93 | 0.91 |
| | | 46 | 0.08 | 0.42 | 0.29 | 0.54 | 0.22 |
| Open/ close GOP | | 1 | 0.96 | 0.98 | 0.99 | 0.99 | 0.99 |
| | | 2 | 0.99 | 0.99 | 0.99 | 0.99 | 0.99 |
| Intra-period | | 8 | 0.97 | 0.98 | 0.98 | 0.99 | 1.00 |
| | | 16 | 0.99 | 0.98 | 1.00 | 0.99 | 0.99 |
| | | 32 | 0.97 | 0.98 | 0.99 | 0.99 | 0.98 |
| | | 64 | 0.97 | 0.99 | 0.99 | 0.99 | 0.99 |
| Slice Arg. | | 0 | 0.98 | 0.98 | 1.00 | 0.99 | 0.99 |
| | | 2 | 0.99 | 0.98 | 1.00 | 0.99 | 1.00 |
| | | 4 | 0.99 | 0.98 | 0.99 | 0.99 | 0.98 |
| | | 1500 | 0.98 | 0.99 | 0.99 | 0.99 | 0.99 |

***Idea:*** The model prediction performance can be characterized by the *data-C*CI of the validation data. The smaller the CI, the better the model. Taking into consideration the training process, the smaller the CI on the training data, the better the model. Finally, taking into consideration the interaction between the training and the validation, the larger the intersection between the CI, the better the fitting.

***Process:*** Train the model on the training data. Calculate the area of the *data-C*CI corridor similar to Subsection IV-B1 in order to obtain the area $b$. Perform the same operation without retraining on the validation data in order to obtain $r$. Calculate the intersection between the two corridors in order to obtain $i$.

***Reporting:*** The goodness value is reported as $P^{(\mathrm{b,r,i})}_{\mathrm{GData}}=\frac{i}{\max(b,r)^2}$, e.g. $P^{(0.5,0.4,0.3)}_{\mathrm{GData}}= 1.2$

***Interpretation:*** The higher the value, the better the model's performance and the data on which the model was trained. The calculation is divided into two terms, the first one being $\frac{i}{\max(b,r)}$ which achieves its maximum value 1 if the intersection covers exactly the larger area, whereas it is equal to zero in case of no overlap. The second term is $\frac{1}{\max(b,r)}$ which becomes larger as the CI areas become smaller. The measure was designed to provide a reasonable tradeoff between these goals. The behavior of this measure can be seen in Figure 10.

***Example:*** Figure 11 shows 4 sub-figures, the first column is related to $P^{(\mathrm{b,r,i})}_{\mathrm{GData}}$ for the features-based NR VQA and bitstream-based NR VQA. In both NR VQA models, the quality/bitrate-based dataset has the largest $P^{(\mathrm{b,r,i})}_{\mathrm{GData}}$ value, whereas the content-based subsets rank third and fifth in both models, respectively.

#### *2) Goodness of two datasets for determining linear model parameters:*

***Purpose:*** evaluate the model's stability when using different datasets.

***Idea:*** A good model should provide a stable linear relation to any given dataset. This is similar to Subsection IV-C1 but exchanging *data-C*with *model-C*

***Process:*** Train the model on the training data. Perform a linear fitting on the training data and calculate the area $b$=$P_{\mathrm{MCI_T}}$as explained in Subsection IV-B2. Perform a linear fitting on the validation data and calculate the area $r$ similarly. Calculate the intersection between the two areas $i$.

***Reporting:*** The goodness value is reported as $P^{(\mathrm{b,r,i})}_{\mathrm{GModel}}= \frac{i}{\max(b,r)^2}$, e.g. $P^{(0.5,0.4,0.3)}_{\mathrm{GModel}}= 1.2$

***Interpretation:*** The higher the value, the better. The same explanation as in Subsection IV-C1 holds but, in this case, the stability of the model to predict different datasets is analyzed.

***Example:*** The second column of Figure 11 is related to $P^{(\mathrm{b,r,i})}_{\mathrm{GModel}}$ for the features-based NR VQA and bitstream-based NR VQA. In both NR VQA models, the quality/bitrate-based dataset has the largest $P^{(\mathrm{b,r,i})}_{\mathrm{GModel}}$(b,r,i) value, whereas the content-based subsets rank third and second in both models, respectively.

### *D. Comparing the performance of HRC subsets*

As discussed and observed in the previous sections, all the aforementioned performance measures yield different results for different HRC subsets. Therefore, in this subsection, all the results are put together in order to judge the HRC subsets. A rank-order technique is applied in order to get a final score for each HRCs set. Since we have 5 HRC subsets, each HRCs set will have an order number for each performance measure discussed in this paper and then a comparison between the two NR VQA measures is presented. Table VI shows all HRC subsets ranks for each performance measure and for the pixel-based and the bit-stream-based NR VQA measures. From the table, it can be surmised that the systematic way of selecting the HRC set to be used for the experiments performs better than the random selection that covers different ranges of bitrate and quality. The key advantage of using such technique in evaluation is that when

TABLE IV: Correlation analysis, expressed as a percentage, for the NR VQA models using XGBoost

| | | PCC | | | | | | Difference to train PCC | | | | | Absolute mean difference |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $HRC_1$ | $HRC_2$ | $HRC_3$ | $HRC_4$ | $HRC_5$ | Average | $HRC_1$ | $HRC_2$ | $HRC_3$ | $HRC_4$ | $HRC_5$ | |
| Pixel-based | $HRC_1$ | 96.5 | 95.8 | 96.9 | 97.2 | 97.0 | 96.7 | 0 | 0.69 | -0.40 | -0.78 | -0.58 | 0.27 |
| | $HRC_2$ | 94.6 | 96.4 | 96.3 | 97.0 | 96.8 | 96.2 | 1.76 | 0 | 0.08 | -0.59 | -0.47 | **0.20** |
| | $HRC_3$ | 98.2 | 98.5 | 99.9 | 99.4 | 99.4 | **98.9** | 1.72 | 1.38 | 0 | 0.46 | 0.46 | 1.00 |
| | $HRC_4$ | 87.1 | 89.5 | 92.0 | 95.9 | 95.1 | 90.9 | 8.76 | 6.35 | 3.89 | 0 | 0.77 | 4.94 |
| | $HRC_5$ | 82.9 | 85.5 | 90.3 | 93.7 | 94.3 | 88.1 | 11.35 | 8.74 | 4.01 | 0.59 | 0 | 6.17 |
| Bitstream-based | $HRC_1$ | 98.0 | 95.9 | 97.0 | 96.5 | 97.0 | 96.6 | 0 | 2.13 | 1.02 | 1.46 | 1.03 | **1.41** |
| | $HRC_2$ | 97.9 | 100.0 | 98.8 | 98.5 | 98.9 | **98.5** | 2.10 | 0 | 1.21 | 1.47 | 1.06 | 1.46 |
| | $HRC_3$ | 97.2 | 97.6 | 99.6 | 98.7 | 98.4 | 98.0 | 2.47 | 1.98 | 0 | 0.94 | 1.18 | 1.64 |
| | $HRC_4$ | 97.1 | 97.8 | 98.8 | 99.9 | 99.0 | 98.2 | 2.80 | 2.14 | 1.12 | 0 | 0.97 | 1.76 |
| | $HRC_5$ | 94.5 | 96.9 | 97.4 | 98.2 | 99.7 | 96.7 | 5.25 | 2.89 | 2.37 | 1.58 | 0 | 3.02 |

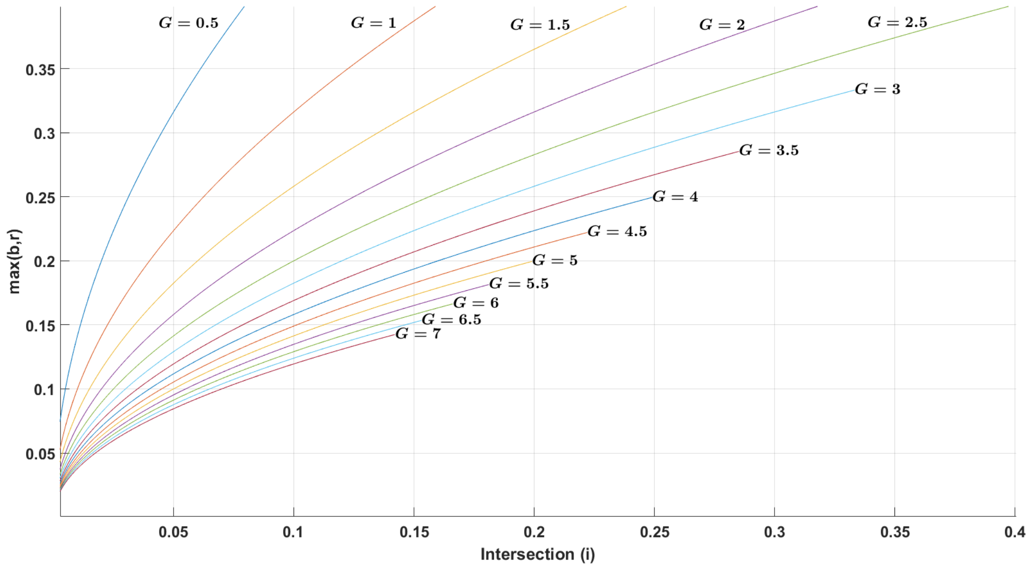
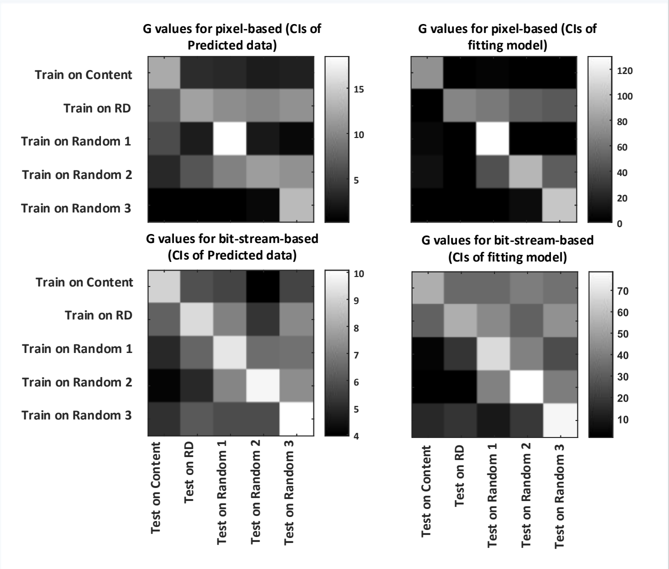


Fig. 10: The behavior of G with different $max(b, r)$ and $i$ values.


G values for pixel-based (CIs of Predicted data)

G values for pixel-based (CIs of fitting model)

G values for bit-stream-based (CIs of Predicted data)

G values for bit-stream-based (CIs of fitting model)

Fig. 11: The G values for the CIs analysis for the pixel-based and bit-stream-based NR VQA models

a given performance measure cannot give clear indications about which model is better than others, another performance measure can.

### E. Detailed Analysis of Support Vectors

Support vector (SV) based machine learning is one of the widespread methodologies for regression fitting. Important insight can be gained from support vectors because they are actual data points from the training data set. In most cases this means that some of the created conditions (resulting in PVSs) are deemed of foremost importance for representing the whole training data set.

***Purpose:*** Evaluate the efficiency of the distribution and the weighting of the selected support vectors with respect to the ground-truth quality.

***Idea:*** Because the SVs are training data points, each support vector is assigned to one ground truth quality score. The machine learning should choose SVs that equally spread over the predicted quality range. In other words, if the training chooses SVs in a small quality subrange, the prediction may get unstable if compared with conditions outside that particular quality range and the chosen SVs may be redundant. The weighting of the SVs needs to be taken into consideration.

***Process:*** Train the algorithm on the training data and extract the SVs and their weights. Identify which training data point corresponds to each SV. Retrieve the ground-truth quality score (i.e., the ones on which the algorithm was trained).

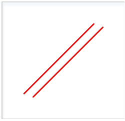
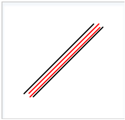
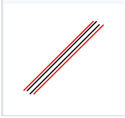
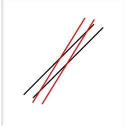
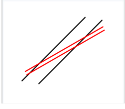
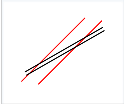
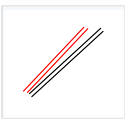
***Reporting:*** Visualize the data in one or several scatterplots: on the x-axis the ground-truth quality score and on the y-axis the main parameter(s) of the condition for the SV (e.g., bitrate). The size of the dots indicates the weight.

***Interpretation:*** The more widespread the data points over the quality range, the better and the more stable the training result. Higher density of data points and/or higher weights in certain quality ranges should be analyzed. They may either indicate redundant or overrepresented training data points or shortcomings of the algorithm in distinguishing between these closely related conditions. In order to further distinguish between such conditions, additional factors (e.g., quality indicators) may need to be added to the prediction algorithm.

***Example:*** Figure 12 is a typical result of this analysis. It shows the same prediction algorithm trained on three different training data sets (from left to right: $HRC_{1,2,3}$). Here, VQM is used as the ground-truth quality score and the "main parameter of the condition for the SV" is the quality control parameter being either QP (if<52) or bitrate (if>52). A major problem can be observed in the case on the right side. It is evident that the density of SV is higher on the low subrange and on the high subrange of the VQM scores. This is an indication of redundant HRCs in the set. On the other hand, the density of support vectors in the content-based (left) and the quality/bitrate-based (middle) training subsets is mostly uniform over the VQM scores.

The example shows the results for the pixel-based NR-VQA model. For the bit-stream-based NR VQA model, this strong difference cannot be observed, i.e. the SVs for each HRC subset cover different ranges of VQM score levels. This may be due to the usage of different factors, i.e. indicators, in the prediction algorithm. Notably the QP value is included as one of the quality indicators for the prediction algorithm. This leads to the effect that the output of the prediction algorithm is mostly determined by QP or at least it is more stable when the QP value is similar. Therefore, the SVM may avoid selecting redundant SVs.

TABLE V: List of interesting cases for analysis of the *data*-*C*CI, the cases for *model*-*C*CI are similar. Black lines indicate the CI on the training data. For simplicity it is assumed that these are fixed which is true in most practical cases. Red lines indicate the CI on the validation data.

| Case | Icon | Note |
|---|---|---|
| 1 |  | **Condition** $b = r = i$**:** Typical case for validating on the training data, this is considered the perfect fitting, i.e. all three areas are identical. Refer for example to the main diagonal $X(n, n)$ in Fig. 6. In this case, $G = \frac{1}{\max(b,r)}$. To compare between different models or data, the lower the $\max(b, r)$, i.e. the smaller the larger CI, the better. |
| 2 |  | **Condition** $r = i$**:** The validation data is better predicted than the training data and the CI lie completely within the boundaries of the trained model. This is likely to be a default of the validation data and thus goodness is reduced compared to Case 1. In this case, $G = \frac{r}{b^2}$. |
| 3 |  | **Condition** $b = i$**:** The validation data is less well predicted than the training data but the validation CI covers completely the training CI. This is considered a case of overfitting of the model and should thus be penalized compared to case 1. In this case, $G = \frac{b}{r^2}$. |
| 4 |  | **Condition** $b \approx r$**:** This is the typical case of slight deviation between training and validation. The goodness depends mainly on the intersection area. In this case, $G = \frac{i}{\max(r,b)^2}$. |
| 5 |  | **Condition** $b >> r$ **and** $b << r$**:** These cases indicate a larger misalignment either of the training CI or the validation CI with respect to the model fit and thus are a combination of Case 4 with the Cases 2 and 3 respectively. In these cases, the smaller intersection penalizes the goodness compared to Case 4 as the value of $i$ is smaller in $G = \frac{i}{\max(r,b)^2}$ |
| 6 |  | |
| 7 |  | **Condition** $i = 0$**:** This is the worst case, the validation data does not succeed in being predicted by the model, thus $G = 0$. Please note that this may also be an indication of a missing alignment between the training and validation data. An additional alignment step may be required in particular for models that were trained on different conditions (e.g. different video encoder). |

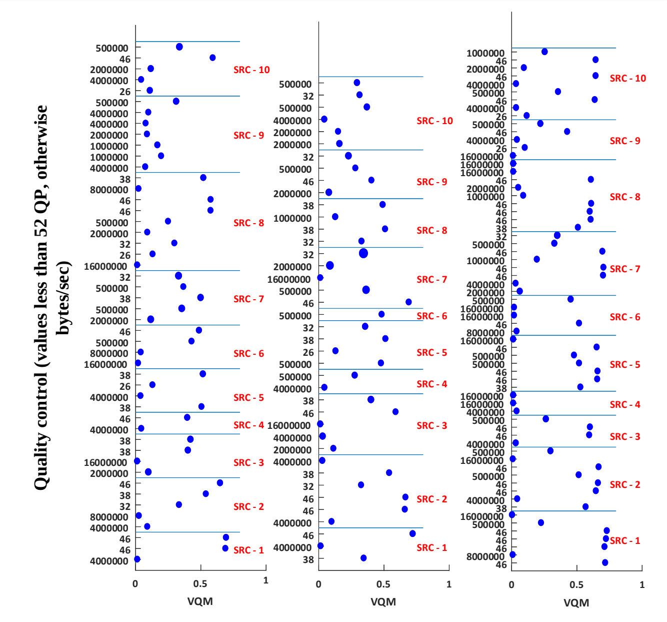


Fig. 12: The VQM quality score and the quality control parameter that are assigned to each SV of the following models: (left) $HRC_1$, (middle) $HRC_2$, and (right) $HRC_3$. The size of the dots indicates the weight of each SV.

## V. NO-REFERENCE IMAGE QUALITY MEASURE

We first present the methodology on an large-scale database that is evaluated objectively using the VQM algorithm. This ensures that sufficiently large subsets of the database can be extracted without database alignment issues. Then, in this Section we present a sample application on a typical quality assessment case: the performance evaluation on several subjective datasets that can be seen as subsets of possible images/videos.

As discussed in the introduction, in this section the aim is not to compare two NR IQA measures, but to see how they work with different image datasets that are different in size, content, and image distortion types and levels. In [27], a machine-learning-based NR IQA measure is introduced. It uses Single Value Decomposition (SVD) based features as input for the machine learning algorithm. Here, the machine learning can be seen as a feature pooling technique; 256 features are extracted from the distorted images. In [28], natural scene statistic (NSS) based features are extracted from patches that correspond to original images. The resulting Natural Image Quality Evaluator (NIQE) is an opinion and distortion unaware model. These NR IQA measures are trained with different datasets that differ in content, size, and number of distortions. These datasets are: the TID database [39] (68 reference and 1700 distorted images), the IVC database [40] (10 reference and 185 distorted images), the Toyama database [41] (14 reference and 168 distorted images), and the WIQ database [42] (7 reference and 80 distorted images). TID images are distorted using 17 distortion types. IVC database uses four distortion types while the Toyama database uses two distortion types. Finally, WIQ uses four distortion types.

TABLE VI: All HRC subsets ranks for each performance measure and for the pixel-based and the bit-stream-based NR VQA measures.

| **Performance measure** | **Pixel-based NR VQA (Proposed)** | | | | | **Bit-stream-based NR VQA** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Content** | **RD** | **Rand 1** | **Rand 2** | **Rand 3** | **Content** | **RD** | **Rand 1** | **Rand 2** | **Rand 3** |
| PCC Cross-dataset | 3 | 1 | 4 | 2 | 5 | 4 | 1 | 3 | 5 | 2 |
| PCC Leave-one-out | 2 | 1 | 5 | 3 | 4 | 2 | 4 | 3 | 5 | 1 |
| PCC Challenging HRCs | 2 | 1 | 3 | 4 | 5 | 1 | 2 | 3 | 5 | 4 |
| RMSE Cross-dataset | 3 | 1 | 4 | 2 | 5 | 5 | 1 | 4 | 3 | 2 |
| RMSE Leave-one-out | 2 | 1 | 4 | 3 | 5 | 3 | 5 | 4 | 1 | 2 |
| RMSE Challenging HRCs | 1 | 2 | 3 | 4 | 5 | 1 | 3 | 5 | 2 | 4 |
| $P^{\text{SRC}}_{\text{RPCA\_T}}(\frac{n}{m}, m)$, $P^{\text{SRC}}_{\text{RPCA\_V}}(\frac{n}{m}, m)$ | 3 | 1 | 4 | 2 | 5 | 1 | 2 | 1 | 1 | 3 |
| $P_{\text{DCI\_V}}(\delta, n)=\frac{i}{o}$ | 3 | 1 | 4 | 2 | 5 | 2 | 1 | 3 | 5 | 4 |
| $P^{\text{(b,r,i)}}_{\text{GModel}}$ | 3 | 1 | 4 | 2 | 5 | 2 | 1 | 3 | 4 | 5 |
| $P^{\text{(b,r,i)}}_{\text{GData}}$ | 3 | 1 | 4 | 2 | 5 | 5 | 1 | 2 | 3 | 4 |
| **Average** | **2.5** | **1.1** | 3.9 | 2.6 | 4.9 | **2.6** | **2.1** | 3.10 | 3.4 | 3.10 |

### *A. Evaluation methods*

The predicted quality is fitted using a 5-parameter logistic function as recommended in [43], and the correlation and the RMSE measures are size-weighted to calculate the average correlation and RMSE.

*1) PCC:* Table VII and Table VIII show 16 experiments and the corresponding PCC and RMSE. In the SVD-based and NIQE models, the TID dataset performs better than the other datasets, i.e., the WIQ, IVC, and Toyama datasets, respectively. It seems that the WIQ dataset is not large enough to be used to build a NIQE model since it shows a very low correlation when it is tested on the training data. Hence, this dataset will be excluded in the final ranking order in the overall ranking Table IX. Regarding the distortion types that are challenging for the models, TID and the WIQ datasets show that there are, indeed, distortion types and levels that are challenging for other models. This result is expected since TID has very different distortion types and WIQ has different distortions than the ones in the IVC and Toyama databases.

*2) RMSE:* With the RMSE performance measure in SVD-based IQA, WIQ performs better than the others. Then TID, IVC and Toyama come next in order. For the case of the NIQE model, the WIQ comes first, and then IVC, Toyama, and TID come next in order. Regarding the distortion types that are challenging for the models, IVC contains distortion types that are often challenging for other models.

*3) $P^{SRC}_{RPCA\_T}$ and $P^{SRC}_{RPCA\_V}$:* This measure is not applied here since the distortions are not similar in terms of distortion type or distortion level.

*4) $P^{(b,r,i)}_{GData}$:* As discussed in Section IV-C1, this measure tries to provide an absolute number for the prediction performance of a trained model on a validation dataset taking into consideration the training dataset. The first column of Figure 13 shows the $P^{\text{(b,r,i)}}_{\text{GData}}$ of both image quality assessment. In the SVD-model and the NIQE-model, IVC and TID datasets have the ability to predict the quality within the confidence intervals corridors respectively.

*5) $P^{(b,r,i)}_{GModel}$:* This measure reports the model's stability when using different datasets, Section IV-C2. This is done by measuring the overlap between the two models ($G$), Table V. As it can be seen in the second column of Figure 13, Toyama and IVC datasets have a higher stability in SVD and NIQE models respectively. In the SVD-based, the $G$ value of the WIQ model is very small compared to the others: this is due to the bad fitting model for the training data, i.e. the black area is very large. Therefore, this dataset is not suitable for training for both models.

*6) Performance comparisons:* As discussed and observed in the previous sections, all the aforementioned performance measures give different results for different image datasets. In this subsection, all of them are put together in order to judge the image datasets. A rank-order technique has been applied in order to get a final score for each image dataset. Since we have 3 image datasets (WIQ has been excluded), each set has an order number for each performance measure discussed in this paper and then a comparison between the two NR IQA measures is presented. Table IX shows all the ranks for each performance measure and for the SVD-based and the

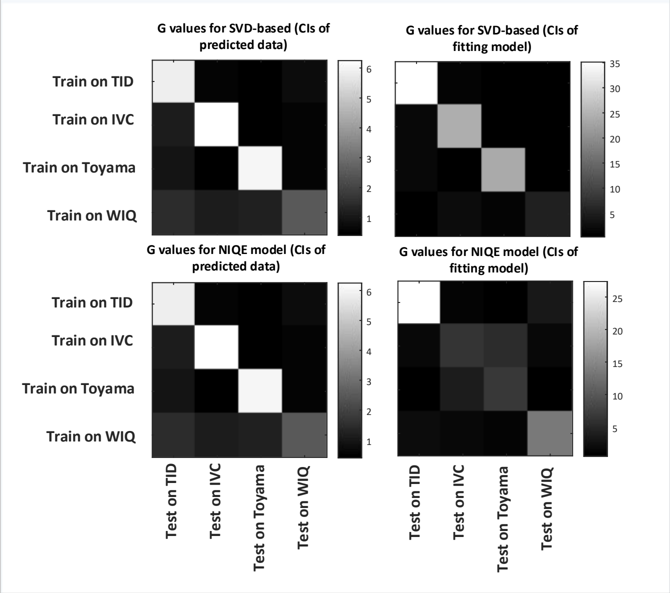


Fig. 13: The G values for the CIs analysis for the SVD-based and NIQE-model NR IQA models.

TABLE VII: The PCC and the RMSE for the 16 experiments that are trained and tested with source 4 image datasets of SVD-based NR IQA measure

| | | Tested on | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PCC | | | | RMSE | | | |
| | | TID | IVC | Toyama | WIQ | TID | IVC | Toyama | WIQ |
| Trained on | TID | 0.95 | 0.81 | 0.72 | 0.7 | 0.05 | 0.17 | 0.21 | 0.16 |
| | IVC | 0.68 | 0.99 | 0.62 | 0.57 | 0.11 | 0.04 | 0.25 | 0.19 |
| | Toyama | 0.53 | 0.74 | 0.99 | 0.47 | 0.13 | 0.2 | 0.04 | 0.2 |
| | WIQ | 0.62 | 0.58 | 0.47 | 0.86 | 0.12 | 0.25 | 0.28 | 0.12 |

TABLE VIII: The PCC and the RMSE for the 16 experiments that are trained and tested with source 4 image datasets of NSS-based NR IQA measure

| | | Tested on | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | PCC | | | | RMSE | | | |
| | | TID | IVC | Toyama | WIQ | TID | IVC | Toyama | WIQ |
| Trained on | TID | 0.4 | 0.62 | 0.82 | 0.15 | 0.14 | 0.24 | 0.18 | 0.22 |
| | IVC | 0.5 | 0.69 | 0.71 | 0.22 | 0.13 | 0.22 | 0.22 | 0.22 |
| | Toyama | 0.4 | 0.62 | 0.81 | 0.15 | 0.14 | 0.24 | 0.18 | 0.23 |
| | WIQ | 0.57 | 0.62 | 0.77 | 0.17 | 0.12 | 0.24 | 0.2 | 0.23 |

NIQE model NR IQA measures. As can be seen from the table, there is no clear indication about which dataset can be used as a generalized dataset. In the SVD-based model, the TID dataset is the winner, while in the NIQE model the IVC dataset is the winner. This observation already considers the exclusion of the WIQ dataset due to its limitation in size and in types of distortions that are not in common with the other datasets. In conclusion, we recommend that different datasets should be tested with different performance measures when a new objective NR IQA tool is introduced.

TABLE IX: All image datasets ranks for each performance measure and for the SVD-based and NIQE NR VQA measures.

| Evaluation | SVD-based NR IQA | | | NIQE NR IQA | | |
|---|---|---|---|---|---|---|
| | TID | IVC | Toyama | TID | IVC | Toyama |
| PCC Cross-dataset | 1 | 2 | 3 | 1 | 2 | 3 |
| PCC Challenging HRCs | 1 | 3 | 2 | 1 | 2 | 3 |
| RMSE Cross-Dataset | 1 | 2 | 3 | 3 | 1 | 2 |
| RMSE Challenging HRCs | 1 | 3 | 2 | 3 | 1 | 2 |
| $P^{(b,r,i)}_{\mathrm{GModel}}$ | 3 | 2 | 1 | 2 | 1 | 3 |
| $P^{(b,r,i)}_{\mathrm{GData}}$ | 3 | 1 | 2 | 1 | 2 | 3 |
| Average | **2.17** | 3.00 | 3.00 | 2.33 | **2.00** | 3.5 |

## VI. CONCLUSION

In this paper we discussed the effects of different training and validation data sets on the performance of objective quality measurement algorithms. As an example study, we used five subsets for training and validation; two were targeted towards different goals, three were random. In the study, two NR VQA algorithms with typical quality indicators were trained by SVR.

We analyzed the outcome of this widespread approach with state-of-the-art performance measures and identified important shortcomings. We therefore proposed several novel performance measures in three categories: The first category analyzes the residual errors to find the systematic redundancies in the training and evaluation subsets. The second category provides insight on the training by using the confidence intervals of models fitting and the confidence interval of the predicted data. The third category is specific to SVR and analyzes the density of SV over the quality range.

An example study on image quality databases with subjective scores illustrates the usefulness of the performance measures.

The newly proposed performance measures are presented such that they can easily be reproduced. It would be very beneficial to report such measures in future proposals of video quality assessment algorithms in order to enable an in-depth analysis and a comparison across the proposals of different authors in the domain who often use varying datasets for training and validation.

Further performance measures may be required, in particular when training with other machine learning algorithms such as deep-learning.

## ACKNOWLEDGMENT

This work is supported by the Marie Skłodowska-Curie under the PROVISION (PeRceptually Optimised VIdeo CompresSION) project bearing Grant Number 608231 and Call Identifier: FP7-PEOPLE-2013-ITN. Some of the computational resources were provided by HPC@POLITO (http://www.hpc.polito.it).

## REFERENCES

[1] K. Gu, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 9, pp. 1480–1494, 2015.

[2] K. Gu, M. Liu, G. Zhai, X. Yang, and W. Zhang, "Quality assessment considering viewing distance and image resolution," *IEEE Transactions on Broadcasting*, vol. 61, no. 3, pp. 520–531, 2015.

[3] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, "Objective video quality assessment methods: A classification, review, and performance comparison," *IEEE Transactions on Broadcasting*, vol. 57, no. 2, pp. 165–182, 2011.

[4] G. Zhai, J. Cai, W. Lin, X. Yang, W. Zhang, and M. Etoh, "Cross-dimensional perceptual quality assessment for low bit-rate videos," *IEEE Transactions on Multimedia*, vol. 10, no. 7, pp. 1316–1324, 2008.

[5] X. Min, K. Gu, G. Zhai, M. Hu, and X. Yang, "Saliency-induced reduced-reference quality index for natural scene and screen content images," *Signal Processing*, vol. 145, pp. 127–136, 2018.

[6] J. Ostermann, J. Bormans, P. List, D. Marpe, M. Narroschke, F. Pereira, T. Stockhammer, and T. Wedi, "Video coding with H.264/AVC: tools, performance, and complexity," *IEEE Circuits and Systems Magazine*, vol. 4, no. 1, pp. 7–28, First 2004.

[7] M. Yuen and H. Wu, "A survey of hybrid MC/DPCM/DCT video coding distortions," *Signal Processing*, vol. 70, no. 3, pp. 247 – 278, 1998. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0165168498001285

[8] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2049–2062, 2018.

[9] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 525–535, 2012.

[10] M. Naccari, M. Tagliasacchi, and S. Tubaro, "No-reference video quality monitoring for H.264/AVC coded video," *IEEE Transactions on Multimedia*, vol. 11, no. 5, pp. 932–946, 2009.

[11] L. Anegekuh, L. Sun, E. Jammeh, I.-H. Mkwawa, and E. Ifeachor, "Content-based video quality prediction for HEVC encoded videos streamed over packet networks," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1323–1334, 2015.

[12] A. Hameed, R. Dai, and B. Balas, "A decision-tree-based perceptual video quality prediction model and its application in FEC for wireless multimedia communications," *IEEE Transactions on Multimedia*, vol. 18, no. 4, pp. 764–774, 2016.

[13] Y. Li, A. Markopoulou, J. Apostolopoulos, and N. Bambos, "Content-aware playout and packet scheduling for video streaming over wireless links," *IEEE Transactions on Multimedia*, vol. 10, no. 5, pp. 885–895, 2008.

[14] Y.-H. Sung and J.-C. Wang, "Fast mode decision for H.264/AVC based on rate-distortion clustering," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 693–702, 2012.

[15] N. Bouten, S. Latré, J. Famaey, W. Van Leekwijck, and F. De Turck, "In-network quality optimization for adaptive video streaming services," *IEEE Transactions on Multimedia*, vol. 16, no. 8, pp. 2281–2293, 2014.

[16] M. Barkowsky, E. Masala, G. Van Wallendael, K. Brunnström, N. Staelens, and P. Le Callet, "Objective video quality assessment – towards large scale video database enhanced model development," *IEICE Transactions on Communications*, vol. E98-B, no. 1, pp. 2–11, Jan. 2015.

[17] A. Aldahdooh, E. Masala, G. Van Wallendael, and M. Barkowsky, "Framework for reproducible objective video quality research with case-study on PSNR implementations," *Digital Signal Processing*, vol. 77, pp. 195–206, 2018.

[18] M. Shahid, A. Rossholm, B. Lövström, and H.-J. Zepernick, "No-reference image and video quality assessment: a classification and review of recent approaches," *EURASIP Journal on Image and Video Processing*, vol. 2014, no. 1, p. 40, 2014. [Online]. Available: http://dx.doi.org/10.1186/1687-5281-2014-40

[19] M. A. Aabed and G. AlRegib, "No-reference quality assessment of HEVC videos in loss-prone networks," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2015–2019.

[20] J. Guo, K. Zheng, G. Hu, and L. Huang, "Packet layer model of HEVC wireless video quality assessment," in *2016 11th International Conference on Computer Science Education (ICCSE)*, Aug 2016, pp. 712–717.

[21] M. Shahid, J. Panasiuk, G. V. Wallendael, M. Barkowsky, and B. Lövstöm, "Predicting full-reference video quality measures using HEVC bitstream-based no-reference features," in *Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, May 2015, pp. 1–2.

[22] K. Izumi, K. Kawamura, T. Yoshino, and S. Naito, "No reference video quality assessment based on parametric analysis of HEVC bitstream," in *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, Sept 2014, pp. 49–50.

[23] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Transactions on Image Processing*, vol. 26, no. 11, pp. 5462–5474, 2017.

[24] A. Aldahdooh, E. Masala, O. Janssens, G. Van Wallendael, and M. Barkowsky, “Comparing simple video quality measures for loss-impaired video sequences on a large-scale database,” in *Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6.

[25] Y. Rai, A. Aldahdooh, S. Ling, M. Barkowsky, and P. Le Callet, “Effect of content features on short-term video quality in the visual periphery,” in *Multimedia Signal Processing (MMSP), 2016 IEEE 18th International Workshop on*. IEEE, 2016, pp. 1–6.

[26] A. Aldahdooh, M. Barkowsky, and P. Le Callet, “Proof-of-concept: role of generic content characteristics in optimizing video encoders,” *Multimedia Tools and Applications*, vol. 77, no. 13, pp. 16 069–16 097, Jul 2018. [Online]. Available: https://doi.org/10.1007/s11042-017-5180-1

[27] M. Narwaria and W. Lin, “SVD-based quality metric for image and video using machine learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 2, pp. 347–364, April 2012.

[28] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ”completely blind” image quality analyzer,” *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 209–212, March 2013.

[29] A. Aldahdooh, M. Barkowsky, and P. L. Callet, “Content-aware adaptive multiple description coding scheme,” in *2016 IEEE International Conference on Multimedia Expo Workshops (ICMEW)*, July 2016, pp. 1–6.

[30] H. Yi, D. Rajan, and L.-T. Chia, “A new motion histogram to index motion content in video segments,” *Pattern Recognition Letters*, vol. 26, no. 9, pp. 1221–1231, 2005.

[31] M. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Transaction on Broadcasting*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[32] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for support vector machines,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[33] C. Zhang, C. Liu, X. Zhang, and G. Almpanidis, “An up-to-date comparison of state-of-the-art classification algorithms,” *Expert Systems with Applications*, vol. 82, pp. 128 – 150, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417417302397

[34] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[35] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 785–794.

[36] G. Vandewiele, P. Colpaert, J. Van Herwegen, O. Janssens, R. Verborgh, E. Mannens, F. Ongenae, and F. De Turck, “Predicting train occupancies based on query logs and external data sources,” in *Proc. of the 26th International World Wide Web Conference: 7th International Workshop on Location and the Web*, 2017.

[37] O. Janssens, N. Noppe, C. Devriendt, R. Van de Walle, and S. Van Hoecke, “Data-driven multivariate power curve modeling of offshore wind turbines,” *Engineering Applications of Artificial Intelligence*, vol. 55, pp. 331–338, 2016.

[38] X. He, J. Pan, O. Jin, T. Xu, B. Liu, T. Xu, Y. Shi, A. Atallah, R. Herbrich, S. Bowers *et al.*, “Practical lessons from predicting clicks on ads at Facebook,” in *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*. ACM, 2014, pp. 1–9.

[39] N. Ponomarenko, V. Lukin, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, “Color image database for evaluation of image quality metrics,” in *IEEE 10th Workshop on Multimedia Signal Processing*, Oct 2008, pp. 403–408.

[40] P. Le Callet and F. Autrusseau, “Subjective quality assessment IRCCyN/IVC database,” 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.

[41] “MICT image quality evaluation database.” [Online]. Available: http://mict.eng.u-toyama.ac.jp/mictdb.html

[42] U. Engelke, M. Kusuma, H.-J. Zepernick, and M. Caldera, “Reduced-reference metric design for objective perceptual quality assessment in wireless imaging,” *Signal Processing: Image Communication*, vol. 24, no. 7, pp. 525 – 547, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596509000836

[43] VQEG, “Final report from the video quality experts group on the validation of objective models of video quality assessment, Phase II,” 2003, available: http://www.vqeg.org.