BMJ Open

Deep learning-based automatic image quality assessment in ultra-widefield fundus photographs

Richul Oh,[1,2] Un Chul Park,[1,2] Kyu Hyung Park [ID] ,[1,2] Sang Jun Park [ID] ,[1,3] Chang Ki Yoon [ID] [1,2]

[1]Department of Ophthalmology, Seoul National University College of Medicine, Jongno-gu, Korea (the Republic of)
[2]Department of Ophthalmology, Seoul National University Hospital, Jongno-gu, Korea (the Republic of)
[3]Department of Ophthalmology, Seoul National University Bundang Hospital, Seongnam, Korea (the Republic of)

Correspondence to
Dr Chang Ki Yoon;
syst18@gmail.com and
Dr Sang Jun Park;
sangjunpark@snu.ac.kr

ABSTRACT

Objective With a growing need for ultra-widefield fundus (UWF) fundus photographs in clinics and AI development, image quality assessment (IQA) of UWF fundus photographs is an important preceding step for accurate diagnosis and clinical interpretation. This study developed deep learning (DL) models for automated IQA of UWF fundus photographs (UWF-IQA model) and investigated intergrader agreements in the IQA of UWF fundus photographs.

Methods and analysis We included 4749 UWF images of 2124 patients to set the UWF-IQA dataset. Three independent board-certified ophthalmologists manually assessed each UWF image on four grading criteria (field of view, peripheral visualisation, details of posterior pole and centring of the image) and a final IQA grading using a five-point scale. The UWF-IQA model was developed to predict IQA scores with EfficientNet-B3 as the backbone model. For the test dataset, Cohen's quadratic weighted kappa score was calculated to evaluate intergrader agreements and agreements between predicted IQA scores and manual gradings.

Results Development and test dataset consisted of 3790 images from 1699 patients and 959 images of 425 patients, respectively, without statistical differences in IQA gradings. The average agreement between the UWF-IQA model and manual graders was 0.731, while the average of intergrader agreements among manual graders was 0.603 (Cohen's weighted kappa score). Posterior pole grading showed the highest average agreements (0.838) between the UWF-IQA model and manual graders, followed by final grading (0.788), centring of the image (0.754), peripheral visualisation (0.754) and field of view (0.535).

Conclusion Predicted IQA scores using the UWF-IQA model showed better agreements with manual graders compared with intergrader agreements. The automated UWF-IQA model offers robust and efficient IQA predictions with the final and subcategory gradings.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ We developed automated image quality assessment (IQA) models for UWF fundus photographs, predicting subcategories and final grading based on assessments from three graders.
⇒ Agreement between the UWF-IQA model and manual graders was higher than agreement among manual graders, demonstrating robust and efficient IQA predictions.
⇒ This study supports the development and inference of DL models by providing IQAs for UWF fundus photographs.
⇒ The study included images from only one device, which may limit generalisability.
⇒ Our manual IQA system is newly proposed and has not yet been validated in other studies.

INTRODUCTION

The development of imaging technology and devices has aided ophthalmologists in diagnosing diseases and interpreting disease progression. However, not all images are of good quality in real-world clinical settings[1]. Image quality assessment (IQA) of various imaging modalities is an essential preceding step for clinical implication. Artificial intelligence (AI) models require good image quality for stable and meaningful prediction[2] and are highly dependent on the datasets used during their development. AI model developed with only good-quality images is likely to show degraded performance in real-world clinical environments.[3 4] To generalise AI models to real-world clinics, researchers have focused on IQA in the ophthalmology field.[5 6]

Datasets for AI model development typically include high-quality images to minimise confusion during labelling and improve model accuracy. To prevent issues arising from data shifts and to generalise the models, it is essential to ensure that datasets include images of varying quality, particularly when creating generalised models that include low-quality images. IQA can be used not only for collecting high-quality images but also for demonstrating the limitations caused by low-quality images during inference. However, manual image quality grading by human experts is highly time-consuming and inconsistent, especially in

the era of big data. Thus, AI models for IQA can play crucial roles in such domains.

Ultra-widefield fundus (UWF) fundus photographs can visualise up to 200° and provide more information about the peripheral retina than colour fundus photographs (CFP). With a growing need for UWF fundus photographs in clinics and AI development, researchers and clinicians need IQA of UWF fundus photographs for further implication. Recently, Cui *et al* developed an automated IQA deep learning (DL) model and reported that the poor quality of UWF fundus photographs has reduced the performance of the DL models, indicating the importance of IQA in UWF fundus photographs.[7] One grader classified image quality into two classes, poor and good, for the IQA model development. However, considering that IQA is subjective and can vary between graders in CFP,[8] one grader for ground truth may not be sufficient as intergrader agreement in IQA has not been studied for UWF fundus photographs. Therefore, in this study, we investigated inter-grader agreements between three graders in the IQA of UWF fundus photographs and developed DL models for automated IQA of UWF fundus photographs (UWF-IQA model).

MATERIAL AND METHODS
Ethical statements
This study was approved by the Institutional Review Board of Seoul National University Hospital (SNUH; No. H-2202-069-1299). All procedures were conducted in compliance with the principles of the Declaration of Helsinki. The Institutional Review Board waived the need for written informed consent due to the retrospective design of the study and the complete anonymisation of patient information. We removed all patient-specific information (eg, patient identification number and name) after extracting information.

Patient and Public Involvement Statement
Patients or the public were not involved in the design, conduct, reporting, or dissemination plans of our research.

Dataset
We retrospectively enrolled all patients who visited the ophthalmology clinic at SNUH between September 2018 and December 2021. The dataset was constructed as previously reported.[9] The UWF-Master dataset included patients with ocular biometric measurements and UWF fundus photographs that were taken within 3 months before the ocular biometric measurement. The UWF fundus photographs were obtained using the Optos California retinal imaging systems (Optos Inc, Dunfermline, United Kingdom). The ocular biometric measurement device was an IOLMaster 700 (Carl Zeiss Meditec, Jena, Germany), a swept source OCT-based ocular biometer.

We used automatically reconstructed pseudocolour UWF fundus photographs. Data acquisition was performed by the big data centre of SNUH. The image files were originally in the 'DICOM' format. The files were subsequently converted into 'png' format. The original size of the images was 3900×3072 pixels. All the red, green and blue (RGB) channels from the pseudocolour images were used.
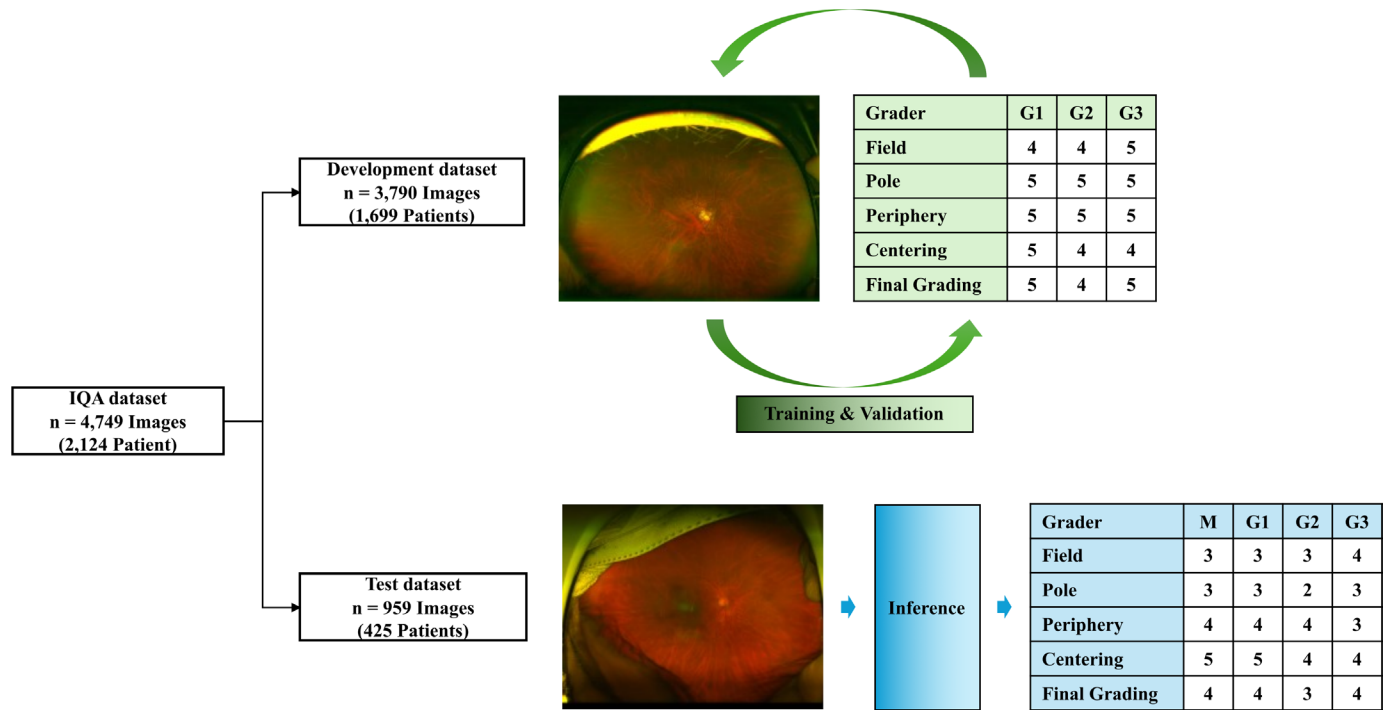
The UWF-Master dataset consisted of 9426 UWF fundus photographs of 3954 patients. Due to the resource limitation, we randomly selected 4729 images of 2124 patients for the IQA dataset. For each image in the IQA dataset, four subcategory grading (field of view, details of posterior pole, peripheral visualisation and centring of images) and one final grading were manually assessed by three ophthalmologists (Grader 1, 2 and 3) using a Likert scale from 0 (worst) to 4 (best). Online supplemental data 1 shows the definition and explanation of each category.

Development of DL models
Figure 1 shows an overview of the development and inference process. During the development process, inputs of the DL model were UWF fundus photographs and outputs of the DL model were IQA gradings of three graders. The IQA dataset was partitioned into development dataset and test dataset at the patient level in a 4:1 ratio. Images from the same patient did not belong to both datasets. Using the development dataset, we applied fivefold cross-validation, and a DL model was developed for each fold. During the inference process, we averaged the prediction values of three outputs to determine the predicted IQA. Then, we averaged the prediction values of the folds for each category. For each category, IQA grading with the highest average prediction values was determined as the predicted IQA.

We used the following augmentations: random crop (lower bounds 8% of the whole image and upper bounds 100%) and resizing the cropped patches to 300×300, random horizontal flipping, and image normalisation. We selected EfficientNet-B3[10] as a backbone of the DL model. EfficientNet is a publicly available lightweight convolutional neural network architecture introduced in 2019. Among the EfficientNet models ranging from B0 to B7, with B7 being the largest, EfficientNet-B3 was selected as the baseline architecture due to limited computing resources. We added customised layers instead of the last fully connected layer (FC layer) of the model. The customised layers consisted of FC layer (output: 1×1024), rectified linear unit (ReLU) layer, dropout layer (dropout rate: 0.2), FC layer (output: 1×128), ReLU layer, dropout layer (dropout rate: 0.2), FC layer (output: 1×64), ReLU layer, dropout layer (dropout rate: 0.2) and FC layer (output: 3×5). The final layers of the classifiers above are softmax layers. Figure 2 illustrates the architecture of the models. We used a transfer learning strategy by initialising the backbone model using pre-trained weights from the ImageNet dataset and fine-tuned based on our dataset.

We used the batch size of 32 and AdamW optimiser (weight decay: 0.05). A total of 50 training epochs were set, with the initial 10 epochs designated for the warm-up

Grader	G1	G2	G3
Field	4	4	5
Pole	5	5	5
Periphery	5	5	5
Centering	5	4	4
Final Grading	5	4	5

Grader	M	G1	G2	G3
Field	3	3	3	4
Pole	3	3	2	3
Periphery	4	4	4	3
Centering	5	5	4	4
Final Grading	4	4	3	4

Figure 1 Overview of the development and inference process. IQA: image quality assessment.

phase (the learning rate gradually increased from 0 to 1×10^{-3}), followed by cosine annealing scheduling (the learning rate gradually descended from 1×10^{-3} to 1×10^{-8}). After each epoch, the model was evaluated using a validation set. The model weights with the lowest average-weighted cross-entropy loss in the validation set were preserved as the model checkpoint. The weight for

cross-entropy loss was set as a ratio of 2:2:1:1:1, which corresponded to Grade 0, 1, 2, 3, and 4, respectively, to increase the sensitivity of low-quality images. After the training phase, the best model in the model checkpoint was selected for the inference phase of the test set.

Pytorch (V.1.13.0) and Pytorch-lightning (V. 1.9.0) were used to develop the fine-tuned models. All development and inference processes were performed on a private server equipped with NVIDIA GeForce RTX 4090 GPU (24 GB) with CUDA V. 12.2, powered by an Intel Xeon Silver 4210R Processor CPU (13.75 MB Cache, 2.40 GHz) in the Ubuntu 22.04 system environment with 192 GB of memory.

Analysis of results
We compared manual gradings of three graders and predicted IQA gradings of the UWF-IQA model in the internal test set using Cohen's quadratic weighted kappa score was used for evaluating agreements between the gradings.

RESULTS
The IQA dataset consisted of 4749 images from 2124 patient. The development dataset consisted of 3790 images from 1699 patients and test dataset consisted of 959 images of 425 patients. Distribution of total number of IQA gradings in the development data set and test dataset showed no statistical differences (table 1, all p>0.05).

Table 2 presents the results of intergrader agreements between manual graders and between the UWF-IQA model and manual graders in the test dataset. The Cohen's quadratic weighted kappa scores indicated
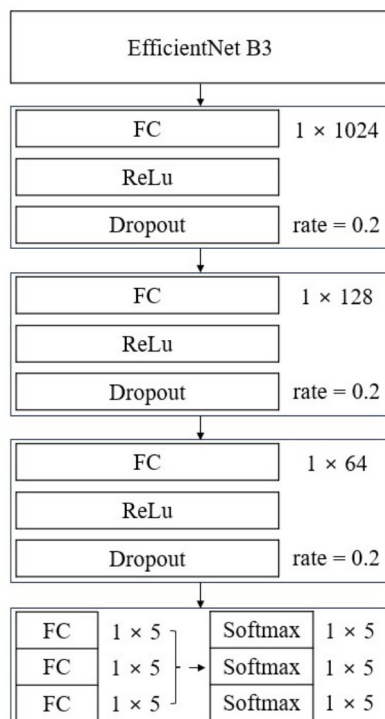
Figure 2 Architecture of IQA model. FC, fully-connected layer; ReLu, rectified linear unit.

3

Table 1 Distributions of total number of IQA gradings in development set and test set

	Development set (n=3790 × 3 graders) (Grading 4/3/2/1/0)	Test set (n=1699 × 3 graders) (Grading 4/3/2/1/0)	P values*
Field of view	5907/3148/1651/335/329 (52.0 % / 27.7 % / 14.5 % / 2.9 % / 2.9%)	1524/761/416/88/88 (53.0 % / 26.5 % / 14.5 % / 3.1 % / 3.1%)	0.726
Details of posterior pole	8222/1177/861/692/418 (72.3 % / 10.4 % / 7.6 % / 6.1 % / 3.7%)	2058/295/216/193/115 (71.5 % / 10.3 % / 7.5 % / 6.7 % / 4.0%)	0.682
Peripheral retina visualisation	6140/2838/1534/648/210 (54.0 % / 25.0 % / 13.5 % / 5.7 % / 1.8%)	1592/677/385/174/49 (55.3 % / 23.5 % / 13.4 % / 6.0 % / 1.7%)	0.490
Centering of image	8173/885/1483/538/291 (71.9 % / 7.8 % / 13.0 % / 4.7 % / 2.6%)	2028/230/402/161/56 (70.5 % / 8.0 % / 14.0 % / 5.6 % / 1.9%)	0.054
Final grading	4388/3798/2,272/738/174 (38.6 % / 33.4 % / 20.0 % / 6.5 % / 1.5%)	1097/951/590/195/44 (38.1 % / 33.1 % / 20.5 % / 6.8 % / 1.5%)	0.936

*χ^2 test

moderate agreement among the three manual graders, with an average score of 0.603 (range: 0.105–0.825). Grader 2 and Grader 3 had the highest average kappa score of 0.678, followed by Grader 1 and Grader 2 (0.609), and Grader 1 and Grader 3 (0.523). The average kappa score for final grading among manual graders was 0.673. For subcategory gradings, details of posterior pole had the highest average kappa score of 0.798, followed by centring of the image (0.638), peripheral retina visualisation (0.636), and field of view (0.272).

The UWF-IQA model demonstrated a higher average kappa score (0.731) with manual graders compared with the intergrader scores. The UWF-IQA model achieved the highest average kappa score with Grader 2 (0.803), followed by Grader 3 (0.737) and Grader 1 (0.652). The average kappa score for final grading between the UWF-IQA model and manual graders was 0.783. Among the subcategories, details of posterior pole had the best average kappa score of 0.814, followed by peripheral retina visualisation (0.756), centring of the image (0.736) and field of view (0.555). In the subgroup analysis using 693 phakic eye images, the UWF-IQA model demonstrated a lower average kappa score (0.693) with manual graders compared with the whole group. However, the model showed a higher average kappa score with manual graders compared with the intergrader scores (0.554). (table 3)

Figure 3 shows representative images of predicted IQA gradings. The total time spent for predicting all IQA gradings using the UWF-IQA model was 40 min and 18 s, averaging 2.52 s per image.

DISCUSSION

In this study, we developed automated IQA models for UWF images predicting subcategories and final grading. Agreements of IQA gradings between the UWF-IQA model and manual graders were greater than agreements between manual graders. To the best of our knowledge, this was the first to evaluate IQA gradings of UWF images with subcategories.

Many clinical studies using fundus images have an exclusion process of poor-quality images. However, due to the increasing amount of data, IQA needs a time-intensive and labour-intensive process if manually applied. This preliminary step limits the development and clinical application of DL models. To overcome this limitation, several IQA algorithms for CFPs have been introduced previously.[5 6 11–15] König et al reported automated IQA for CFP and fluorescein angiography images using DL. Their

Table 2 Intergrader agreements between manual graders and agreements between UWF-IQA model and manual graders in the test dataset using Cohen's weighted kappa score

	Inter-grader agreements				Agreements between model and graders			
	G1 vs G2	G1 vs G3	G2 vs G3	Average	M vs G1	M vs G2	M vs G3	Average
Field of view	0.433	0.105	0.280	0.272	0.541	0.669	0.455	0.555
Details of posterior pole	0.781	0.795	0.820	0.798	0.760	0.844	0.840	0.814
Peripheral retina visualisation	0.599	0.555	0.754	0.636	0.683	0.800	0.786	0.756
Centering of image	0.532	0.556	0.825	0.638	0.562	0.847	0.827	0.745
Final grading	0.701	0.604	0.715	0.673	0.714	0.857	0.777	0.783
Average	0.609	0.523	0.678	0.603	0.652	0.803	0.737	0.731

G, grader; IQA, image quality assessment; M, model; UWF, ultra-widefield fundus.

Table 3 Intergrader agreements between manual graders and agreements between UWF-IQA model and manual graders in 693 phakic eye images

	Intergrader agreements				Agreements between model and graders			
	G1 vs G2	G1 vs G3	G2 vs G3	Average	M vs G1	M vs G2	M vs G3	Average
Field of view	0.376	0.072	0.128	0.192	0.543	0.598	0.293	0.478
Details of posterior pole	0.718	0.712	0.770	0.733	0.701	0.815	0.805	0.774
Peripheral retina visualisation	0.540	0.498	0.690	0.576	0.622	0.757	0.750	0.710
Centring of image	0.568	0.629	0.843	0.680	0.602	0.856	0.874	0.777
Final grading	0.632	0.500	0.638	0.590	0.639	0.830	0.710	0.726
Average	0.567	0.482	0.614	0.554	0.621	0.771	0.684	0.693

G, grader; IQA, image quality assessment; M, model; UWF, ultra-widefield fundus.

model predicted IQA scores for 3 and 4 modality-specific categories and overall quality scores together with an uncertainty score.[5] FundusQ-Net provided a regression quality assessment ranging from 1 to 10 grading instead of binary classification as in most studies.[16]

While there has been extensive research on the IQA in CFPs using open-source databases, IQA for UWF images has undergone relatively less research because there are currently no open-source databases available for this modality. Li *et al* developed a DL-based image filtering

system using 40 562 UWF images.[17] The DL model for the classification of poor-quality UWF images achieved AUCs of 0.994, 0.996 and 0.997 for three datasets. Our study differed from the previous study in several aspects. First, we hypothesised that there is no definite ground truth for image quality as the manual grading process is highly subjective. We analysed agreements between the AI model and manual graders instead of analysing AUC or accuracy, in which the definite ground truths for IQA were set. Second, we assessed image qualities with four
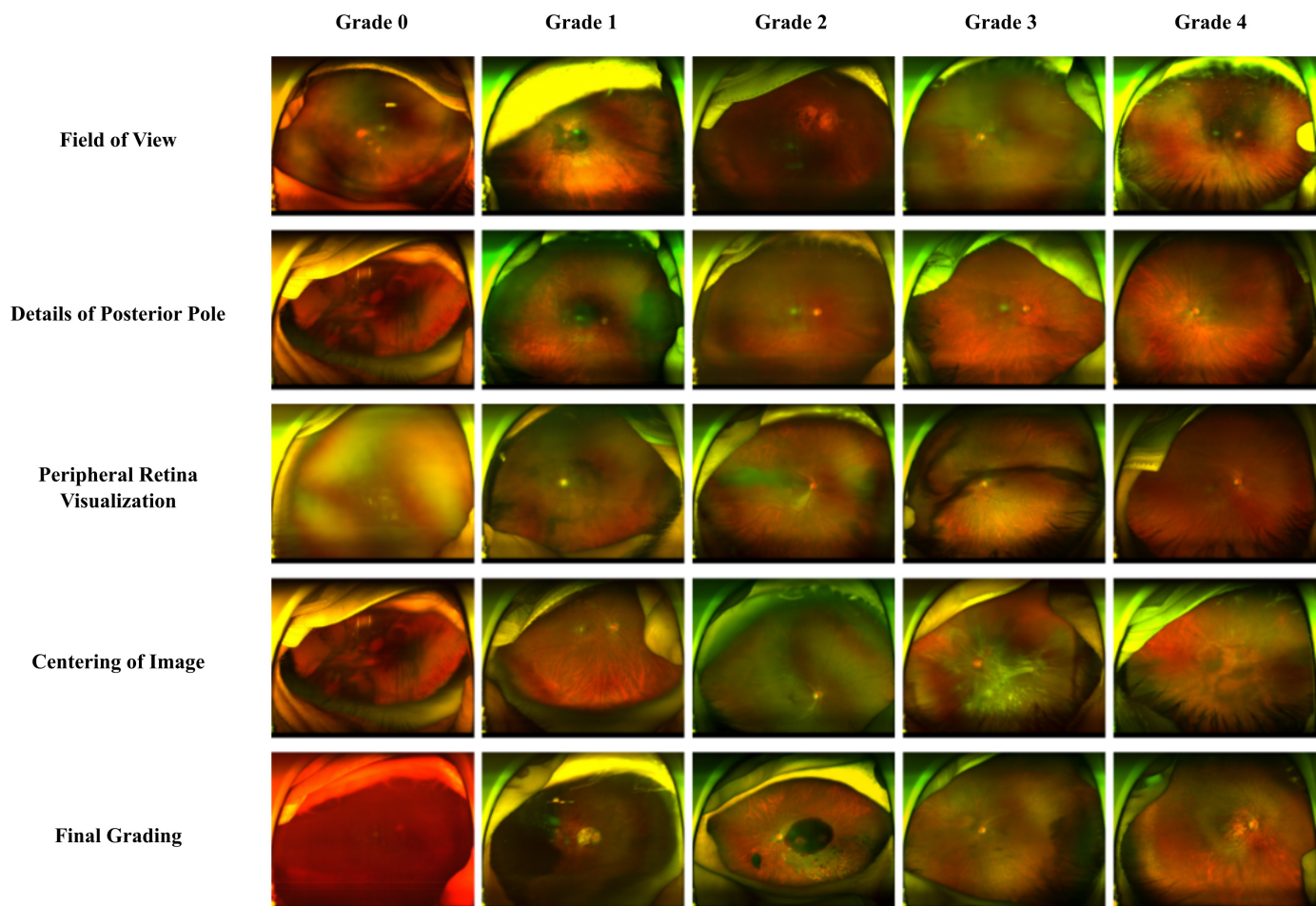
Figure 3 Representative images of IQA gradings predicted by the UWF-IQA model. IQA, image quality assessment; UWF, ultra-widefield fundus

subcategories and one final grading for each image. Third, we assessed image quality with five-scale gradings instead of binary classification.

In the domain of precise diagnostics, the expertise of a highly skilled and experienced professional can be regarded as a definitive ground truth. However, in the subjective area of IQA, there is considerable variability and disagreement, even among experienced ophthalmologists. Laurik-Feuerstein *et al* showed that IQA in CFPs showed moderate inter-rater agreement (Cohen's weighted kappa score of 0.564) when using four scales. Agreements between graders with medical background (0.590) were higher than those between non-medical background graders (0.554); however, there was still moderate inter-rater agreement.[8] Image quality assessed by a single grader, even an experienced professional, can be highly biased and is not appropriate for DL model development. Our study included three ophthalmologists with varying experiences in IQA of UWF fundus photographs. The average inter-grader agreement was 0.603, which was similar to the previous study using CFPs.[8] To minimise subjective discrepancies and enhance the generalisability of evaluations, we set our loss function as an average of categorical cross-entropy loss for three IQA grades. This approach resulted in a more robust model with average agreement between the DL model and graders (0.731), which was greater than inter-grader agreements (0.603). Agreements of each subcategory and final grading showed similar tendencies as well.

We evaluated image quality with a final grading and four subcategories for each image. These subcategories can facilitate various clinical applications and approaches. For instance, diagnosing glaucoma or macular diseases may require a higher level of detail in the posterior pole compared with diagnosing retinal detachment or peripheral retinal diseases. For images with the same final grading, the subcategory gradings play a crucial role in the specific disease categories.

We used five-scale gradings for each category, offering several advantages over binary classification in the previous studies.[18 19] Binary classification requires selecting an arbitrary cut-off to distinguish between good-quality and poor-quality images, which can be highly subjective and influenced by the grader's experience. In contrast, a five-scale grading enables us to establish a less biased cut-off for determining acceptable quality. In research using fundus images, there may be cases where including poor-quality images is necessary for real-world studies, while other scenarios might involve only high-quality images. To incorporate these diverse tasks, different thresholds are needed. Binary classification lacks intermediate stages, which can have moderate qualities, making it unavailable for applying various thresholds. However, with a five-scale system, researchers can set thresholds according to their specific goals, thereby broadening the scope of data usage.

The UWF-IQA model demonstrated superior agreement compared with manual graders across all categories,

exhibiting consistently better agreements on average. This suggests that the UWF-IQA model gives more robust predictions than human grading, suggesting the UWF-IQA model be used in further studies. Additionally, the average inference time is 2.52s per image, making it significantly more efficient than manual grading, especially in large-scale studies using UWF images.

A single DL model which performs various prediction tasks simultaneously may not be efficient. We suggest that a staged approach or parallel approach be set together to perform various prediction tasks. In both development and inference processes, the UWF-IQA model is crucial. In the development process, the model can exclude poor-quality images to achieve better performance. In the inference process, the model can provide the image qualities of the images as well,[7 20] from which clinicians can independently evaluate the reliability of the findings. The UWF-IQA model can be particularly beneficial in AI-assisted diagnostics and telemedicine by providing warnings for low-quality images, aiding both clinicians and users. In the subgroup analysis of phakic eyes, where cataracts are a major cause of low-quality images, the model demonstrated higher agreement with graders than inter-grader agreement. Additionally, its performance did not significantly decline compared with the overall test set, suggesting that it may perform well in real-world datasets with a high prevalence of phakic eyes.

The limitations of our study should be noted. First, the study included only one device (Optos California retinal imaging system) from a single institution. However, it is one of the most widely used commercially available UWF devices. External validation using other devices and other institutions could be considered for generalisation. Second, our manual IQA system is newly proposed and, thus, has not been validated in other studies. Third, the dataset size was smaller than the previous study using UWF fundus photographs.[17] However, every image was assessed by three independent graders. Intergrader variability has been considered in the model development. Finally, data augmentation techniques were selected based on the study's purpose. Cropping and horizontal flipping were used to maintain anatomical consistency while enhancing model generalisation, whereas rotation and brightness adjustments were excluded to preserve clinical interpretability. Since low-quality images were relatively rare, a weight ratio was applied to address this imbalance. Different data augmentation strategies could impact model performance.

In conclusion, we developed the UWF-IQA model to predict the IQA score of UWF fundus photographs. Predicted IQA scores using the UWF-IQA model showed better agreements with manual graders compared with inter-grader agreements. The automated UWF-IQA model offers robust and efficient IQA predictions with the final and subcategory gradings.

ORCID iDs
Kyu Hyung Park http://orcid.org/0000-0002-5516-8121
Sang Jun Park http://orcid.org/0000-0003-0542-2758
Chang Ki Yoon http://orcid.org/0000-0003-4637-8044

REFERENCES

1 Piyasena MP, Hewage SN, Banduthilake K. Retinal image quality assessment in diabetic-retinopathy screening: Real world evidence from a lower-middle income country. *LAJO* 2024;7:8.
2 Maier K, Zaniolo L, Marques O. Image quality issues in teledermatology: A comparative analysis of artificial intelligence solutions. *J Am Acad Dermatol* 2022;87:240–2.
3 Finlayson S, SubbaswamyA, Singh K, *et al*. The Clinician and Dataset Shift in Artificial Intelligence. *N Engl J Med* 2021;385:283–6.
4 Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. *Biostatistics* 2020;21:345–52.
5 König M, Seeböck P, Gerendas BS, *et al*. Quality assessment of colour fundus and fluorescein angiography images using deep learning. *Br J Ophthalmol* 2024;108:98–104.
6 Wang X, Zhang S, Liang X, *et al*. A CNN-BASED RETINAL IMAGE QUALITY ASSESSMENT SYSTEM FOR TELEOPHTHALMOLOGY. *J Mech Med Biol* 2019;19:1950030.
7 Cui T, Lin D, Yu S, *et al*. Deep Learning Performance of Ultra-Widefield Fundus Imaging for Screening Retinal Lesions in Rural Locales. *JAMA Ophthalmol* 2023;141:1045–51.
8 Laurik-Feuerstein KL, Sapahia R, Cabrera DeBuc D, *et al*. The assessment of fundus image quality labeling reliability among graders with different backgrounds. *PLoS ONE* 2022;17:e0271156.
9 Oh R, Lee EK, Bae K, *et al*. Deep Learning-based Prediction of Axial Length Using Ultra-widefield Fundus Photography. *Korean J Ophthalmol* 2023;37:95–104.
10 Tan M. EfficientNet: rethinking model scaling for convolutional neural networks. 2019.
11 Gonçalves MB, Nakayama LF, Ferraz D, *et al*. Image quality assessment of retinal fundus photographs for diabetic retinopathy in the machine learning era: a review. *Eye (Lond)* 2024;38:426–33.
12 Zago GT, Andreão RV, Dorizzi B, *et al*. Retinal image quality assessment using deep learning. *Comput Biol Med* 2018;103:64–70.
13 Bolla M, Biswas S, Palanisamy R. Deep Learning Based Quality Prediction of Retinal Fundus Images. *Curr Dir Biomed Eng* 2023;9:706–9.
14 Pérez AD, Perdomo O, González FA. A lightweight deep learning model for mobile eye fundus image quality assessment. In: Brieva J, Romero E, Lepore N, eds. 15th International Symposium on Medical Information Processing and Analysis; Medelin, Colombia. 10.1117/12.2547126 Available: https://doi.org/10.1117/12.2547126
15 Chalakkal RJ, Abdulla WH, Thulaseedharan SS. Quality and content analysis of fundus images using deep learning. *Comput Biol Med* 2019;108:317–31.
16 Abramovich O, Pizem H, Van Eijgen J, *et al*. FundusQ-Net: A regression quality assessment deep learning algorithm for fundus images quality grading. *Comput Methods Programs Biomed* 2023;239:107522.
17 Li Z, Guo C, Nie D, *et al*. Deep learning from 'passive feeding' to 'selective eating' of real-world data. *NPJ Digit Med* 2020;3:143.
18 Costa P, Campilho A, Hooi B, *et al*. EyeQual: accurate, explainable, retinal image quality assessment. 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA); Cancun, Mexico. 10.1109/ICMLA.2017.0-140 Available: https://doi.org/10.1109/ICMLA.2017.0-140
19 Saha SK, Fernando B, Cuadros J, *et al*. Automated Quality Assessment of Colour Fundus Images for Diabetic Retinopathy Screening in Telemedicine. *J Digit Imaging* 2018;31:869–78.
20 Chuter B, Huynh J, Bowd C, *et al*. Deep Learning Identifies High-Quality Fundus Photographs and Increases Accuracy in Automated Primary Open Angle Glaucoma Detection. *Transl Vis Sci Technol* 2024;13:23.