Ultra-Widefield Fundus Imaging for Diabetic Retinopathy - MICCAI Challenge 2024 - Team [ailsjku2024] Submission

Berthold Scheuringer[1], Moritz Haderer[1], Martin Marinschek[1], Oleksandra Menzatiuk[1,2], and Vera Pils[1,3,4]

[1] Johannes Kepler University, 4040 Linz, Austria
[2] Silicon Austria Labs, 4040 Linz, Austria
[3] BOKU Core Facility Bioinformatics, BOKU University, 1180 Vienna, Austria
[4] Institute of Statistics, Department of Landscape, Spatial and Infrastructure Sciences, BOKU University, 1180 Vienna, Austria

miccai.challenge.jku@gmail.com

Abstract. Diabetic Retinopathy (DR) is a leading cause of preventable blindness, affecting over 100 million adults globally, with prevalence expected to rise significantly by 2045. Early detection of DR and diabetic macular edema (DME) is crucial for timely treatment. This study addresses the challenges of DR diagnosis using ultra-widefield (UWF) fundus images, which provide a 200-degree retinal view We participated in the MICCAI 2024 UWF4DR challenge, tackling three tasks: image quality assessment, DR detection, and DME identification, using AutoMorph, ShuffleNet, and EfficientNetB0 models, respectively. Our approach secured competitive leaderboard rankings, highlighting the potential of deep learning models for scalable, automated DR diagnosis in clinical settings.

Keywords: Diabetic Retinopathy (DR) · Ultra-Widefield (UWF) fundus imaging · Diabetic Macular Edema (DME) · Deep Learning.

1 Introduction

As of 2020, an estimated 103 million adults were affected by Diabetic Retinopathy (DR), and this number is projected to rise to 161 million by 2045 [1]. DR is a major complication of diabetes mellitus and is one of the leading causes of preventable blindness among working-aged individuals worldwide [2].

Early signs of DR [3] include microaneurysms, which manifest as small, sharply defined red dots on the retina due to abnormal leakage from retinal blood vessels. When these vessels swell and rupture, they are hemorrhages. Continued leakage from these capillaries leads to the formation of exudates, which are fluid deposits rich in protein and cellular material, often seen as yellow spots near the outer retinal layers, a condition known as diabetic macular edema (DME).

Manual diagnosis of DR is time-consuming and resource-intensive. However, computer-aided diagnosis can greatly reduce both time and costs. In recent years,

machine learning and deep learning have proven to be highly effective tools in this domain [4], [5], [6]. Traditionally, the gold standard for DR classification has been standard color fundus photography, which provides a 30 to 50-degree field of view encompassing the macula and optic nerve. However, ultra-widefield (UWF) fundus imaging, which offers a broader 200-degree view of the retina, has emerged as an advantageous alternative.

Despite various methods proposed for detecting DR in retinal fundus images, the task remains challenging [3]. Key difficulties stem from the spherical shape of the eye, leading to uneven lighting—brighter in the center and darker at the periphery—as well as low image contrast and the small size of lesions such as microaneurysms. Further, normal retinal structures share visual similarities with pathological features like exudates and hemorrhages or DR-related features can sometimes be distorted by image pre-processing techniques, making detection even more complex.

2 Details on the MICCAI UWF4DR 2024 Challenge

The ultra-widefield fundus images for diabetic retinopathy (UWF4DR) challenge is associated with the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)[5]. In this challenge, the objective was to classify ultra-widefield (UWF) fundus images and consisted of three tasks. Task 1 involved determining image quality, distinguishing between poor quality (label 0) and good quality (label 1). Task 2 required classifying images based on the presence of DR (label 1) or absence (label 0). Task 3 focused on identifying the presence (label 1) or absence of DME (label 0).

Participants were provided with a training dataset, which was utilized for model development. Additionally, a separate validation dataset was used for performance evaluation and a public leaderboard score. The validation data remained unknown to the participants throughout the competition (Table 1). In the end, the participants could run their proposed model on the test set with a hidden leaderboard. Despite our limited GPU resources, our models achieved 6th place on the public leaderboard for all the tasks.

Table 1. Different Tasks in the MICCAI Challenge.

Task	Name	Training Set	Validation Set
1	Image quality assessment for ultra-widefield fundus	434	61
2	Identification of referable diabetic retinopathy	201	50
3	Identification of diabetic macular edema	167	45

To evaluate the performance of the models comprehensively, the Area Under the Curve (AUC) and the average inference time per image was used as

[5] https://codalab.lisn.upsaclay.fr/competitions/18605

evaluation metrics. The ranking score is calculated as:

$$\text{Ranking Score} = \text{AUC} - 5\% \times \text{Time (seconds)}$$

Although the inference time provided by Codalab (challenge platform) serves as a reference, it may fluctuate. Therefore to ensure fairness, an average out of 100 iterations is calculated. A higher ranking score indicates better performance. Auxiliary metrics such as AUPRC, sensitivity, and specificity will be used as tie-breakers if needed.

3 Dataset

The UWF fundus images were provided as colored JPG files with a resolution of 800×1016 pixels. The dataset for task 1 consisted of 434 UWF fundus images. To enhance the dataset, we incorporated additional public data from the Deep-DRID dataset [7]. Label 5 in the Deep-DRID dataset indicated bad quality and we relabeled them as 0. DR is indicated as having different severity levels, which uses a scale of 0 to 4. We changed the label 1 for all the good-quality cases. An example of both datasets can be found in the Appendix (Supplemental Fig. 11).

The dataset used for task 2 consists of UWF4DR and Deep-DRID. Images labeled with 0 indicate no presence of DR, while those labeled with 1 indicate the presence of DR. As we had a binary task, we also had to change the Deep-DRID labels. Zero was kept 0 for no DR, and 1-4 was changed into 1, for DR. Example images for both labels can be seen in Fig. 1.

For task 3, we utilized a dataset consisting of 167 UWF fundus images as provided: 77 images with DME (labeled as 1), and 90 images without DME (labeled as 0) (see Appendix Supplemental Fig. 12). Table 2 presents the distribution of image labels for every task in both datasets.

Table 2. Label Distribution Across UWF4DR and Deep-DRID Datasets.

Task	Label	UWF4DR	Deep-DRID	**Total**
Task 1	0	205	4	**209**
	1	229	198	**427**
	Total	**434**	**202**	**636**
Task 2	0	89	60	**149**
	1	112	138	**250**
	Total	**201**	**198**	**399**
Task 3	0	90	-	**90**
	1	77	-	**77**
	Total	**167**	-	**167**

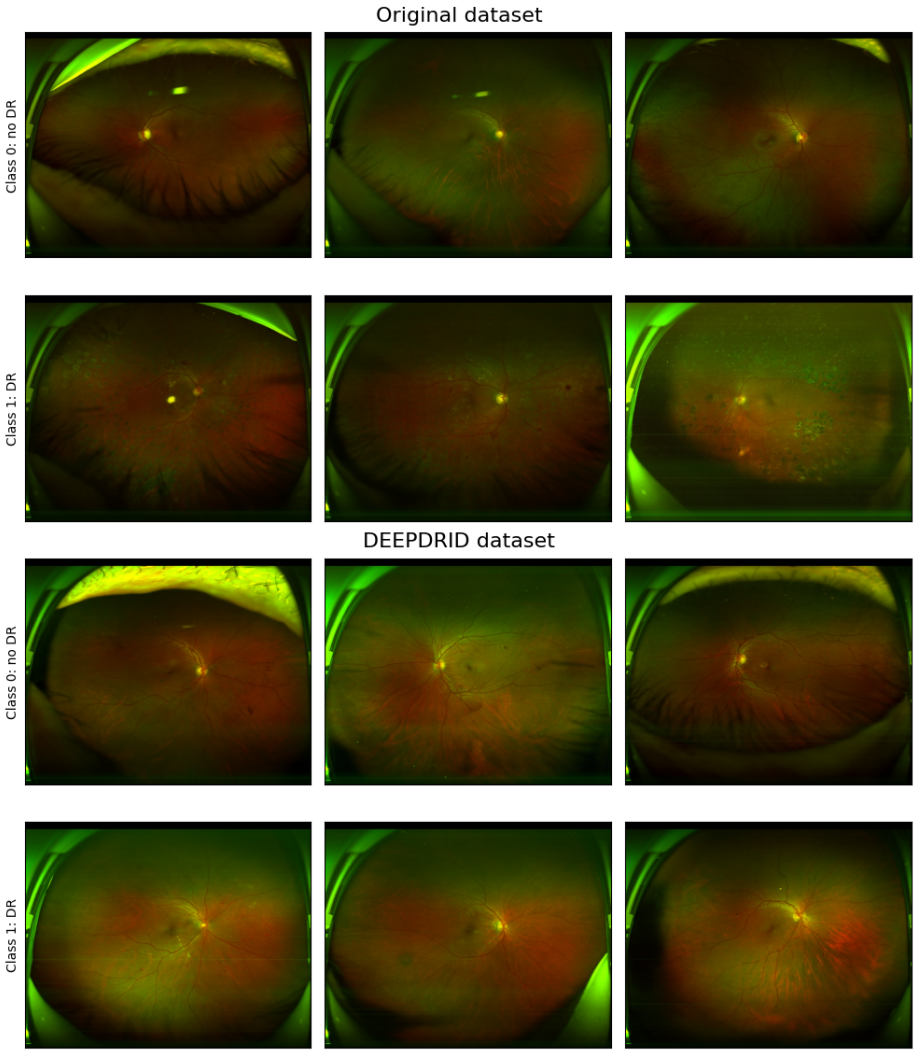

Fig. 1. UWF fundus images (Task 2). Patients that have DR versus not. Additionally, we also included images from Deep-DRID.

4 Method

We used different data augmentations for preprocessing and a different model for each task as different models seem to be superior at each task. We used an Automorph model for defining the quality of image, ShuffleNet for identifying DR and an EfficientNet for DME.

4.1 Preprocessing

Our data augmentation methods spanned from general computer vision techniques to more specialized medical imaging approaches. For the quality assessment, we resized the images to 800x800 using zero padding to preserve the aspect ratio. Additionally, we applied random horizontal and vertical flips to augment the data.

To focus the model on relevant retinal structures for detecting DR, we used GIMP to apply an elliptical mask, setting areas of the image outside the pupil - specifically those showing parts of the imaging device - to 0. This step effectively eliminates unnecessary background information. Each image is masked before entering the augmentation pipeline. For data augmentation, we first applied a residual Gaussian blur with 0.5 probability, which adds each image to its own blurred version. Next, Contrast Limited Adaptive Histogram Equalization (CLAHE) was used to enhance local contrast - again with probability 0.5. Furthermore, each image had a 50% chance of undergoing a horizontal flip and a one-third chance of being randomly rotated by an angle uniformly distributed between -10° and +10°, introducing variation to enhance model robustness.

A series of pre-processing steps were applied to help the model focus on features relevant to DME. Initially, we applied histogram equalization to half of the images, followed by geometric transformations (random horizontal flips).

All images were normalized using the ImageNet mean and standard deviation. These steps aimed to enhance the model's ability to identify DR and DME while minimizing false positives due to artifacts or noise. It should be mentioned, that we had limited hardware resources and therefore focused on constructing a pipeline consisting of several hardware-resource efficient neural architectures, which we could train with our limited number of available GPUs.

4.2 Automorph Model: Image Quality Assessment

We utilized the AutoMorph model [8] for the image quality assessment. AutoMorph is based on the EfficientNetB4 architecture [9] and features a fully connected head for feature extraction and a final output layer that produces a single scalar value. AutoMorph, a publicly available model, is designed for automated analysis of retinal vascular morphology on fundus images, aiding research in both ophthalmic and systemic diseases. AutoMorph was pretrained on two large image quality grading datasets: EyePACS-Q [10] (28,792 fundus images) and DDR-test [11] (13,673 fundus images), which allowed us to leverage the model's robust feature extraction capabilities for our image quality assessment task. The method pipeline for the task 1 can be seen in Fig. 2.

4.3 ShuffleNet for Diabetic Retinopathy

We implemented the ShuffleNet architecture [12], a computation-efficient convolutional neural network (CNN) that leverages group convolutions, channel

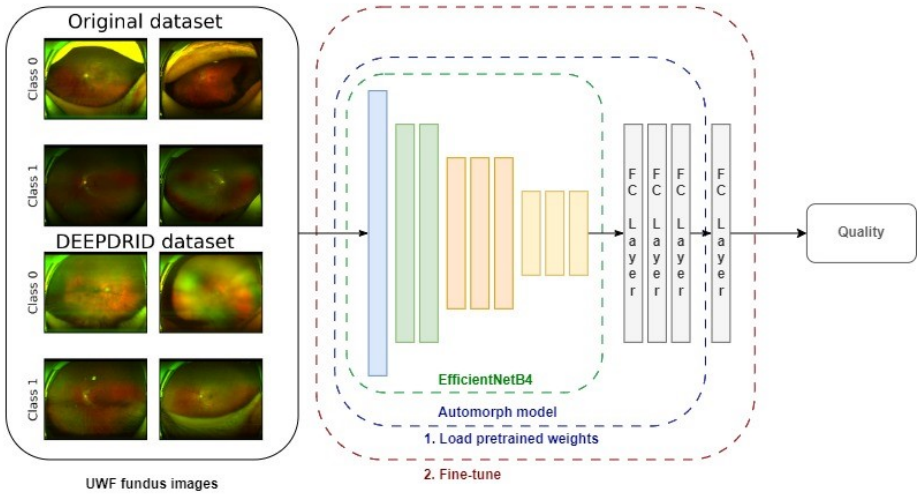

Fig. 2. The framework of our proposed method for image quality assessment.

shuffling, and depthwise separable convolutions to reduce the number of parameters and computational cost while maintaining performance. After loading the pretrained model, we fine-tuned it on our dataset, modifying the final fully connected layer to output a single scalar value corresponding to the binary classification task. The ShuffleNet model was initialized with weights pretrained on the ImageNet dataset. Pretraining on ImageNet allows the model to leverage robust feature extraction capabilities developed from a large and diverse dataset (Fig. 3).

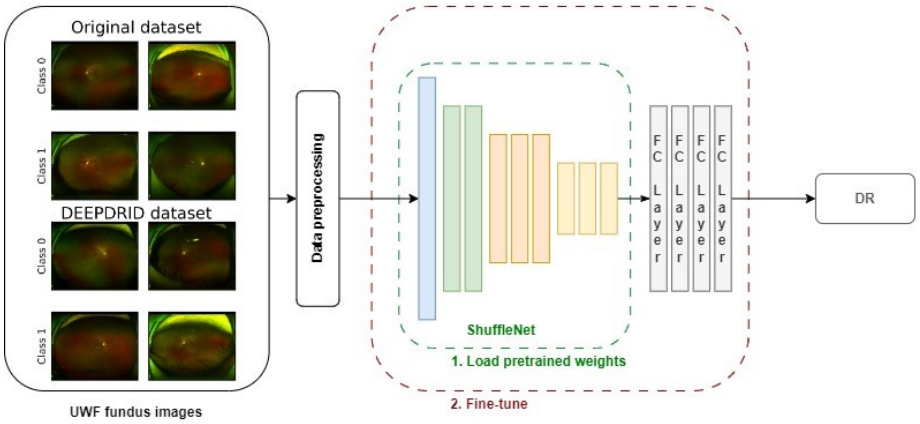

Fig. 3. The framework of our proposed method for DR identification.

4.4 EfficientNetB0 for Diabetic Macular Edema

We proposed the use of EfficientNetB0 [9], pre-trained on ImageNet [13]. EfficientNet combines the concept of compound scaling, mobile inverted bottleneck convolutional layers, and squeeze-and-excitation modules to enhance both accuracy and efficiency, specifically reducing CPU time measured on the leaderboard. The final output of the model is a binary classification indicating the presence or absence of DME. We employed transfer learning by initializing the model with weights pre-trained on the ImageNet dataset. This pre-training provided a strong foundation for feature extraction, improving the model's ability to recognize relevant patterns. The method pipeline for the task 3 can be seen in Fig. 4.

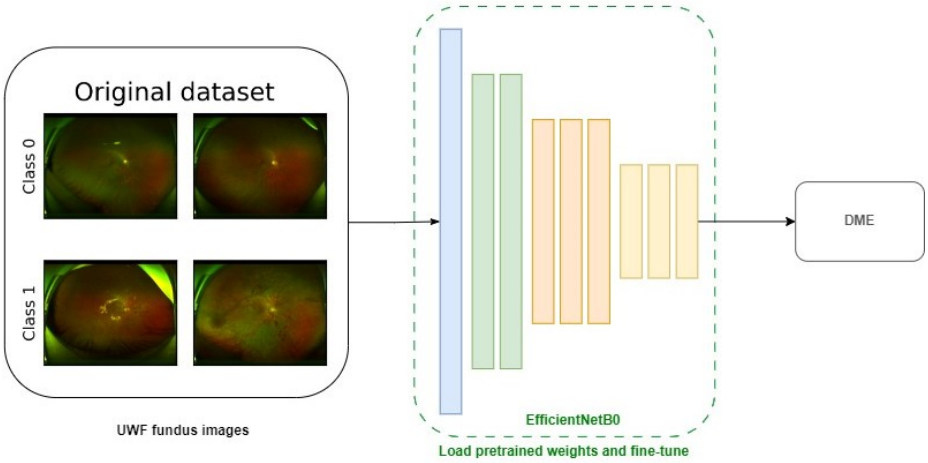

Fig. 4. The framework of our proposed method for DME classification.

5 Results

After training (for training details see Appendix A.1), model predictions were evaluated on the public leaderboard during the validation phase, which served as a test scenario to evaluate our methods and ensure that our code submission was in the correct format. The best models in the validation phase were then evaluated in the test phase with a hidden leaderboard, to assess generalization to unseen data.

Our model performed well in assessing the quality in our validation set and the public leaderboard with a score of 0.9326 (Table 3). The successful classification of image quality is essential for filtering out suboptimal images and ensuring that only the best-quality data is used for DR and DME analysis. What stands

out, is a lot of images were determined as good quality but were in fact bad quality (false positive) (Table 4), which is also obvious with the specificity or true negative rate.

Early detection of DR can prevent progression to more severe stages and complications such as DME. When evaluating our model on the leaderboard, we achieved strong results in detecting DR, with a classification score of 0.9770 during the validation phase and 0.9842 in the test phase (Table 3). In the training phase this had the lowest false positive and false negative amount of images. These results highlight the robustness of our model suggesting that the model generalizes well to unseen data.

As DME can lead to significant vision impairment if left untreated, its accurate diagnosis is essential in managing DR. Our method demonstrated strong performance in detecting DME, as evidenced by the consistent results across the validation and leaderboard datasets with 0.9643 and 0.9820, respectively (Table 3). As with DR we have more false negative images, which is worrisome because the disease might not be detected in a clinical setting.

These findings confirm the effectiveness of our approach in identifying both DR and its complications, offering a reliable tool for automated diagnosis.

Table 3. Performance metrics of the best submission on our validation sets as well as on the public leaderboard

	Task 1			Task 2			Task 3		
Metric	Training	Leaderboard		Training	Leaderboard		Training	Leaderboard	
		Validation Phase	Test Phase		Validation Phase	Test Phase		Validation Phase	Test Phase
AUROC	0.9898	0.9326	0.9326	1.0000	0.9770	0.9842	1.0000	0.9643	0.9820
AUPRC	0.9949	0.9546	0.9546	1.0000	0.9881	0.9814	1.0000	0.9687	0.9800
Sensitivity	0.9767	0.9322	0.9322	1.0000	0.9310	0.9483	1.0000	0.8750	0.9000
Specificity	0.9762	0.8000	0.8000	1.0000	1.0000	0.9577	1.0000	0.9524	1.0000
CPU Time	-	0.5590	0.4853	-	0.1146	0.0778	-	0.3159	0.3630
Total	-	0.9326	0.9326	-	0.9770	0.9842	-	0.9643	0.9820

Table 4. Confusion Matrix

	Training			Validation		
Task	1	2	3	1	2	3
True Positive	42	23	16	37	29	22
True Negative	39	15	18	13	22	11
False Positive	2	3	0	12	1	1
False Negative	4	0	0	4	3	5

Fig. 5. Prototypes for poor quality (Validation Phase). Images that are blurred, over- or underexposed or where the person was blinking were classified correctly by our model.

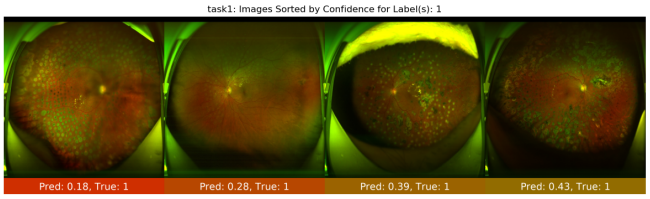

Fig. 6. Criticism for Quality Control (Training Phase): Some images exhibit artifacts that were incorrectly classified as poor quality, even though they should have been identified as good quality. This misclassification represents a false negative.

6 Discussion

Catching DR in its earlier stages allows for treatment that can slow or halt progression before it reaches a more severe state, reducing the risk of complications such as DME or proliferative diabetic retinopathy.

We first focused on image quality assessment, which is essential to ensure that the images used in diagnosis are clear and usable. One of the most common issues leading to poor-quality images is blinking during image capture, resulting in images obscured by the eyelid or eyelashes. Additionally, other factors like blurring, under- or overexposure, and the presence of artifacts can severely affect image quality. These issues can compromise the model's ability to make accurate predictions and highlight the importance of filtering out suboptimal images before analysis (Fig. 5).

In the confusion matrix (Fig. 4), there were more false positives than false negatives identified for the training set. When we look at the images, which were falsely classified as poor quality (false negatives), we could see green holes/spots, which could be artifacts, or laser photocoagulation after intervention(Fig. 7). Whereas, looking at misclassified as good quality (false positive), the culprit is harder to find. The Automorph model probably could not identify large black segments in the lower part of the picture as negative.

During our training, we observed that for poor-quality images, the model, using the Grad-CAM [14] visualization technique (Fig. 8), focused primarily around the periphery of the macula in cases where the eyes were only partially closed due to blinking. Despite the challenges, the model correctly labeled these images as class 0. Notably, there was a prominent line in the lower region of the

Fig. 7. Criticism for Quality Control (Validation Phase): Instances of false positives, where images of poor quality were mistakenly classified as good quality.

image where the model placed significant importance, which appeared to resemble lens-related artifacts. It is possible that these artifacts, along with blinking, served as clear indicators for the Automorph model to classify the images as poor quality during training. However, these were not the case for the validation phase (people blinking, dirty lens), and the model consistently classified such images as positive.

Next, we focused on identifying DR. The Grad-CAM images (Fig. 9, and Supplemental Fig. 13) showed that the model was attending to the correct regions of the macula during its decision-making process. Small, localized areas of attention were consistent with known pathological features such as microaneurysms and hemorrhages. These findings suggest that the model successfully learned to detect these early signs of DR, which is crucial for accurate screening.

In the confusion matrix (Fig. 4), we observed a higher number of false positives compared to false negatives during training. From a medical perspective, this is a favorable outcome. In screening applications, it is generally better for a model to overestimate the presence of disease (false positives) rather than underestimate it (false negatives). Missing a case of DR (a false negative) could lead to delayed diagnosis and treatment, increasing the risk of visual impairment. False positives, while requiring further examination, are less harmful, as they ensure that no cases of DR are missed, providing an extra layer of caution in clinical practice. The validation phase showed a shift toward more false negatives. This change is concerning because, in clinical practice, false negatives pose a greater risk. The change in error patterns between the training and validation phases might suggest potential overfitting. In the training phase, the model may have been more conservative (leading to more false positives), but the validation phase shows an increase in false negatives, which indicates the model might not generalize well to new, unseen data.

Lastly, classifying DME showed a different pattern of attention in the Grad-CAM visualizations. Here, the model focused on flat areas where exudates, which are lipid or protein deposits, tend to accumulate. The model's ability to concentrate on these regions indicates that it could effectively identify DME based on key retinal features (Fig. 10 and Supplemental Fig. 14).

One challenge we encountered during the training process was early overfitting of the model. To address this, we employed several strategies tailored to the dataset. These included using cross-validation, applying various data augmentation techniques, and incorporating the Deep-DRID dataset for both image

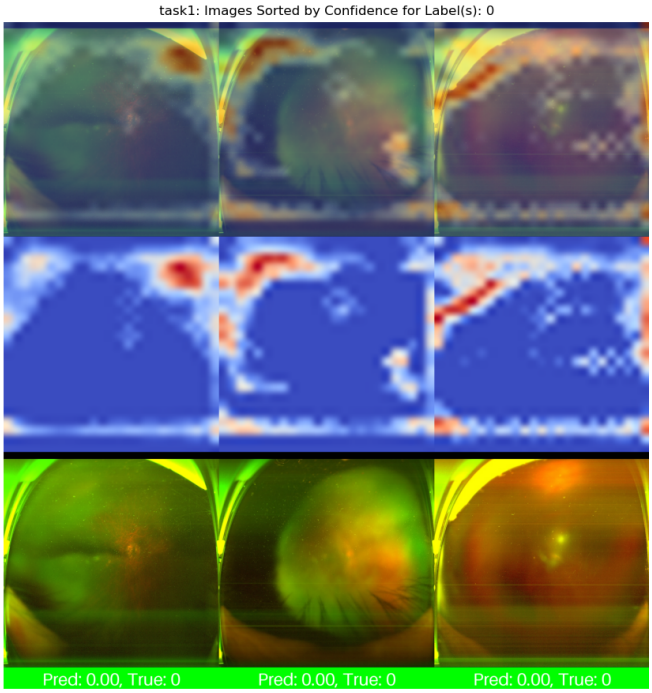

Fig. 8. Training Phase: Grad-CAM of poor quality images: overlay, grad-CAM, and original image (from top to bottom).

quality and identification of DR to increase data diversity. We also experimented with different learning rate schedulers, such as ReduceOnPlateau, CosineAnnealingLR, and StepLR, as well as multiple loss functions, including BCELoss with class weights, FocalLoss, and SmoothL1 Loss.

Additionally, we utilized pre-trained models, such as those trained on ImageNet or Automorph, and explored reducing the model complexity to prevent overfitting. The results presented in this paper represent the best outcomes from this extensive screening process. One avenue we have yet to explore is ensemble learning, which could further improve model robustness and performance. This remains an important direction for future work. Further validation on larger datasets and in diverse clinical settings remains also necessary to confirm the generalizability of our approach.

In conclusion, our results highlight the importance of ensuring high-quality images for accurate diagnosis, while the Grad-CAM visualizations provide insights into the model's decision-making process across different tasks. The balance between false positives and false negatives in medical models is crucial, and our model's tendency to generate more false positives aligns with the overarching goal of minimizing missed diagnoses in DR screening. While our current model demonstrates effectiveness in detecting changes in DR, there is significant

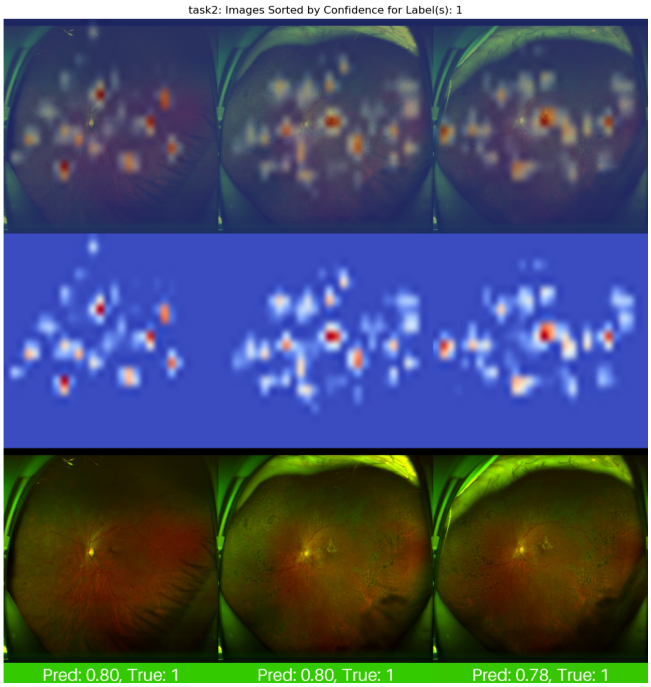

Fig. 9. Validation Phase: Grad-CAM of DR assessment: overlay, grad-CAM, and original image (from top to bottom).

potential for enhancement. The challenges we encountered and the alternative approaches we explored provide valuable insights for future research directions in this critical area of ophthalmological image analysis.

7 Link to Public Code Repository

The implementation of our method, along with the pre-trained models and evaluation scripts, is available at the following public repository:
`https://github.com/moritsih/AILS-MICCAI-UWF4DR-Challenge.git`

Acknowledgments. This research was supported by the Institute for Machine Learning of the Johannes Kepler University Linz. We are specifically grateful to Andreas Mayr and Niklas Schmidinger for their continued feedback and support and to Christian Huber for his valuable feedback and insightful suggestions.

Disclosure of Interests. The authors declare no competing interests.

References

1. Zhen Ling Teo, Yih Chung Tham, Marco Yu, Miao Li Chee, Tyler Hyungtaek Rim, Ning Cheung, Mukharram M. Bikbov, Ya Xing Wang, Yating Tang, Yi Lu, Ian Y.

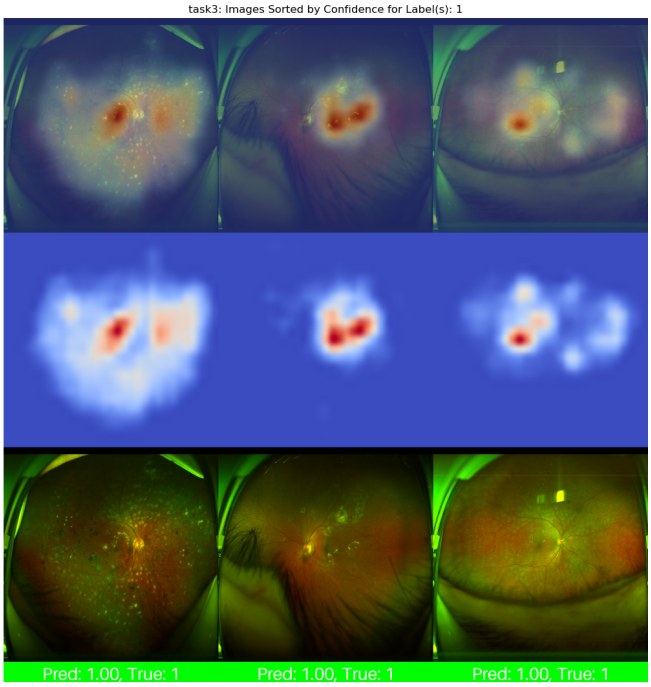

Fig. 10. Validation Phase: Grad-CAM of DME assessment: overlay, grad-CAM, and original image (from top to bottom).

Wong, Daniel Shu Wei Ting, Gavin Siew Wei Tan, Jost B. Jonas, Charumathi Sabanayagam, Tien Yin Wong, and Ching Yu Cheng. Global prevalence of diabetic retinopathy and projection of burden through 2045: Systematic review and meta-analysis. *Ophthalmology*, 128:1580–1591, 11 2021.

2. Ning Cheung, Paul Mitchell, and Tien Yin Wong. Diabetic retinopathy. *Lancet (London, England)*, 376:124–136, 2010.
3. Mohamed Chetoui and Moulay A. Akhloufi. Explainable end-to-end deep learning for diabetic retinopathy detection across multiple datasets. *Journal of medical imaging (Bellingham, Wash.)*, 7, 8 2020.
4. Zineb Farahat, Nabila Zrira, Nissrine Souissi, Yasmine Bennani, Soufiane Bencherif, Safia Benamar, Mohammed Belmekki, Mohamed Nabil Ngote, and Kawtar Megdiche. Diabetic retinopathy screening through artificial intelligence algorithms: A systematic review. *Survey of ophthalmology*, 2024.
5. Daniel Shu Wei Ting, Carol Yim Lui Cheung, Gilbert Lim, Gavin Siew Wei Tan, Nguyen D. Quang, Alfred Gan, Haslina Hamzah, Renata Garcia-Franco, Ian Yew San Yeo, Shu Yen Lee, Edmund Yick Mun Wong, Charumathi Sabanayagam, Mani Baskaran, Farah Ibrahim, Ngiap Chuan Tan, Eric A. Finkelstein, Ecosse L. Lamoureux, Ian Y. Wong, Neil M. Bressler, Sobha Sivaprasad, Rohit Varma, Jost B. Jonas, Ming Guang He, Ching Yu Cheng, Gemmy Chui Ming Cheung, Tin Aung, Wynne Hsu, Mong Li Lee, and Tien Yin Wong. Development and validation of a deep learning system for diabetic retinopathy and related eye dis-

eases using retinal images from multiethnic populations with diabetes. *JAMA*, 318:2211–2223, 12 2017.
6. Duoru Lin, Jianhao Xiong, Congxin Liu, Lanqin Zhao, Zhongwen Li, Shanshan Yu, Xiaohang Wu, Zongyuan Ge, Xinyue Hu, Bin Wang, Meng Fu, Xin Zhao, Xin Wang, Yi Zhu, Chuan Chen, Tao Li, Yonghao Li, Wenbin Wei, Mingwei Zhao, Jianqiao Li, Fan Xu, Lin Ding, Gang Tan, Yi Xiang, Yongcheng Hu, Ping Zhang, Yu Han, Ji Peng Olivia Li, Lai Wei, Pengzhi Zhu, Yizhi Liu, Weirong Chen, Daniel S.W. Ting, Tien Y. Wong, Yuzhong Chen, and Haotian Lin. Application of comprehensive artificial intelligence retinal expert (care) system: a national real-world evidence study. *The Lancet Digital Health*, 3:e486–e495, 8 2021.
7. Ruhan Liu, Xiangning Wang, Qiang Wu, Ling Dai, Xi Fang, Tao Yan, Jaemin Son, Shiqi Tang, Jiang Li, Zijian Gao, Adrian Galdran, J. M. Poorneshwaran, Hao Liu, Jie Wang, Yerui Chen, Prasanna Porwal, Gavin Siew Wei Tan, Xiaokang Yang, Chao Dai, Haitao Song, Mingang Chen, Huating Li, Weiping Jia, Dinggang Shen, Bin Sheng, and Ping Zhang. Deepdrid: Diabetic retinopathy—grading and image quality estimation challenge. *Patterns*, 3:100512, 6 2022.
8. Yukun Zhou, Siegfried K. Wagner, Mark A. Chia, An Zhao, Peter Woodward-Court, Moucheng Xu, Robbert Struyven, Daniel C. Alexander, and Pearse A. Keane. Automorph: Automated retinal vascular morphology quantification via a deep learning pipeline. *Translational Vision Science Technology*, 11, 7 2022.
9. Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:10691–10700, 5 2019.
10. Huazhu Fu, Boyang Wang, Jianbing Shen, Shanshan Cui, Yanwu Xu, Jiang Liu, and Ling Shao. Evaluation of retinal image quality assessment networks in different color-spaces. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 48–56, Cham, 2019. Springer International Publishing.
11. Tao Li, Yingqi Gao, Kai Wang, Song Guo, Hanruo Liu, and Hong Kang. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501:511–522, 2019.
12. Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6848–6856, 7 2017.
13. Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pages 248–255, 2009.
14. Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.
15. Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017.

A Appendix

A.1 Training Details

We randomly split the dataset of the training phase into an 80:20 ratio for training and validation. The model accuracy and generalization capabilities were

tested on an independent but identical distributed validation dataset, those results were then evaluated on a public test server. The best models from the validation phase could be submitted in the test phase, where they were evaluated on a hidden leaderboard.

To avoid overfitting, we selected the model state that yielded the lowest loss on the validation split. For image quality assessment, training was initialized with a learning rate of 1×10^{-4}. The model was trained for 40 epochs with a batch size of 4, achieving the best validation loss at epoch 9. A learning rate scheduler was employed to reduce the learning rate by a factor of 0.5 if no improvement in validation loss was observed for five consecutive epochs, helping to prevent overfitting by fine-tuning the model during periods of stagnation.

For assessment of DR, the loss function combined Binary Cross-Entropy (BCE) loss and Smooth L1 loss, with equal weights of 0.5 for each. The initial learning rate was set to 1×10^{-3}, with a reduction factor of 0.5 when the validation loss plateaued. However, the model continued to improve, so the learning rate did not reduce. The training was carried out with a batch size of 32 for 25 epochs. To address class imbalance, we applied an oversampling strategy to equalize the classes in the dataset. The best validation loss was achieved at epoch 24, with most images being classified correctly.

For identifying DME, our best submission utilized EfficientNetB0. The model was trained for 20 epochs with a batch size of 8, achieving optimal performance at epoch 12. A learning rate scheduler adjusted the learning rate when validation loss stopped improving, reducing it from 1×10^{-4} to 5×10^{-5} as training progressed. We employed BCEWithLogitsLoss, which combines a Sigmoid layer with BCELoss, as the loss function. The AdamW [15] optimizer was used.

A.2 Supplemental Materials

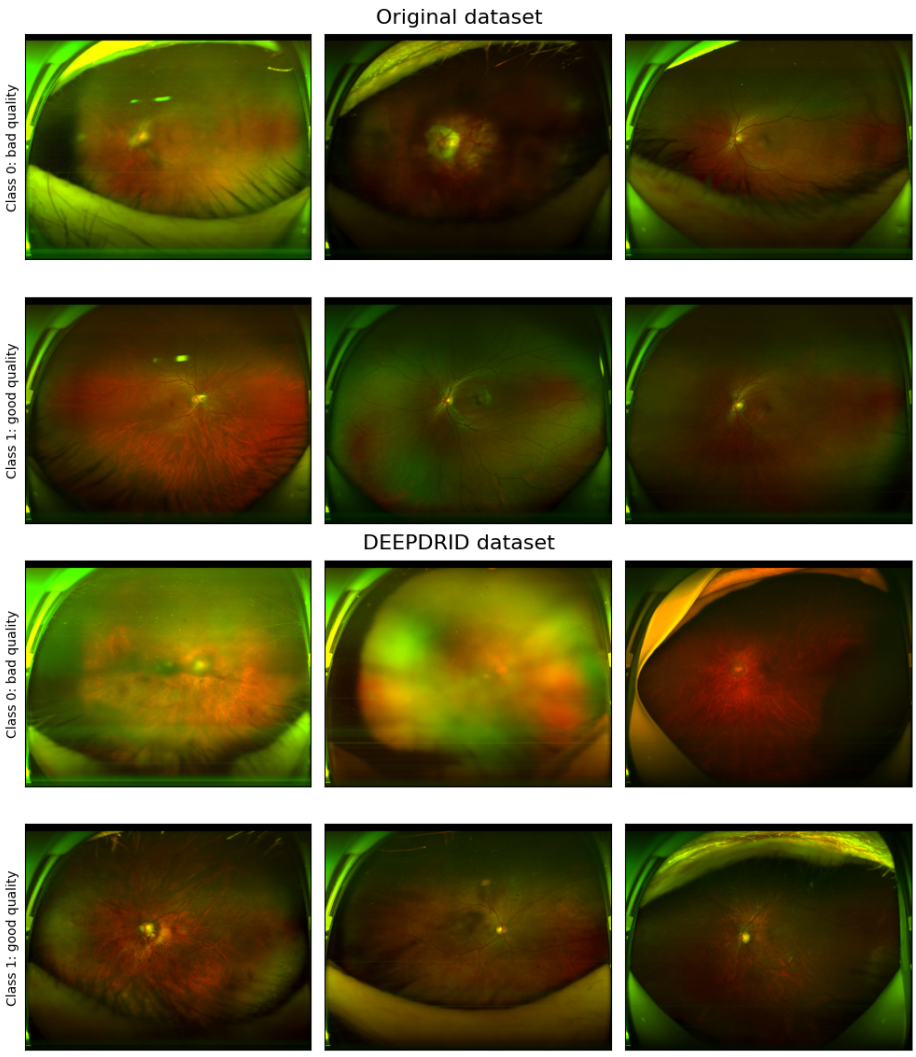

Fig. 11. UWF fundus image from Task 1.

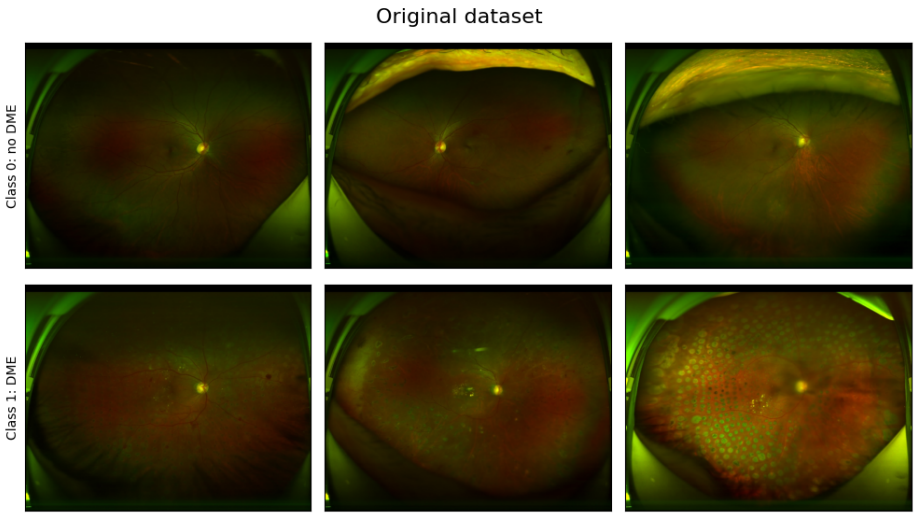

Fig. 12. UWF fundus image from Task 3.

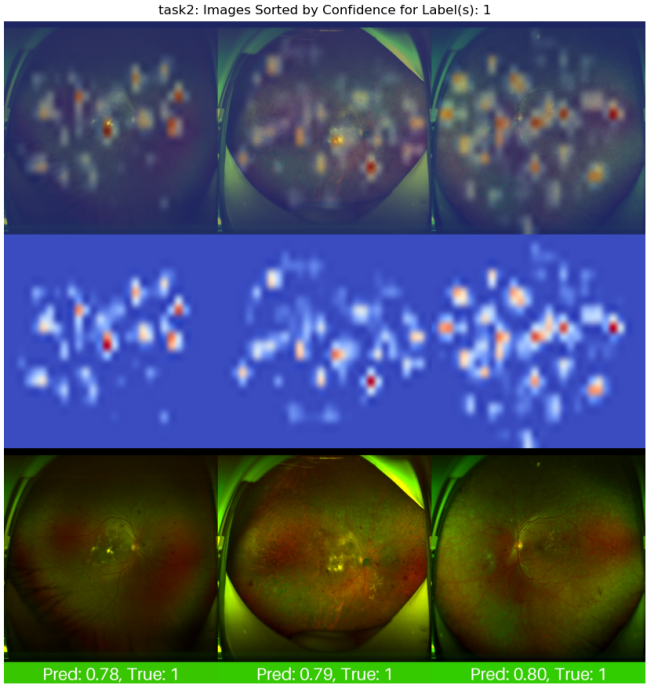

Fig. 13. Training Phase: Grad-CAM of Task 2 - DR (labeled 1): overlay, grad-CAM, and original image (from top to bottom).

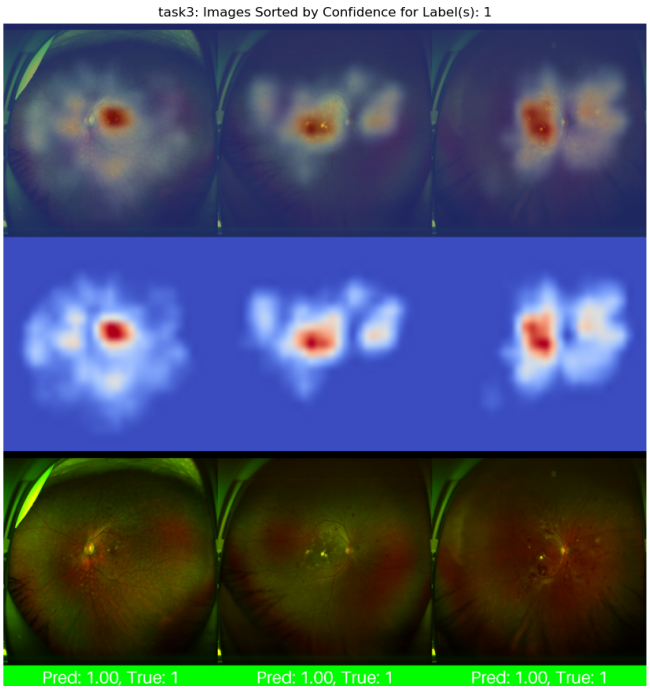

Fig. 14. Training Phase: Grad-CAM of Task 3 - DME (labeled 1): overlay, grad-CAM, and original image (from top to bottom).