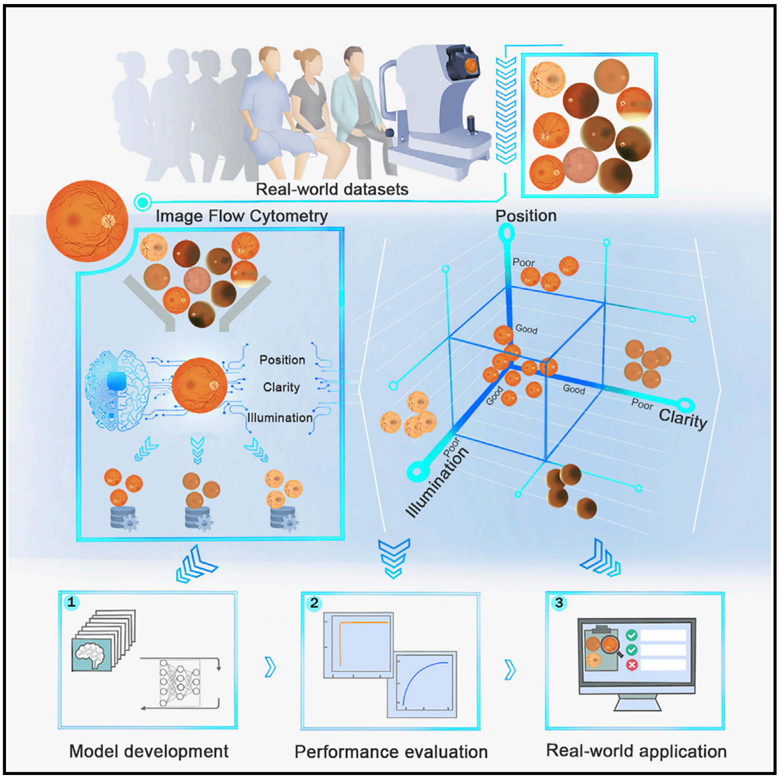
Cell Reports
Medicine

Article

DeepFundus: A flow-cytometry-like image quality classifier for boosting the whole life cycle of medical artificial intelligence

Graphical abstract



Highlights

- DeepFundus is a deep-learning-based, flow-cytometry-like fundus image classifier
- DeepFundus performs robustly in multidimensional and high-throughput image sorting
- DeepFundus can be used to improve intelligent diagnostics in real-world application
- DeepFundus provides a systematic solution to data quality issues in medical AI

Authors

Lixue Liu, Xiaohang Wu, Duoru Lin, ..., Yuzhong Chen, Yizhi Liu, Haotian Lin

Correspondence

wxhang@mail2.sysu.edu.cn (X.W.),
liuyizhi@gzzoc.com (Y.L.),
linht5@mail.sysu.edu.cn (H.L.)

In brief

Liu et al. develop a deep-learning-based, flow-cytometry-like image quality classifier that enables automated, high-throughput, and multidimensional classification of fundus image quality and has significant implications for the whole life cycle of medical artificial intelligence (AI).

Liu et al., 2023, Cell Reports Medicine *4*, 100912
February 21, 2023 © 2022 The Authors.
https://doi.org/10.1016/j.xcrm.2022.100912

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

DeepFundus: A flow-cytometry-like image quality classifier for boosting the whole life cycle of medical artificial intelligence

Lixue Liu,[1,28] Xiaohang Wu,[1,28,*] Duoru Lin,[1] Lanqin Zhao,[1] Mingyuan Li,[1] Dongyuan Yun,[1] Zhenzhe Lin,[1] Jianyu Pang,[1] Longhui Li,[1] Yuxuan Wu,[1] Weiyi Lai,[1] Wei Xiao,[1] Yuanjun Shang,[1] Weibo Feng,[1] Xiao Tan,[1] Qiang Li,[2] Shenzhen Liu,[2] Xinxin Lin,[2] Jiaxin Sun,[2] Yiqi Zhao,[2] Ximei Yang,[2] Qinying Ye,[3] Yuesi Zhong,[4] Xi Huang,[4] Yuan He,[5] Ziwei Fu,[5] Yi Xiang,[6] Li Zhang,[6] Mingwei Zhao,[7] Jinfeng Qu,[7] Fan Xu,[8] Peng Lu,[8] Jianqiao Li,[9] Fabao Xu,[9] Wenbin Wei,[10] Li Dong,[10]

(Author list continued on next page)

[1]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangdong Provincial Key Laboratory of Ophthalmology and Vision Science, Guangdong Provincial Clinical Research Center for Ocular Diseases, Guangzhou, Guangdong, China
[2]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong, China
[3]Department of Ophthalmology, Second Affiliated Hospital, Guangdong Medical University, Zhanjiang, Guangdong, China
[4]Department of Ophthalmology, Third Affiliated Hospital, Sun Yat-sen University, Guangzhou, Guangdong, China
[5]Department of Ophthalmology, The Second Affiliated Hospital of Xi'an Medical University, Xi'an, Shaanxi, China
[6]Department of Ophthalmology, Central Hospital of Wuhan, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei, China
[7]Department of Ophthalmology, People's Hospital of Peking University, Beijing, China
[8]Department of Ophthalmology, People's Hospital of Guangxi Zhuang Autonomous Region, Nanning, Guangxi, China
[9]Department of Ophthalmology, Qilu Hospital, Shandong University, Jinan, Shandong, China
[10]Beijing Tongren Eye Center, Beijing Key Laboratory of Intraocular Tumor Diagnosis and Treatment, Beijing Ophthalmology & Visual Sciences Key Lab, Medical Artificial Intelligence Research and Verification Key Laboratory of the Ministry of Industry and Information Technology, Beijing Tongren Hospital, Capital Medical University, Beijing, China
[11]He Eye Specialist Hospital, Shenyang, Liaoning, China
[12]School of Public Health, He University, Shenyang, Liaoning, China
[13]The Eye Hospital of Wenzhou Medical University, Wenzhou, Zhejiang, China
[14]Department of Ophthalmology, Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China
[15]Department of Ophthalmology, Tianjin Medical University General Hospital, Tianjin, China
[16]Department of Ophthalmology, Xiang'an Hospital of Xiamen University, Xiamen, Fujian, China

(Affiliations continued on next page)

SUMMARY

Medical artificial intelligence (AI) has been moving from the research phase to clinical implementation. However, most AI-based models are mainly built using high-quality images preprocessed in the laboratory, which is not representative of real-world settings. This dataset bias proves a major driver of AI system dysfunction. Inspired by the design of flow cytometry, DeepFundus, a deep-learning-based fundus image classifier, is developed to provide automated and multidimensional image sorting to address this data quality gap. DeepFundus achieves areas under the receiver operating characteristic curves (AUCs) over 0.9 in image classification concerning overall quality, clinical quality factors, and structural quality analysis on both the internal test and national validation datasets. Additionally, DeepFundus can be integrated into both model development and clinical application of AI diagnostics to significantly enhance model performance for detecting multiple retinopathies. DeepFundus can be used to construct a data-driven paradigm for improving the entire life cycle of medical AI practice.

INTRODUCTION

Artificial intelligence (AI) has long been expected to facilitate clinical workflows, improve patient outcomes, and transform current modes of healthcare services.[1] Although AI-based models have generally performed well in experimental conditions, this capability cannot be sustained in real-world studies, where multiple socioenvironmental hurdles impact data quality and downstream analysis, producing a notable decline in model performance, patient experience, and clinical workflow efficiency.[2–4] This data quality gap has been recognized as one of the greatest barriers at all stages of medical AI research from

Guangzheng Dai,[11] Xingru He,[12] Wentao Yan,[13] Qiaolin Zhu,[13] Linna Lu,[14] Jiaying Zhang,[14] Wei Zhou,[15] Xiangda Meng,[15] Shiying Li,[16] Mei Shen,[16] Qin Jiang,[17] Nan Chen,[17] Xingtao Zhou,[18] Meiyan Li,[18] Yan Wang,[19] Haohan Zou,[19] Hua Zhong,[20] Wenyan Yang,[20] Wulin Shou,[21] Xingwu Zhong,[22] Zhenduo Yang,[22] Lin Ding,[23] Yongcheng Hu,[24] Gang Tan,[25] Wanji He,[26] Xin Zhao,[26] Yuzhong Chen,[26] Yizhi Liu,[1,*] and Haotian Lin[1,22,27,29,*]

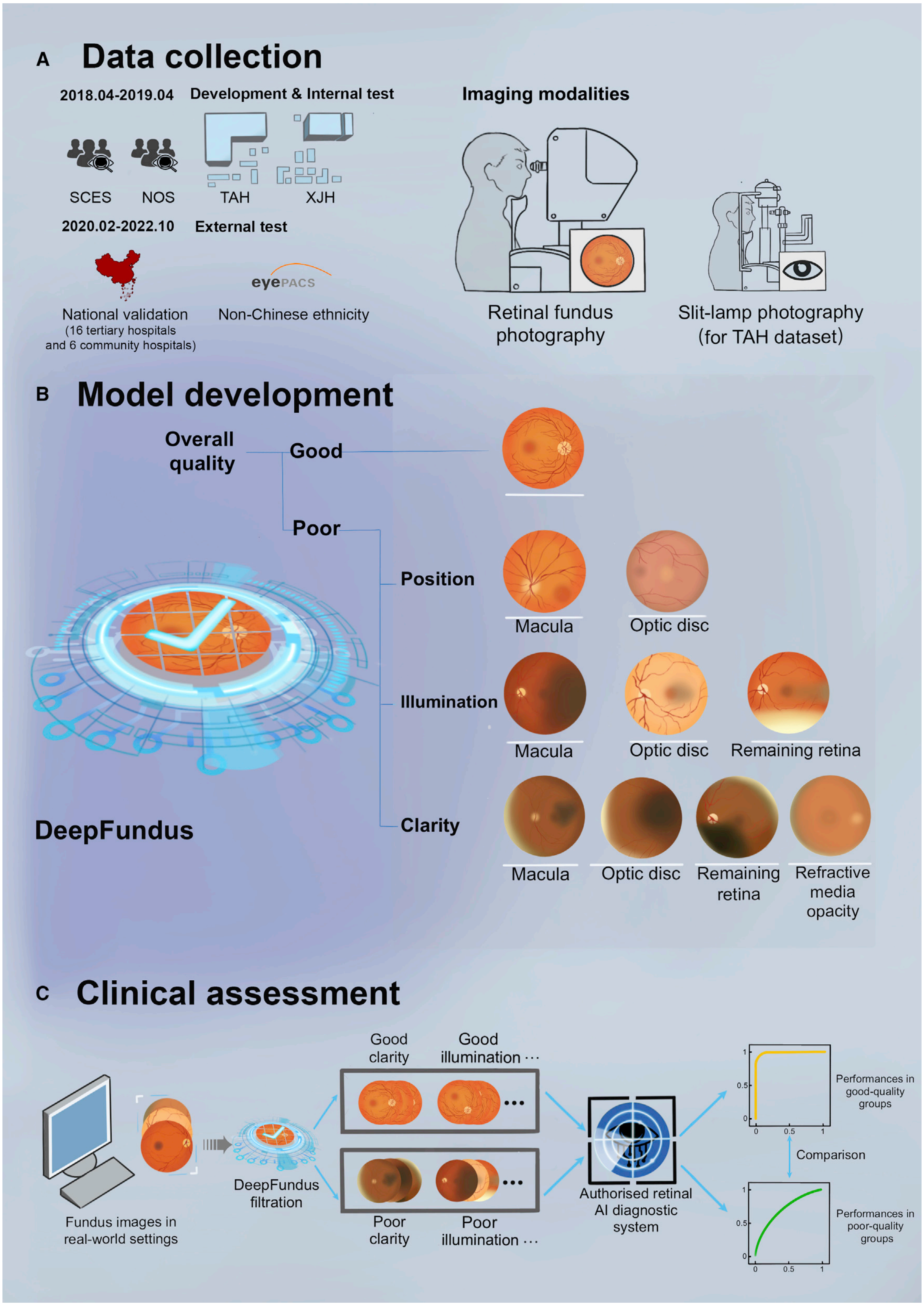
[17]The Affiliated Eye Hospital of Nanjing Medical University, Nanjing, Jiangsu, China
[18]Department of Ophthalmology, Eye and ENT Hospital, Fudan University, Shanghai, China
[19]Tianjin Eye Hospital, Tianjin Key Lab of Ophthalmology and Visual Science, Tianjin Eye Institute, Nankai University, Tianjin, China
[20]Department of Ophthalmology, The First Affiliated Hospital of Kunming Medical University, Kunming, Yunnan, China
[21]Jiaxing Chaoju Eye Hospital, Jiaxing, Zhejiang, China
[22]Hainan Eye Hospital and Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Haikou, Hainan, China
[23]Department of Ophthalmology, People's Hospital of Xinjiang Uygur Autonomous Region, Urumqi, Xinjiang, China
[24]Bayannur Xudong Eye Hospital, Bayannur, Inner Mongolia, China
[25]Department of Ophthalmology, The First Affiliated Hospital, Hengyang Medical School, University of South China, Hengyang, Hunan, China
[26]Beijing Airdoc Technology Co., Ltd., Beijing, China
[27]Center for Precision Medicine and Department of Genetics and Biomedical Informatics, Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou, Guangdong, China
[28]These authors contributed equally
[29]Lead contact
*Correspondence: wxhang@mail2.sysu.edu.cn (X.W.), liuyizhi@gzzoc.com (Y.L.), linht5@mail.sysu.edu.cn (H.L.)
https://doi.org/10.1016/j.xcrm.2022.100912

model development to clinical deployment, and current solutions rely on already overburdened medical practitioners to perform additional data classification tasks.[5,6] This method, however, is both subjective and labor intensive, presenting a major impediment for medical AI to achieve sustainable development. Accordingly, automating high-throughput and systematic data quality classification is essential for AI-driven health interventions to thrive properly and become a new standard of care.

As an image-centric specialty, ophthalmology has become one of the frontiers of deep-learning systems (DLSs) for healthcare applications, with IDx-DR, a software program designed to perform diabetic retinopathy (DR) screening based on fundus photography, being the first authorized autonomous diagnostic AI system in any field of medicine in both China and the US.[7,8] Fundus photography is one of the most commonly used modalities for the evaluation of both systemic diseases that affect the eye (e.g., diabetes and hypertension) and primary ocular diseases (e.g., age-related macular degeneration [AMD]).[9] It has become an ideal candidate for telehealth services supported by DLSs to promote the early diagnosis and timely treatment of various disorders.[10–14] To allow the implementation and adoption of DLSs for clinical care, Lin et al. conducted a national study to further validate their performance using prospective, real-world data.[8] From a broader application perspective, however, according to a study led by Google Health in 11 clinics in Thailand, 21% of retinal images collected in real-world clinical workflows could not be identified by a validated DLS.[3] These ungradable images usually result from operator-dependent factors, camera-related issues, and ocular media opacities, which can incur referrals and associated costs, such as travel or time off work for patients.[15] To alleviate the uncertainty from ungradable images and reduce unnecessary referrals, a fundus image quality classifier is urgently needed to facilitate the clinical implementation of DLSs.

Several automated image quality assessment systems have been proposed for the identification of ungradable fundus images, but none of them have achieved large-scale deployment.[16–20] One possible reason is that these systems only detect poor-quality images without providing detailed information about specific quality defects, which leads to poor explainability in real practice. For fundus image quality assessment, multiple aspects should be considered, including clarity, illumination, and position. Discrimination of these factors is essential for operators since they require distinct adjustments on site.[21] Another limitation of these studies is that they only focused on images whose ineligibility was caused by technical factors. While technical factors, such as inappropriate illumination, positional deviation, and lack of focus, can usually be corrected by repeated image acquisition, ocular media opacity, a major cause of obscured images, is indicative of the presence of ophthalmic diseases that cannot be corrected by image recaptures, including cataracts, vitreous hemorrhage, and corneal edema. Therefore, accurate recognition of different quality factors will not only avoid unnecessary recaptures but also allow eligible patients to receive timely referrals. Moreover, it remains largely unknown how the implementation of image quality analysis will alter the performance of AI-based systems in real-world clinical settings. The extent of performance alterations and how these changes are associated with different quality factors can offer great guidance for both healthcare providers and algorithm engineers.

Inspired by the design of flow cytometry, which provides high-throughput cell sorting according to multiple biomarkers, we developed DeepFundus, a deep-learning-based fundus image classifier, to provide automated, real-time image sorting according to multidimensional quality properties and externally tested its performance on both nationwide and non-Chinese datasets collected from different scenarios. Then, DeepFundus was integrated into a recently certified AI diagnostic system to remove ungradable images before the detection of AMD, DR, and optic disc edema in real-world prospective cohorts to demonstrate its constructive role in the clinical deployment of medical AI. To make AI models aware of this data quality gap upstream from the clinical implementation phase, we identified a group of



(legend on next page)

Table 1. Performance of DeepFundus on the national validation dataset

Model	Sensitivity (95% CI)	Specificity (95% CI)	AUC (95% CI)
Overall quality	0.946 (0.944, 0.948)	0.842 (0.838, 0.845)	0.949 (0.947, 0.951)
Position	0.958 (0.956, 0.96)	0.957 (0.955, 0.959)	0.987 (0.986, 0.988)
Position, macula	0.828 (0.824, 0.831)	0.995 (0.994, 0.996)	0.988 (0.987, 0.989)
Position, optic disc	0.815 (0.812, 0.819)	0.985 (0.984, 0.987)	0.966 (0.965, 0.968)
Illumination	0.852 (0.848, 0.855)	0.966 (0.964, 0.967)	0.971 (0.969, 0.972)
Illumination, macula	0.829 (0.825, 0.832)	0.977 (0.976, 0.979)	0.978 (0.976, 0.979)
Illumination, optic disc	0.862 (0.859, 0.865)	0.992 (0.992, 0.993)	0.987 (0.986, 0.988)
Illumination, retina	0.829 (0.825, 0.832)	0.971 (0.969, 0.972)	0.987 (0.986, 0.989)
Clarity	0.836 (0.832, 0.839)	0.965 (0.964, 0.967)	0.964 (0.962, 0.965)
Clarity, macula	0.922 (0.919, 0.924)	0.934 (0.932, 0.937)	0.948 (0.946, 0.950)
Clarity, optic disc	0.825 (0.821, 0.828)	0.993 (0.992, 0.994)	0.960 (0.959, 0.962)
Clarity, remaining retina	0.914 (0.911, 0.917)	0.943 (0.941, 0.946)	0.975 (0.974, 0.977)
Refractive media opacity	0.897 (0.736, 0.964)	0.859 (0.760, 0.922)	0.953 (0.950, 0.957)

AUC, area under the receiver operating characteristic curve.

low-quality images with sufficient diagnostic certainty and proved that adding these images by a proper percentage to training datasets during model development could enhance model robustness. These procedures can be used to establish a data-driven operating paradigm for boosting the entire life cycle of medical AI practice.

RESULTS

Characteristics of the datasets

From 2018 to 2022, a total of 65,851 fundus images were collected from 27 distinct cohorts to develop and evaluate DeepFundus, which consisted of 13 quality classification models (Figures 1A and 1B). After image annotation, our study included 39,348 images of good overall quality and 26,503 images of poor overall quality for various reasons. Examples and distribution of fundus photographs in each quality aspect are shown in Figure S1 and Table S1. For images collected at the Third Affiliated Hospital of Sun Yat-sen University (TAH), 1,655 fundus images were labeled as poor clarity. Following additional reference to their corresponding slit-lamp images, 514 were considered blurred due to refractive media opacity and 1,141 due to technical reasons.

Performance of DeepFundus on the internal test dataset

In the 1,906-image internal test set, DeepFundus achieved an under the receiver operating characteristic curve (AUC) of 0.975 (95% confidence interval [CI]: 0.968–0.982) in detecting poor overall quality images and AUCs of 0.970–0.985, 0.967–0.979, and 0.909–0.960 in classifying poor-quality images concerning position, illumination, and clarity, respectively. In the TAH internal test dataset consisting of 248 images, DeepFundus achieved an AUC of 0.955 (0.946–0.965) in distinguishing poor-quality fundus images caused by refractive media opacity from those whose poor quality was caused by technical reasons. Details on the performance of the 13 models above are displayed in Table S2.

Performance of DeepFundus on national validation and non-Chinese population

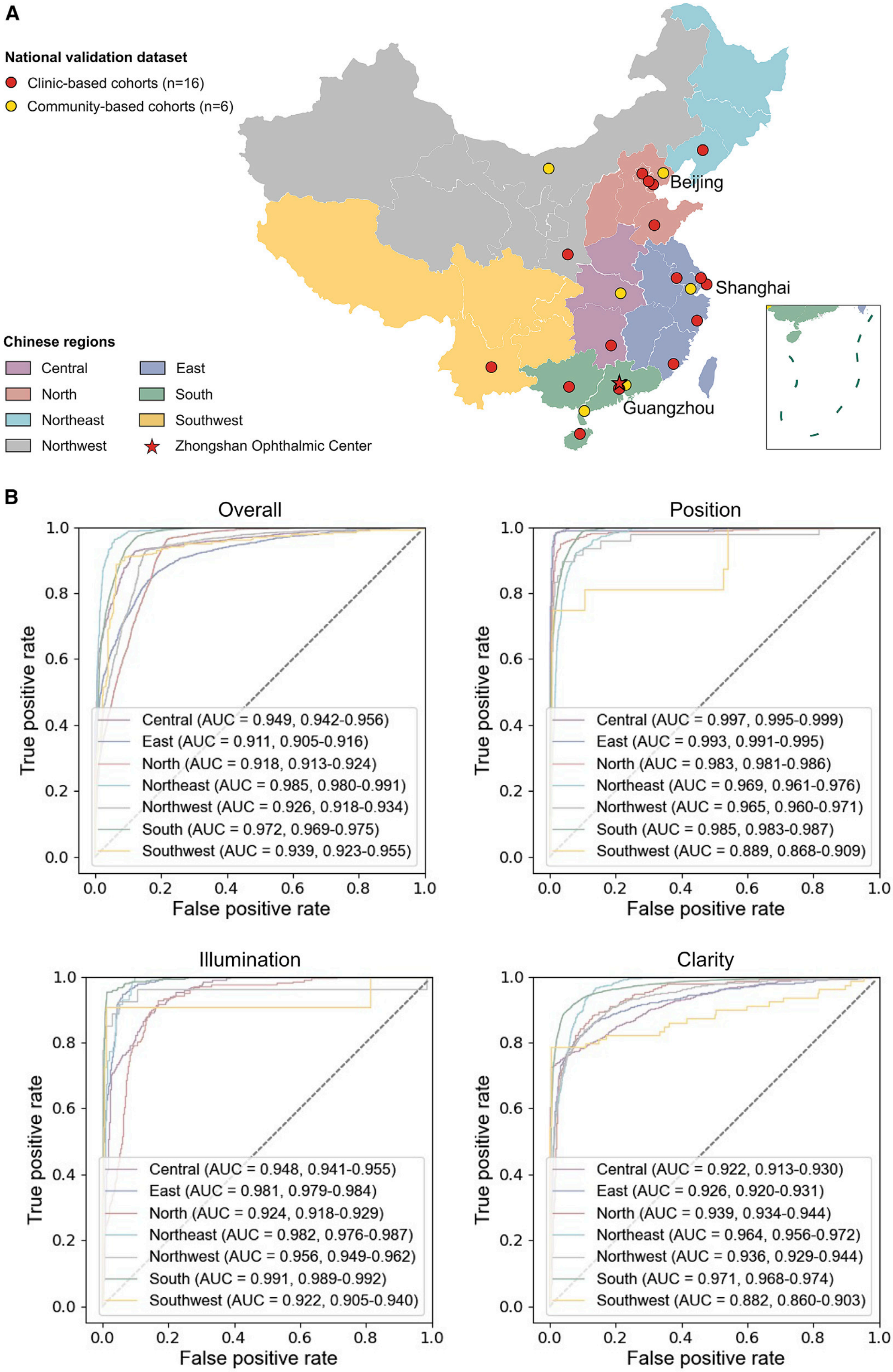
In the national validation dataset consisting of 44,712 fundus images, DeepFundus achieved an AUC of 0.949 (95% CI: 0.947–0.951) for detecting images of poor overall quality (Table 1). For classifying images of poor position, the AUCs were 0.987 (0.986–0.988), 0.988 (0.987–0.989), and 0.966 (0.965–0.968) for the overall image, macular area, and optic disc, respectively. For identifying images of poor illumination, the AUCs were 0.971 (0.969–0.972), 0.978 (0.976–0.979), 0.987 (0.986–0.988), and 0.987 (0.986–0.989) for the overall image, macular area, optic disc, and the remaining retina, respectively. For detecting images of poor clarity, the AUCs were 0.964 (0.962–0.965), 0.948 (0.946–0.950), 0.960 (0.959–0.962), and 0.975 (0.974–0.977) for the overall image, macular area, optic disc, and the remaining retina, respectively. Further information about its performance in 7 separate Chinese regions is presented in Figures 2 and S2 and Data S1 and S2. For detecting blurred fundus images caused by refractive media opacity, DeepFundus obtained an AUC of 0.953 (0.950–0.957) in the external test dataset (Table 1). When tested using the Kaggle dataset, DeepFundus obtained AUCs of

Figure 1. Overall study design

(A) Data collection, including retinal fundus images and slit-lamp images (obtained only from TAH), was conducted from 27 distinct cohorts for 4 years.

(B) Fundus images are used to develop DeepFundus, a deep-learning-based system for the identification of fundus images with 13 different types of quality defects.

(C) DeepFundus is integrated into an AI diagnostic system for removing unqualified fundus images before diagnosis, and its performances in good-quality and poor-quality image groups are compared. SCES, South China Eye Screening Program; NOS, Guangdong Neuro-ophthalmology Study; TAH, The Third Affiliated Hospital, Sun Yat-sen University; XJH, The People's Hospital of Xinjiang Uygur Autonomous Region; AI, artificial intelligence.



(legend on next page)

Table 2. Performance of the certified AI diagnostic system in the detection of drusen in different quality groups

Quality filters	Metrics (95% CI)	GQ group	PQ group	Difference	p value[a]
Overall quality	sensitivity	0.853 (0.830, 0.874)	0.494 (0.430, 0.557)	0.360 (0.289, 0.430)	<0.001
	specificity	0.975 (0.968, 0.981)	0.981 (0.972, 0.987)	−0.005 (−0.015, 0.005)	0.344
	accuracy	0.938 (0.929, 0.946)	0.914 (0.900, 0.926)	0.024 (0.008, 0.040)	0.002
Position	sensitivity	0.791 (0.768, 0.813)	0.567 (0.392, 0.726)	0.225 (0.029,0.421)	0.006
	specificity	0.978 (0.972, 0.982)	0.974 (0.941, 0.989)	0.004 (−0.022, 0.029)	0.947
	accuracy	0.930 (0.923, 0.937)	0.920 (0.877, 0.949)	0.011 (−0.028, 0.049)	0.630
Illumination	sensitivity	0.812 (0.788, 0.834)	0.489 (0.389, 0.582)	0.327 (0.221, 0.434)	<0.001
	specificity	0.977 (0.971, 0.982)	0.982 (0.966, 0.991)	−0.005 (−0.019, 0.009)	0.574
	accuracy	0.934 (0.926, 0.941)	0.900 (0.873, 0.921)	0.034 (0.008, 0.061)	0.003
Clarity	sensitivity	0.837 (0.814, 0.858)	0.471 (0.398, 0.545)	0.366 (0.285, 0.447)	<0.001
	specificity	0.975 (0.968, 0.981)	0.983 (0.974, 0.989)	−0.008 (−0.018, 0.003)	0.183
	accuracy	0.935 (0.926, 0.942)	0.917 (0.901, 0.930)	0.018 (0.001, 0.036)	0.031
Macula	sensitivity	0.851 (0.828, 0.872)	0.513 (0.449, 0.575)	0.339 (0.269, 0.408)	<0.001
	specificity	0.976 (0.968, 0.981)	0.980 (0.972, 0.986)	−0.005 (−0.015, 0.005)	0.407
	accuracy	0.937 (0.928, 0.945)	0.917 (0.903, 0.929)	0.020 (0.004, 0.036)	0.009
Optic disc	sensitivity	0.826 (0.802, 0.848)	0.606 (0.541, 0.668)	0.220 (0.149, 0.291)	<0.001
	specificity	0.976 (0.969, 0.981)	0.981 (0.971, 0.987)	−0.005 (−0.016, 0.005)	0.395
	accuracy	0.934 (0.925, 0.942)	0.920 (0.904, 0.933)	0.014 (−0.003, 0.031)	0.090
Retina	sensitivity	0.854 (0.831, 0.874)	0.478 (0.413, 0.543)	0.376 (0.308, 0.448)	<0.001
	specificity	0.977 (0.970, 0.982)	0.979 (0.969, 0.985)	−0.002 (−0.012,0.009)	0.839
	accuracy	0.940 (0.932, 0.948)	0.908 (0.892, 0.921)	0.032 (0.016, 0.049)	<0.001

This experiment was conducted in a community-based, age-related macular degeneration study consisting of 1,234 images with drusen and 3,732 images without drusen. p <0.05 was considered significantly different. All patients were diagnosed through comprehensive imaging examinations. GQ, good quality. PQ, poor quality.

[a] p values were calculated between the good-quality and poor-quality groups using the two-proportion Z-test.

0.764–0.987 to identify retinal images with different quality defects (Table S3).

Heatmap analysis

To visualize the areas contributing most to the 13 models of DeepFundus, we generated a heatmap that superimposed a visualization layer on the original images. The quality defects corresponding to poor clarity, illumination, and position were explicitly highlighted in the heatmaps (Figure S3). Notably, the heatmap of poor overall quality highlighted both poor-clarity and poor-illumination areas. Furthermore, heatmaps for structural quality analysis models (e.g., clarity, optic disc) can primarily highlight corresponding retinal areas despite other blurred structures. Typical examples of the heatmaps for other models are presented in Figure S3.

Clinical application of DeepFundus

To facilitate the implementation of DeepFundus in clinical settings, we designed a standardized workflow to properly arrange the 13 models of DeepFundus. Each fundus photograph collected in real-world settings will undergo DeepFundus classification, and real-time feedback will be provided to improve clinical application efficiency, as shown in Figure S4.

Then, we integrated DeepFundus into a certified AI retinal diagnostic system as a preprocessing function to investigate its effects on diagnostic performance (Figure 1C). After DeepFundus image filtration, the diagnostic metrics for detecting AMD, referable DR, and optic disc edema in the good quality (GQ) and poor quality (PQ) groups were compared in Tables 2, S1–S4, and S2–S4, respectively. In a community-based cohort for AMD screening, AI achieved sensitivities of 0.853 (95% CI: 0.830–0.874) and 0.494 (0.430–0.557); specificities of 0.975 (0.968–0.981) and 0.981 (0.972–0.987); and accuracies of 0.938 (0.929–0.946) and 0.914 (0.900–0.926) in detecting drusen in the overall GQ and PQ groups, respectively; the GQ group demonstrated significantly higher sensitivity and accuracy values. Similar results were also observed in the other 6 quality

Figure 2. Performance of DeepFundus on the national validation dataset

(A) After model development, DeepFundus was externally tested using a national validation dataset prospectively collected from 16 clinic-based and 6 community-based cohorts across China.

(B) On the national validation dataset, DeepFundus achieved AUCs of 0.911–0.985 for detecting images of poor overall quality; AUCs of 0.965–0.997 for detecting images of poor position concerning the overall image; AUCs of 0.924–0.991 for detecting images of poor illumination concerning the overall image; and AUCs of 0.922–0.971 for detecting images of poor clarity concerning the overall image in 7 different Chinese regions. AUC, area under the receiver operating characteristic curve.

Table 3. Performance of established models in the detection of diabetic retinopathy using different model architectures

Model Architecture	Metrics (95% CI)	Model 1	Model 2	Model 3	Model 4	P1	P2	P3	P4	P5	P6
InceptionV3	sensitivity	0.775 (0.734, 0.816)	0.925 (0.899, 0.951)	0.775 (0.734, 0.816)	0.725 (0.681, 0.769)	0.117	1	0.796	0.117	0.039	0.796
	specificity	1 (1, 1)	0.986 (0.975, 0.998)	0.992 (0.983, 1)	0.992 (0.983, 1)	0.073	0.247	0.247	0.722	0.722	1
	accuracy	0.978 (0.963, 0.992)	0.980 (0.966, 0.994)	0.970 (0.953, 0.987)	0.965 (0.947, 0.983)	1	0.658	0.397	0.497	0.280	0.842
Inception ResnetV2	sensitivity	0.675 (0.629, 0.721)	1 (1, 1)	0.725 (0.681, 0.769)	0.750 (0.708, 0.792)	<0.001	0.807	0.621	0.001	0.002	1
	specificity	1 (1, 1)	0.969 (0.953, 0.986)	0.997 (0.992, 1)	0.997 (0.992, 1)	0.002	1	1	0.009	0.009	1
	accuracy	0.968 (0.950, 0.985)	0.973 (0.957, 0.989)	0.970 (0.953, 0.987)	0.973 (0.957, 0.989)	0.836	1	0.836	1	1	1
Densenet	sensitivity	0.525 (0.476, 0.574)	0.725 (0.681, 0.769)	0.700 (0.655, 0.745)	0.725 (0.681, 0.769)	0.106	0.169	0.106	1	1	1
	specificity	1 (1, 1)	1 (1, 1)	0.992 (0.983, 1)	0.989 (0.979, 0.999)	1	0.247	0.133	0.247	0.133	1
	accuracy	0.953 (0.932, 0.973)	0.973 (0.957, 0.989)	0.963 (0.944, 0.981)	0.963 (0.944, 0.981)	0.193	0.599	0.599	0.550	0.550	1

P1 indicates the p value calculated between model 1 and model 2 using the two-proportion Z-test. P2 indicates the p value calculated between model 1 and model 3 using the two-proportion Z-test. P3 indicates the p value calculated between model 1 and model 4 using the two-proportion Z-test. P4 indicates the p value calculated between model 2 and model 3 using the two-proportion Z-test. P5 indicates the p value calculated between model 2 and model 4 using the two-proportion Z-test. P6 indicates the p value calculated between model 3 and model 4 using the two-proportion Z-test.

filters. In another community-based dataset for DR screening, the system showed remarkably better performance in the GQ groups for 2 quality filters (illumination and clarity). In a clinic-based dataset for optic disc edema analysis, the application of DeepFundus yielded significantly higher sensitivities in 5 quality filters and markedly higher accuracy in the optic disc quality filter.

To evaluate the robustness of DeepFundus among different retinal disease types and grades, we performed subgroup analysis in the aforementioned AMD and DR datasets (Table S5). In both datasets, DeepFundus demonstrated consistent performance between the two groups. Additionally, typical examples of failure cases in DeepFundus analysis were provided to facilitate better understanding and application of DeepFundus (Figure S5).

Effects of adequate-quality fundus images on model development

After image annotation and dataset construction, the distribution of diagnosis classifications in each dataset is summarized in Table S6. For each task, the performances of the 4 designed models using 3 types of selected model architectures are demonstrated in Tables 3, S1–S7, and S2–S7 and Figure 3. In the detection of optic disc edema, model 3 achieved the best accuracy for all types of model architecture. In the detection of drusen, the accuracy of model 2 was the best in the InceptionV3 and InceptionResNetV2 architectures and second only to model 4 by a narrow gap with the DenseNet architecture. However, model 4 with the DenseNet architecture had an extremely unbalanced distribution of sensitivity (0.490) and specificity (0.967). In the detection of DR, model 2 achieved the best accuracy and sensitivity with all types of model architecture.

DISCUSSION

Currently, the majority of AI applications are model driven and focus on designing empirical tests to develop the best model architecture and training procedure to improve the performance of the model. These applications, however, fail to perform properly in real-world settings where data are at the core of every decision-making process, necessitating a paradigm shift from model-driven to data-driven approaches. Data-driven approaches involve systematically improving datasets to enhance the performance of AI applications, which is important but often neglected because it is traditionally regarded as a tedious, low-skill job. To facilitate such data-driven approaches, this study introduced DeepFundus, an automated deep-learning-based fundus image quality classifier involving a total of 13 models using the InceptionResNetV2 technique.

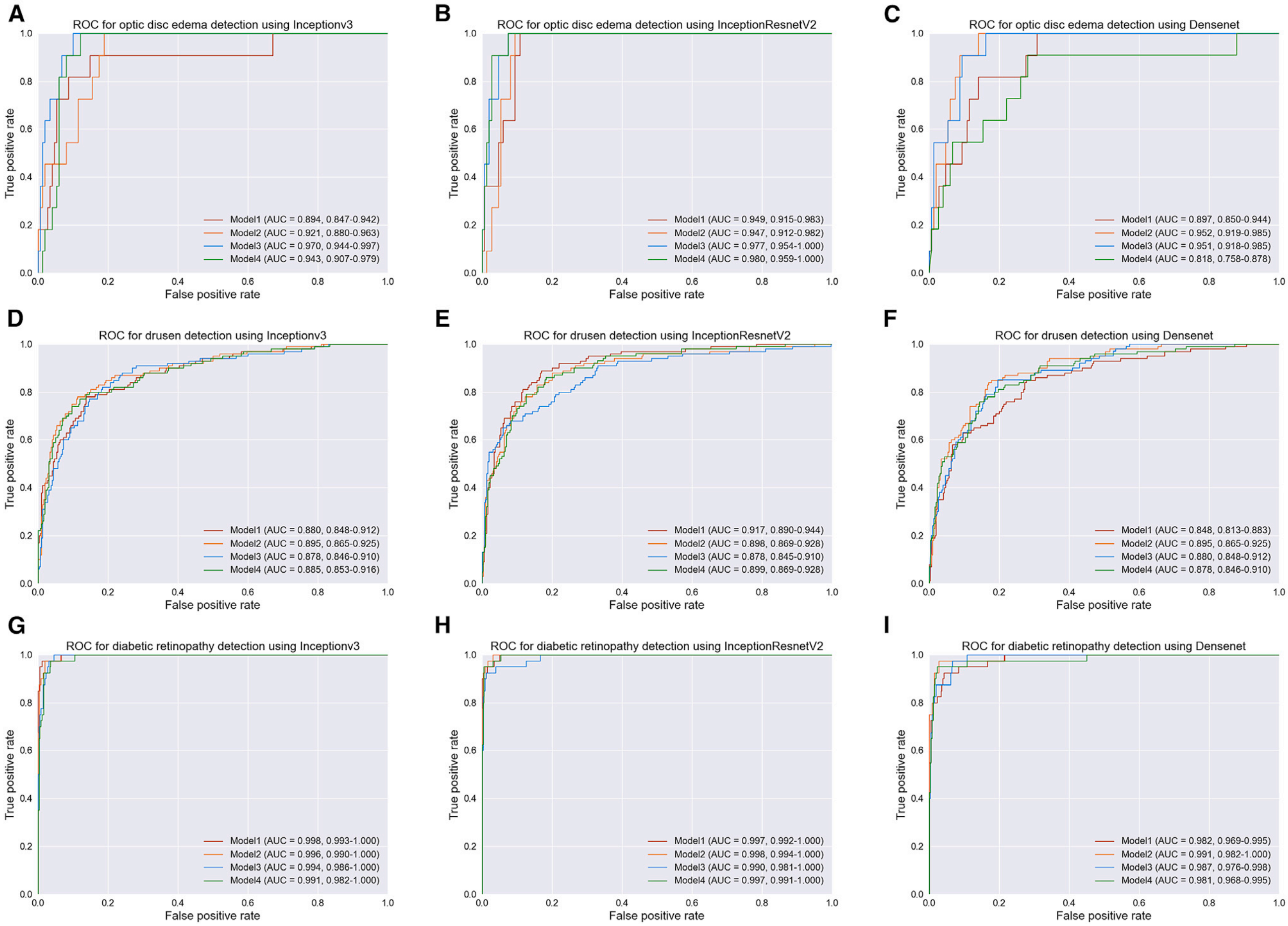


Figure 3. Performance of established models in the detection of retinal diseases using different model architectures

(A–C) ROCs for the detection of optic disc edema using InceptionV3, InceptionResNetV2, and DenseNet.

(D–F) ROCs for the detection of drusen using InceptionV3, InceptionResNetV2, and DenseNet.

(G–I) ROCs for the detection of diabetic retinopathy using InceptionV3, InceptionResNetV2, and DenseNet.

ROC, receiver operating characteristic; AUC, area under the receiver operating characteristic curve.

DeepFundus was designed to assess fundus images in terms of overall quality, clinical quality factors, affected retinal structures, and refractive media opacity. In both the internal test and national validation datasets, DeepFundus achieved AUCs >0.9 for all quality aspects. DeepFundus also demonstrated generalizability to non-Chinese ethnicities in the Kaggle dataset, presenting AUCs of 0.764–0.987 in different quality aspects. Moreover, in clinical application, our results demonstrated the positive effects of DeepFundus in removing inadequate fundus images and enhancing the real-world performance of established AI diagnostics in the detection of multiple retinopathies. During the development phase of the medical AI models, we proposed another group of quality analysis criteria and enhanced model robustness by adding an appropriate proportion of adequate-quality images to the training datasets.

Several studies have reported on the use of deep-learning algorithms for image quality assessment. Mahapatra et al. proposed a system using convolutional neural networks (CNNs) to assess fundus image quality for the first time and obtained an accuracy of 97.9% in distinguishing gradable from ungradable fundus images.[16,22] Afterward, Li et al. developed another AI system to evaluate fundus images in terms of both clarity and location.[23] This system achieved robust performance in a 6,200-image external test dataset. Compared with these studies, our study demonstrated several important features. First, previous studies focused on the classification of retinal images disqualified by different technical factors. However, a large proportion of real-world retinal images are ungradable due to refractive media opacity, especially cataracts.[24] Ignorance of refractive media opacity will not only cause unnecessary image recaptures without quality improvement but also lead to the delayed diagnosis and treatment of ophthalmic diseases. To the best of our knowledge, our study established the first automated system for differentiating between images obscured by refractive media opacity and those obscured due to technical factors, which can be of great assistance to both technicians and patients. Second, our system included both generic quality factors (position, illumination, and clarity) and affected retinal

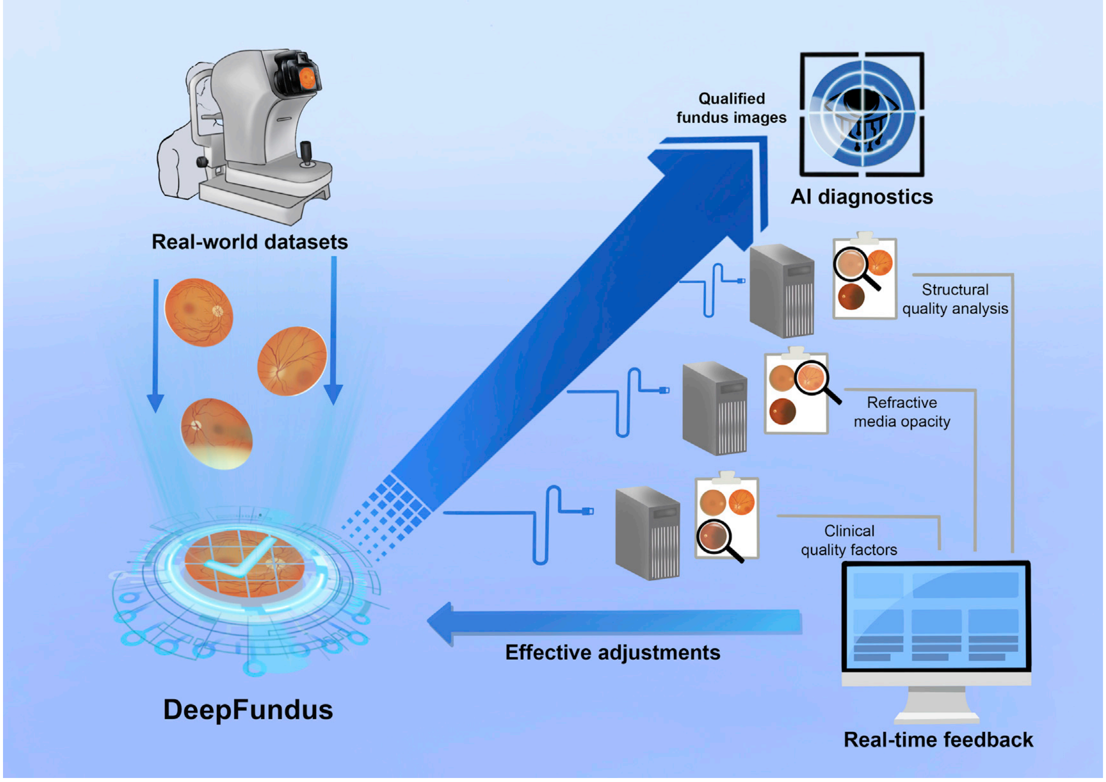


Figure 4. Clinical application of DeepFundus

Each fundus photograph collected in real-world settings will receive DeepFundus classification in terms of clinical quality factors, refractive media opacity, and structural quality analysis before entering downstream analysis. This system can also provide effective adjustments in real time for image acquisition based on quality analysis. These functions allow DeepFundus to serve as a data management tool in the whole life cycle of medical AI.

structures (optic disc, macula, and other retinal areas) in the quality assessment, thus evaluating fundus image quality from a more specific and clinically relevant perspective. Third, the datasets used to validate DeepFundus were acquired via various types of digital fundus cameras from multiethnic populations in different clinical settings and were therefore more representative of the real world. These features could enable DeepFundus to serve as a practical and fundamental tool in medical AI research, such as flow cytometry in biological research. To facilitate automated, multidimensional quality analysis of fundus images and improve interinstitutional collaboration, we built a website-based platform freely available at http://www.myopiaprediction.com/modelstore/#/model/list. After registration, users can upload retinal images from a file. Then, the analysis results concerning different quality aspects and recommended adjustments will be presented on the website for downstream utilization, as presented in Figure 4.

Data quality issues have aroused great concerns in the field of medical AI, but solutions remain to be investigated.[5,6,15] Recently, Dai et al. integrated on-site image quality assessment into DeepDR, a deep-learning-based DR screening system, and demonstrated that real-time quality feedback can improve DR diagnosis using DeepDR.[25] Still, it remains to be explored how image quality analysis will alter AI diagnosis in other retinal diseases and how these changes are associated with different quality factors. To further elucidate this problem, DeepFundus was embedded into a certified AI diagnostic system to remove PQ images prior to the detection of three types of retinal abnormalities: drusen, referable DR, and optic disc edema. In the identification of drusen, the AI system achieved significantly higher sensitivities and accuracies in the GQ groups than in the PQ groups for most quality filters. In the detection of referable DR and optic disc edema, DeepFundus seemed to produce slightly distinct impacts on its diagnostic performance. There are a number of possible reasons for these differences in performance. While drusen are delicate features and thereby vulnerable to degraded image quality, referable DR is a relatively obvious characteristic and can still be recognized in some unqualified images; this probably accounts for the comparatively minor effects of DeepFundus on AI detection of referable DR. Compared with macular areas, where drusen are mostly located, optic discs are less likely to be affected by inadequate illumination and opacity. Accordingly, the accuracy of the AI-based diagnosis of optic disc edema may not benefit as much from the application of DeepFundus. Nevertheless, it should be noted that in the detection of optic disc edema, the AI system attained markedly higher sensitivities in the GQ groups than in the PQ groups for most quality filters. Since optic disc edema usually indicates severe vision-threatening or even life-threatening conditions such as multiple sclerosis, optic neuritis, and intracranial hypertension, an increase in sensitivity can be of substantial importance to these patients, particularly in large-scale disease screening.

To solve data quality issues upstream from the clinical implementation phase, we extended our quality annotation criteria to cater to data classification tasks during model development, where levels of quality defects are more relevant than the types of quality factors. We demonstrated that adding adequate-quality images to training datasets could contribute to better model performance in the test datasets, where both excellent-quality and adequate-quality images exist. In addition, the best models were generally trained on datasets whose data quality distribution was the closest to that of the test dataset. Accordingly, to guarantee model robustness in real-world settings, where multiple data quality defects can arise, including an appropriate proportion of low-quality images in the training datasets is necessary.

Thus, DeepFundus exhibited robust performance in the systematic quality classification of fundus images, constituting a practical data management system for medical AI. Based on its model framework, DeepFundus can provide specific instructions for operators and patients on site, contributing to standardization of retinal image datasets, accurate detection of both ophthalmic and systemic disorders, and improved real-world application of AI diagnostics. Furthermore, we demonstrated that the concept of data quality classification can also be applied to the model development stage to enhance model robustness, indicating the great potential of this system to be integrated into a data-driven workflow to deploy and maintain DLSs more reliably and flexibly. Considering the ubiquity of data quality issues in medical imaging, irrespective of the field, the concept and design of DeepFundus could be considered when developing an image-based AI diagnostic system.

Limitations of the study

The study findings should be interpreted with several limitations. First, DeepFundus was designed to classify ineligible images based on a single fundus image focusing on the posterior pole. Consequently, this system cannot be used for the retinal examination of different fields based on multiple fundus images, such as mydriatic 7-field imaging. Second, only three major retinal diseases were included in the clinical assessment. The effects of DeepFundus filtration on the diagnostic performance of other retinal diseases, such as glaucoma, central serous chorioretinopathy, and retinitis pigmentosa, will be investigated in our future studies. Third, the generalizability of DeepFundus in other real-world datasets, such as DeepDRiD[26] and Messidor,[27] also remains to be explored in the future.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
  - ○ Study design and participants
  - ○ Image quality annotation criteria
  - ○ Development and internal test of DeepFundus
  - ○ External test of DeepFundus
  - ○ Heatmap generation
  - ○ Clinical assessment of DeepFundus
  - ○ Experimental design for quality analysis during model development
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2022.100912.

ACKNOWLEDGMENTS

This project was supported by the National Natural Science Foundation of China (NSFC; grant nos. 82171035 and 82201237), the Science and Technology Program of Guangzhou (grant no. 202201020337), Guangzhou Basic and Applied Basic Research Project (grant no. 202201011301), and the Hainan Province Clinical Medical Center. The authors also acknowledge Chong Guo for his technical support for this project.

AUTHOR CONTRIBUTIONS

Conceptualization, H.L., Y.L., X.W., and L. Liu; methodology, X.W., L. Liu, and D.L.; software, Mingyuan Li and Z.L.; validation, Mingyuan Li, Q.Y., Y. Zhong, X. Huang, Y. He, Z.F., Y.X., L. Zhang, M.Z., J.Q., Fan Xu, P.L., J.L., Fabao Xu, W.W., L. Dong, G.D., X. He, W. Yan, Q.Z., L. Lu, J.Z., W.Z., X.M., S.L., M.S., Q.J., N.C., X. Zhou, Meiyan Li, Y. Wang, H. Zou, H. Zhong, W. Yang, W.S., X. Zhong, Z.Y., L. Ding, Y. Hu, G.T., W.H., X. Zhao, and Y.C.; formal analysis, L. Zhao; investigation, L. Liu, X.W., W.L., W.X., Y.S., W.F., and X.T.; resources, H.L. and Y.L.; data curation, L. Liu, Y.S., Q.L., S. Liu, X.L., J.S., Y. Zhao, and X. Yang; writing – original draft, L. Liu and X.W.; writing – review & editing, H.L., Y.L., D.L., Mingyuan Li, D.Y., J.P., L. Li, and Y. Wu; visualization, L. Liu, Mingyuan Li, and Y.W.; funding acquisition, H.L., X.W., and W.X.; supervision, H.L., Y.L., and X.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 21, 2022
Revised: November 1, 2022
Accepted: December 26, 2022
Published: January 19, 2023

REFERENCES

1. Denny, J.C., and Collins, F.S. (2021). Precision medicine in 2030-seven ways to transform healthcare. Cell *184*, 1415–1419. https://doi.org/10.1016/j.cell.2021.01.015.
2. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., and Folk, J.C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. *1*, 39. https://doi.org/10.1038/s41746-018-0040-6.
3. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., and Vardoulakis, L.M. (2020). A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy, pp. 1–12.

4. Lin, H., Li, R., Liu, Z., Chen, J., Yang, Y., Chen, H., Lin, Z., Lai, W., Long, E., Wu, X., et al. (2019). Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine *9*, 52–59. https://doi.org/10.1016/j.eclinm.2019.03.001.

5. Maier, K., Zaniolo, L., and Marques, O. (2022). Image quality issues in teledermatology: a comparative analysis of artificial intelligence solutions. J. Am. Acad. Dermatol. *87*, 240–242. https://doi.org/10.1016/j.jaad.2021.07.073.

6. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., and Saria, S. (2021). The clinician and dataset shift in artificial intelligence. N. Engl. J. Med. *385*, 283–286. https://doi.org/10.1056/NEJMc2104626.

7. US Food and Drug Administration (2018). FDA Permits Marketing of Artificial Intelligence-Based Device to Detect Certain Diabetes-Related Eye Problems. https://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm604357.htm.

8. Lin, D., Xiong, J., Liu, C., Zhao, L., Li, Z., Yu, S., Wu, X., Ge, Z., Hu, X., Wang, B., et al. (2021). Application of Comprehensive Artificial intelligence Retinal Expert (CARE) system: a national real-world evidence study. Lancet. Digit. Health *3*, e486–e495. https://doi.org/10.1016/S2589-7500(21)00086-8.

9. Liu, Y., Wu, F., Lu, L., Lin, D., and Zhang, K. (2015). Videos in clinical medicine. Examination of the retina. N. Engl. J. Med. *373*, e9. https://doi.org/10.1056/NEJMvcm1308125.

10. Ting, D.S.W., Cheung, C.Y.L., Lim, G., Tan, G.S.W., Quang, N.D., Gan, A., Hamzah, H., Garcia-Franco, R., San Yeo, I.Y., Lee, S.Y., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. JAMA *318*, 2211–2223. https://doi.org/10.1001/jama.2017.18152.

11. Burlina, P.M., Joshi, N., Pekala, M., Pacheco, K.D., Freund, D.E., and Bressler, N.M. (2017). Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. JAMA Ophthalmol. *135*, 1170–1176. https://doi.org/10.1001/jamaophthalmol.2017.3782.

12. Li, Z., He, Y., Keel, S., Meng, W., Chang, R.T., and He, M. (2018). Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. Ophthalmology *125*, 1199–1206. https://doi.org/10.1016/j.ophtha.2018.01.023.

13. Poplin, R., Varadarajan, A.V., Blumer, K., Liu, Y., McConnell, M.V., Corrado, G.S., Peng, L., and Webster, D.R. (2018). Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat. Biomed. Eng. *2*, 158–164. https://doi.org/10.1038/s41551-018-0195-0.

14. Mitani, A., Huang, A., Venugopalan, S., Corrado, G.S., Peng, L., Webster, D.R., Hammel, N., Liu, Y., and Varadarajan, A.V. (2020). Detection of anaemia from retinal fundus images via deep learning. Nat. Biomed. Eng. *4*, 18–27. https://doi.org/10.1038/s41551-019-0487-z.

15. Ruamviboonsuk, P., Tiwari, R., Sayres, R., Nganthavee, V., Hemarat, K., Kongprayoon, A., Raman, R., Levinstein, B., Liu, Y., Schaekermann, M., et al. (2022). Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. Lancet. Digit. Health *4*, e235–e244. https://doi.org/10.1016/S2589-7500(22)00017-6.

16. Mahapatra, D., Roy, P.K., Sedai, S., and Garnavi, R. (2016). A CNN based neurobiology inspired approach for retinal image quality assessment. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. *2016*, 1304–1307. https://doi.org/10.1109/EMBC.2016.7590946.

17. Shao, F., Yang, Y., Jiang, Q., Jiang, G., and Ho, Y.-S. (2018). Automated quality assessment of fundus images via analysis of illumination, naturalness and structure. IEEE Access *6*, 806–817. https://doi.org/10.1109/access.2017.2776126.

18. Zago, G.T., Andreão, R.V., Dorizzi, B., and Teatini Salles, E.O. (2018). Retinal image quality assessment using deep learning. Comput. Biol. Med. *103*, 64–70. https://doi.org/10.1016/j.compbiomed.2018.10.004.

19. Chalakkal, R.J., Abdulla, W.H., and Thulaseedharan, S.S. (2019). Quality and content analysis of fundus images using deep learning. Comput. Biol. Med. *108*, 317–331. https://doi.org/10.1016/j.compbiomed.2019.03.019.

20. Shen, Y., Sheng, B., Fang, R., Li, H., Dai, L., Stolte, S., Qin, J., Jia, W., and Shen, D. (2020). Domain-invariant interpretable fundus image quality assessment. Med. Image Anal. *61*, 101654. https://doi.org/10.1016/j.media.2020.101654.

21. Public Health England (2013). Diabetic Eye Screening Programme: Pathway for Adequate or Inadequate Images and where Images Cannot Be Taken. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/403107/Pathway_for_adequate_inadequate_images_and_where_images_cannot_be_taken_v1_4_10Apr13.pdf.

22. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., and Sánchez, C.I. (2017). A survey on deep learning in medical image analysis. Med. Image Anal. *42*, 60–88. https://doi.org/10.1016/j.media.2017.07.005.

23. Li, Z., Jiang, J., Zhou, H., Zheng, Q., Liu, X., Chen, K., Weng, H., and Chen, W. (2021). Development of a deep learning-based image eligibility verification system for detecting and filtering out ineligible fundus images: a multicentre study. Int. J. Med. Inf. *147*, 104363. https://doi.org/10.1016/j.ijmedinf.2020.104363.

24. Scanlon, P.H., Foy, C., Malhotra, R., and Aldington, S.J. (2005). The influence of age, duration of diabetes, cataract, and pupil size on image quality in digital photographic retinal screening. Diabetes Care *28*, 2448–2453. https://doi.org/10.2337/diacare.28.10.2448.

25. Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., Liu, R., Wang, X., Hou, X., Liu, Y., et al. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nat. Commun. *12*, 3242. https://doi.org/10.1038/s41467-021-23458-5.

26. Liu, R., Wang, X., Wu, Q., Dai, L., Fang, X., Yan, T., Son, J., Tang, S., Li, J., Gao, Z., et al. (2022). DeepDRiD: diabetic retinopathy-grading and image quality estimation challenge. Patterns (N Y) *3*, 100512. https://doi.org/10.1016/j.patter.2022.100512.

27. Abràmoff, M.D., Folk, J.C., Han, D.P., Walker, J.D., Williams, D.F., Russell, S.R., Massin, P., Cochener, B., Gain, P., Tang, L., et al. (2013). Automated analysis of retinal images for detection of referable diabetic retinopathy. JAMA Ophthalmol. *131*, 351–357. https://doi.org/10.1001/jamaophthalmol.2013.1743.

28. EyePACS (2022). Why EyePACS. http://www.eyepacs.com/why-eyepacs.

29. Artificial Intelligent Ophthalmology Group, Intelligent Medicine Special Committee of China Medicine Education Association; National Key Research and Development Program of China “Development and Application of Ophthalmic Multimodal Imaging and Artificial Intelligence Diagnosis and Treatment System” Project Team (2019). guidelines for artificial intelligent diabetic retinopathy screening system based on fundus photography. Chinese Journal of Experimental Ophthalmology *37*, 593–598.

30. Christian Szegedy, S.I., Vincent, V., and Alemi, A. (2017). Inception-v4, Inception-Resnet and the Impact of Residual Connections on Learning (Thirty-first AAAI conference on artificial intelligence).

31. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, *22-29*, pp. 618–626.

32. Fundus Image Asisted Diagnostic Software Products for Diabetic Retinopacy Appproved (2020). https://www.nmpa.gov.cn/zhuanti/ypqxgg/gggzjzh/20200810093435157.html?type=pc&m=.

33. Milea, D., Najjar, R.P., Zhubo, J., Ting, D., Vasseneix, C., Xu, X., Aghsaei Fard, M., Fonseca, P., Vanikieti, K., Lagrèze, W.A., et al. (2020). Artificial intelligence to detect papilledema from ocular fundus photographs. N. Engl. J. Med. *382*, 1687–1695. https://doi.org/10.1056/NEJMoa1917130.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python 3.6.13	Python	https://www.python.org
TensorFlow 1.13.1	TensorFlow	https://www.tensorflow.org
Keras 2.3.1	Keras	https://keras.io

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Haotian Lin (linht5@mail.sysu.edu.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The confidential medical records data reported in this study cannot be deposited in a public repository. To request access, contact the Lead Author. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Study design and participants

The overall study design is shown in Figure 1. We formed the development and internal test datasets using fundus images collected from 4 different cohorts from April 2018 to April 2019 to develop DeepFundus. Then, fundus images from another 22 hospitals nationwide and the Kaggle dataset (EyePACS LLC, San Jose, CA, USA) featuring different geographical and ethnic compositions were used to externally test its performance.[28] Detailed information about the datasets is described in Table S8.

Image quality annotation criteria

All retinal images were labeled according to three-level annotation criteria: overall quality, clinical quality factors, and structural quality analysis. Detailed definitions and examples of each quality aspect and their logical relations are demonstrated in Figure S1. The proposed quality factors (position, illumination, and clarity) are well-established image quality assessment aspects in clinical practice.[29] For each quality factor, the affected retinal areas were also labeled to constitute structural quality analysis. All images were independently labeled by two ophthalmologists licensed in China (each with >5 years of experience). For each photograph, if all labels were the same, they were regarded as the ground truth. Otherwise, arbitration was performed by a retinal expert with >10 years of experience in practice.

For patients enrolled in TAH, both slit-lamp photography and retinal images were acquired. When the patients' retinal images were labeled poor clarity, slit-lamp photographs of the corresponding eyes were inspected by 2 senior ophthalmologists (each with >10 years of experience) to determine whether poor clarity was caused by refractive media opacities (e.g., cataracts). Accordingly, this dataset (TAH) included additional classification labels of obscured fundus images (Figure S1).

Development and internal test of DeepFundus

DeepFundus was designed to consist of 12 models to classify retinal images concerning different quality aspects and 1 model for detecting refractive media opacity from blurred fundus images. Images included from April 2018 to April 2019 were randomly split in a 7:1.5:1.5 ratio into the training set, internal validation set, and internal test set; no images overlapped among these sets.

Image standardization was performed prior to model construction. All images were downsized to 512 × 512 pixels, and the pixel values were normalized to an interval between 0 and 1. Data augmentation was used to increase image heterogeneity of the training dataset and thus reduce the chance of overfitting during the deep learning process. The new samples were obtained through simple transformations of the original images and corresponded to "real-world" acquisition conditions. Random horizontal and vertical flipping, random rotations up to 90 degrees around the center of the image, and random brightness shifts within the range of 0.8 to 1.2 were applied to the images of the training set in real time during training.

A state-of-the-art deep CNN architecture, InceptionResNetV2, which incorporates the architectural features of the Inception family with residual connections, was used to construct the model.[30] Python 3.6.13, TensorFlow 1.13.1, and Keras 2.3.1 were used to train and test the models with an initial learning rate of 0.001. All models were trained up to 500 epochs. In the training process, the validation loss was assessed after each epoch for model selection. Early stopping was employed, and when the validation loss did not improve over 120 consecutive epochs, the training process was stopped. The model state with the lowest loss was saved as the final state of the model. Each model had one input and one output; the input of the model was a retinal image, and the output function was a standard binary task for determining whether the quality of the input image was poor in the targeted aspect or whether the blurred image was caused by refractive media opacity. The development environment was based on Ubuntu 16.04.6 with NVIDIA Tesla V100 PCIe 32GB.

External test of DeepFundus

After model development, DeepFundus was externally tested using images collected from February 2020 to October 2022, including a national validation dataset from 22 hospitals of different levels across China and a non-Chinese ethnicity dataset from Kaggle.

Heatmap generation

To combat the black-box effect in DLSs, we used improved Grad-CAM (Gradient-weighted Class Activation Mapping) to enhance the interpretability of DeepFundus.[31] The Grad-CAM algorithm can produce a class-specific activation heatmap where each activation value represents the importance of classifying to that class. Redder regions indicate more significant features. Using this method, a heatmap was generated to display the location on which the decision of DeepFundus was based.

Clinical assessment of DeepFundus

To evaluate the effectiveness of DeepFundus in real-world implementation, DeepFundus was integrated into an established AI diagnostic system for removing unqualified fundus images before diagnosis. We previously developed this AI diagnostic system to detect multiple retinopathies using fundus images,[8] and its DR diagnosis module was recently designated among the first batch of class III AI-based devices by the National Medical Products Administration (NMPA).[32] This experiment was conducted using 8783 retinal images collected in 3 prospective cohorts (from May 2020 to December 2020) for the analysis of AMD, DR, and optic disc edema. The definitions for judgment of these included retinopathies were based on previously established criteria.[10,11,33] For each quality filter, fundus images were classified into GQ or PQ groups. The performance of the AI diagnostic system in the GQ groups was compared to that in the corresponding PQ groups.

Experimental design for quality analysis during model development

To investigate the impact of data quality on the model development phase, we annotated fundus image gradability according to a new set of standards including 3 quality categories: excellent, adequate and inadequate. "Excellent" means no noticeable quality defects and all targeted retinopathy lesions gradable; "Adequate" means noticeable quality defects and all targeted retinopathy lesions gradable; "Inadequate" means severe quality defects and some targeted retinopathy lesions ungradable. These criteria were applied to three cohorts to establish DLSs separately depicting optic disc edema, referable DR, and drusen. For each task, 4 datasets with the same size but different quality distributions were constructed, and we used a common test set to evaluate model performance. A total of 3 types of model architectures, including InceptionV3, DenseNet, and InceptionResNetV2, were tested on these datasets. For each CNN architecture, the hyperparameters were fixed to explore how different quality distributions influence model performance.

QUANTIFICATION AND STATISTICAL ANALYSIS

The performance of DeepFundus in detecting poor-quality images in terms of each quality aspect was evaluated by calculating the sensitivity and specificity with 95% confidence intervals (CIs). We plotted a receiver operating characteristic (ROC) curve to show the ability of the system to evaluate image quality. The ROC curve was created by plotting the ratio of true positive cases (sensitivity) against the ratio of false-positive cases (1-specificity). The larger the area under the ROC curve (AUC), the better the performance was inferred to be. All statistical tests were 2-sided with a significance level of 0.05. Statistical analyses were conducted using Python 3.6.13.

ADDITIONAL RESOURCES

This study was registered with ClinicalTrials.gov (NCT04289064) and approved by the Institutional Review Board of Zhongshan Ophthalmic Center at Sun Yat-sen University (IRB-ZOC-SYSU). All procedures followed the tenets of the Declaration of Helsinki. All patients were informed of the study and signed consent forms before inclusion.