Blind Image Quality Assessment of Smartphone-captured Images in the Wild

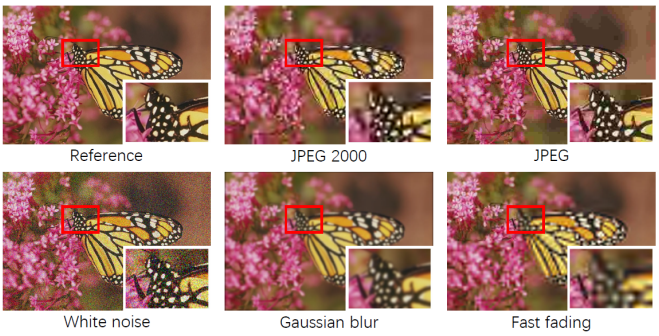
Tejas Chandra Mohan
University of Twente, M-Emsys
Enschede, The Netherlands

***Abstract*— Real-world images captured using an imaging device suffers from distortion while capturing, processing, or storage. These distortions in images affect their visual quality, rendering them unusable for further processing. This thesis concentrates on images captured by a smartphone from behind a car's windshield. The objective is to classify these images into good quality and bad quality employing deep learning models focusing on Image Quality Assessment. This paper provides an overview of recent developments in Blind Image Quality Assessment (BIQA) using deep learning and the available standard datasets.**

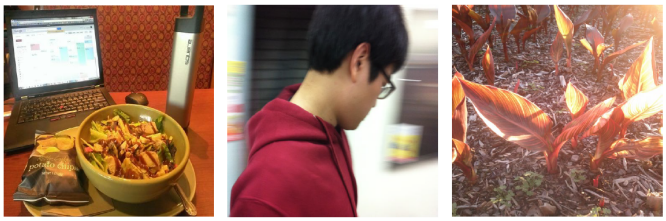
Specifically, three recent BIQA models are selected to evaluate these images and quantify them as good and bad based on their image quality. Further research is conducted on an ensemble of these BIQA models for the same task.

Later, a classification approach is explored consisting of three transfer learning models to classify the images as good quality and bad quality. An ensemble comprising of these models is built. The test results show that the ensemble combination comprising of two BIQA models delivers the highest accuracy towards rightly classifying images as good quality and bad quality.

***Index Terms*— Blind image quality assessment, smartphone images, convolutional neural networks, deep learning, multi-model techniques, ensemble, transfer learning, multi-class classification, binary classification, good image quality, bad image quality**



(a) Synthetic distortion in images



(b) Authentic distortion in images

Fig. 1: Examples of Synthetic and Authentic Distorted Images [2]

I. INTRODUCTION

This thesis work is in collaboration with 'CamenAI' [1], a start-up in Utrecht, The Netherlands. CamenAI is a young start-up focusing on computer vision techniques to make the environment a cleaner, better place. They gather images utilizing a camera attached to the car's windshield. Once the images are gathered, they are manually inspected for image quality. The images that are considered good quality are retained. The retained images are subjected to anonymization where sensitive information contained in the images, such as the human face(s) or license plates of vehicles, are blurred. The set of anonymized images is inspected to form the dataset for algorithms such as garbage detection, crack detection in asphalts, over-grown plant maintenance, and damaged traffic signs. The results from these algorithms are sent to the concerned party who monitors the environment. During image acquisition, a vehicle drives through a given city on a given day and captures images of the environment. The images are gathered in hundreds when a vehicle drives through inspection areas. The quality of gathered images depends on several factors such as time of the day (day or night), weather (sunny or cloudy), reflections on the car's dashboard, etc. These affect an image's visual quality.

The task of image quality assessment becomes vital for two main reasons (a) Since there are many images of the same inspection area, it is essential to filter them into good quality and bad quality. Manual inspection proves cumbersome in this real-time setup. Thus, an automation approach is preferred (b) The set of images having good quality can be retained and be subjected to anonymization. The anonymized images can be used to form the dataset to the algorithms highlighted above. Thus, image quality assessment becomes a necessary precursor.

The quality of an image is highly determined by the

distortions it contains. Typically, these distortions [2] can be categorized as (a) Synthetic Distortions and (b) Authentic Distortions. Synthetic distortions are laboratory-induced distortions using a clean reference image to study the image quality with respect to the clean reference image. Images are synthetically distorted by adding one type of distortion such as White Noise, Gaussian Blur, etc. On the other hand, authentic distortions are more realistic in real-world captured images due to overexposure, underexposure, motion blurring, framing, etc. Thus, authentic distortions form the main element that affects image quality in real-world captured images. An example is illustrated in figure1.

The process of Image quality assessment [19] branches into two categories (a) Subjective image quality assessment and (b) Objective image quality assessment. Subjective image quality assessment involves large viewers to gather human opinion scores for the images. This large group of viewers provide each image with a rating. The mean of ratings provided forms the basis of a score that corresponds to the quality of that image. Employing this approach to rate images frequently in a real-time setup is time-consuming and deemed expensive.

Objective image quality assessment [19] is the process of extracting features in an image, analyzing them and measuring the degree of distortion. It is an algorithm-driven approach. Such algorithms typically result in an image quality score as its output. Objective image quality assessment is further categorized into three types, depending on the availability of a reference image [19]. They are (a) Full-Reference Image Quality Assessment (FR-IQA), (b) Reduced-Reference Image Quality Assessment (RR-IQA) and (c) No-Reference or Blind Image Quality Assessment (NR-IQA or BIQA). As the name suggests, in FR-IQA, the algorithm considers the 'full' information available of both the reference image and distorted image to predict the quality of an image based on the differences between them. On the contrary, due to the unavailability of a clean reference image, BIQA employs algorithms to predict the quality of an image based on the available information of the distorted image only. RR-IQA is an intermediate where algorithms operate based on the 'partial' information of the reference image and 'full' information of the distorted image.

This thesis concentrates on Blind Image Quality Assessment (BIQA) due to the non-availability of reference images. A smartphone shoots the images at different locations in the Netherlands at different times of the day. Hence, fixing a reference image is not feasible. The objective is to perform Blind Image Quality Assessment (BIQA) on the CamenAI [1] dataset to classify images into good and bad based on their image quality. The idea being the set of bad quality images will be deleted, and the set of good quality images will be retained. A detailed overview that determines whether an image is a good quality or bad quality is highlighted in subsection IV-B. Typically, an image is considered good quality if its quality score lies on the higher end of the annotation scale. Three recent BIQA models - NIMA [3], DBCNN [5], UNIQUE [4]- were selected for evaluation on the CamenAI dataset. An ensemble employing different combinations of the three BIQA models is proposed and evaluated on the CamenAI dataset. The BIQA models are regression-based deep learning models, pre-trained on existing standard BIQA datasets. Later, a classification approach is explored towards classifying CamenAI dataset images into good and bad quality. To this end, three transfer learning models bearing the same backbone as NIMA, DBCNN and UNIQUE has been selected, respectively. These models are fine-tuned on the CamenAI dataset to perform classification. An ensemble of these models is built to study its effect on classifying images into good and bad quality. A multi-class classification, as well as binary classification, is explored. Multi-class classification is employed because images may lie on the borderline of good quality and bad quality. It is unknown whether such images are classified as good quality or bad quality in such a case.

A. Research Questions

This thesis will address five research questions as follows:

1) What are the available BIQA models and training datasets used to benchmark the performance of BIQA models?
2) How does the CamenAI [1] dataset compare to the standard datasets in terms of annotations, resolutions, and score distribution?
    a) In what manner can the CamenAI dataset be annotated for the task of BIQA?
3) To what extent does cropping the images for dashboard interference affect the score predicted by BIQA regression models such as NIMA [3], UNIQUE [4]and DBCNN [5]?
4) What effect does different ensembling combinations of NIMA [3], DBCNN [5] and UNIQUE [4] have on the accuracy of their performances on the CamenAI dataset?
    a) How does NIMA, DBCNN and UNIQUE perform individually on the CamenAI dataset concerning accuracy?
    b) How do we normalize NIMA, DBCNN and UNIQUE results to ensure the same output score range across these models?
    c) What is the trade-off between the number of models combined in the ensemble concerning accuracy and processing time on the CamenAI dataset?
    d) What is the best and worst combination of the ensemble surrounding accuracy on the CamenAI dataset?
5) What effect does different ensembling combinations of transfer learning models have on the accuracy of their performances on the CamenAI dataset?
    a) What are the selected transfer learning models?
    b) How do these models perform for multi-class classification and binary classification on the CamenAI dataset concerning accuracy?

c) What is the trade-off between the number of models combined in the ensemble concerning accuracy and processing time on the CamenAI dataset for a multi-class classification?
d) What is the trade-off between the number of models combined in the ensemble concerning accuracy and processing time on the CamenAI dataset for a binary classification?

B. Scientific and Technical Contributions

The CamenAI [1] dataset comprising of 4,780 images have been annotated independently by the author. An ensemble approach comprising of different combinations of the three BIQA models is proposed. Results show that the ensemble combination of NIMA [3] and UNIQUE [4] deliver the highest accuracy towards classifying the images into good and bad quality.

II. Background

This section provides an overview of the existing BIQA models and the datasets available for the task of BIQA. When we consider the standard datasets available for the task of image quality assessment, they fall under two categories [2]:

- Synthetic distortion dataset
- Authentic distortion dataset

Images in the synthetic distortion datasets are built by introducing one type of noise to an available high-quality image in a controlled laboratory setup. For instance, considering a high-quality image 'A', we can make three copies of this image by adding one type of distortion such as White Noise, Gaussian Blur, and Fading. However, these images do not correctly model authentic distortions since real-world images captured by a smartphone contain a combination of multiple synthetic distortions. Hence, this section first describes the available, standard datasets that significantly concentrate on authentic distortions. Following this will be a comparison of different BIQA models, focusing on performance, evaluation metrics, and datasets used by these respective models to benchmark their performances.

A. Authentic Distortion datasets

The datasets available for authentically distorted images are as follows:

1) **LIVE Challenge**[6]: This dataset comprises 1,162 authentically distorted images taken by various smartphones. These images do not contain any reference images. This dataset covers a wide range of distortions images undergo such as motion blurring, overexposure, underexposure, noise, and JPEG compression. The images are rated in the range [0,100] employing crowdsourcing of 8100 humans. The types of images in this dataset include human faces, animals, natural scenes, man-made objects, close-up shots, wide-angle shots, and shots without the object of interest.
2) **KONIQ-10K** [7]: It is the largest dataset for image quality assessment, comprising 10,073 images. The images are captured using imaging devices such as DSLR and smartphones. A total of 1,459 crowd workers were responsible for providing image quality ratings or annotations to this dataset. The images are annotated with human Mean Opinion Score (MOS) in the range [1,5]. The images found in this dataset are similar to the LIVE WILD dataset, as discussed above.
3) **BID** [8] - A relatively more minor dataset that contains 586 realistic blurred images of varying resolution ranges: 1280×960 to 2272×1704, acquired using a single DSLR. The images are rated in a laboratory setup, and the scores range between [0,5]. These images highlight realistic scenarios.
4) **CID2013** [9] - This dataset consists of 480 authentically distorted images, captured by 79 different imaging devices such as smartphones, DSC, and DSLR. Crowd-sourcing is employed to evaluate these images, ranging between [0,100].
5) **SPAQ** [10] - This dataset is introduced by Fang, Yuming, et al [20]. It consists of 11,125 smartphone captured images by different smartphones. In addition, each image is accompanied by an EXIF tag that contains details about the captured scene.

The summary of the compared datasets for authentic distortions is found in table I.

B. Existing BIQA models

Here, we present different BIQA models and highlight their novel features with respect to image quality assessment.

NIMA [3]: A novel approach of IQA, introduced by Google. Here, a CNN is used to predict the distribution of human-opinion scores. Earth mover's distance is used as a loss function to operate on the distribution of the ground-truth scores and the distribution of the predicted image quality scores.

Meta-IQA[11]: This approach employs the concept of 'deep-meta learning' for IQA. Here, a model is pre-trained to operate on known distortions, and meta-learning is used to gain knowledge about these distortions. The resulting model is refined to operate on images containing unknown distortions.

DB-CNN[5]: It utilizes two CNNs that handle synthetic and authentic distortions. The features resulting from these neural networks are pooled in a bi-linear fashion to represent the final image quality score.

BIQA: Self-adaptive hyper network[12]: A BIQA model is proposed aimed at authentic distorted image dataset. The architecture follows a 'self-adaptive hyper network' for quality prediction parameters. In addition, it uses a 'local distortion module' to capture distortions.

NRIQA using Contrast Enhancement[13]: It employs a BIQA model that is tested on two contrast-distortion dataset i.e., CID2013 and CCID2014. The approach generates an enhanced image (from the original image) that acts as a reference. Based on this, SSIM, Histogram based entropy, and cross-entropy is obtained. The resulting features are used by a regression module to predict the final score.

TABLE I: Available Authentic Distortion datasets

Database	#images	#cameras	Type of cameras	Subjective environment	Annotation	Range
LIVE Challenge [6]	1,162	15	DSLR/DSC/Smartphones	Crowd-sourcing	MOS, SD	[0,100]
KONIQ-10K [7]	10,073	N/A	DSLR/DSC/Smartphones	Crowd-sourcing	MOS, SD	[1,5]
BID [8]	585	1	DSLR	Laboratory	MOS, SD	[0,5]
CID2013 [9]	480	79	DSLR/DSC/Smartphones	Laboratory	MOS, SD	[0,100]
SPAQ [10]	11,125	66	Smartphones	Laboratory	MOS, SD	[0,100]

TABLE II: Recent BIQA models and their performances on the standard datasets

Study	Core Features	datasets used		Performance Results	
		Synthetic	Authentic	SRCC	PLCC
NIMA [3] Talebi et al	Distribution of human opinion scores	AVA TID2013	LIVE-C	0.637	0.698
Meta-IQA [11] Zhu et al	Deep Meta-learning	TID2013 KADID-10K	CID-2013 LIVE-C KONIQ-10K	0.766 0.802 0.850	0.784 0.835 0.887
DB-CNN [5] Zhang et al	Deep bi-linear network	LIVE CSIQ TID2013 LIVE-MD	LIVE-C	0.851	0.869
BIQA: Self-adaptive hyper network [12] Shaolin et al	Self-adaptive hyper network	LIVE CSIQ	LIVE-C KONIQ-10K BID	0.859 0.906 0.869	0.882 0.917 0.878
NRIQA using Contrast enhancement [13] Yan et al	Contrast Enhancement	CSIQ TID2013	CID2013 CCID2014	0.8934 0.8363	0.8960 0.8675
NRIQA using multi-pooled inception features [14] Varga et al	Extract visual features by multi-pooling	N/A	KONIQ-10K LIVE-C	0.925 0.826	0.928 0.857
Perceptual NRIQA [20] Fang et al	SPAQ dataset + EXIF data for all images	CID2013	BID LIVE-C KONIQ-10K SPAQ	0.926	0.932
NRIQA using global statistical features [15] Varge et al	Global statistical features in images	CSIQ MDID KADID-10K	LIVE-C KONIQ-10K	0.595 0.752	0.618 0.784
Uncertainity-Aware BIQA [4] Zhang et al	IQA dataset combination + Cross-distortion dataset tackling	CSIQ LIVE KADID-10K	LIVE-C KONIQ-10K BID	0.854 0.896 0.858	0.890 0.901 0.873
BIQA using high-level semantics [16] Li et al	Exploits high-level semantics during feature extraction	LIVE TID2008	LIVE-C BID	0.8130 0.8269	0.8313 0.8401
BIQA using controllable list-wise ranking [17] Zhao et al	Extending Live-C dataset + Controllable list-wise ranking	LIVE CSIQ TID2013	LIVE-C	0.779	0.828
DeepRN [21] Varga et al	Content preserving architecture + Pyramid pooling (Images can be fed directly without resizing)	LIVE	LIVE-C KONIQ-10K	0.91 0.92	0.93 0.95
Learning for BIQA [18] Zhang et al	BIQA model for cross-distortion scenario + Continuous quality annotation	LIVE CSIQ TID2013	LIVE-C KONIQ-10K BID	0.851 0.894 0.863	

NRIQA using Multi-Pooled Inception Features[14]: Another approach aimed at authentic distorted image dataset that utilizes 'Global Average Pooling (GAP)' to extract image features. The whole image is directly fed into the CNN (and not patches of images).

Perceptual NRIQA[10]: It proposes the 'SPAQ' image dataset, uses EXIF data of images to build a BIQA model for better prediction of image quality. Again, this approach is significantly aimed towards authentic distorted image dataset.

NRIQA using Global Statistical Features[15]: In this approach, the focus is given towards extracting features from the images globally. It employs a 132-Dimensional feature vector to extract information from images, which in turn is used to train a BIQA model.

Uncertainty-aware BIQA[4]: Here, images from multiple datasets are sampled into image pairs to drive the BIQA model. The BIQA model is optimized for its performance by enforcing fidelity loss over the sampled pairs of images.

BIQA using high-level semantics[16]: In this approach, the input image is decomposed into multiple overlapping patches and high-level features of each patch are extracted. The features extracted from various image patches are pooled. A linear regression model is used to predict the image quality score.

BIQA using Controllable list-wise ranking[17]: This approach simulates the images in LIVE-Challenge dataset. A controllable list-wise ranking function is proposed to achieve consistency between predicted scores and the ground truth.

DeepRN[21]: It uses a fine-tuned residual deep learning network, along with Pyramid pooling. Attention is given to predict output quality as a distribution of scores (and not just MOS). This approach handles input images of any given size.

Learning for BIQA[18]: This approach tackles the cross-distortion scenario faced while testing the BIQA model over synthetic and authentic distortions. Image pairs are created from individual databases. It uses continuous quality annotation, where Mean Opinion Score (MOS) and the standard deviation are used to give the probability score.

Existing BIQA models are evaluated on Spearman's Rank Order Correlation Coefficient (SROCC) and Pearson's Linear Correlation Coefficient (PLCC). This metric was considered to shortlist the BIQA models. However, in the evaluation of experiments, this metric will not be used.

SROCC [22] measures prediction monotonicity and PLCC [23] measures prediction accuracy. These values typically lie between [-1,1]. Closer the value to 1, the better is the correlation between the two variables. SROCC tells us how good the relationship between the two sets of data is, while PLCC tells us the linear correlation between the two sets of data (prediction ground-truth).

III. Design Choices

This thesis takes into consideration certain design choices highlighted in this section.

A. Shortlisted BIQA Models

Among the different BIQA models discussed in subsection II-B, NIMA [3], DBCNN [5] and UNIQUE [4] have been shortlisted to be tested on the CamenAI [1] dataset. These models are regression-based models. They output either a score or a score with its standard deviation. The motivation behind shortlisting is explained below.

NIMA is the state-of-the-art BIQA model by Google. This approach is the first to predict image quality score as a distribution of image ratings since it outputs an image quality score and a standard deviation from this score. The output score range of NIMA lies in the range [1,10]. An image with a score 10 indicates the image with the highest image quality.

DBCNN stands out in architecture than the other discussed BIQA models. It comprises two CNNs to handle synthetic distortion and authentic distortion, respectively. The final image quality score results from bi-linear pooling from these two CNNs. The output score range for DBCNN is [0,10], where an image with a score of 10 indicates the image with the highest image quality. DBCNN results in only an image quality score as its output.

On the other hand, UNIQUE employs a training strategy unique to other BIQA models. UNIQUE is pre-trained on six image quality assessment datasets with an equal distribution of 3 synthetic and authentic datasets. The output score range for UNIQUE is [-3,3], where an image with a score of 3 indicates the image with the highest image quality. UNIQUE results in an image quality score and standard deviation as its output.

A summary that highlights the novel features of the considered BIQA models is shown in table III. In addition, code accessibility and python implementation were also considered while shortlisting these 3 BIQA models. This ensures feasibility to integrate them into the corporate pipeline if needed.

TABLE III: Shortlisted BIQA models and their salient features

BIQA Model	Salient Features
NIMA [3]	Employs 3 baseline architectures (MobileNet, VGG16, InceptionNet-v2) Predicts distribution of ratings of an image State-of-the-art by Google
UNIQUE [4]	Trained equally on 6 IQA datasets (3 synthetic and 3 authentic datasets) Employs pairwise ranking in training
DBCNN [5]	Deep bilinear architecture using 2 CNNs 1 CNN for synthetic distortion 1 CNN for authentic distortion Bilinear pooling from 2 CNNs to predict image quality

TABLE IV: Round-off from Mean Opinion Scores to Opinion Scores

Mean Opinion Scores (MOS) Range	Opinion Scores
[0.5x - 1.4x]	1
[1.5x - 2.4x]	2
[2.5x - 3.4x]	3
[3.5x - 4.4x]	4
[4.5x - 5.4x]	5

B. Annotation Technique for CamenAI dataset

The different authentic datasets presented in table I follow an annotation technique where a large group of people provide image quality ratings for each image. This, in turn, forms the basis of 'Mean Opinion Scores (MOS)' and Standard deviation (SD) for each image. MOS can take different ranges and is a free choice to select the range as [1,5] or [1,10] etc.

The annotation MOS range selected for the CamenAI [1] dataset is [1,5]. Due to the absence of a large group to annotate the CamenAI dataset, the author has annotated the images independently. Since it is one person annotating the images, it is impossible to represent the ratings in MOS and SD. Thus, the images have been annotated as 'Opinion Scores' only. This means the MOS range [1,5] has been rounded-off to [1,2,3,4,5], representing the opinion scores.

For example, in the MOS range [1,5], the range [1,2) indicates all images with ratings of 1.0x until 1.9x that are considered to be poor quality. In the CamenAI dataset, poor quality images take the opinion score as '1'. The round-off from the MOS range [1,5] to opinion scores [1,2,3,4,5] to annotate the CamenAI dataset is shown in table IV. For the opinion scores [1,2,3,4,5], images with scores 1 and 2 are considered poor image quality that should be discarded. Images with scores 3 or 4 or 5 are considered good image quality that should be retained for further processing.

The shortlisted BIQA regression models have different output ranges. When inference of a BIQA pre-trained regression model on the CamenAI dataset, the predicted score is first scaled to [1,5] and then rounded to [1,2,3,4,5] to compare the score predicted with the annotation for that image. By doing so, we go from regression to classification.

C. Transfer Learning Classification Models

The CamenAI [1] dataset follows the annotation as opinion scores [1,2,3,4,5] to compensate for the absence of a large group to annotate the images as MOS. The purpose behind selecting three transfer learning models is to evaluate this dataset from a classification point of view rather than an image quality assessment point of view. To this end, the opinion scores [1,2,3,4,5] is now interpreted as five classes—the end goal being classifying the images as good image quality and poor image quality. Images belonging to class 1,2 represent poor image quality that should be discarded, and images belonging to class 3,4,5 represent good image quality that should be retained.

To this end, three transfer learning models bearing the same backbone as the shortlisted BIQA regression models are selected. The selected models are pre-trained on the ImageNet dataset. These models are fine-tuned on the CamenAI dataset for (a) Multi-class classification among five classes and (b) Binary classification among two classes. The dataset for binary classification is modified where images belonging to classes 1 and 2 are now grouped in a single class that represent images of bad quality. The images belonging to class 3,4,5 belong to a single class representing images with good image quality. The backbone architecture of the BIQA regression models discussed are as follows:

- **NIMA** [3]: MobileNet, VGG16, InceptionNet-v2
- **UNIQUE** [4]: ResNet-34
- **DBCNN** [5]: VGG16

The selected transfer learning models are MobileNet [24], VGG16 [25] , and ResNet-50 [26]. Since these models follow a TensorFlow implementation, it was impossible to import ResNet-34 from Keras applications. Thus, ResNet-50 is used rather than ResNet-34. Furthermore, four additional layers are added to each transfer learning model for the task for multi-class classification and binary classification. This is summarized in table V.

D. Performance Evaluation Criteria

The F1-scores metric [27] is used to evaluate the accuracy of experiments carried out. F1-scores is defined as the harmonic mean of precision and recall.

$$F1 = 2 * \frac{(Recall * Precision)}{(Recall + Precision)}$$

$$Precision = \frac{(TruePositives)}{(TruePositives + FalsePositives)}$$

$$Recall = \frac{(TruePositives)}{(TruePositives + FalseNegatives)}$$

A high precision value indicates a small number of false positives are found. A high recall value indicates a small number of false negatives. A high f1 score indicates a small number of false positives and false negatives.

IV. Methodology

A. Proposed Method

The proposed method is composed of several stages of evaluation as highlighted below.

- **Inference**:
  - The shortlisted BIQA models are first evaluated on the CamenAI dataset. These BIQA models are not fine-tuned for the CamenAI dataset. Instead, they are pre-trained on existing standard BIQA datasets. As a result, the image quality score range observed is different for the three BIQA models. Thus, a normalization process in the form of scaling and rounding-off is performed to yield the same output score range across the three BIQA models to compare them individually concerning the ground truth.
  - The three transfer learning models considered are independently fine tuned on the CamenAI dataset for (a) Multi-class classification and (b) Binary classification. This results in three classification models that perform multi-class classification and three classification models that perform binary classification. Post fine tuning, each model is independently evaluated on the CamenAI dataset to compare the predicted class with the ground truth.
- **Majority Voting and Aggregation**:
  - As means of determining ways to improve the image quality score accuracy concerning the ground truth, the normalized scores across the three BIQA models is subjected to two tests. These two tests are termed majority voting and aggregation. In the case of majority voting, if 2 out of 3 BIQA models predict the same score, we consider this to be the final prediction. However, majority voting will fail if the predictions of the 3 BIQA models are different. In order to tackle this, we perform aggregation on the predictions of the 3 BIQA models. This final score is compared with the ground truth. For other cases where the aggregation result yields a regression score, they are rounded to their nearest integer.
  - Majority voting is also applied to the predicted class across the three classification models that perform multi-class classification and the ones performing binary classification. The final predicted class resulting from majority voting is compared with the ground truth.
- **Neural Network Ensemble**:
  - In this approach, the output image quality score obtained across the three BIQA models are fed to a Multilayer Perceptron (MLP) [28] neural network that results in a final score. This score is compared with the ground truth. The motivation behind this approach is to study the effect of different ensembling combinations of the three BIQA models on

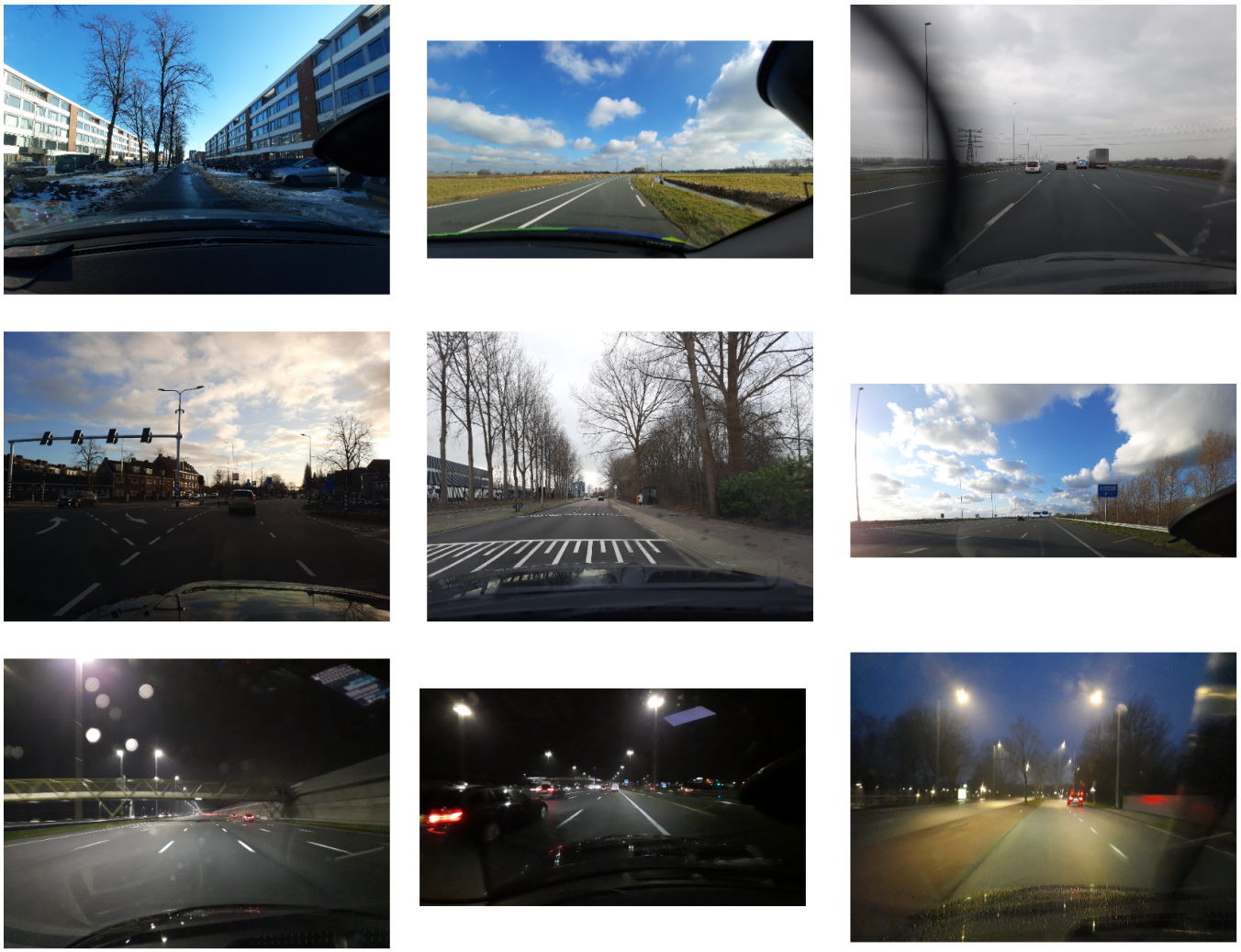


Fig. 2: Example of Images from the CamenAI dataset [1]

TABLE V: Additional layers added for transfer learning classification models

Transfer Learning Model	Multi-class Classification	Binary Classification
MobileNet [24] VGG16 [25] ResNet-50 [26]	GlobalAveragePooling2D Dense layer with Relu activation Dropout Layer Dense layer with Softmax activation	GlobalAveragePooling2D Dense layer with Relu activation Dropout Layer Dense layer with Sigmoid activation

the final predicted image quality score concerning the ground truth.

– Similarly, the predicted class across the three multi-class classification models is fed to an MLP neural network for different ensemble combinations of the multi-class classification models. The same is repeated for the three binary classification models.

B. CamenAI dataset

The dataset [1] considered comprises smartphone-captured images of roads and the natural environment from the car's windshield. These images are collected across different times of day where daylight plays a vital role because images captured during the night have poor visibility of the road and are prone to reflections by street lights. CamenAI dataset is an unbalanced real-time dataset (subjected to new, unstructured data every day) that contains images with varying image quality. An example of our dataset is illustrated in figure 2. All images contained in the CamenAI dataset are in 'jpeg' format, unlike other standard datasets that possess images in 'jpeg' or 'png' or 'bmp'.

In addition, this dataset is governed by several factors that affect the visibility of the image, as described below:

1) Rain droplets - Images contain rain droplets present on the windshield that affect the road's visibility.
2) Dust on the windshield - Images contain windshield dust in addition to rain droplets that affect the visibility of the image.
3) Motion blurring - Since the images collected involve roads; there are different vehicles that pass on the

road. This motion by a vehicle causes some part of the image to appear blurred (Stationary camera v/s moving object).

4) Reflections - It is observed that images are prone to reflections of the car's dashboard onto the car's windshield due to direct sunlight on the car's windshield.

CamenAI dataset is unlabelled and contains varying resolutions of images. The smallest resolution observed is 2560x1440, and the highest resolution is 4000x3000. The resolution of an image significantly depends on how the user captures an image. This dataset is gathered using two smartphones: Android and iPhone. Unlike other standard datasets as discussed, this dataset is significantly about roads and the natural environment.

For the task of classifying images as good or bad quality, a total of 4,780 images are selected and annotated. After annotation, these images are divided randomly into two sets representing the training and testing sets.

- Training Set: 2,790 images
- Testing Set: 1990 images

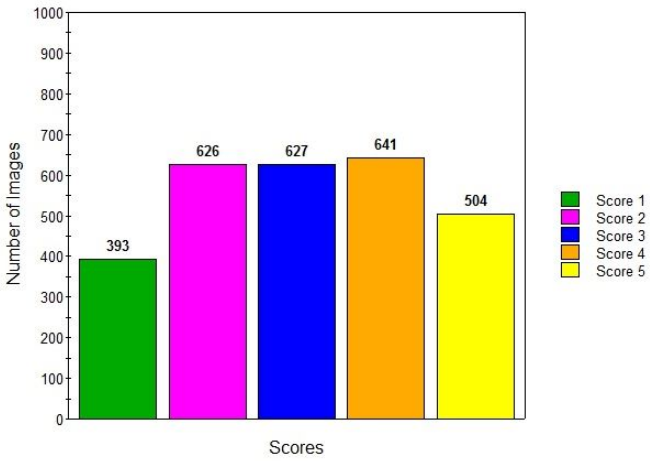


Fig. 3: Distribution of scores in training set (histogram created online [29])

A histogram highlighting the images in the training set and testing set is shown in figure 3 and figure 4.

As seen in the standard datasets in table I, they are annotated by a large group of people who provide ratings for these images, also called scores. The score is a floating-point value since it is the mean of the opinions provided by the group. The images in the standard datasets primarily have two parameters in their annotations: Mean Opinion Score and Standard Deviation of Mean Opinion Score. With the CamenAI dataset, the images are annotated by a single user (the author), thus making it difficult to produce a mean opinion score and a standard deviation from this mean score. In a crowd-sourcing experiment, [20], humans were asked to rate the quality of smartphone images based on the following parameters. These parameters are considered to annotate the CamenAI dataset into image quality ratings or opinion scores

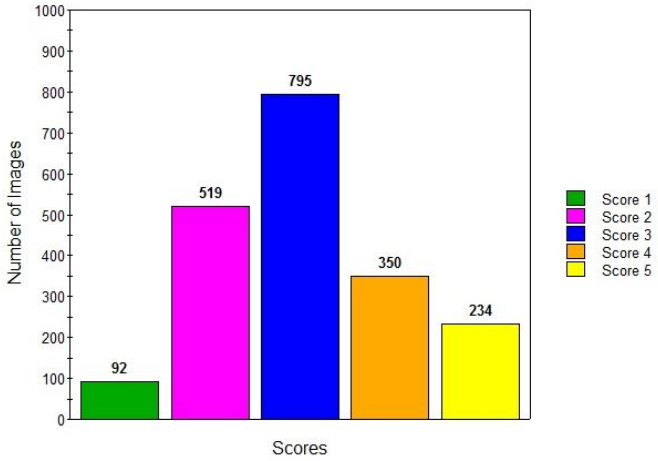


Fig. 4: Distribution of scores in testing set (histogram created online [29])

as ['1','2','3','4','5'].

1) Brightness - This parameter is used to evaluate the entire image to check how light or bright the image appears to be.
2) Sharpness - This parameter evaluates the entire image and highlights how clean and focused the image or objects in the image turn out.
3) Colorfulness - This parameter is used to evaluate a part of an image to see if the perceived region of interest appears bright or dull.
4) Contrast - This parameter is used to evaluate objects found in the image to check whether they are distinguishable to the human eye or not.
5) Noise - It is challenging to estimate the type of noise and the magnitude by visual inspection only, especially for a single annotator. Hence this parameter is not considered.

For annotating the images, five quality levels are employed [20], i.e., Bad, Poor, Fair, Good, Excellent. The image is evaluated on the parameters highlighted above and given a suitable quality level by means of manual inspection.

The images are given the scores [1,2,3,4,5] based on the quality level assigned. A summary of this annotation style is shown in table VI. For images having an outlier concerning the assigned quality level, manual inspection is done to re-assess the image. An example set of images post annotation is shown in figure 5.

Images having the score of '1' or '2' are images with bad quality and should be discarded. Images with scores '3', '4', '5' are images with good quality and should be retained. While annotating the CamenAI dataset, it was observed that for a single user, it is easy to annotate an image that is of bad quality as ['1','2'] and an image with good quality as ['4', '5']. However, it becomes slightly tricky to annotate while dealing with images between good and bad quality. Thus, crowd-sourcing annotation plays a crucial role to ensure the stability of the image quality ratings. .

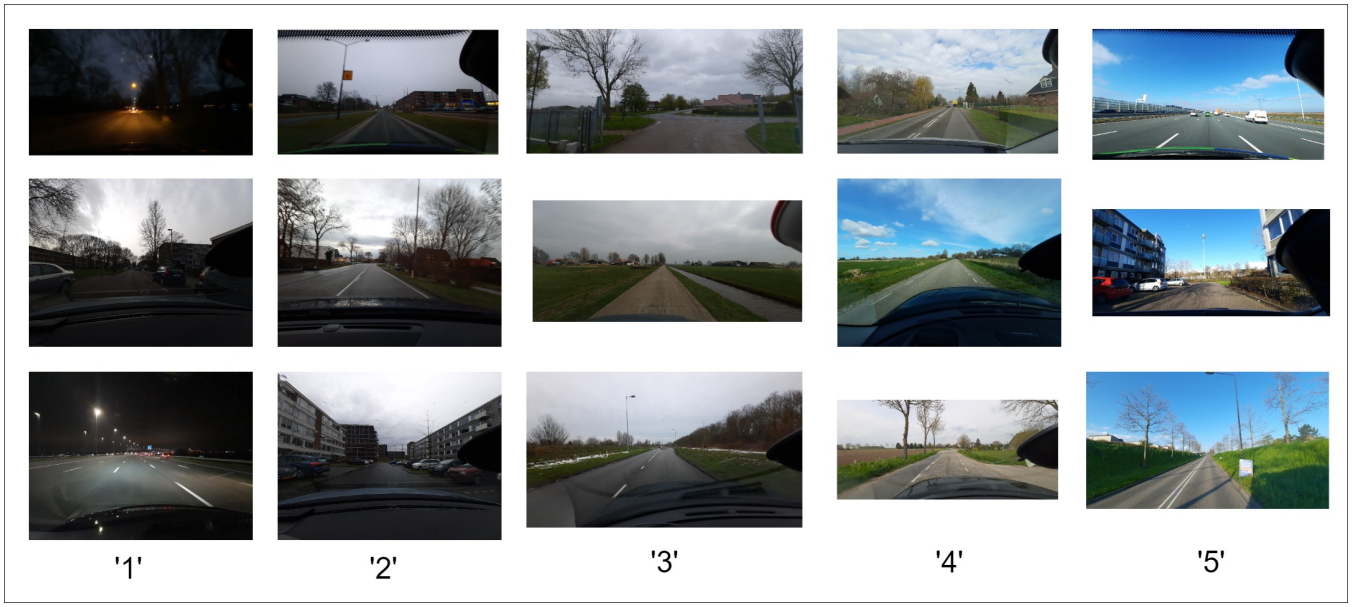


Fig. 5: Example of images in the CamenAI dataset annotated as opinion scores [1]

TABLE VI: Annotation technique employed for the CamenAI dataset based on five quality levels

Annotation	Quality Level
1	Brightness and Sharpness = Bad Contrast and Colorfulness = Poor or Bad
2	Brightness and Sharpness = Poor Contrast and Colorfulness = Poor or Bad
3	Brightness and Sharpness = Fair Contrast and Colorfulness = Fair or Poor
4	Brightness and Sharpness = Good Contrast and Colorfulness = Good or Fair
5	Brightness and Sharpness = Excellent Contrast and Colorfulness = Excellent or Good

C. Cropping for Dashboard Interference on the CamenAI dataset

This subsection presents case study results that are disjoint from the proposed methodology's evaluation. A majority of the images in the CamenAI [1] dataset is prone to dashboard interference that obstructs the natural scene. This dashboard interference is not the region of interest. This section highlights a case study to infer if cropping the dashboard interference in the images of the CamenAI dataset play a role in the score predicted by the BIQA regression model.

To this end, 200 original images and their cropped version is used to study the score predicted by the BIQA model. An example of an image containing dashboard interference and its cropped version are shown in figures 6a and 6 respectively.

The set of original images is independently evaluated by three BIQA models - NIMA [3], DBCNN [5] and UNIQUE [4] . The score predicted by each of these three models is noted. Following this, the set of 200 cropped versions of the original images is again independently evaluated by the three models to observe the score for the cropped images. Since the score obtained across the three BIQA models for

(a) Example of an image from the CamenAI dataset containing dashboard interference

(b) Example of CamenAI dataset image cropped for dashboard interference

Fig. 6: Example Image in CamenAI dataset - Original image & its Cropped version [1]

original images and cropped images lie in a different range, the scores for the three BIQA model is scaled to [1,5].

The absolute difference [30] is calculated for the scaled score obtained for each original image and the scaled score obtained for its cropped version. Later, the mean of the absolute difference across all images is noted to conclude whether dashboard interference indeed affects the score predicted by the BIQA model. The absolute difference does not provide a clear conclusion whether cropping images for dashboard interference increases or decreases the score predicted by the BIQA model. It provides the degree of score change observed across the set of original images and cropped images.

- NIMA [3] - The histogram showing the overlap of scores across original images and cropped images evaluated by NIMA is shown in figure 7. The calculated absolute difference is 0.0314

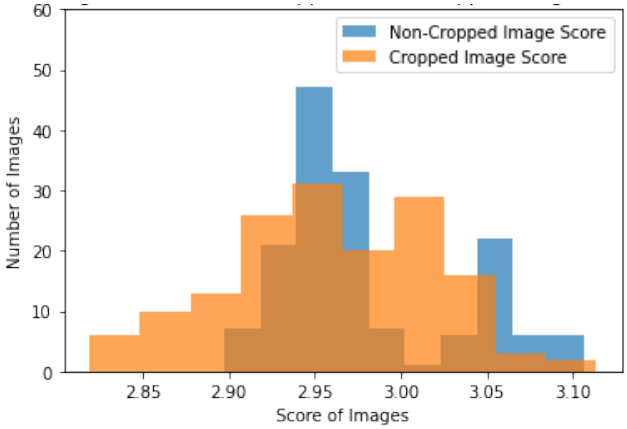


Fig. 7: Histogram of NIMA for scores across original and cropped images (histogram created using matplotlib [31])

- DBCNN [5] - The histogram showing the overlap of scores across original images and cropped images evaluated by DBCNN is shown in figure 8. The calculated absolute difference is 0.5541

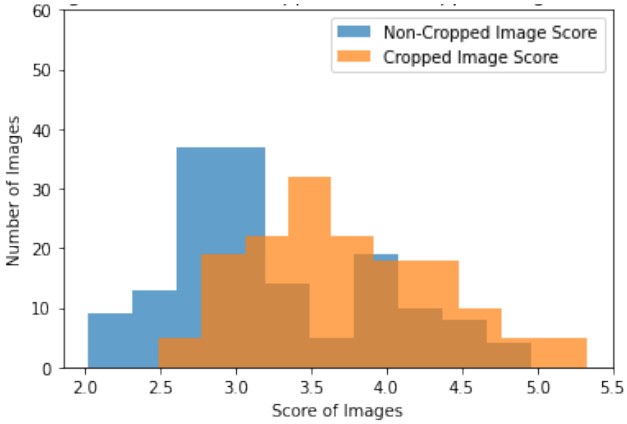


Fig. 8: Histogram of DBCNN for scores across original and cropped images (histogram created using matplotlib [31])

TABLE VII: Summary of absolute difference between scores of original images to that of cropped images for dashboard interference

Evaluation	Mean Absolute Difference
NIMA [3]	0.0314
DBCNN [5]	0.5541
UNIQUE [4]	0.0566

- UNIQUE [4]- The histogram showing the overlap of scores across original images and cropped images evaluated by DBCNN is shown in figure 9. The calculated absolute difference is 0.0566

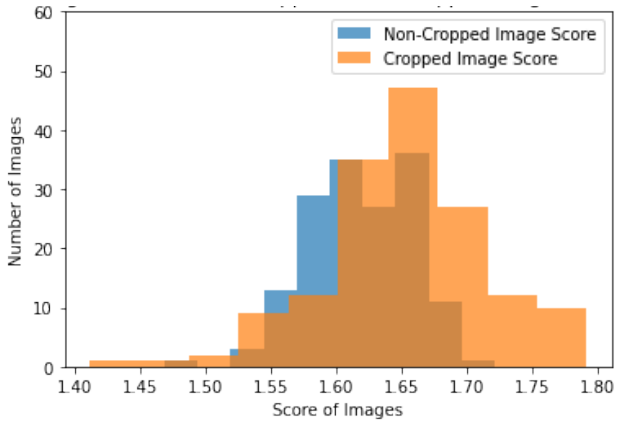


Fig. 9: Histogram of UNIQUE for scores across original and cropped images (histogram created using matplotlib [31])

The summary of the calculated absolute difference by the three BIQA models is shown in table VII. From this table, it is clear that cropping images for dashboard interference does not significantly affect the scores predicted by NIMA [3] and UNIQUE [4]. In the case of DBCNN [5], a large absolute difference is observed, indicating cropping images for dashboard interference does affect the score predicted by DBCNN.

V. Experimentation Setup

This section highlights the experiment set-up for research questions 4 and 5.

A. *Experimentation set-up 1*

The experiment set-up 1 is illustrated in figure 10. The distribution of the dataset comprising of 1894 images is highlighted in figure 11. Initially, this dataset is used to perform inference on the pre-trained BIQA models to generate predictions. The generated predictions are subjected to an 80%-20% split. Since the BIQA models are pre-trained on existing respective datasets, the output score range of these BIQA models is different.

The score of the BIQA models lie in the range:

- NIMA [3] = [1,10]
  Minima Score = 1; Maxima Score = 10

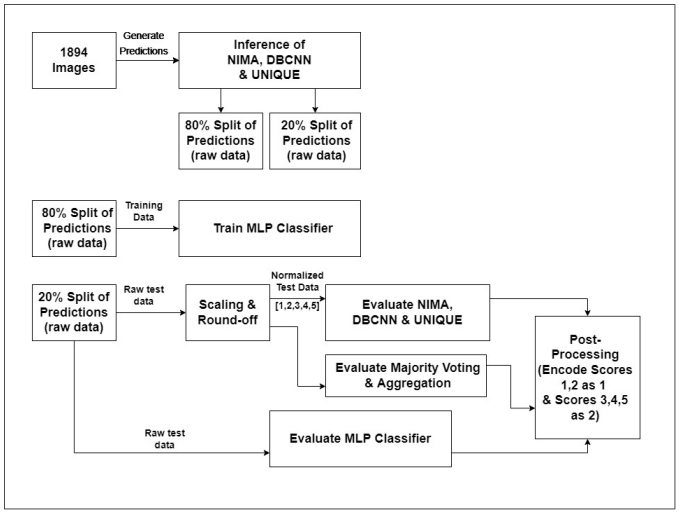


Fig. 10: Experimentation set-up for BIQA regression models, Majority Voting, Aggregation, Ensemble of BIQA Models (Diagram created using draw.io [32])

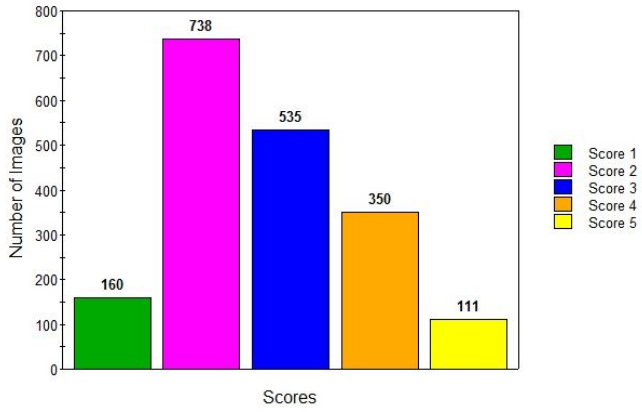


Fig. 11: Distribution of scores for images considered in experiment set-up 1 (Histogram created online [29])

- DBCNN [5] = [0,10]
  Minima Score = 0; Maxima Score = 10
- UNIQUE [4]= [-3,3]
  Minima Score = -3; Maxima Score = 3

The 80% split of NIMA, DBCNN and UNIQUE predictions is used to train the MLP classifier [28] neural network for different combinations of the ensemble. This MLP classifier is trained with respect to the ground truth that contains the annotations of the images in the score range [1,2,3,4,5].

The 20% split of the generated raw predictions from NIMA, DBCNN and UNIQUE is subjected to scaling and rounding-off. Scaling is performed to bring the different score ranges of the BIQA models to a common score range of [1,5] using the formula,

$$1 + \left(\frac{ScorePredicted}{MaximaScore - MinimaScore} \right) * 4$$

Next, the scaled scores are rounded-off to their nearest integer values to yield the score range [1,2,3,4,5]. This normalized data is used to evaluate the performance of NIMA, DBCNN, UNIQUE, majority voting and aggregation tests. Furthermore, the 20% split of the generated raw predictions is used to evaluate the MLP classifier. The output of the MLP classifier represents the final image quality score that lies in the range [1,2,3,4,5].

As already stated, images with scores 3,4,5 represent good quality images, and images with scores 1,2 represent bad quality images. Thus, a post-processing technique is introduced to encode the final image quality score. Scores 1,2 are encoded as score 1, and scores 3,4,5 are encoded as score 2. Similarly, the annotations are also encoded. The F1-score classification report is generated for the encoded score and the encoded annotations across the three BIQA models, majority voting, aggregation and MLP classifier. This results in two classes - 1 (bad quality) and 2 (good quality). The F1-score [27] of these two classes is inspected, and the approach that results in the maximum F1-score is concluded as the best approach to retain a maximum number of images in the usable category.

B. Experiment set-up 2

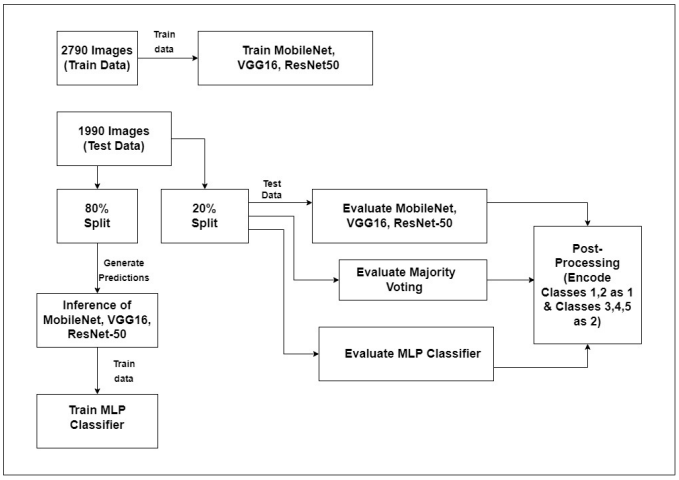
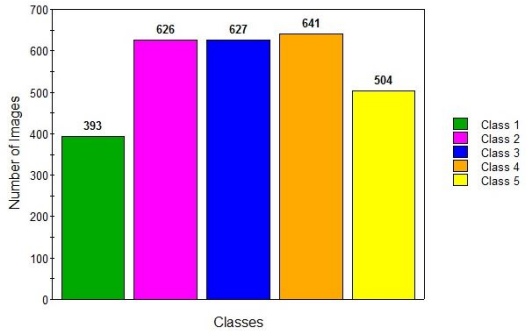


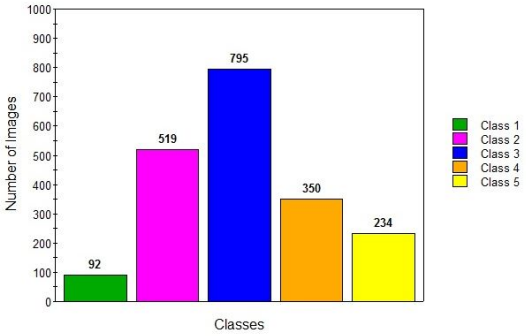
Fig. 12: Experimentation set-up for classification models, Majority Voting, Ensemble of Classification Models (Diagram created using draw.io [32])

The experiment set-up is illustrated in figure 12. The dataset is now treated as five classes - 1,2,3,4,5. This experiment is carried out to determine if images in the CamenAI [1] dataset can be treated as classes or not. This experiment investigates a classification-based approach to retaining good quality in the CamenAI dataset. Here, a multi-class classification approach is employed. Classes 3,4,5 represent images of good quality. Classes 1 and 2 represent images of bad quality.

Figures 13a and 13b represent the distribution of classes in train data and test data, respectively, for this experiment. The train data is used to train the classification models



(a) Distribution of classes for training data



(b) Distribution of classes for testing data

Fig. 13: Distribution of classes for train and test data in experiment set-up 2 (Histogram created online [29])

(a) Distribution of classes for training data

(b) Distribution of classes for testing data

Fig. 14: Distribution of classes for train and test data in experiment set-up 3 (Histogram created online [29])

(MobileNet [24], VGG16 [25], ResNet-50 [26]). The test data is subjected to an 80%-20% split. The 80% split of the test data is used to perform inference on the trained classification models to generate predictions (class the image belongs to). The predictions serve as train data for the MLP classifier [28]. The 20% split of test data is used as a common test set to evaluate the performance of the classification models individually, evaluate the performance of majority voting and lastly to evaluate the MLP classifier.

A post-processing technique is introduced where classes 1,2 are encoded as class 1 and 3,4,5 are encoded as class 2. Similarly, the annotations are also encoded. The F1-score [27] classification report is generated for the encoded score and the encoded annotations across the three classification models, majority voting and MLP classifier. This results in two classes - 1 (bad quality) and 2 (good quality). The F1-score of these two classes is inspected, and the approach that results in the maximum F1-score is concluded as the best approach to retain a maximum number of images in the usable category.

C. Experiment set-up 3

The experiment set-up 3 is similar to experiment set-up 2 highlighted in figure 12. As seen in experiment set-up 2, a multi-class classification approach is employed over the fives classes. Experiment set-up 3 performs the same task as experiment set-up 2, but for a binary classification approach. Thus, post-processing is not necessary. Since the final goal is to inspect whether the images are classified as good quality or bad quality, a binary classification approach is employed upfront.

Figure 14 represents the distribution of classes in train data for the individual classification models. The train data in figure 13a is modified such that images of classes 1 and 2 are now grouped as one class and images of classes 3,4,5 grouped as another class. A balance is maintained across the two classes for the train data. The test data in figure 13b is also modified. The distribution of train data and test data for experiment set-up 3 is illustrated in figures 14a and 14b.

VI. Evaluation

Across the three experiment set-ups, there exist two classes - Class 1 and Class 2 (or Score 1 and Score 2 in case of experiment set-up 1). The chosen performance evaluation metrics - Precision, Recall and F1-score- evaluates performance of each experiment set-ups towards rightly classifying the images to their respective classes.

The F1-score [27] of a class is the harmonic mean of precision and recall of that class. The precision of a class highlights the ratio of correctly predicted positive observations to the total predicted positive observations. The recall

TABLE VIII: Summary of Evaluation of BIQA models, Majority Voting and Aggregation tests

	Class-1			Class-2		
Evaluation	Precision	Recall	F1	Precision	Recall	F1
NIMA [3]	1.00	0.04	0.08	0.56	1.00	0.72
DBCNN [5]	0.49	0.48	0.49	0.58	0.59	0.59
UNIQUE [4]	0.45	1.00	0.62	0.00	0.00	0.00
Majority Voting	0.50	0.49	0.50	0.59	0.59	0.59
Aggregation	0.50	0.49	0.50	0.59	0.59	0.59

TABLE IX: Summary of evaluation for different combination of the ensemble NIMA, DBCNN and UNIQUE

Ensemble using MLP Classifier	**#Hidden Layers**	**#Neurons per Hidden Layer**	**Class-1**			**Class-2**		
			Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
NIMA[3]-DBCNN[5]-UNIQUE[4]	4	3000,2000,1000,500	0.67	0.76	0.71	0.78	0.69	0.73
	3	3000,2000,1000	0.51	0.88	0.65	0.75	0.30	0.43
	2	2000,1000	0.57	0.84	0.68	0.78	0.47	0.59
	1	2000	0.66	0.79	0.72	0.79	0.66	0.72
NIMA[3]-DBCNN[5]	4	3000,2000,1000,500	0.67	0.77	0.71	0.78	0.68	0.73
	3	3000,2000,1000	0.68	0.75	0.71	0.77	0.70	0.73
	2	2000,1000	0.61	0.87	0.72	0.83	0.54	0.65
	1	2000	0.62	0.88	0.73	0.84	0.55	0.67
NIMA[3]-UNIQUE[4]	4	3000,2000,1000,500	0.54	0.95	0.69	0.89	0.33	0.48
	3	3000,2000,1000	0.56	0.91	0.69	0.84	0.42	0.56
	2	2000,1000	0.73	**0.67**	**0.70**	0.74	**0.79**	**0.77**
	1	2000	0.60	0.84	0.70	0.80	0.54	0.64
DBCNN[5]-UNIQUE[4]	4	3000,2000,1000,500	0.49	0.80	0.60	0.64	0.30	0.41
	3	3000,2000,1000	0.47	0.85	0.60	0.60	0.18	0.28
	2	2000,1000	0.47	0.81	0.60	0.61	0.25	0.36
	1	2000	0.47	0.86	0.60	0.61	0.18	0.28

TABLE X: Average Processing time for different combination of NIMA, DBCNN and UNIQUE models combined in the ensemble

Evaluation	**Average Processing Time (in seconds)**
NIMA [3]	27 (A)
DBCNN [5]	45 (B)
UNIQUE [4]	5 (C)
NIMA[3]-DBCNN[5]-UNIQUE[4]-MLP Classifier[28]	A + B + C + 0.2748
NIMA[3]-DBCNN[5]-MLP Classifier[28]	A + B + 0.0133
NIMA[3]-UNIQUE[4]-MLP Classifier[28]	A + C + 0.0790
DBCNN[5]-UNIQUE[4]-MLP Classifier[28]	B + C + 0.0224

is the ability of the model to find all relevant classes within a dataset.

Images belonging to Class 1 indicates images with bad image quality that should be discarded, whereas images belonging to Class 2 indicate images with good image quality that should be retained. Thus, the focus of interest is the F1-score parameter for both the classes. A similar and large F1-score value is desired for the two classes (at least 0.6). It is also desired to have a similar and large recall value for both the classes (at least 0.6). A low recall value for a class indicates misclassifications of that class. The values of precision, recall and F1-score range from [0,1].

A. Evaluation for experiment set-up 1

The evaluation summary of the individually tested BIQA models, majority voting and aggregation results are summarized in table VIII. Across the three BIQA models - NIMA [3], DBCNN [5] and UNIQUE [4], it is observed that there exists a large quantity of misclassification of class-1 images as class-2 and vice-versa.

NIMA exhibits a low F1-score and a low recall value for class 1. This means most bad quality images are predicted as good quality images. The F1-score and recall value for DBCNN is centred at 0.5, indicating a misclassification of both classes. For UNIQUE, the F1-score and recall value for class-2 is 0, indicating all good image quality images are predicted as bad image quality images. The three BIQA models render a poor performance since they are not fine-tuned on the CamenAI dataset. Applying the principle of majority voting and aggregation on the output score yielded by the BIQA models results in similar behaviour to that of DBCNN.

The evaluation summary of the MLP classifier [28] towards ensembling different combinations of NIMA [3], DBCNN [5] and UNIQUE [4] is shown in table IX. It is observed that the ensemble comprising of DBCNN and UNIQUE performed poorly in comparison to other ensemble combinations. This ensemble delivers a low F1-score and a low recall value for class-2, indicating that good image quality images are majorly misclassified into class-1. The

TABLE XI: Best validation loss obtained across training MobileNet, VGG16 and ResNet-50 for multi-class classification

Model	**Final validation loss obtained during training (multi-class classification)**
MobileNet [24]	0.88225
VGG16 [25]	0.90468
ResNet-50 [26]	0.94138

TABLE XII: Summary of results for Multi-class classification encoded to binary classification for MobileNet, VGG16 ResNet-50 and majority voting

	Class-1			Class-2		
Evaluation	Precision	Recall	F1	Precision	Recall	F1
MobileNet [24]	0.34	0.60	0.43	0.64	0.37	0.47
VGG16 [25]	0.55	0.03	0.06	0.66	0.99	0.79
ResNet-50 [26]	0.34	0.98	0.50	0.00	0.00	0.00
Majority Voting	0.33	0.60	0.43	0.64	0.37	0.47

TABLE XIII: Summary of evaluation for different combination of the ensemble MobileNet, VGG16 and ResNet-50 for multi-class classification encoded as binary classification

Ensemble using MLP Classifier	**#Hidden Layers**	**#Neurons per Hidden Layer**	**Class-1**			**Class-2**		
			Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MobileNet[24]-VGG16[25]-ResNet-50[26]	4	3000,2000,1000,500	0.65	0.12	0.20	0.68	0.97	0.80
	3	3000,2000,1000	0.72	0.09	0.16	0.67	0.98	0.80
	2	2000,1000	0.78	0.09	0.16	0.67	0.99	0.80
	1	2000	0.72	0.09	0.16	0.67	0.98	0.80
MobileNet[24]-VGG16[25]	4	3000,2000,1000,500	0.64	0.04	0.08	0.66	0.99	0.79
	3	3000,2000,1000	0.57	0.08	0.14	0.67	0.97	0.79
	2	2000,1000	0.55	0.08	0.14	0.67	0.97	0.79
	1	2000	0.55	0.08	0.14	0.67	0.97	0.79
MobileNet[24]-ResNet-50[26]	4	3000,2000,1000,500	0.93	0.06	0.12	0.67	1.00	0.80
	3	3000,2000,1000	0.93	0.06	0.12	0.67	1.00	0.80
	2	2000,1000	0.93	0.06	0.12	0.67	1.00	0.80
	1	2000	0.93	0.06	0.12	0.67	1.00	0.80
VGG16[25]-ResNet-50[26]	4	3000,2000,1000,500	0.63	0.09	0.16	0.67	0.97	0.79
	3	3000,2000,1000	0.63	0.09	0.16	0.67	0.97	0.79
	2	2000,1000	0.63	0.09	0.16	0.67	0.97	0.79
	1	2000	0.59	0.09	0.16	0.67	0.97	0.79

TABLE XIV: Average Processing time for different combination of MobileNet, VGG16 and ResNet-50 combined in the ensemble for multi-class classification

Evaluation (multiclass)	**Average Processing Time (in seconds)**
MobileNet [24]	0.3302 (A)
VGG16 [25]	0.8149 (B)
ResNet-50 [26]	0.4786(C)
MobileNet[24]-VGG16[25]-ResNet-50[26]-MLP Classifier[28]	A + B + C + 0.3619
MobileNet[24]-VGG16[25]-MLP Classifier[28]	A + B + 0.0291
MobileNet[24]-ResNet-50[26]-MLP Classifier[28]	A + C + 0.0417
VGG16[25]-ResNet-50[26]-MLP Classifier[28]	B + C + 0.0235

ensemble of NIMA and DBCNN exhibits stability in F1-score and recall value across different hidden layers and neurons per hidden layer. This indicates a majority of the images are rightly classified into correct classes. This ensemble delivers an F1-score of 0.71 and recall of 0.75 for class-1 and an F1-score of 0.71, and a recall of 0.73 for class-2. The ensemble comprising of NIMA, DBCNN, UNIQUE achieves the same F1-score and recall value of the ensemble NIMA and DBCNN. The ensemble of NIMA and UNIQUE delivers the highest F1-score and recall value for the MLP classifier comprising of two hidden layers. This indicates a majority of the images are rightly classified into correct classes. It is interesting to note the effect the number of hidden layers in the MLP classifier has on delivering a large F1-score and a large recall value. This is viewed as a future work where hyper-parameter tuning can be employed on the number of hidden layers and number of neurons per layer to determine the right combination that delivers a large F1-score and recall value. The processing time for each combination of the ensemble is shown in table X.

B. Evaluation for experiment set-up 2

The final validation loss obtained while training MobileNet [24], VGG16 [25] and ResNet-50 [] on the CamenAI dataset [1] is shown in table XI. The evaluation summary of MobileNet, VGG16, ResNet-50 trained and evaluated towards a multiclass classification approach is shown in table XII.

TABLE XV: Best validation loss obtained across training MobileNet, VGG16 and ResNet-50 for binary classification

Model	**Final validation loss obtained during training (binary classification)**
MobileNet [24]	0.38975
VGG16 [25]	0.40308
ResNet-50 [26]	0.41971

TABLE XVI: Summary of results for binary classification for MobileNet, VGG16 ResNet-50 and majority voting

	Class-1			Class-2		
Evaluation	Precision	Recall	F1	Precision	Recall	F1
MobileNet[24]	0.38	0.31	0.34	0.71	0.77	0.74
VGG16 [25]	0.30	0.08	0.12	0.69	0.92	0.79
ResNet-50 [26]	0.31	0.98	0.47	0.00	0.00	0.00
Majority Voting	0.31	1.00	0.47	0.00	0.00	0.00

TABLE XVII: Summary of evaluation for different combination of the ensemble MobileNet, VGG16 and ResNet-50 for binary classification

Ensemble using MLP Classifier	**#Hidden Layers**	**#Neurons per Hidden Layer**	**Class-1**			**Class-2**		
			Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
MobileNet[24]-VGG16[25]-ResNet-50[26]	4	3000,2000,1000,500	1.00	0.02	0.03	0.69	1.00	0.82
	3	3000,2000,1000	1.00	0.02	0.03	0.69	1.00	0.82
	2	2000,1000	1.00	0.02	0.03	0.69	1.00	0.82
	1	2000	1.00	0.02	0.03	0.69	1.00	0.82
MobileNet[24]-VGG16[25]	4	3000,2000,1000,500	0.00	0.00	0.00	0.69	1.00	0.82
	3	3000,2000,1000	0.00	0.00	0.00	0.69	1.00	0.82
	2	2000,1000	0.00	0.00	0.00	0.69	1.00	0.82
	1	2000	0.00	0.00	0.00	0.69	1.00	0.82
MobileNet[24]-ResNet-50[26]	4	3000,2000,1000,500	1.00	0.02	0.03	0.69	1.00	0.82
	3	3000,2000,1000	1.00	0.02	0.03	0.69	1.00	0.82
	2	2000,1000	1.00	0.02	0.03	0.69	1.00	0.82
	1	2000	1.00	0.02	0.03	0.69	1.00	0.82
VGG16[25]-ResNet-50[26]	4	3000,2000,1000,500	1.00	0.02	0.03	0.69	1.00	0.82
	3	3000,2000,1000	1.00	0.02	0.03	0.69	1.00	0.82
	2	2000,1000	1.00	0.02	0.03	0.69	1.00	0.82
	1	2000	1.00	0.02	0.03	0.69	1.00	0.82

TABLE XVIII: Average Processing time for different combination of MobileNet, VGG16 and ResNet-50 combined in the ensemble

Evaluation (binary)	**Average Processing Time (in seconds)**
MobileNet [24]	0.2942(A)
VGG16 [25]	0.7525 (B)
ResNet-50 [26]	0.3239(C)
MobileNet[24]-VGG16[25]-ResNet-50[26]-MLP Classifier [28]	A + B + C + 0.0151
MobileNet[24]-VGG16[25]-MLP Classifier[28]	A + B + 0.0145
MobileNet[**?**]-ResNet-50[26]-MLP Classifier[28]	A + C + 0.1995
VGG16[25]-ResNet-50[26]-MLP Classifier[28]	B + C + 0.0131

MobileNet exhibits a low F1-score and low recall value for Class-2. This indicates a large quantity of good image quality images is misclassified as poor image quality images. VGG16 exhibits an opposite behaviour to MobileNet, where class-1 images are misclassified as class-2. In the case of ResNet-50, all images are classified into class-1, thereby an F1-score and recall value of 0 for class-2. Applying the principle of majority voting, it is observed the results obtained is the same as results for MobileNet.

These individual models, although fine-tuned on the CamenAI dataset, perform poorly while classifying images into respective classes. This can be an indication that the CamenAI dataset should not be interpreted as classes since there exist no distinguishable features across the five classes for the model to learn and detect.

The evaluation summary of the MLP classifier towards ensembling different combinations of MobileNet, VGG16 and ResNet-50 is shown in table XIII. It is observed that across all ensemble combination there exists a large misclassification of Class-1 images as Class-2 for different hidden layers incorporated in the MLP classifier.

The processing time for each combination of the ensemble is shown in table XIV.

C. Evaluation for experiment set-up 3

The evaluation summary of MobileNet [24], VGG16 [25], ResNet-50 [26] trained and evaluated towards a binary classification approach on the CamenAI dataset [1] is shown in table XVI. The final validation loss obtained while training these models on the CamenAI dataset is shown in table XV.

It is observed that MobileNet and VGG16 produce a large F1-score and recall value for class-2 but a significantly low F1-score and recall value for Class-1. This indicates there exist many misclassifications of class-1 images. ResNet-50 and majority voting produce the same F1-score and recall value where all images of Class-2 are misclassified as Class-1.

The evaluation summary of the MLP classifier [28] towards ensembling different combinations of MobileNet, VGG16 and ResNet-50 is shown in table XVII. It is observed that all combination of the ensemble deliver a high F1-score and recall value for Class-2 images and perform significantly poorly for images of Class-1. There exists a large quantity of misclassification of Class-1 images into Class-2.

This further proves that the CamenAI dataset should not be interpreted as classes towards retaining good quality images because the model cannot detect the features required to classify the images into their respective classes.

The processing time for each combination of the ensemble is shown in table XVIII.

VII. Results and Discussion

This section discusses the results obtained across the three experiment setups.

In experiment setup 1, three BIQA regression models (NIMA [3], DBCNN [5], UNIQUE[4]) are considered to perform inference on the CamenAI dataset [1] that follows the annotation [1,2,3,4,5]. The annotations are interpreted as scores. The predicted image quality scores of these BIQA models on the CamenAI dataset lie in different ranges. This is because the BIQA models are pre-trained on different standard BIQA datasets that take different ranges for Mean Opinion Scores. The scores predicted by these models are first scaled to the range [1,5] and rounded to align with the CamenAI dataset annotations. Post-processing in the form of encoding is employed where the scores 3,4,5 are encoded as score 2 and scores 1,2 are encoded as score 1. This is done since images with scores 3,4,5 are considered good image quality images that should be retained whereas images with scores 1,2 are considered poor image quality images that should be discarded. To facilitate better visualization and analysis of results, encoding is employed. It is observed that the BIQA models performed poorly. NIMA exhibits a considerable misclassification of images with scores 1 as score 2 and vice versa in the case of UNIQUE. DBCNN, majority voting and aggregation tests lead to a similar result where a misclassification occurs for images with score 1 and score 2. This is because these BIQA models are not fine-tuned on the CamenAI dataset due to the lack of annotators to represent the annotation in the form of Mean Opinion Scores. It is interesting to note the performance of these BIQA models once fine-tuned on the CamenAI dataset.

Later, an evaluation of the MLP classifier [28] for different ensemble combinations of the BIQA models is explored for different hidden layer sizes and neurons per layer in the MLP classifier. It is observed that the ensemble of NIMA and DBCNN rendered a stabilized performance in F1-score and recall value across different combinations of hidden layer size and neurons per layer. The ensemble of DBCNN and UNIQUE exhibits a low F1-score and recall value for score 2, indicating a large misclassification of images with scores 2 as score 1. Thus a major chunk of good image quality images are lost. However, the highest F1-score and recall value was achieved for the ensemble of NIMA and UNIQUE for a hidden layer size of 2. The ensemble comprising of three BIQA models eventually achieved the same F1-score and recall value as the ensemble of NIMA and DBCNN. The number of BIQA models combined in the ensemble and their respective processing time is calculated.

Due to the lack of annotators, the CamenAI dataset is annotated as opinion scores rather than mean opinion scores. The annotated opinion scores follow the range [1,2,3,4,5]. In experiment setup 2, the CamenAI dataset is interpreted as five classes. The classes 1,2 represent images of bad quality that should be discarded and classes 3,4,5 represent good quality images. This experiment follows a classification approach towards rightly classifying images as good quality and bad image quality. To this end, transfer learning models bearing the backbone as MobileNet [24], VGG16 [25] and ResNet-50 [26] is trained on the CamenAI dataset for a multiclass classification task. This experiment employs a post-processing where classes 1,2 are encoded as class 1 and classes 3,4,5 are encoded as class 2. Class 1 now represents all images exhibiting bad quality, and class 2 represents all images with good quality. Later an MLP classifier neural network comprising of different ensemble combinations of MobileNet, VGG16 and ResNet-50 is explored to determine the precision, recall and F1-score towards classifying images as good quality and bad quality. The results show that the performance of these individual models is poor. MobileNet results in a significant misclassification of class 2 images as class 1. VGG16 results in a significant misclassification of class 1 images as class 2. In the case of ResNet-50, all images of class 2 are incorrectly classified as class 1. The results of majority voting are similar to the results of MobileNet. This shows that these models are incapable of learning and detecting the features to classify them as good quality and bad quality rightly. The evaluation of the MLP classifier for different ensemble combinations explored shows a significant misclassification of class 1 images as class 2.

In order to better conclude to discard the idea of employing a classification approach on the CamenAI dataset, experiment setup 3 follows a binary classification approach. The individual models - MobileNet, VGG16 and ResNet-50 are now trained for a binary classification of two classes 1 2, where class 1 represents images of bad quality and class 2 represents images of good quality. The performance of these models are individually evaluated, followed by an evaluation of the MLP classifier for different ensemble combination of these models. Results show that these models perform poorly. MobileNet and VGG16 exhibits a large misclassification of class 1 images as class 2. In the case of ResNet-50 and majority voting, all images of class 2 are misclassified as class 1. The results of the MLP classifier

follows a similar trend where almost all images belonging to class 1 are misclassified as class 2 across different combinations of the ensemble. This shows that the images of the CamenAI dataset annotated as [1,2,3,4,5] may tend to have inconsistency across the images since a single person annotated them. In addition, images in the CamenAI dataset contain finer features to interpret them as classes. As a result, the model (or the ensemble) is incapable of learning and detecting the classes' features to classify them rightly. Thus, a classification approach should not be employed towards classifying images as good quality and bad quality. Rather, a BIQA approach is desired which results in better classification of images.

VIII. Conclusion

This thesis evaluates existing deep learning regression models for BIQA on the CamenAI dataset. Due to the non-availability of a large group, the CamenAI dataset was annotated by a single person. The research proposes an ensemble NIMA-UNIQUE [3], [4] for a hidden layer size of 2 in the MLP classifier [28] as the best method towards rightly classifying images as good and bad quality. Although other BIQA ensemble combinations yield an F1-score and recall value of greater than 0.5, the ensemble of NIMA-UNIQUE achieves the highest F1-score and recall value of 0.77 and 0.79 class-2 & 0.70 and 0.67 for class-1.

In the later part, classification models do not perform well on this dataset because the five classes are challenging to represent the behaviour of the images while annotating. In addition, this dataset has finer features to be represented as classes that may prove the model is incapable of learning and detecting features across the classes—also, owing to the fact that this dataset is unbalanced since data is gathered in real-time. Hence annotating them into classes should not be done; rather crowdsourcing annotations must be employed to represent the image annotation as mean opinion score and standard deviation. This is one of the outcomes of this study.

All the different ensemble combinations across the classification models did not perform poorly on the CamenAI dataset exhibiting a significant misclassification of class-1 images as class-2.

This section provides answers to the research questions proposed.

1) What are the available BIQA models and training datasets used to benchmark the performance of BIQA models?
   - The recently available BIQA models include NIMA [3], MetaIQA [11], DBCNN [5], UNIQUE [4], BIQA using self-adaptive hyper network [12], NRIQA using contrast enhancement [13], NRIQA using multi-pooled inception features [14], Perceptual NRIQA, NRIQA using global statistical features [20], BIQA using high-level semantics [16], BIQA using controllable list-wise ranking [17], BIQA using DeepRN [21], Learning for BIQA.
   - The standard authentic datasets used to benchmark the performance of BIQA models include: Live Challenge [6], CID2013 [9], Koniq-10k [7], BID [8]
2) How does the **CamenAI dataset** compare to the standard datasets in terms of annotations, resolutions, and score distribution?
   - Existing standard datasets are annotated as mean opinion scores and standard deviation. The mean opinion score for the standard datasets range from [0,100], [1,5],[0,5]. The range of opinion scores for the CamenAI dataset is [1,2,3,4,5]. The image resolution across the standard dataset is 500x500 pixels in the Live Challenge dataset, 512x384 pixels in the case of the Koniq-10k dataset. In the case of the BID dataset, the resolutions range from 1280x960 pixels to 2272x1704 pixels. The lowest and largest image resolution in the CamenAI dataset is 2560x1440 pixels and 4000x3000 pixels, respectively.
   a) In what manner can the CamenAI dataset be annotated for the task of BIQA? - Due to the absence of a large group of annotators to annotate images in the CamenAI dataset, it is impossible to annotate as mean opinion scores and standard deviation. The author annotates the CamenAI dataset as opinion scores that take the form [1,2,3,4,5], where a score 1 indicates the poorest image quality and a score 5 indicates the highest image quality.
3) To what extent does cropping the images for dashboard interference affect the score predicted by BIQA regression models such as NIMA [3], UNIQUE [4] and DBCNN [5]?
   - Absolute difference between the scores obtained for the set of original images and their cropped version is calculated across the three BIQA models. Cropping the images for dashboard interference does affect the score predicted by the BIQA model. In the case of NIMA and UNIQUE, the degree of score change observed are 0.0314 and 0.0566, respectively. The absolute difference for DBCNN is 0.5541 indicating a larger degree of score change
4) What effect does different ensembling combinations of NIMA [3], DBCNN [5] and UNIQUE [4] have on the accuracy of their performances on the CamenAI dataset?
   a) How does NIMA, DBCNN and UNIQUE perform individually on the CamenAI dataset concerning accuracy?
   NIMA delivers a low F1-score and a low recall value for Score-1, indicating that most images considered bad quality are misclassified as good quality. In the case of UNIQUE, the F1-score and recall value for Score-2 is 0, indicating all images considered good quality are misclassified as bad quality. DBCNN exhibits an F1-score and recall value centred at 0.5 for Score-1 and Score-2, indicating an equal amount of misclassification

of good quality images as bad quality and vice-versa.

b) How do we normalize NIMA, DBCNN and UNIQUE results to ensure the same output score range across these models?
- The output scores from NIMA, DBCNN and UNIQUE are normalized by performing scaling and round-off. The formula to scale the output scores across NIMA, DBCNN and UNIQUE is: $1 + \left(\frac{ScorePredicted}{MaximaScore - MinimaScore}\right) * 4$

c) What is the trade-off between the number of models combined in the ensemble concerning accuracy and processing time on the CamenAI dataset?
-The ensemble combination of DBCNN and UNIQUE performs poorly than other ensemble combinations. This combination exhibits an F1-score and recall value of 0.41 and 0.30 for Score-2 images, indicating a large misclassification as score-1. The ensemble of NIMA and DBCNN delivers a stabilized performance with F1-score and recall values of 0.73 and 0.70, respectively, for Score-2. The ensemble comprising NIMA, DBCNN and UNIQUE achieves the same F1-score and recall value as the ensemble combination NIMA and DBCNN. The ensemble of NIMA and UNIQUE delivers the highest F1-score and recall value of 0.77 and 0.79 for score-2 respectively and 0.70 and 0.67 for score-1 respectively. The processing time (in seconds) for NIMA, DBCNN and UNIQUE are 27,45 and 5, respectively. In addition to their processing time, the processing time for the MLP classifier for different ensemble combinations, NIMA-DBCNN-UNIQUE, NIMA-DBCNN, NIMA-UNIQUE and DBCNN-UNIQUE, are 0.2748s,0.0133s,0.0790s and 0.0224s, respectively.

d) What is the best and worst combination of the ensemble surrounding accuracy on the CamenAI dataset?
-The best ensemble combination is NIMA-UNIQUE that delivers an F1-score and recall value of 0.77 and 0.79 for score-2 & F1-score and recall value of 0.70 and 0.67 for score-1, respectively. The worst ensemble combination is DBCNN-UNIQUE, which delivers an F1-score and recall value of 0.41 and 0.30 for score-2 & 0.60 and 0.80 for score-1, respectively.

5) What effect does different ensembling combinations of transfer learning models have on the accuracy of their performances on the CamenAI dataset?

a) What are the selected transfer learning models?
-The selected transfer learning models are MobileNet [24], VGG16 [25] and ResNet-50 [26]

b) How do these models perform for multiclass classification and binary classification on the CamenAI dataset concerning accuracy?
-In the case of multiclass classification encoded as binary classification, MobileNet exhibits a large misclassification of class-2 images as class-1. It delivers a low F1-score and recall value of 0.47 and 0.37 for class-2 & 0.43 and 0.60 for class-1. In the case of VGG16, nearly all images of class-1 are misclassified as class-2. It exhibits an F1-score and recall value of 0.06 and 0.03 for class-1, respectively. ResNet-50 misclassifies all class-2 images as class-1, thereby delivering an F1-score and recall value of 0.00 for class-2.
-In the case of binary classification, MobileNet delivers a low F1-score and recall value of 0.34 and 0.31 for class-1. This indicates a large misclassification of class-1 images as class-2. On the other hand, VGG16 delivers even a larger misclassification than MobileNet. Here, the F1-score and recall value for class-1 are 0.12 and 0.08, respectively. ResNet-50 misclassifies all class-2 images as class-1. It delivers an F1-score and recall value of 0.00 for class-2

c) What is the trade-off between the number of models combined in the ensemble concerning accuracy and processing time on the CamenAI dataset for a multiclass classification?

-The F1-score and recall value across different combinations of MobileNet, VGG16 and ResNet-50 exhibit a similar and lower value for class-1 (less than 0.2) indicating a significant misclassification of class-1 images as class-2. The processing time (in seconds) for MobileNet, VGG16 and ResNet-50 are 0.3302s, 0.8149s and 0.4786s respectively. In addition to their individual processing time, the processing time for the MLP classifier for different ensemble combinations, MobileNet-VGG16-ResNet-50, MobileNet-VGG16, MobileNet-ResNet-50 and VGG16-ResNet-50 are 0.3619s, 0.029s, 0.041s and 0.023s respectively.

d) What is the trade-off between the number of models combined in the ensemble concerning accuracy and processing time on the CamenAI dataset for a binary classification?

-The F1-score and recall value across different combinations of MobileNet, VGG16 and ResNet-50 for different hidden layer sizes in the MLP classifier delivered a significantly low F1-score and recall value of less than 0.05 for class-1. This indicates a significant misclassification of class-1 images as class-2. The processing time for MobileNet, VGG16 and ResNet-50 are 0.0294s, 0.7525s

and 0.3239s respectively. In addition to their individual processing time, the processing time for the MLP classifier across different ensemble combinations, MobileNet-VGG16-ResNet-50, MobileNet-VGG16, MobileNet-ResNet-50 and VGG16-ResNet-50 are 0.3619s, 0.029s, 0.0.041s and 0.0.023s respectively.

IX. Future Work

The images of the CamenAI [1] dataset must be subjected to crowdsourcing annotations. This opens the door to fine-tune the BIQA regression models on the CamenAI dataset. It becomes interesting to study the effect of the ensemble containing the fine tuned BIQA models. This scenario may deliver even better results on the CamenAI dataset. Exploration towards other BIQA regression models should be made that behaves better with the CamenAI dataset. In this way, more classifiers can be used in the ensemble to produce higher accuracy. Since one of the models in the ensemble is affected by dashboard interference, a study must be made to crop this interference across all images and then evaluated by the BIQA model.

The results of the MLP classifier [28] for different ensemble combinations of the BIQA models tend to deliver a higher accuracy with respect to the number of hidden layers and neurons per layer used in the MLP classifier. Thus, a hyperparameter tuning approach should be used to determine the right number of hidden layers, and neurons per layer desired that offers the highest accuracy for the ensemble combination.

References

[1] CamenAI - In Control of the Pubic Space, https://www.camenai.com/en/

[2] Ou, Fu-Zhao, et al. "Controllable List-wise Ranking for Universal No-reference Image Quality Assessment." arXiv preprint arXiv:1911.10566 (2019)

[3] Talebi, Hossein, and Peyman Milanfar. "NIMA: Neural image assessment." IEEE Transactions on Image Processing 27.8 (2018): 3998-4011.

[4] Zhang, Weixia, et al. "Uncertainty-aware blind image quality assessment in the laboratory and wild." IEEE Transactions on Image Processing 30 (2021): 3474-3486.

[5] Zhang, Weixia, et al. "Blind image quality assessment using a deep bilinear convolutional neural network." IEEE Transactions on Circuits and Systems for Video Technology 30.1 (2018): 36-47.

[6] LIVE WILD Image Quality Dataset, https://live.ece.utexas.edu/research/ChallengeDB/index.html

[7] Koniq-10K Image Quality Dataset, http://database.mmsp-kn.de/koniq-10k-database.html

[8] Koniq-10K Image Quality Dataset, https://live.ece.utexas.edu/research/quality/subjective.htm

[9] Koniq-10K Image Quality Dataset, https://zenodo.org/record/2647033#.Yb-8NmjMKUk

[10] Fang, Yuming, et al. "Perceptual quality assessment of smartphone photography." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[11] Zhu, Hancheng, et al. "MetaIQA: deep meta-learning for no-reference image quality assessment." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[12] Su, Shaolin, et al. "Blindly assess image quality in the wild guided by a self-adaptive hyper network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[13] Yan, Jia, Jie Li, and Xin Fu. "No-reference quality assessment of contrast-distorted images using contrast enhancement." arXiv preprint arXiv:1904.08879 (2019).

[14] Varga, Domonkos. "Multi-pooled inception features for no-reference image quality assessment." Applied Sciences 10.6 (2020): 2186. 1987, pp. 740–741 [Dig. 9th Annu. Conf. Magnetics Japan, 1982, p. 301].

[15] Varga, Domonkos. "No-Reference Image Quality Assessment with Global Statistical Features." Journal of Imaging 7.2 (2021): 29.

[16] Li, Dingquan, Tingting Jiang, and Ming Jiang. "Exploiting high-level semantics for no-reference image quality assessment of realistic blur images." Proceedings of the 25th ACM international conference on Multimedia. 2017.

[17] Ou, Fu-Zhao, et al. "Controllable List-wise Ranking for Universal No-reference Image Quality Assessment." arXiv preprint arXiv:1911.10566 (2019).

[18] Zhang, Weixia, et al. "Learning to blindly assess image quality in the laboratory and wild." 2020 IEEE International Conference on Image Processing (ICIP). IEEE, 2020.

[19] Guan, Xiaodi, et al. "Quality Assessment on Authentically Distorted Images by Expanding Proxy Labels" https://www.mdpi.com/2079-9292/9/2/252

[20] Fang, Yuming, et al. "Perceptual quality assessment of smartphone photography." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

[21] Varga, Domonkos, Dietmar Saupe, and Tamás Szirányi. "DeepRN: A content preserving deep architecture for blind image quality assessment." 2018 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2018.

[22] Spearman's Rank Order Correlation Coefficient, https://en.wikipedia.org/wiki/Spearman27s_rank_correlation_coefficient

[23] Pearson's Linear Correlation Coefficient, https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

[24] TensorFlow Instantiation of MobileNet, https://www.tensorflow.org/api_docs/python/tf/keras/applications/mobilenet/MobileNet

[25] TensorFlow Instantiation of VGG16, https://www.tensorflow.org/api_docs/python/tf/keras/applications/vgg16/VGG16

[26] TensorFlow Instantiation of ResNet-50, https://www.tensorflow.org/api_docs/python/tf/keras/applications/resnet50/ResNet50

[27] Accuracy, Precision, Recall & F1 score interpretation of performance measures, https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of\-performance-measures/

[28] Multilayer Perceptron Classifier (MLP) using sklearn, https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

[29] Learning with NCES - Create a Bar Graph, https://nces.ed.gov/nceskids/graphing/classic/bar_pie_data.asp?ChartType=bar

[30] Absolute Difference, https://en.wikipedia.org/wiki/Absolute_difference

[31] Matplotlib Python Library, https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.hist.html

[32] draw.io - Flowchart Maker & Online Diagram Software, https://draw.io