AMERICAN ACADEMY OF OPHTHALMOLOGY®

Diabetic Retinopathy Assessment through Multitask Learning Approach on Heterogeneous Fundus Image Datasets

Hongkang Wu, MS,[1,] Kai Jin, MD, PhD,[1,*] Yiyang Jing, MD,[2] Wenyue Shen, MD,[1] Yih Chung Tham, PhD,[3,4] Xiangji Pan, PhD,[1] Victor Koh, MD, PhD,[3] Andrzej Grzybowski, MD, PhD,[5] Juan Ye, MD, PhD[1]*

Objective: To develop and validate an artificial intelligence (AI)-based system, Diabetic Retinopathy Analysis Model Assistant (DRAMA), for diagnosing diabetic retinopathy (DR) across multisource heterogeneous datasets and aimed at improving the diagnostic accuracy and efficiency.

Design: This was a cross-sectional study conducted at Zhejiang University Eye Hospital and approved by the ethics committee.

Subjects: The study included 1500 retinal images from 957 participants aged 18 to 83 years. The dataset was divided into 3 subdatasets: color fundus photography, ultra-widefield imaging, and portable fundus camera. Images were annotated by 3 experienced ophthalmologists.

Methods: The AI system was built using EfficientNet-B2, pretrained on the ImageNet dataset. It performed 11 multilabel tasks, including image type identification, quality assessment, lesion detection, and diabetic macular edema (DME) detection. The model used LabelSmoothingCrossEntropy and AdamP optimizer to enhance robustness and convergence. The system's performance was evaluated using metrics such as accuracy, sensitivity, specificity, and area under the curve (AUC). External validation was conducted using datasets from different clinical centers.

Main Outcome Measures: The primary outcomes measured were the accuracy, sensitivity, specificity, and AUC of the AI system in diagnosing DR.

Results: After excluding 218 poor-quality images, DRAMA demonstrated high diagnostic accuracy, with EfficientNet-B2 achieving 87.02% accuracy in quality assessment and 91.60% accuracy in lesion detection. Area under the curves were >0.95 for most tasks, with 0.93 for grading and DME detection. External validation showed slightly lower accuracy in some tasks but outperformed in identifying hemorrhages and DME. Diabetic Retinopathy Analysis Model Assistant diagnosed the entire test set in 86 ms, significantly faster than the 90 to 100 minutes required by humans.

Conclusions: Diabetic Retinopathy Analysis Model Assistant, an AI-based multitask model, showed high potential for clinical integration, significantly improving the diagnostic efficiency and accuracy, particularly in resource-limited settings.

Financial Disclosure(s): The author(s) have no proprietary or commercial interest in any materials discussed in this article. *Ophthalmology Science 2025;5:100755*

Supplemental material available at www.ophthalmologyscience.org.

Diabetic retinopathy (DR) is the leading cause of preventable adult blindness all around the world. It is suggested that DR will affect around an estimated 160.5 million people worldwide by 2045.[1] In China, where the diabetic population is the largest in the world, the reported prevalence of DR among these individuals is 16.3%.[2,3] Early detection allows for interventions like glycemic control, photocoagulation, or intravitreal injections to prevent vision loss.[4] Consequently, annual dilated eye examinations for regular DR and diabetic macular edema (DME) screening are essential for individuals with diabetes.[5] Fundus photography, an efficient and noninvasive method, is commonly and extensively utilized for DR screening.[6] However, in developing nations such as China, the scarcity of ophthalmologists significantly hinders the ability to meet the growing need for DR screening and treatment.[7]

Artificial intelligence (AI) and deep learning algorithms empower computers to surpass human performance in specific domains by leveraging extensive datasets.[8] Notably, recent models on DR image recognition have demonstrated exceptional sensitivity and specificity, leading to their integration into clinical settings.[9–12] Moreover, human–AI synergy can manifest in various forms,

https://doi.org/10.1016/j.xops.2025.100755
ISSN 2666-9145/25

optimizing performance and resource allocation to enhance efficiency and reduce costs.[13–15] However, there are still several challenges in terms of the applications of AI in the clinical practice.[16,17] Most of the studies primarily concentrate on classifying or detecting images from a single device.[18–20] In addition, their scopes often extend solely to specific facets of the diagnostic process, such as image quality,[21] diagnosis,[22] or grading,[23] and not the entire clinical diagnostic procedure. These limitations hinder their applicability in factual clinical settings. Multitask learning enables the model to learn shared representations that are beneficial across multiple related tasks, such as lesion detection, grading, and quality assessment.[24] This shared learning enhances the model's ability to generalize across different image types and patient populations, ultimately improving diagnostic accuracy and efficiency in factual clinical settings. For instance, by simultaneously learning from tasks like lesion detection and image quality assessment, the model becomes more robust, as it can leverage information from multiple sources to make more informed decisions, reducing the risk of overfitting to a single task.[25]

In our prior work, we introduced a multisource heterogeneous fundus dataset,[26] capturing intricate clinical scenarios encompassing 3 distinct fundus image types. Color fundus photography (CFP) is a widely adopted method for screening various ocular pathologies. Ultra-widefield imaging (UWF) is an advanced fundus imaging technique capable of capturing intricate and comprehensive images. The portable fundus camera (PC), a handheld device, proves to be convenient for deployment in rural settings and plays a pivotal role in advancing telemedicine. In this study, we have developed an AI-based automated system for DR diagnosis. We evaluated its effectiveness as a tool for accurately classifying 3 types of images and assisting in multitask intelligent diagnosis of DR. We posit that our study holds the potential to support ophthalmologists during specific stages of DR consultation and bears generalization in diverse clinical settings.

Methods

Resource Availability

Materials Availability. This study did not generate new unique reagents.

Data and Code Availability. All data reported in this paper will be shared by the lead contact upon request. This paper does not report the original code. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

Study Design and Data Collection

This study was conducted with the approval of the Zhejiang University Eye Hospital (ZUEH) Ethics Committee (ClinicalTrials.gov identifier: NCT04718532). All procedures adhered to the principles outlined by the Declaration of Helsinki (No. Y2023-1073). Data for this study were retrospectively collected from the electronic medical records of patients who had previously provided informed consent for their medical data to be used in research. All images were preprocessed to ensure privacy prior to the commencement of the study. The workflow of the comprehensive study is shown in Figure 1.

The primary dataset used for training, validation, and testing was mainly sourced from the image quality study conducted by ZUEH in 2023[26] and supplemented with corresponding data. Participants were eligible for the study if they were >18 years of age and had been diagnosed with DR between January 2020 and November 2023. External validation cohorts were prospectively assembled from 4 independent ophthalmology centers, establishing a multicenter validation framework reflecting diverse clinical settings. Participants diagnosed with other fundus disorders or with cataracts, refractive media clouding, or other conditions that affect image quality were excluded. Images that met clinically acceptable quality (clear refractive media, entire retinal area, and blood vessels easily recognizable) were included. Included patients were identified by retinal specialists based on appropriate fundus characteristics. At the time of assessment, the majority of the patients were diagnosed with different stages of DR, ranging from mild nonproliferative DR to proliferative DR. The presence of laser scars on some images indicated previous panretinal photocoagulation treatments. Finally, we retrospectively collected 1500 images from 957 patients and categorized them into 3 distinct subdatasets: CFP, UWF, and PC. Each subdataset featured a distinct type of image. The CFP subdataset comprised images obtained from patients diagnosed with DR at ZUEH. The images were captured using the desktop TRC-NW8 fundus camera (Top-Con Medical Systems), offering a field of view of 50° and a resolution of 1924 × 1556 pixels. The UWF subdataset included images of DR patients, which were acquired using the Optos UWF System (Optos Plc Fife) at ZUEH. This imaging system boasts a field of view of 200° and a resolution of 1924 × 1556 pixels. The PC subdataset contained images of healthy volunteers from diverse locations. Data collection was carried out by ZUEH using a Handheld Retinal Camera (Mocular Medical), providing a field of view of 60° and a resolution of 2560 × 1960 pixels.

Annotation

We established a labeling team of 3 ophthalmologists, including 2 certificated ophthalmologists (with >5 years of clinical experience) and 1 retina specialist (with >15 years of clinical experience). They all underwent standardized training and testing before engaging in categorical labeling. The training involved multiple sessions where they reviewed sample retinal images and were guided through the diagnostic criteria for DR and its related features. After the training, a testing phase was conducted where each participant was required to annotate a set of 100 retinal images. Their performance was evaluated based on consistency, accuracy, and adherence to diagnostic guidelines. Discrepancies in annotation were addressed through group discussions and additional feedback from the retina specialist to ensure alignment with standardized diagnostic practices. Two certificated ophthalmologists independently label all images to maintain consistent standards. If both ophthalmologists' labels for an image agree, that labeling is deemed the ground truth. When labels differ, we consult a seasoned retinal specialist for arbitration, and the specialist's assessment becomes the final ground truth.

In the simulation of a clinical treatment process, we established 11 label categories. Image quality was assessed using a comprehensive set of criteria to ensure consistency and reliability across all images. Specifically, the criteria included factors such as focus clarity, illumination, contrast, and the presence of artifacts. Images were categorized as "poor," "medium," or "good" based on these criteria. An image was considered "poor" if it exhibited significant blur, low contrast, or severe artifacts that could interfere with lesion detection. "Medium" quality images showed minor issues that

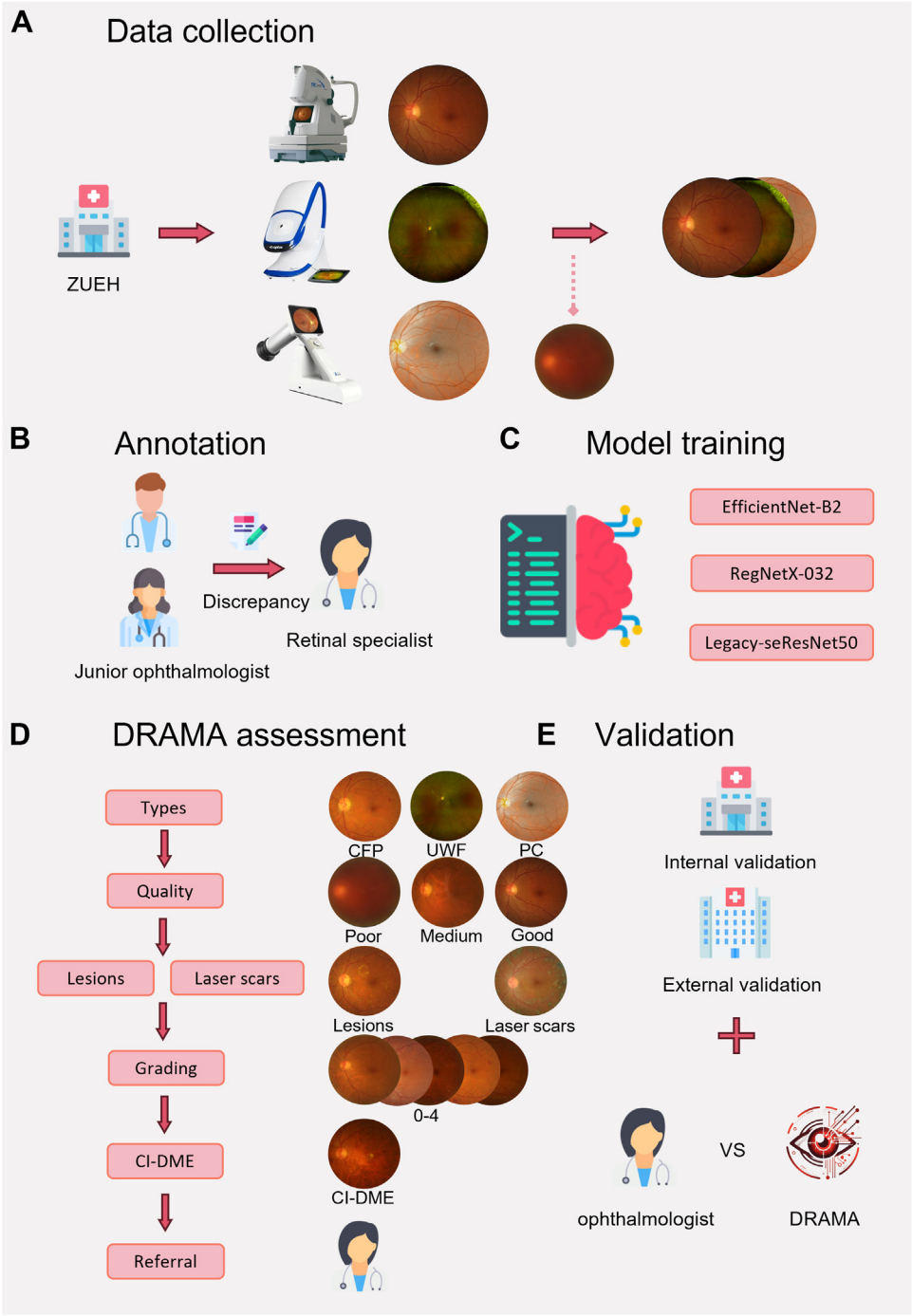

Figure 1. The workflow of the overall study. **A,** Images were collected from 3 different devices and screened for quality. **B,** Annotations were performed by 2 ophthalmologists, and any discrepancies were resolved by a retinal specialist. **C,** Three CNN networks were used for training. **D,** Workflow of DRAMA. **E,** Internal and external validation of models and human–machine comparisons. CFP = color fundus photography; CI-DME = center-involved diabetic macular edema; CNN = convolutional neural network; DRAMA = Diabetic Retinopathy Analysis Model Assistant; PC = portable fundus camera; UWF = ultra-widefield imaging; ZUEH = Zhejiang University Eye Hospital.

might affect the interpretation of subtle features but were generally sufficient for diagnosis. "Good" quality images were those that displayed clear focus, proper illumination, and high contrast, with no significant artifacts. Recognition of ocular laterality is based on the relative position of the optic disc and macula. Lesion identification involves defining a solitary red dot as microaneurysms (MAs) and multiple red dot-like lesions in an area as "hemorrhage (HE)." "Hard exudate (EX)" denotes deposits with a white or yellowish-white color, distinct margins, and a shiny appearance. In contrast, "soft exudate (SE)" features a grayish-

white color, blurred margins, and a cottony texture. Prominent features in fundus photos, like "neovascularization (NV)," "proliferative membranes," and various HEs, correspond to proliferative diabetic retinopathy (PDR) staging. In subsequent sections of the article, for ease of presentation, we will use NV to refer to these serious lesions. We grouped these lesions for classification. "Laser scars" serves as a unique label for images of eyes treated with photocoagulation. Adhering to the standard International Clinical Diabetic Retinopathy Disease Severity Scale,[27] we sorted the images into 5 severity categories: "normal (no apparent retinopathy)," "mild nonproliferative diabetic retinopathy (NPDR)," "moderate NPDR," "severe NPDR," and "PDR." For center-involved DME, a positive identification corresponds to focal EX deposits in the macula's central depression.[28] In addition, we have established that referral is necessary for images lacking laser scars and those showing severe NPDR and PDR or center-involved DME.

The Development of DRAMA

The images were randomly divided into training, validation, and test sets in an 8:1:1 ratio, ensuring that there was no overlap of images between these sets. Diabetic Retinopathy Analysis Model Assistant (DRAMA) was engineered to concurrently perform 11 multilabeling tasks: identifying image types, assessing image quality, determining laterality, detecting lesions (MA, HE, EX, SE, and NV), identifying laser scars, grading, and detecting center-involved DME. The tasks "types," "quality," and "grading" involved multiple categories, while all other tasks were binary. EfficientNet-B2 was selected as the backbone model for our multitask learning system due to its optimal balance between performance and computational efficiency.[29] Compared with 2 additional models, RegNetX-032[30] and LegacyseResNet50,[31] EfficientNet-B2 offers a more efficient scaling of network depth, width, and resolution, which leads to better accuracy with fewer parameters. We conducted a comparative analysis where we trained EfficientNet-B2, RegNetX-032, and LegacyseResNet50 on the same dataset for the tasks of lesion detection, grading, and image quality assessment. The results, as shown in Table 1, demonstrated that EfficientNet-B2 achieved the highest accuracy across all tasks while requiring less computational power and memory, making it a more suitable choice for deployment in clinical settings where computational resources may be limited.

To address the challenge of collecting data in the task, we utilized 14 197 122 sheets of data from the Imagenet dataset[32] to pretrain the backbone network. EfficientNet was adapted for multitask learning by modifying its architecture to support multiple output branches corresponding to the different tasks. Specifically, after the global feature extraction through the EfficientNet backbone, the model was extended to include separate task-specific output layers for each task, such as lesion detection, image quality assessment, and grading. Each of these branches consists of fully connected layers that are independently trained to optimize for their respective tasks. The shared backbone allows the model to learn generalized features from the input images, which are then fine-tuned by the task-specific layers to meet the requirements of each individual task. We then incorporated randomly initialized multitasking subheads, including multi-branching and multilabeling classifiers, to process multiple tasks simultaneously. Images were processed by a backbone network, followed by the extraction of global features via an adaptive average pooling layer. Finally, various information were classified through multiple branches.

To ensure that each task is effectively combined in the multitask learning framework, we used the LabelSmoothingCrossEntropy loss function, which enhances model robustness and generalization through label smoothing.[33] Specifically, for each task, we independently calculated the loss values. For example, cross-entropy loss was used for classification tasks such as lesion detection and grading, while other appropriate loss functions were selected based on the nature of each task, such as for image quality assessment. These individual losses were then combined into the final total loss function by assigning weights to each task, reflecting their relative importance and difficulty. The weighted combination is expressed as follows:

$$L_{\text{final}} = \sum_{i=1}^{n} \lambda_i L_i$$

Additionally, we employed AdamP as our optimizer—an enhanced version of the Adam optimizer.[34] AdamP introduces features like weight decay ratio, Nesterov momentum, and gradient centralization, effectively preventing model overfitting and accelerating convergence to the optimal solution.[35]

By combining these methods, we were able to manage up to 11 tasks within a shared EfficientNet backbone, ensuring that the model performed well across all tasks. The model structure is shown in the Figure S1 (available at www.ophthalmologyscience.org).

Statistical Analysis

The performance of DRAMA in detecting DR across various types of images was demonstrated by calculating the accuracy, sensitivity, and specificity. We plotted the receiving operating characteristic curve, with a larger area under the curve (AUC) indicative of superior model performance. Furthermore, the utilization of confusion matrices for improving the visual representation of false-positive and false-negative rates will be incorporated. Statistical analysis was conducted using the specialized statistical software, Python 3.12 (64-bit).

Heatmap

Heatmap analysis was conducted to visualize the regions of interest that contributed most to the model's decision-making process. To generate these heatmaps, we employed gradient-weighted class activation mapping, a technique that calculates the gradients of the target class scores with respect to the feature maps of the final convolutional layer. These gradients are pooled and weighted to highlight the areas of the image that had the greatest influence on the model's predictions. For each input image, we extracted the feature maps from the final convolutional layer and computed the corresponding gradients. These gradients were then applied to the feature maps to create a heatmap that emphasized the most relevant regions for the model's decision.[36]

Once generated, the heatmaps were overlaid on the original images to visually inspect where the model was focusing its attention. We analyzed these heatmaps to assess whether the model was accurately identifying lesions and other critical features, such as the optic disc and macula. In cases where the model appeared to be distracted by irrelevant features, such as image borders or artifacts, the heatmaps were instrumental in refining the model. This refinement process involved adjusting the preprocessing steps or retraining the model with additional data to ensure that the model concentrated on clinically relevant regions. The insights gained from the heatmap analysis were crucial for understanding and improving the model's decision-making process.

External Test Sets

Fundus photographs of diabetic patients were prospectively collected from 4 distinct ophthalmology centers to ensure comprehensive representation of factual scenarios. These images

Table 1. The Performance of 3 CNNs on 11 Tasks

Task	CNNs	Accuracy	Sensitivity	Specificity	AUC
Quality	Efficientnet_b2	0.8702	*	*	0.9622
	Regnetx_032	0.8855	*	*	0.9615
	Legacy_seresnet50	0.8473	*	*	0.9482
Laterality	Efficientnet_b2	1.0000	1.0000	1.0000	1.0000
	Regnetx_032	1.0000	1.0000	1.0000	1.0000
	Legacy_seresnet50	0.9695	0.9692	0.9692	0.9979
Lesions	Efficientnet_b2	0.9160	0.9190	0.9190	0.9702
	Regnetx_032	0.8931	0.8957	0.8957	0.9466
	Legacy_seresnet50	0.8779	0.8912	0.8912	0.9385
Microaneurysms	Efficientnet_b2	0.9313	0.9328	0.9328	0.9622
	Regnetx_032	0.9008	0.9084	0.9084	0.9485
	Legacy_seresnet50	0.8702	0.8758	0.8758	0.9393
Hemorrhage	Efficientnet_b2	0.9237	0.9222	0.9222	0.9724
	Regnetx_032	0.9084	0.9111	0.9111	0.9485
	Legacy_seresnet50	0.8855	0.8834	0.8834	0.9464
Hard exudate	Efficientnet_b2	0.8626	0.8628	0.8628	0.9473
	Regnetx_032	0.8779	0.8797	0.8797	0.9343
	Legacy_seresnet50	0.8550	0.8573	0.8573	0.9068
Soft exudate	Efficientnet_b2	0.9008	0.7428	0.7428	0.9097
	Regnetx_032	0.8779	0.5795	0.5795	0.8808
	Legacy_seresnet50	0.8931	0.6383	0.6383	0.8070
Neovascularization	Efficientnet_b2	0.9695	0.8000	0.8000	0.9264
	Regnetx_032	0.9847	0.9459	0.9459	0.9868
	Legacy_seresnet50	0.9618	0.8876	0.8876	0.9893
Laser scars	Efficientnet_b2	0.9542	0.9401	0.9401	0.9967
	Regnetx_032	0.9618	0.9457	0.9457	0.9854
	Legacy_seresnet50	0.9466	0.9412	0.9412	0.9474
Grading	Efficientnet_b2	0.8015	0.7098	0.9275	0.9319
	Regnetx_032	0.7252	0.6530	0.9132	0.9102
	Legacy_seresnet50	0.6870	0.6155	0.9039	0.8560
Center-involved diabetic macular edema	Efficientnet_b2	0.8702	0.8360	0.8360	0.9278
	Regnetx_032	0.8473	0.7727	0.7727	0.9045
	Legacy_seresnet50	0.8092	0.7279	0.7279	0.8639

AUC = area under the curve; CNN = convolutional neural network.
*Not calculable.

were collected by investigators at the study sites during a specific time period and represent the DR cohort within their respective regions. All hospitals participating in this study adhered to a uniform image acquisition protocol consistent with that employed for the subdataset. All data were filtered and labeled in the same way as the internal dataset. Specifically, datasets were obtained from the following sources:

1. Anhui Provincial Hospital, China: This dataset included 150 images collected using the TRC-NW8 fundus camera with a resolution of 1924 × 1556 pixels. The dataset comprised patients with various stages of DR.
2. Institute for Research in Ophthalmology, Poland: This dataset consisted of 133 fundus images, acquired using the same TRC-NW8 system, ensuring consistency in image quality and acquisition methods with the Anhui Provincial Hospital dataset.
3. The Second Affiliated Hospital of Zhejiang University School of Medicine, China: This dataset included 86 images captured using a handheld retinal camera with a resolution of 2560 × 1960 pixels.
4. The Affiliated People's Hospital of Ningbo University, China: This dataset contained 100 ultra-widefield images, obtained using the Optos UWF System, which provides a field of view of 200° at a resolution of 1924 × 1556 pixels.

All patient-related information has been anonymized to protect privacy. All images were annotated using a methodology consistent with that of the internal dataset.

Comparison between Human and DRAMA

To evaluate the performance of the DRAMA system against human expertise, we recruited 2 ophthalmologists who were not involved in the initial annotation process to diagnose the images in the test set. One of the ophthalmologists had <5 years of clinical experience, while the other was a retina specialist with >10 years of experience. The test set comprised a representative subset of images covering various stages of DR and varying levels of image quality. Each ophthalmologist independently reviewed and diagnosed the images without any prior knowledge of the patients' clinical histories or the DRAMA system's predictions. The diagnostic process was conducted in a controlled environment to ensure consistency, and the time taken for each diagnosis was recorded using standardized timing methods. The diagnostic labels provided by the ophthalmologists were then compared with the labels generated by the DRAMA system. The comparison was based on standard metrics including accuracy, sensitivity, and specificity. Additionally, the time required for diagnosis by each ophthalmologist

was documented and compared to the time taken by the DRAMA system to analyze the same images. To ensure a comprehensive comparison, the accuracy of each diagnosis was evaluated against pre-existing ground truth labels, which had been verified by a panel of experienced ophthalmologists during the initial dataset annotation phase. The results of the comparison, including diagnostic accuracy and time efficiency, were visually represented using histograms and other relevant statistical charts, allowing for a clear interpretation of the DRAMA system's performance relative to human experts.

Results

Datasets and Annotation

In this study, a total of 1500 images were retrospectively collected from 957 patients, with 218 images excluded due to "poor" quality that precluded labeling. Table 2 shows the characteristics of the primary and external test datasets for the study participants. For image quality, in the CFP and UWF datasets, the percentage of "good" images was higher than the "medium" and lower quality images in PC. Ocular laterality distribution was even. For lesions, the percentage of CFP and PC images exhibiting lesions was 53 and 41, respectively, whereas in UWF, this percentage reached 98. Among all lesions, MA had the highest positivity rate (the probability of occurrence), followed by HE, EX, SE, and NV, which corresponded to the progression of DR. For laser scars, the positivity rates for CFP and PC were 5% and 4%, respectively, while UWF had a positivity rate of 73.5%. For grading, similar to lesions, the overall number decreased as the staging increased. For center-involved DME, CFP and PC had lower positivity rates of 14% and 11%, respectively, while

Table 2. Summary of Internal and External Datasets for Training, Validating, and Testing the DRAMA

Labels	ZUEH			APH	IRO	SAHZJU	APHNU
	CFP	*UWF*	*PC*				
Quality							
Poor	41	2	175	0	0	0	2
Medium	156	95	223	15	9	26	29
Good	303	403	102	26	124	60	69
Laterality							
OD	247	253	240	15	64	25	52
OS	248	247	236	26	69	61	47
Lesions							
Positive	244	487	134	41	57	22	87
Negative	215	11	191	0	76	64	12
Microaneurysms							
Positive	229	471	101	40	56	19	85
Negative	230	27	224	1	77	67	14
Hemorrhage							
Positive	164	418	105	39	17	20	79
Negative	295	80	220	2	116	66	20
Hard exudate							
Positive	132	326	85	36	16	11	60
Negative	327	172	240	5	117	75	39
Soft exudate							
Positive	43	125	7	9	0	0	8
Negative	416	373	318	32	133	86	91
Neovascularization							
Positive	25	88	31	9	1	4	12
Negative	434	410	294	32	132	82	87
Laser scars							
Positive	22	366	13	6	3	11	38
Negative	437	132	311	35	130	75	61
Grading							
0	215	11	191	0	76	64	12
1	56	42	10	0	34	2	4
2	94	203	57	10	14	8	34
3	69	154	35	22	8	8	36
4	25	88	31	9	1	4	12
Center-involved diabetic macular edema							
Positive	64	200	36	24	8	5	32
Negative	395	298	288	17	125	81	67

APH = Anhui Provincial Hospital; APHNU = The Affiliated People's Hospital of Ningbo University; CFP = color fundus photography; DRAMA = Diabetic Retinopathy Analysis Model Assistant; IRO = Institute for Research in Ophthalmology; OD = oculus dexter; OS = oculus sinister; PC = portable fundus camera; SAHZJU = The Second Affiliated Hospital of Zhejiang University School of Medicine; UWF = ultra-widefield imaging; ZUEH = Zhejiang University Eye Hospital.

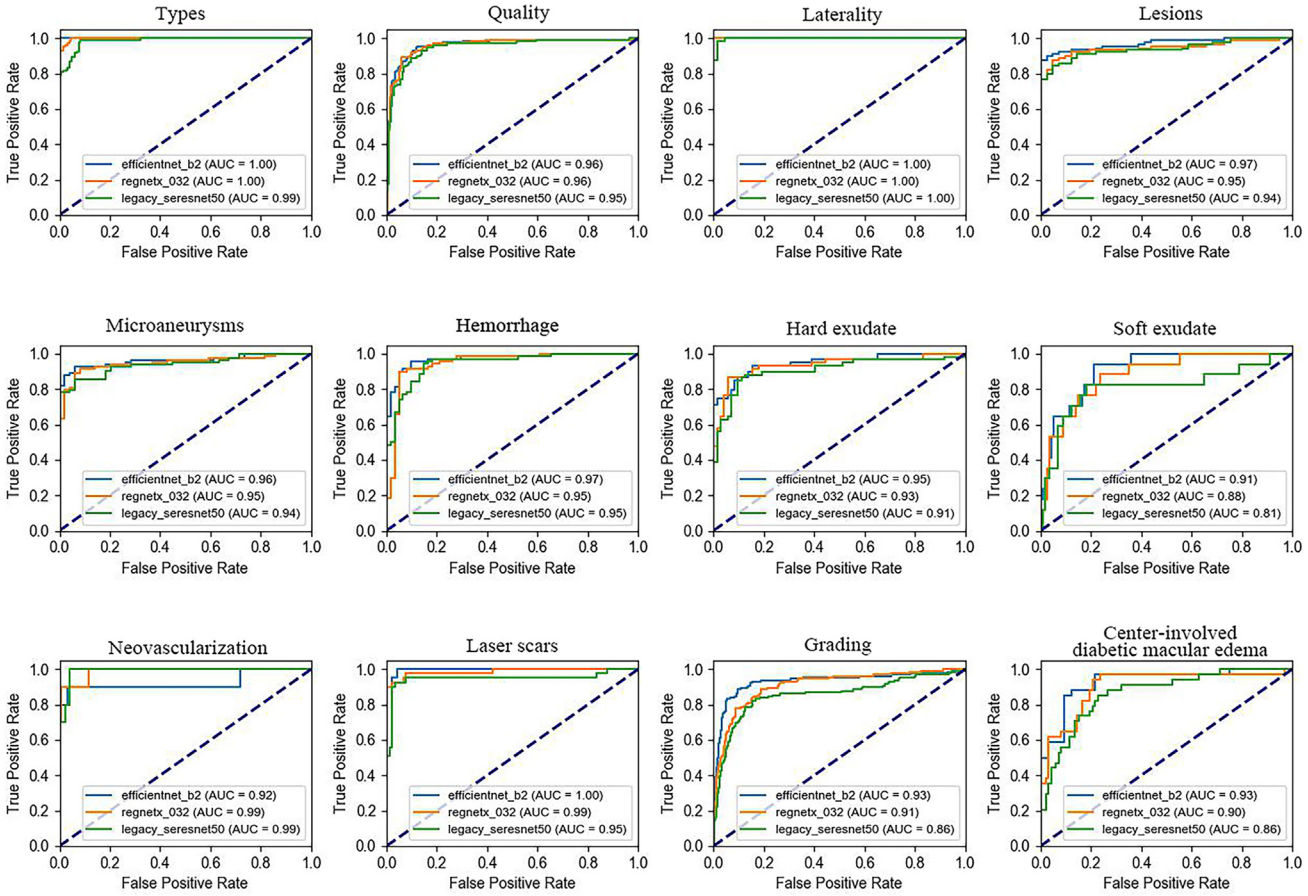

Figure 2. Receiver operating characteristic curve for each task of 3 CNNs. AUC = area under the curve; CNN = convolutional neural network; ROC = receiver operating characteristic.

UWF had a higher rate of 40%. Finally, the proportion of images requiring referral was 23.3%, 24.3%, and 18.1%, respectively.

Performance of DRAMA

The 3 convolutional neural networks we used and their accuracy, sensitivity, specificity, and AUC are shown in Table 1. The model can perform 12 classification tasks simultaneously. Among the convolutional neural network models developed, EfficientNet-B2 demonstrated the best performance. In the "quality" task, EfficientNet-B2 achieved an accuracy of 87.02%, which is comparable to the other 2 models. In both the "laterality" and "laser scars" tasks, all 3 models achieved approximately 95% accuracy. EfficientNet-B2 outperformed the other 2 models in terms of "lesions" accuracy, achieving 91.60%. All 3 models exhibited robust performance in recognizing "MA, HE, SE, NV," with peak accuracy around 95%. Furthermore, all 3 models achieved accuracies exceeding 80% in detecting "center-involved DME." Regarding the "grading" task, the top-performing model, EfficientNet-B2, attained an accuracy rate of 80.15%. The receiver operating characteristic curves for each task were illustrated in Figure 2. Among the 3 convolutional neural networks, EfficientNet-B2 performed best in terms of AUC, with quality, laterality, lesions, and laser scars all having AUCs >0.95 and grading and center-involved DME both having AUCs of 0.93.

The confusion matrix of EfficientNet-B2 containing the results of all tasks is shown in Figure 3. In the "laterality" task, the classification was 100% correct, and in the "quality" task, 22.50% of the "medium" images were classified as "good." In this case, the sensitivity and specificity of "quality" could not be calculated because images of poor quality were excluded from the follow-up study. The performance for "lesions" and for detecting "MA, HE, EX" and laser scars was excellent. However, 47.06% of SE-positive images were identified as negative, and 40.00% of NV-positive images were identified as negative for SE and NV. In the "grading" task, the correct recognition rate was low for images with staging 0 and 4, while the performance was acceptable for images with other staging. Finally, in the "center-involved DME" task, 25.53% of the lesion images were considered negative.

The Assessment and Validation of DRAMA

To visualize the most significant regions in DRAMA, we created a heatmap, overlaying a visualization layer atop the original image. Because the selected images may not

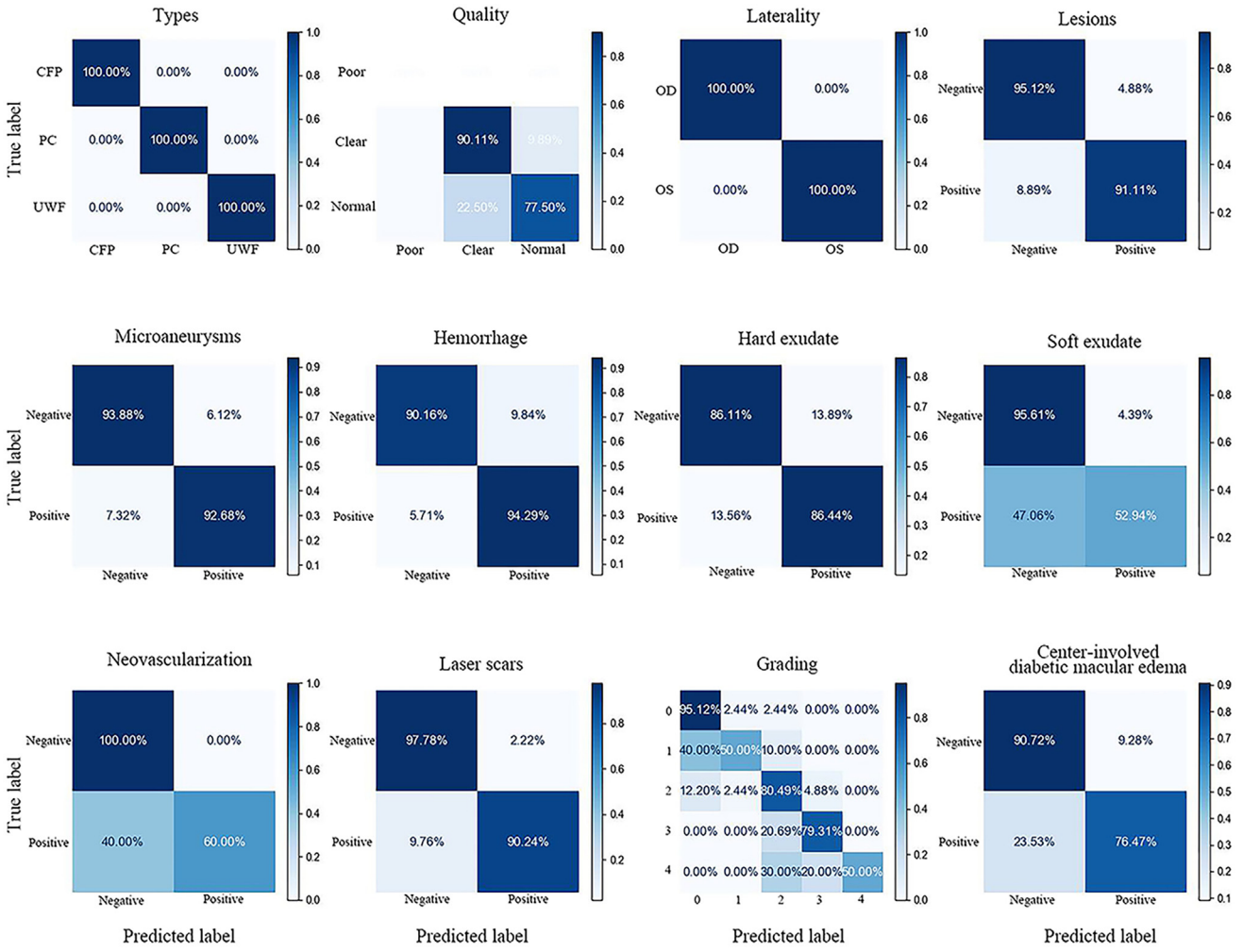

Figure 3. Confusion matrix for each task of DRAMA. CFP = color fundus photography; DRAMA = Diabetic Retinopathy Analysis Model Assistant; OD = oculus dexter; OS = oculus sinister; PC = portable fundus camera; UWF = ultra-widefield imaging.

contain a specific lesion, we only chose the heatmap for the "lesions" task. The heatmap clearly displayed the features employed to differentiate between different types of labels. As shown in Figure 4A, the model located the optic disc to determine the eye's laterality. It also correctly recognized the lesion area and its relevant features for lesion detection and DR classification. The lesion areas in the image were pointed out with arrows. However, some heatmaps were inaccurate because the model was distracted by more salient features (such as eyelids or noises) near the image edge. Moreover, the heatmaps revealed normal fundus structures including the optic disc or macula.

Table 3 displays the performance of DRAMA on multicentric datasets. For the identification of "quality," "laterality," and some lesions, the external datasets performed slightly worse than the ZUEH dataset. However, the external datasets outperformed the ZUEH dataset in the identification of "HE," "SE," and "center-involved DME." The results indicated that the performance on external datasets is good. Notably, the accuracy of the external datasets in the "grading" task is much lower than that of the ZUEH dataset, with a score of 68.16% compared to 80.15%.

The results of the human−computer comparison are shown in Figure 4B. A randomly generated test dataset of 126 images was given independently to a junior ophthalmologist, retinal specialist, and the DRAMA for analysis. The obtained accuracy rates were 83.32%, 83.64%, and 90.91%, respectively. The AI outperformed the ophthalmologists in terms of accuracy in the tasks of "quality" and "grading," as well as in recognizing "MA," "SE," and "NV." The AI performed worse than both ophthalmologists in the tasks of "lesions" and "laser scars." In the other tasks, the AI performance was comparable to that of the ophthalmologists. In addition, junior ophthalmologists and retina specialists are labeled for 100 minutes and 90 minutes respectively. In contrast, DRAMA took only 86 ms. It follows that AI demonstrated significantly higher efficiency compared to the 2 ophthalmologists.

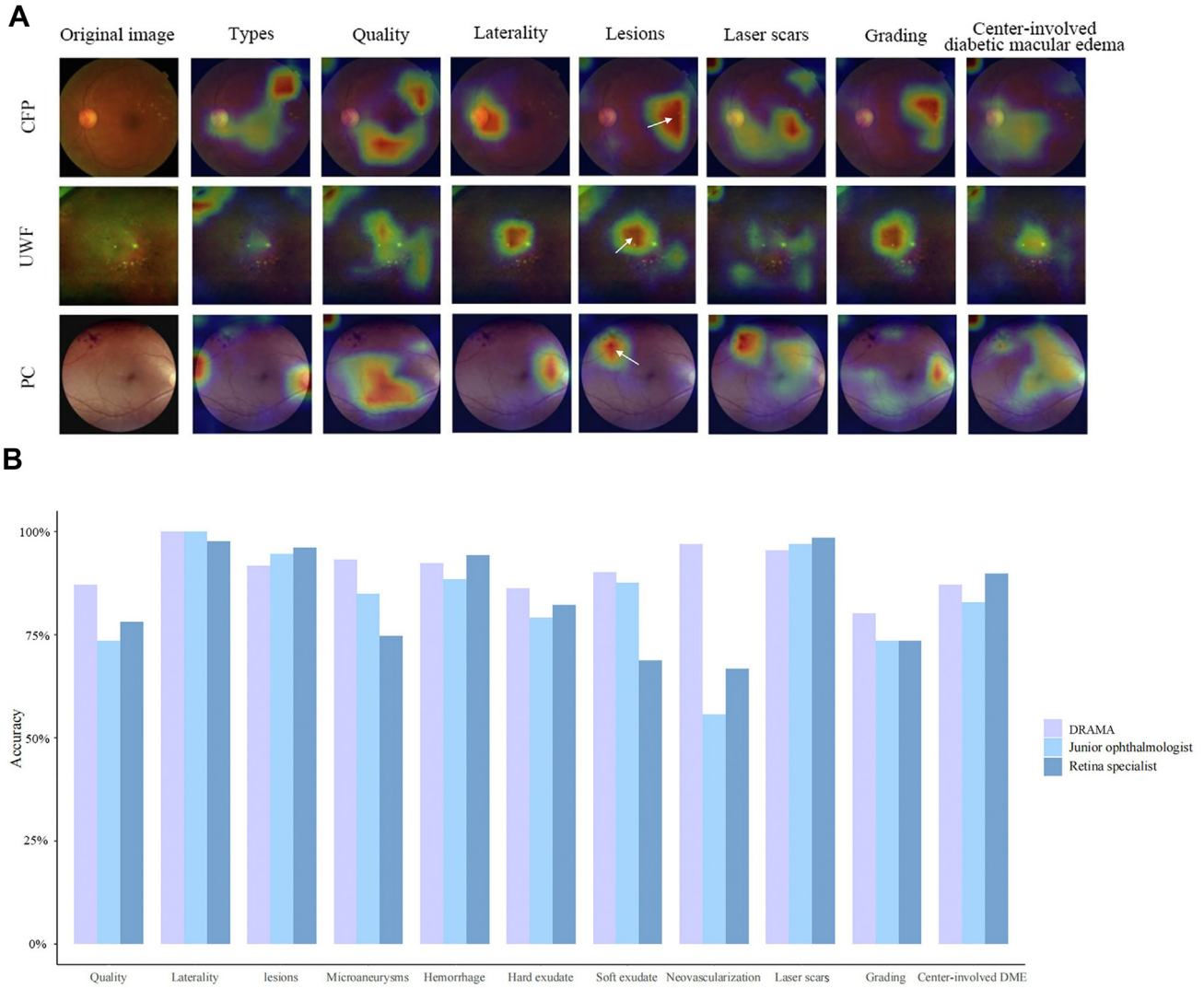

Figure 4. A, The heatmaps of 3 types of images in different tasks. **B,** Human−AI comparisons in the internal test set. AI = artificial intelligence; CFP = color fundus photography; DME = diabetic macular edema; DRAMA = Diabetic Retinopathy Analysis Model Assistant; PC = portable fundus camera; UWF = ultra-widefield imaging.

Discussion

Artificial intelligence has progressed rapidly in ophthalmology, moving from theoretical to practical applications and from imitating diagnosis of single diseases to participating in clinical setting medical process.[37,38] Several previous studies have explored the application of deep learning models in the diagnosis and grading of DR. For instance, Gulshan et al[9] developed a deep learning algorithm that demonstrated high sensitivity and specificity in detecting DR from retinal fundus photographs. Similarly, Ting et al and Dai et al validated deep learning systems for DR and related eye diseases, with robust performance in multiethnic populations.[10,11] Our study builds on these foundations by incorporating multitask learning and heterogeneous datasets to improve generalization across diverse clinical settings. Moreover, the performance of DRAMA in lesion detection, quality assessment, and grading is consistent with the high accuracy levels reported in these prior works. Our study is the first to automatically discriminate and clinically evaluate fundus photographs taken with 3 different devices, solving the generalizability problem that previously plagued AI. Diabetic Retinopathy Analysis Model Assistant was developed on a multisource heterogeneous fundus dataset[26] and performs the entire process of DR image analysis, including image category differentiation, quality assessment, laterality identification, lesion diagnosis, and DR grading. Based on grading, laser scars, and center-involved DME, DRAMA can provide referral recommendations with great potential for patient triage. The model has promising prospects for clinical applications, as it can improve the diagnostic ability of junior ophthalmologists and contribute

Table 3. The Performance of the Model on Multicentric Datasets

Task	Metrics	ZUEH	APH	IRO	SAHZJU	NYPH	Average in External Datasets
Quality	Accuracy	0.8702	0.6829	0.9624	0.8488	0.7347	0.8408
	AUC	0.9622	0.8242	0.9869	0.9125	0.8972	0.9374
Laterality	Accuracy	1.0000	1.0000	0.9925	0.9884	0.9796	0.9888
	AUC	1.0000	1.0000	0.9959	1.0000	0.9962	0.9980
Lesions	Accuracy	0.9160	1.0000	0.8045	0.9419	0.8776	0.8799
	AUC	0.9702	*	0.8153	0.9652	0.7771	0.9071
Microaneurysms	Accuracy	0.9313	0.9756	0.7970	0.9419	0.8980	0.8977
	AUC	0.9622	0.9250	0.8224	0.9819	0.9014	0.9191
Hemorrhage	Accuracy	0.9237	0.9512	0.9398	0.9302	0.8469	0.9134
	AUC	0.9724	0.7949	0.9762	0.9909	0.8474	0.9724
Hard exudate	Accuracy	0.8626	0.9024	0.9248	0.8837	0.7245	0.8575
	AUC	0.9473	0.9944	0.9621	0.9891	0.8680	0.9609
Soft exudate	Accuracy	0.9008	0.8049	0.9925	0.9767	0.8673	0.9330
	AUC	0.9097	0.7674	*	*	0.8694	0.9025
Neovascularization	Accuracy	0.9695	0.7561	0.9925	1.0000	0.9184	0.9469
	AUC	0.9264	0.7812	0.7879	1.0000	0.7326	0.8699
Laser scars	Accuracy	0.9542	0.8293	0.9774	0.9884	0.8878	0.9385
	AUC	0.9967	0.6857	0.7436	1.0000	0.9715	0.9577
Grading	Accuracy	0.8015	0.5610	0.6692	0.8837	0.5714	0.6816
	AUC	0.9319	0.8473	0.8756	0.9681	0.8308	0.8868
Center-involved diabetic macular edema	Accuracy	0.8702	0.7073	0.9850	0.8837	0.8163	0.8827
	AUC	0.9278	0.7819	0.9910	0.8790	0.9143	0.9410

APH = Anhui Provincial Hospital; AUC = area under the curve; IRO = Institute for Research in Ophthalmology; NYPH = The Affiliated People's Hospital of Ningbo University; SAHZJU = The Second Affiliated Hospital of Zhejiang University School of Medicine; ZUEH = Zhejiang University Eye Hospital.
*Not calculable.

to the homogenization of medical standards across different regions.

Our study is a multitask study for fundus image classification. The image type reflects the device used to capture the image, which varies in range, brightness, contrast, and sharpness. These variations pose challenges for clinical work and require accurate identification. Quality evaluation enables screening of images for recapture or further evaluation. This approach reflects the dynamic and diverse nature of clinical settings. Doctors may have subjective judgments of image quality and tolerate some disturbances that do not affect lesion diagnosis, while AI may not. Therefore, a unified and generalizable quality assessment standard for DR fundus images is needed. Ocular laterality is also a basic information of fundus images that can support clinical diagnosis. Lesions are an important component of fundus images, including "MA," "HE," "EX," "SE," and "NV." Microaneurysm is the earliest lesion in DR and plays a crucial role in early screening for DR. The identification of MA facilitates the early diagnosis of DR. Grading determines the treatment of DR; patients with DR staged at severe NPDR or above need interventional therapy. Laser scars are the marks left by retinal photocoagulation. Their presence indicates that the patient has received treatment, which influences the subsequent treatment options. Center-involved DME is a common complication of DR. It significantly impairs vision and requires timely treatment.[39] The detection of lesions not only helps to track disease progression and identify complications such as PDR or DME but also increases the interpretability of the model and reduces the black-box effect. This transparency helps clinicians understand the reasons behind AI decisions and increases trust in the use of AI for referral decisions.

Tables 1 and 3 demonstrate the excellent performance of DRAMA on both internal and external datasets. The external dataset contains images collected from different centers, different races/ethnicities, and different environments, which demonstrates the good generalization of DRAMA. However, DRAMA also exhibits some weaknesses in certain tasks, as shown in Table 2 and Figure 3. For example, it achieves only 80% accuracy in DR grading and confuses the images of stage 1 and stage 4 with other stages. This may be due to the difficulty of the DR grading task, the small difference among different grades, and the requirement to focus on a specific type of characteristic lesions to judge. Future studies can train the AI model with the criteria of DR grading beforehand to improve the accuracy. Upon further analysis, the model demonstrated higher accuracy in detecting NPDR and grading compared to PDR; we attribute this to the small sample size of the PDR image. Furthermore, SE- and NV-positive images are likely (>40%) to be misclassified as negative, which can be explained by 2 factors: (1) the subtle features of these 2 types of lesions, the resemblance between SE and EX on fundus photographs, and the challenge in differentiating some NV lesions from normal blood vessels; and (2) the limited sample sizes of these 2 types of lesions compared to others due to the constraint of the dataset.

Hence, future studies need to increase the sample size to lower the false-positive and false-negative rates. The heatmap in Figure 4A indicates that the model's attention aligns with the lesion-driven grading logic (e.g., MA clusters for mild NPDR, NV foci for PDR). However, the model also pays attention to the black areas of the borders in CFP and PC images, or the eyelid parts of the edges in UWF images, which hinders the model's judgment. Therefore, eliminating these borders in the image preprocessing stage and keeping only the fundus image can enhance the model's performance. The human−computer comparison experiment in Figure 4B also demonstrates that the model outperforms the ophthalmologist in most of the tasks and has high clinical application value.

Fundus photography, the most widely used and cost-effective screening modality for DR, has been a subject of research in AI. In this study, multimodal images are recognized by a single model, which addresses the issue of generalizability in the use of AI. Compared with traditional AI models, it can reduce the cost of integration with hardware in the process of clinical implementation, which follows the new trend of the development of AI technology.

Limitation and Future Consideration

While analyzing the above results, we should also consider the following limitations. Firstly, the image acquisition was challenging due to the inclusion of 3 different types of data, resulting in a small dataset—the inclusion of only Asian and European data. Future studies should collect more relevant data to enlarge the dataset and improve the representativeness of samples across countries and ethnic groups. Secondly, diagnostic accuracy for DME using fundus photography alone was low, and OCT imaging should be incorporated for improved diagnosis. Thirdly, this study focused on image assessment without including clinical information about the patients. Additional data like age, blood glucose, glycated hemoglobin, and disease duration could potentially be integrated into the medical large language model for further analysis.

In conclusion, our study developed and validated DRAMA, a multitask model based on multisource heterogeneous fundus datasets. The external validation study demonstrated that the system performed well in various clinical scenarios. Therefore, DRAMA has great potential to be integrated into clinical hardware for more efficient diagnosis of DR.

Footnotes and Disclosures

Originally received: September 3, 2024.
Final revision: February 15, 2025.
Accepted: February 24, 2025.
Available online: March 11, 2025. Manuscript no. XOPS-D-24-00349.

[1] Zhejiang Provincial Key Laboratory of Ophthalmology, Zhejiang Provincial Clinical Research Center for Eye Diseases, Zhejiang Provincial Engineering Institute on Eye Diseases, Eye Center of Second Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China.

[2] College of Control Science and Engineering, Zhejiang University, Hangzhou, China.

[3] Centre for Innovation and Precision Eye Health, Department of Ophthalmology, National University of Singapore, Singapore.

[4] Ophthalmology and Visual Science Academic Clinical Program, Singapore Eye Research Institute, Singapore National Eye Centre, Duke-NUS Medical School, Singapore.

[5] Institute for Research in Ophthalmology, Foundation for Ophthalmology Development, Poznan, Poland.

*Co-first authors.

Disclosure(s):

All authors have completed and submitted the ICMJE disclosures form.

The author(s) have no proprietary or commercial interest in any materials discussed in this article.

This work was financially supported by the Key Research and Development Program of Zhejiang Province (2024C03204), the Natural Science Foundation of China (grant number 82201195), the National Natural Science Foundation Regional Innovation and Development Joint Fund (U20A20386), the Key Program of the National Natural Science Foundation of China (82330032), and the Clinical Medical Research Center for Eye Diseases of Zhejiang Province (2021E50007).

Support for Open Access publication was provided by the Zhejiang Provincial Key Laboratory of Ophthalmology, Zhejiang Provincial Clinical Research Center for Eye Diseases, Zhejiang Provincial Engineering Institute on Eye Diseases, Eye Center of the Second Affiliated Hospital, and School of Medicine, Zhejiang University.

HUMAN SUBJECTS: Human subjects were included in this study. This study was conducted with the approval of the Zhejiang University Eye Hospital (ZUEH) Ethics Committee (ClinicalTrials.gov identifier: NCT04718532). All procedures adhered to the principles outlined by the Declaration of Helsinki (No Y2023-1073). Data for this study were retrospectively collected from the electronic medical records of patients who had previously provided informed consent for their medical data to be used in research.

No animal subjects were used in this study.

Author Contributions:

Conception and design: Wu, Shen, Jing, Tham, Koh, Grzybowski, Jin, Ye

Data collection: Wu, Shen, Jin

Analysis and interpretation: Wu, Jing, Pan, Jin

Obtained funding: Ye

Overall responsibility: Wu, Shen, Jing, Tham, Koh, Pan, Grzybowski, Jin, Ye

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Juan Ye (yejuan@zju.edu.cn).

Abbreviations and Acronyms:
AI = artificial intelligence; **AUC** = area under the curve; **CFP** = color fundus photography; **DME** = diabetic macular edema; **DR** = diabetic retinopathy; **DRAMA** = Diabetic retinopathy analysis model assistant; **EX** = hard exudate; **HE** = hemorrhage; **MA** = microaneurysms; **NPDR** = nonproliferative diabetic retinopathy; **NV** = neovascularization; **PC** = portable fundus camera; **PDR** = proliferative diabetic retinopathy; **SE** = soft exudate; **UWF** = ultra-widefield imaging; **ZUEH** = Zhejiang University Eye Hospital.

Keywords:
Diabetic retinopathy, Artificial intelligence, Deep learning, Multitask learning, Multisource heterogeneous dataset.

Correspondence:
Juan Ye, MD, PhD, The Second Affiliated Hospital of Zhejiang, University School of Medicine, No.88 Jiefang Road, Hangzhou, Zhejiang 310009, China. E-mail: yejuan@zju.edu.cn; and Andrzej Grzybowski, MD, PhD, Uniwersytet Warminsko-Mazurski w Olsztynieul, Michała Oczapowskiego 2, Olsztyn, WM, Poland. E-mail: ae.grzybowski@gmail.com; and Victor Koh, MD, PhD, NUS Yong Loo Lin School of Medicine, NUHS Tower Block, Level 11, 1E Kent Ridge Road, Singapore City, Singapore. E-mail: ophkohtc@nus.edu.sg.

References

1. Teo ZL, Tham YC, Yu M, et al. Global prevalence of diabetic retinopathy and projection of burden through 2045: systematic Review and meta-analysis. *Ophthalmology*. 2021;128: 1580−1591.
2. Hou X, Wang L, Zhu D, et al. Prevalence of diabetic retinopathy and vision-threatening diabetic retinopathy in adults with diabetes in China. *Nat Commun*. 2023;14:4296.
3. Hu C, Jia W. Diabetes in China: epidemiology and genetic risk factors and their clinical utility in personalized medication. *Diabetes*. 2018;67:3−11.
4. Early photocoagulation for diabetic retinopathy. ETDRS report number 9. Early treatment diabetic retinopathy study research group. *Ophthalmology*. 1991;98:766−785.
5. Solomon SD, Chew E, Duh EJ, et al. Diabetic retinopathy: a position statement by the American diabetes association. *Diabetes Care*. 2017;40:412−418.
6. Das T, Takkar B, Sivaprasad S, et al. Recently updated global diabetic retinopathy screening guidelines: commonalities, differences, and future possibilities. *Eye (Lond)*. 2021;3: 2685−2698.
7. Resnikoff S, Lansingh VC, Washburn L, et al. Estimated number of ophthalmologists worldwide (International Council of Ophthalmology update): will we meet the needs? *Br J Ophthalmol*. 2020;104:588−592.
8. Jin K, Ye J. Artificial intelligence and deep learning in ophthalmology: current status and future perspectives. *Adv Ophthalmol Pract Res*. 2022;2:100078.
9. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*. 2016;316: 2402−2410.
10. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318:2211−2223.
11. Dai L, Wu L, Li H, et al. A deep learning system for detecting diabetic retinopathy across the disease spectrum. *Nat Commun*. 2021;12:3242.
12. Ipp E, Liljenquist D, Bode B, et al. Pivotal evaluation of an artificial intelligence system for autonomous detection of referrable and vision-threatening diabetic retinopathy. *JAMA Netw Open*. 2021;4:e2134254.
13. Xie Y, Nguyen QD, Hamzah H, et al. Artificial intelligence for teleophthalmology-based diabetic retinopathy screening in a national programme: an economic analysis modelling study. *Lancet Digit Health*. 2020;2:e240−e249.
14. Liu H, Li R, Zhang Y, et al. Economic evaluation of combined population-based screening for multiple blindness-causing eye diseases in China: a cost-effectiveness analysis. *Lancet Glob Health*. 2023;11:e456−e465.
15. Gomez Rossi J, Rojas-Perilla N, Krois J, Schwendicke F. Cost-effectiveness of artificial intelligence as a decision-support system applied to the detection and grading of melanoma, dental caries, and diabetic retinopathy. *JAMA Netw Open*. 2022;5:e220269.
16. Tseng R, Gunasekeran DV, Tan SSH, et al. Considerations for artificial intelligence real-world implementation in ophthalmology: providers' and patients' perspectives. *Asia Pac J Ophthalmol (Phila)*. 2021;10:299−306.
17. Li Z, Wang L, Wu X, et al. Artificial intelligence in ophthalmology: the path to the real-world clinic. *Cell Rep Med*. 2023;4:101095.
18. Yan Y, Huang X, Jiang X, et al. Clinical evaluation of deep learning systems for assisting in the diagnosis of the epiretinal membrane grade in general ophthalmologists. *Eye (Lond)*. 2023;38:730−736.
19. Wang Y, Wei R, Yang D, et al. Development and validation of a deep learning model to predict axial length from ultra-wide field images. *Eye (Lond)*. 2023;38:1296−1300.
20. 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;141:1045−1051.
21. Liu L, Wu X, Lin D, et al. DeepFundus: a flow-cytometry-like image quality classifier for boosting the whole life cycle of medical artificial intelligence. *Cell Rep Med*. 2023;4:100912.
22. Cao J, You K, Zhou J, et al. A cascade eye diseases screening system with interpretability and expandability in ultra-wide field fundus images: a multicentre diagnostic accuracy study. *EClinicalMedicine*. 2022;53:101633.
23. Yang Y, Shang F, Wu B, et al. Robust collaborative learning of patch-level and image-level annotations for diabetic retinopathy grading from fundus image. *IEEE Trans Cybern*. 2022;52:11407−11417.
24. SJapa R. An overview of multi-task learning in deep neural networks. *arXiv*. 2017. https://doi.org/10.48550/arXiv.1706.05098.
25. MJapa C. Multi-task learning with deep neural networks: a survey. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2009.09796.
26. Jin K, Gao Z, Jiang X, et al. MSHF: a multi-source heterogeneous fundus (mshf) dataset for image quality assessment. *Sci Data*. 2023;10:286.
27. Wilkinson CP, Ferris 3rd FL, Klein RE, et al. Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*. 2003;110: 1677−1682.
28. Flaxel CJ, Adelman RA, Bailey ST, et al. Diabetic retinopathy preferred practice pattern. *Ophthalmology*. 2020;127: P66−P145.
29. Tan M, Le Q. *Efficientnet: Rethinking model scaling for convolutional neural networks*. Long Beach, CA: PMLR; 2019:6105−6114.
30. Radosavovic I, Kosaraju RP, Girshick R, et al. *Designing network design spaces*. Seattle, WA: IEEE; 2020: 10428−10436.
31. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *IEEE Trans Pattern Anal Mach Intell*. 2020;42:2011−2023.
32. Deng J, Dong W, Socher R, et al. *Imagenet: a large-scale hierarchical image database*. Miami, FL: IEEE; 2009: 248−255.

33. Müller R, Kornblith S, Hinton GE. When does label smoothing help? *arXiv*. 2019:32. https://doi.org/10.48550/arXiv.1906.02629.
34. Heo B, Chun S, Oh SJ, et al. Slowing down the weight norm increase in momentum-based optimizers. *arXiv*. 2020;14. https://doi.org/10.48550/arXiv.2006.08217.
35. Dozat T. Incorporating nesterov momentum into adam. https://openreviewnet/forum?id=OM0jvwB8jIp57ZJjtNEZ; 2016. Accessed October 13, 2023.
36. Selvaraju RR, Cogswell M, Das A, et al. Grad-cam: visual explanations from deep networks via gradient-based localization. *arXiv*. 2017:618−626. https://doi.org/10.48550/arXiv.1610.02391.
37. Zhao X, Lin Z, Yu S, et al. An artificial intelligence system for the whole process from diagnosis to treatment suggestion of ischemic retinal diseases. *Cell Rep Med*. 2023;4:101197.
38. Gao Z, Pan X, Shao J, et al. Automatic interpretation and clinical evaluation for fundus fluorescein angiography images of diabetic retinopathy patients by deep learning. *Br J Ophthalmol*. 2023;107:1852−1858.
39. Wong TY, Sun J, Kawasaki R, et al. Guidelines on diabetic eye care: the international council of ophthalmology recommendations for screening, follow-up, referral, and treatment based on resource settings. *Ophthalmology*. 2018;125:1608−1622.