Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Enhancing AI reliability: A foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis

Yuanyuan Peng,[1,17] Aidi Lin,[2,17] Meng Wang,[3,4,17] Tian Lin,[2] Linna Liu,[5] Jianhua Wu,[5] Ke Zou,[6] Tingkun Shi,[2] Lixia Feng,[7] Zhen Liang,[1,8] Tao Li,[9] Dan Liang,[9] Shanshan Yu,[9] Dawei Sun,[10] Jing Luo,[11] Ling Gao,[11] Xinjian Chen,[12,13] OCT reading group[2] Ching-Yu Cheng,[3,4,14,15] Huazhu Fu,[16,*] and Haoyu Chen[2,18,*]

[1]School of Biomedical Engineering, Anhui Medical University, Hefei, Anhui 230032, China
[2]Joint Shantou International Eye Center, Shantou University and the Chinese University of Hong Kong, Shantou, Guangdong 515041, China
[3]Centre for Innovation & Precision Eye Health, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 119228, Singapore
[4]Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore 117549, Singapore
[5]Wuhan Aier Eye Hospital, Wuhan, Hubei 430063, China
[6]National Key Laboratory of Fundamental Science on Synthetic Vision and the College of Computer Science, Sichuan University, Chengdu, Sichuan 610065, China
[7]Department of Ophthalmology, First Affiliated Hospital of Anhui Medical University, Hefei, Anhui, China
[8]The Affiliated Chuzhou Hospital of Anhui Medical University, First People's Hospital of Chuzhou, Chuzhou, Anhui 239099, China
[9]State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, Guangdong 510000, China
[10]Department of Ophthalmology, The Second Affiliated Hospital of Harbin Medical University, Harbin 150001, China
[11]Department of Ophthalmology, The Second Xiangya Hospital, Central South University, Changsha 410011, China
[12]School of Electronics and Information Engineering, Soochow University, Suzhou, Jiangsu 215006, China
[13]State Key Laboratory of Radiation Medicine and Protection, Soochow University, Suzhou, Jiangsu 215006, China
[14]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Republic of Singapore
[15]Ophthalmology & Visual Sciences Academic Clinical Program (EYE ACP), Duke-NUS Medical School, Singapore, Singapore
[16]Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore
[17]These authors contributed equally
[18]Lead contact
*Correspondence: hzfu@ieee.org (H.F.), drchenhaoyu@gmail.com (H.C.)
https://doi.org/10.1016/j.xcrm.2024.101876

SUMMARY

Inability to express the confidence level and detect unseen disease classes limits the clinical implementation of artificial intelligence in the real world. We develop a foundation model with uncertainty estimation (FMUE) to detect 16 retinal conditions on optical coherence tomography (OCT). In the internal test set, FMUE achieves a higher F1 score of 95.74% than other state-of-the-art algorithms (92.03%–93.66%) and improves to 97.44% with threshold strategy. The model achieves similar excellent performance on two external test sets from the same and different OCT machines. In human-model comparison, FMUE achieves a higher F1 score of 96.30% than retinal experts (86.95%, $p = 0.004$), senior doctors (82.71%, $p < 0.001$), junior doctors (66.55%, $p < 0.001$), and generative pretrained transformer 4 with vision (GPT-4V) (32.39%, $p < 0.001$). Besides, FMUE predicts high uncertainty scores for >85% images of non-target-category diseases or with low quality to prompt manual checks and prevent misdiagnosis. Our FMUE provides a trustworthy method for automatic retinal anomaly detection in a clinical open-set environment.

INTRODUCTION

Retinal diseases are important causes of irreversible blindness.[1] Early diagnosis and timely management are essential to effectively reduce visual impairment and blindness. Over the past three decades, the emergence of optical coherence tomography (OCT) has dramatically altered the landscape of vitreoretinal diagnosis and become one of the most important imaging modalities in ophthalmology. It is a non-invasive technology that can quickly generate high-resolution *in vivo* retinal images, enabling the visualization of the cross-sectional structure.[2] Currently, OCT is widely used in clinics and provides a vital reference for the diagnosis of many retinal diseases, such as vitreomacular interface disorders,[3] diabetic macular edema (DME),[4] and age-related macular degeneration (AMD).[5] However, diagnosing retinal diseases requires well-trained ophthalmologists, whose shortage makes it difficult to cope with the growing number of patients with disease. Furthermore, the distribution of

medical resources is uneven. In rural and underdeveloped regions, the shortage of ophthalmologists is worse. Additionally, interpreting OCT images is quite time-consuming and labor intensive. Therefore, developing an automatic retinal disease detection model based on OCT images shows promise to assist clinical decision-making, reduce the workload of ophthalmologists, and facilitate blindness prevention.

Deep learning (DL) has been applied in retinal imaging, including OCT, to facilitate automatic diagnosis. In 2018, transfer learning was introduced to classify OCT images into normal and three diseases.[6] A DL model was also developed to classify OCT volumes to different referral suggestions and diagnosis probabilities.[7] Recently, RETFound, a foundation model pretrained on large-scale color fundus photography (CFP) and OCT images, has demonstrated great potential in detecting retinal diseases after fine-tuning.[8] However, a significant downside of the standard artificial intelligence (AI) model is that it only gives the prediction results without any information reflecting the reliability of the prediction, which may lead to low credibility of the model in clinical implementation. Furthermore, these traditional AI models are developed with a limited number of disease categories and would encounter unseen diseases (out of distribution, OOD) in real-world implementation and make incorrect predictions. The limitation of these models may lead to misdiagnosis or missed diagnosis and finally affect the clinical outcome of the patients.

In our previous study, we developed an uncertainty-inspired open-set learning (UIOS) model for CFP classification,[9] which can output an uncertainty score in addition to the probabilities of disease categories. When the model encounters OOD data, such as previously unseen diseases, it will generate a high uncertainty score exceeding the threshold, indicating the need for a double-check by an ophthalmologist to prevent misdiagnosis. In the current study, we integrated a fine-tuned foundation model with uncertainty estimation (FMUE) in OCT images, enabling the capability of expressing the level of confidence and reliability of disease classification in open-set clinical implementation. Figure 1 shows the training and inference process of our proposed FMUE framework. We compared our model performance with other recent state-of-the-art DL algorithms and a group of ophthalmologists with varying clinical experience to validate its diagnostic accuracy.

RESULTS

Performance on the internal test set

In the internal test set with 19,655 images obtained from Topcon OCT devices, our FMUE achieved the following average performance metrics: an F1 score of 95.74% (95% confidence interval [CI]: 95.19%–96.02%, Table 1 and Data S1), sensitivity of 96.11% (95% CI: 95.58%–96.38%, Table S1A and Data S1), precision of 95.58% (95% CI: 95.02%–95.87%, Table S1B and Data S1), an area under the curve (AUC) of the receiver operating characteristic curve of 98.93% (95% CI: 98.79%–99.07%, Figure 2A), and overall accuracy of 95.69% (95% CI: 95.41%–95.97%, Table S4A and Data S4). The confusion metrics are shown in Figure 2B. The performance of FMUE was superior to RETFound, Swin transformer (Swin_T), ensemble models (Ensemble), and UIOS; for example, the average F1 score of FMUE (95.74%, 95% CI: 95.19%–96.02%) was higher than that of RETFound (93.34%, 95% CI: 92.99%–93.69%, p = 0.139, Tables 1 and S4B, Data S1), Swin_T (92.85%, 95% CI: 92.49%–93.21%, p = 0.048, Tables 1 and S4B, Data S1), Ensemble (93.66%, 95% CI: 93.32%–94.00%, p = 0.112, Tables S2A and S4B, Data S2), and UIOS (92.03%, 95% CI: 91.29%–92.41%, p = 0.014, Tables S2A and S4B, Data S2).

We also used a threshold strategy to remove samples with high uncertainty. There were 5.19% samples with uncertainty scores above the threshold in FMUE, which was lower than that in Ensemble (52.85%, Tables S4A andS5) and UIOS (10.37%, Tables S4A andS5). Similar to the UIOS model, the distribution of uncertainty scores for the FMUE model in the internal test set was similar to the validation set (Figure 3A). The samples with high uncertainty score were 19.187-fold (95% CI: 16.400–22.449, $p < 0.001$) risk of being misclassified if they were not removed (Table S6A). Furthermore, after filtering the OOD samples with high uncertainty scores, the performance metrics (Tables S2 and S3, Figure S1) and confusion matrix (Figure S2) of FMUE were further improved; for example, the average F1 score improved to 97.44% (95% CI: 97.22%–97.66%, Table S3A and Data S3), which was better than that of Ensemble (86.77%, 95% CI: 86.30%–87.24%, p = 0.109, Tables S3A and S4B, Data S3) and UIOS with threshold strategy (92.47%, 95% CI: 91.75%–92.84%, p = 0.011, Tables S3A and S4B, Data S3).

Performance on the external test sets

The models were also tested on two external test sets from private and public sources. The external-private set obtained from the Topcon OCT device includes the same 16 categories as the training set (Table S5), while the external-public set contains five publicly available OCT datasets, with various types of diseases and different models of OCT instruments (Table S8C). In the external-private set with 5,175 images, our FMUE achieved an average of 94.73% (95% CI: 94.12%–95.34%) for F1 score (Table 1), 93.41% (95% CI: 92.73%–94.09%) for sensitivity (Table S1A), 96.76% (95% CI: 96.28%–97.24%) for precision (Table S1B), 98.53% (95% CI: 98.20%–98.88%) for AUC (Figure 2A), and 96.48% (95% CI: 95.97%–96.98%) for overall accuracy (Table S4A). The confusion metrics are shown in Figure 2B. In the external-public set with 6,182 images, the average F1 score, accuracy, sensitivity, and precision were 77.11% (95% CI: 76.06%–78.16%), 85.02% (95% CI: 84.13%–85.91%), 86.45% (95% CI: 85.60%–87.30%), and 74.93% (95% CI: 73.85%–76.01%), respectively (Tables 1, S4A, and S1). Overall, in the two external test sets, the F1 score of FMUE (94.73% and 77.11%) was higher than that of RETFound (92.17% and 73.27%, p = 0.184 and 0.031), Swin_T (88.89% and 61.18%, p = 0.032 and 0.052), Ensemble (92.72% and 75.68%, p = 0.303 and 0.511), and UIOS (85.48% and 67.04%, p = 0.045 and 0.111) (Tables 1,S2A, and S4B).

Furthermore, after thresholding, the F1 score and accuracy of FMUE, UIOS, and Ensemble improved further (Tables S2A, S3A, and S4A), except for F1 score of UIOS on the external-private set and Ensemble on the internal and external-private set. It was noted that more samples were identified as having high

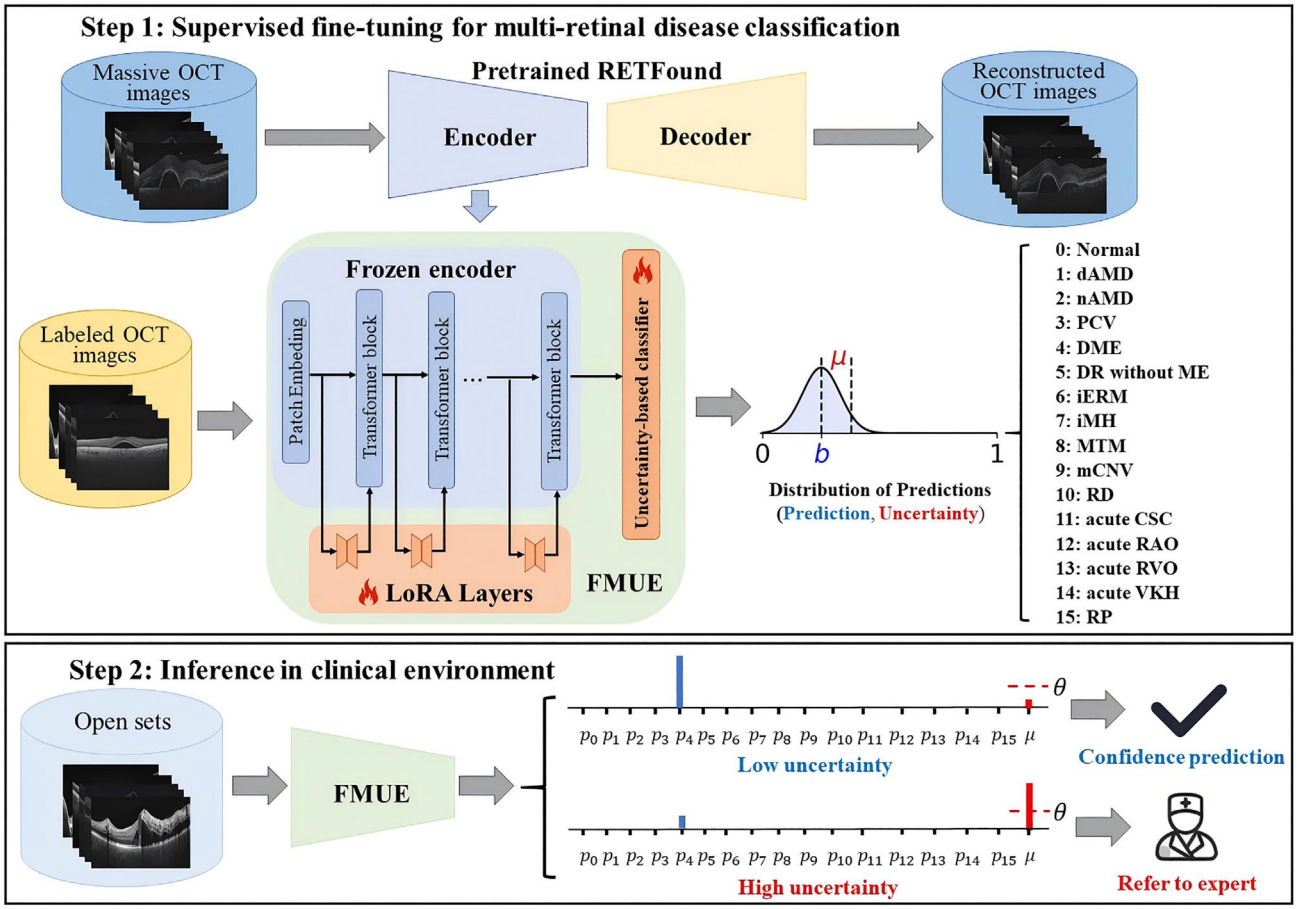


Figure 1. Schematic diagram of our FMUE for clinical work

Step 1 adapts pretrained RETFound to multiple retinal disease classifications on OCT images by means of supervised fine-tuning on data with explicit label. We freeze the image encoder of RETFound (blue area) and insert additional trainable LoRA layers to RETFound for OCT image feature extraction. In addition, to increase the credibility of AI model prediction results, we develop an uncertainty-based classifier to obtain the final prediction result with a corresponding uncertainty score. Step 2 shows the inference process of our FMUE in clinical environment. When the model is fed with an image with obvious features of retinal disease in the training categories, our FMUE will give a diagnosis result with an uncertainty score below the threshold θ to indicate the diagnosis result is reliable. Conversely, when the input image contains ambiguous features or is OOD data, our model will give a high uncertainty score above the threshold θ to indicate the result is unreliable and refer the patient to an experienced ophthalmologist for double-checking.

uncertainty in the external-public sets, compared to in the internal test set and the external-private set (Figure 3A, Table S4A, and Figure S3). This was probably due to the gap existing in the different models of OCT device.

In the two external test sets, the F1 score of FMUE after thresholding (95.74% and 91.47%) was superior to Ensemble (85.81% and 78.43%, p = 0.136 and 0.036) and UIOS (85.07% and 72.26%, p = 0.028 and 0.002) after thresholding (Tables S3A and S4B). The model uncertainty also showed a positive relationship with the misclassification made by FMUE in both the external-private (odds ratio [OR]: 32.628, $p < 0.001$) and the external-public sets (OR: 9.211, $p < 0.001$) (Table S6A). Furthermore, we analyzed the accuracy by excluding different numbers of high uncertainty samples and found that the AUCs of the accuracy vs. percentage of samples were all higher in FMUE (99.75% and 94.38%) than UIOS (98.12% and 87.90%) (Figure 3B). The curves also showed that excluding the samples with high uncertainty improved the accuracy of classification.

OOD detection

We evaluated the OOD detection performance of our FMUE using three OOD datasets with samples of non-target categories (NTCs) and low quality. In these three datasets, the uncertainty rate was calculated based on the uncertainty scores exceeding the threshold, thereby detecting the OOD samples. FMUE detected 89.22%, 85.44%, and 89.03% of samples with high uncertainty scores on NTC-internal, NTC-external, and low-quality OCT datasets, respectively (Table S4C, Figure 3A), which were more than those of Ensemble (84.80%, 85.44%, and 77.30%, respectively) and UIOS (79.35%, 83.01%, and 79.45%, respectively). Furthermore, the distribution of uncertainty scores in the three OOD datasets was more skewed toward higher values in FMUE than UIOS (Figure 3A).

Table 1. F1 score of different methods on the internal and external test sets (%)

	Internal test set				External-private set				External-public set			
Category	RETFound	Swin_T	FMUE	FMUE_T	RETFound	Swin_T	FMUE	FMUE_T	RETFound	Swin_T	FMUE	FMUE_T
Normal	86.92	91.37	97.23	98.86	98.28	98.14	99.16	99.50	95.38	85.02	97.87	99.68
dAMD	71.75	83.97	93.98	96.30	60.67	83.33	84.00	87.01	74.85	76.04	85.49	94.31
nAMD	91.63	92.35	92.58	97.49	93.07	90.32	96.93	97.59	73.12	64.23	78.40	86.03
PCV	89.63	91.47	92.47	97.61	44.44	49.46	82.69	82.93	79.14	75.6	82.02	93.21
DME	93.20	88.29	93.31	95.51	98.78	89.27	94.30	95.67	67.09	56.84	67.76	85.25
DR without ME	88.58	87.14	88.47	95.24	91.36	93.13	89.49	91.22	–	–	–	–
iERM	94.75	80.80	100.00	100.00	94.43	80.91	97.69	99.68	57.26	18.21	59.82	83.97
iMH	100.00	98.23	100.00	100.00	100.00	98.26	98.26	99.56	74.42	77.3	68.70	100.00
MTM	99.70	99.61	99.34	99.34	95.14	97.23	96.57	98.28	–	–	–	–
mCNV	95.58	99.60	98.50	98.48	99.21	81.91	98.79	98.74	–	–	–	–
RD	97.34	96.01	96.01	96.47	91.90	96.87	94.58	97.40	–	–	–	–
Acute CSC	94.04	94.43	91.96	93.42	98.54	95.14	97.35	98.10	44.44	61.02	53.33	81.63
Acute RAO	99.78	99.06	100.00	100.00	100.00	100	100.00	100.00	80.77	23.73	89.76	97.67
Acute RVO	96.23	92.79	96.50	97.99	96.17	97.11	96.70	99.44	86.20	73.8	88.48	92.94
Acute VKH	94.41	93.40	91.58	92.29	92.00	73.68	89.23	91.01	–	–	–	–
RP	99.97	97.13	99.97	100.00	100.00	97.51	100.00	100.00	–	–	–	–
Average	93.34	92.85	95.74	97.44	92.17	88.89	94.73	95.74	73.27	61.18	77.11	91.47

Abbreviations: dAMD, dry age-related macular degeneration; nAMD, neovascular age-related macular degeneration; PCV, polypoidal choroidal vasculopathy; DME, diabetic retinopathy with macular edema; DR without ME, diabetic retinopathy without macular edema; iERM, idiopathic epiretinal membrane; iMH, idiopathic macular hole; MTM, myopic traction maculopathy; mCNV, myopic choroidal neovascularization; RD, retinal detachment; CSC, central serous chorioretinopathy; RAO, retinal artery occlusion; RVO, retinal vein occlusion; VKH, Vogt-Koyanagi-Harada disease; RP, retinitis pigmentosa; RETFound, a foundation model for retinal images; Swin_T, Swin transformer; FMUE, foundation model with uncertainty estimation; FMUE_T, FMUE with threshold strategy.

Performance in human-model comparison

The human-model performance comparison was conducted on the human-model comparison (HMC) set via a customized online image reading system (Figure 4A), the results of which are displayed in Figure 4B and Tables 2 and S4B. The performance of our FMUE (F1 score: 96.30%; sensitivity: 96.25%; precision: 96.57%; accuracy: 96.25%) surpassed that of almost all retinal experts (F1 scores: 80.77%–95.88%; sensitivities: 81.88%–96.25%; precisions: 84.07%–96.85%; accuracies: 81.88%–96.25%). The average F1 scores of retinal experts, senior doctors, and junior doctors were 86.95%, 82.71%, and 66.55%, respectively, which were all significantly lower than that achieved by FMUE (96.30%, $p = 0.004$, $p < 0.001$, and $p < 0.001$, respectively).

As shown in Table S6B, our FMUE obtained high uncertainty scores in 12 images (7.50%), and 6 images were misclassified without thresholding (3.75%) in the HMC set. The model uncertainty was positively associated with the misclassification made by FMUE on the HMC set, at an OR of 16.111 (95% CI: 2.839–91.440, $p = 0.002$) (Table S6A). The high uncertainty and misclassification of FMUE in the HMC set were primarily due to the similar feature sharing in different diseases such as macular edema in DME and acute retinal vein occlusion (RVO), or subtle features such as very small drusen that may be ignored and misclassified as normal condition (Figure S4A and Table S6B).

Furthermore, we compared FMUE with generative pretrained transformer 4 with vision (GPT-4V), which is a multi-modal large model with visual understanding capabilities. As shown in Figure 4B and Table 2, the performance of our FMUE was far superior to GPT-4V, with an average F1 score increase of 60.91% (96.30% vs. 32.39%, $p < 0.001$, Table S4B).

Examples and vision interpretation

Figure 5 displays the heatmaps generated by gradient-weighted class activation mapping (Grad-CAM) to provide visual explanations for decisions made by our FMUE. Figures 5A and 5B are two examples with typical features highlighted in red color and correctly predicted by FMUE with low uncertainty scores. RETFound and UIOS also made correct predictions, although UIOS outputted a high uncertainty score for the first image.

Figures 5C and 5D are two examples of target categories but with ambiguous features; Figures 5E and 5F are two NTC examples while Figure 5G is a low-quality OCT image. The model did not identify the features of these images, as shown in the heatmaps. RETFound made an incorrect prediction without warning of the unreliability. UIOS and FMUE outputted high uncertainty scores to indicate unreliable classification, and a double-check by experienced ophthalmologists was needed.

DISCUSSION

In the current study, we fine-tuned a foundation model and integrated uncertainty estimation for the task of retinal OCT multi-classification and OOD detection. We compared the performance

CellPress
OPEN ACCESS

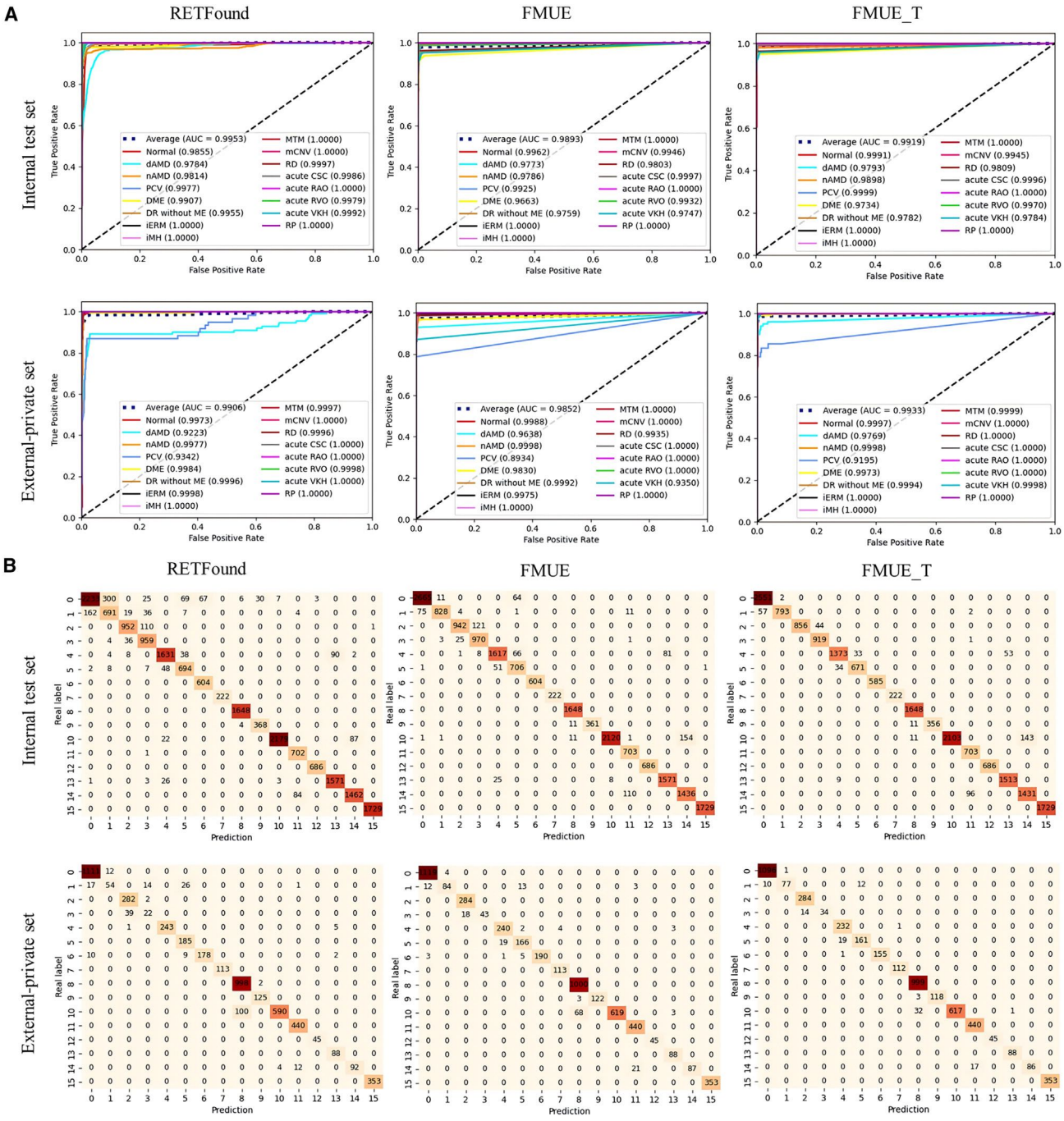


Figure 2. The receiver operating characteristic curves and confusion matrices of the standard AI model RETFound, our FMUE, and FMUE_T on the internal and external-private test set

(A) Receiver operating characteristic (ROC) curves.

(B) Confusion matrices.

of FMUE with RETFound,[8] Swin_T,[10] Ensemble,[11] and UIOS.[9] RETFound and Swin_T were transformer-based model. RETFound was pretrained on large-scale OCT images and focused on disease classification of OCT images, while Swin_T was pretrained on general datasets such as ImageNet. Ensemble and UIOS were uncertainty estimation methods. The results showed that FMUE achieved better performance than RETFound, Swin_T, Ensemble, and UIOS in multi-classification in both internal and external test datasets. FMUE also outperformed Ensemble and UIOS in OOD detection, which was absent

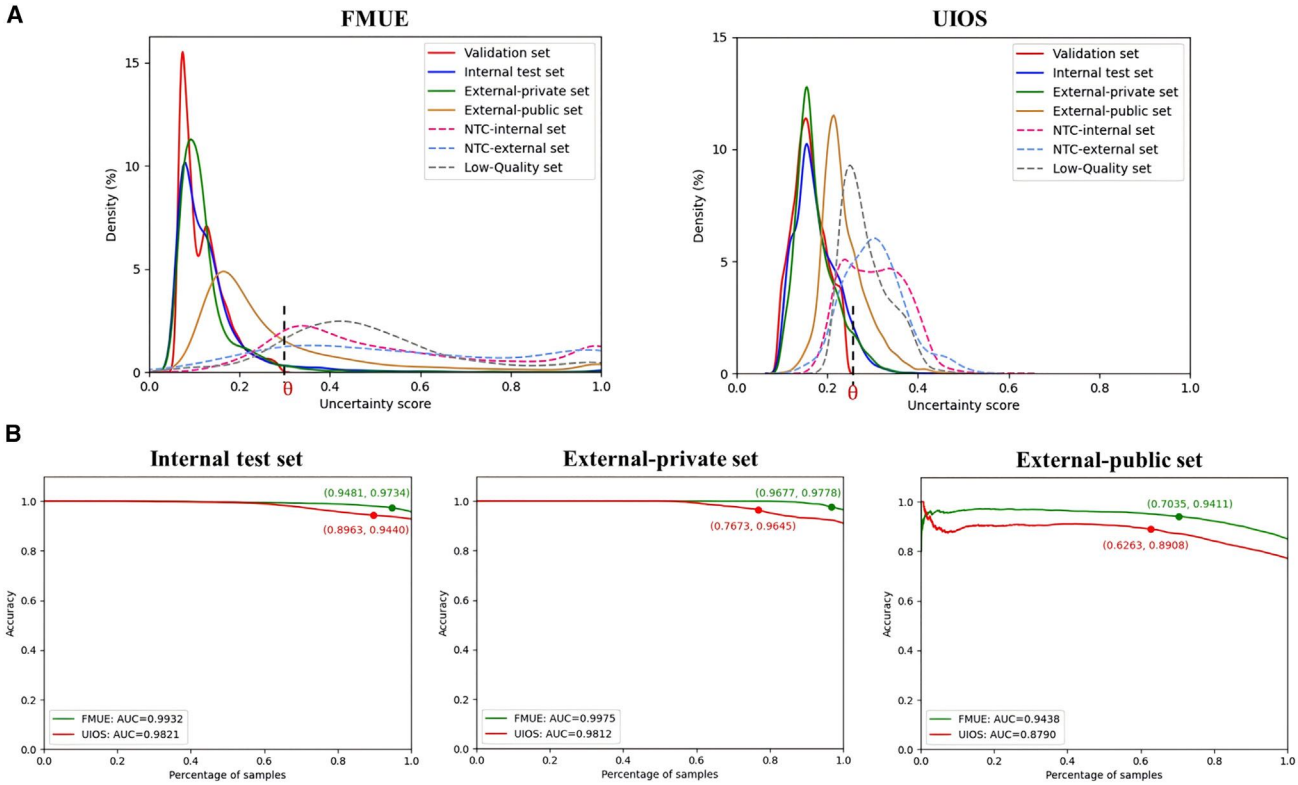


Figure 3. The performance of FMUE and UIOS on different datasets

(A) Uncertainty density distribution for different datasets in FMUE and UIOS. Solid lines indicate validation and test datasets for target categories of retinal diseases, while different colored dashed lines indicate different out-of-distribution datasets. θ, threshold theta; NTCs, non-target categories.

(B) The accuracy of FMUE and UIOS with different percentages of samples remained after excluding the high uncertainty samples on the internal and external test sets. The green and red line curves represent FMUE and UIOS, respectively. The dots on the curves indicate the coordinators of the threshold.

in RETFound and Swin_T. In HMC, FMUE surpassed all ophthalmologists, including experts, and achieved a significantly higher F1 score. The images with high uncertainty scores had a higher risk of misclassification.

The results show that FMUE is superior to RETFound, Swin_T, Ensemble, and UIOS in target disease classification and OOD detection on OCT images. In comparison with RETFound and Swin_T, we integrated uncertainty estimation and had the capability to detect samples with ambiguous features that have higher risk of misclassification. FMUE also has the advantage of detecting low-quality or OOD samples unseen during training. It is worth noting that RETFound performs better than Swin_T, benefiting from its specialized pretraining on large-scale OCT images to learn more relevant medical features. Compared with UIOS, we used a transformer-based foundation model instead of a convolutional neural network (CNN) as the backbone model and fine-tuned it with LoRA, which kept the pretrained weights of the backbone network frozen. Unlike CNN that extracts local features of images through local receptive fields, transformers rely on self-attention mechanisms to capture long-distance global features of the entire image, which is crucial for complex medical images such as OCT. It can be seen from Figure 3A that the density distribution of uncertainty score of UIOS and FMUE on the internal and external-private sets is similar to that on the validation set. FMUE retains more samples after thresholding, and the overall classification performance before and after thresholding is better than that of UIOS. Furthermore, FMUE provides a higher uncertainty score on three OOD datasets and can detect more OOD samples compared to UIOS (Figure 3A, Table S4C). FMUE may benefit from the powerful feature extraction capability of the transformer-based model pretrained on OCT images, even if only a small portion of weights were updated. Furthermore, compared to Ensemble, which generates prediction distributions by training multiple independent DL models on the same input samples and using their mean and variance as the final prediction and uncertainty scores, FMUE directly calculates the belief masses and uncertainty scores of different categories by mapping the features learned by the backbone network to the Dirichlet parameter distribution, thus achieving end-to-end training, easy implementation, and deployment. In addition, it only requires one forward propagation to obtain uncertainty estimates, reducing computational costs. In addition, the performance of our FMUE was far superior to GPT-4V,[12] possibly because GPT-4V has not been fine-tuned on OCT images, which limits its OCT disease classification ability.

The performance of our FMUE exceeded that of all ophthalmologists in identifying retinal diseases from OCT images, as

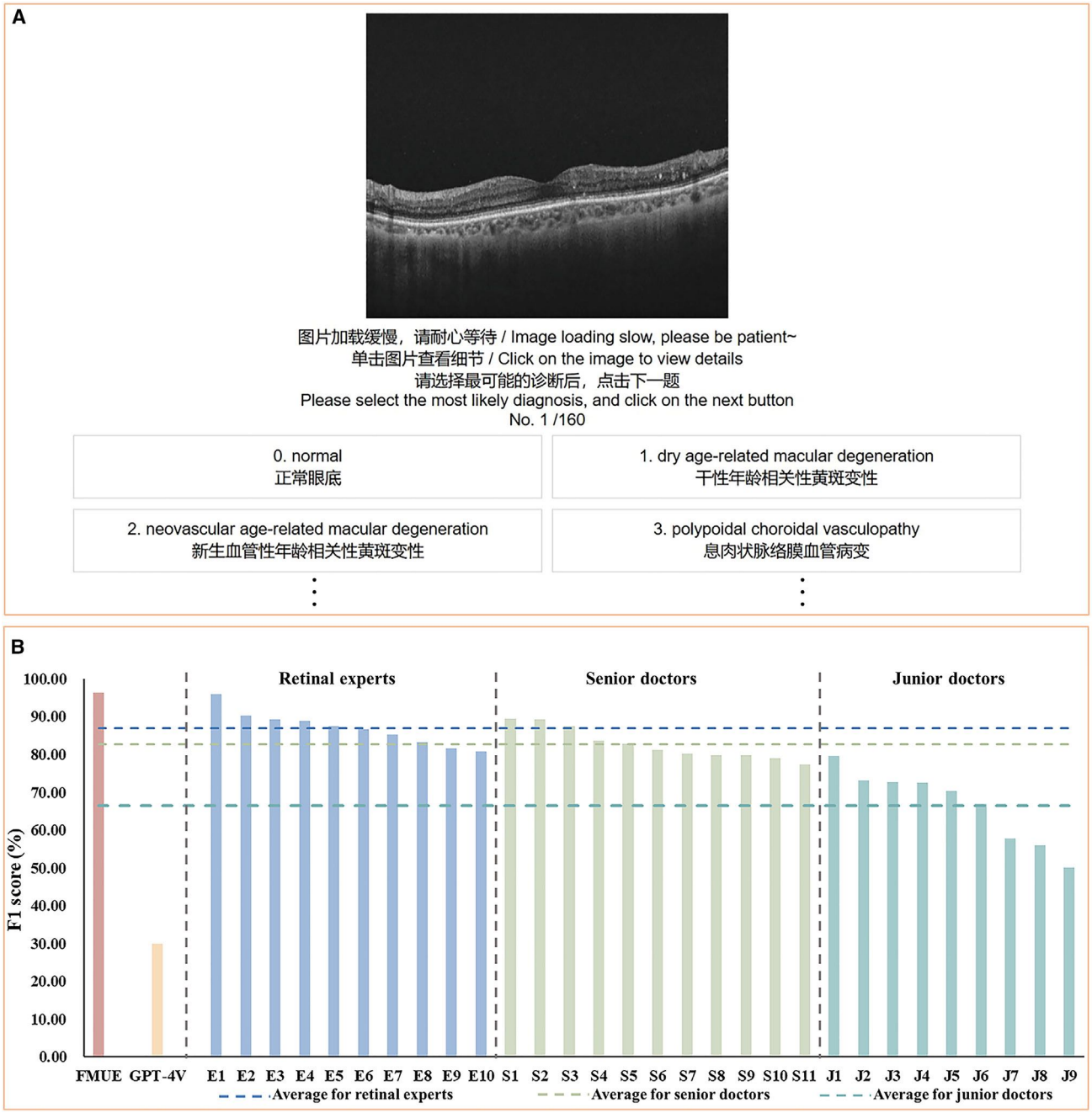


Figure 4. Human-model comparison

(A) Online retinal OCT image reading system that allows ophthalmologists to make the diagnosis based on the image content.

(B) F1 scores achieved by our proposed FMUE model, GPT-4V, and 30 ophthalmologists with varying clinical experience on the HMC set.

shown in Figure 4B. In clinical practice, ophthalmologists usually interpret retinal diseases using 3D volume OCT images.[2] However, evaluating 3D OCT images requires that ophthalmologists be highly focused, and this process is very time-consuming. While making predictions based on 2D OCT images, less experienced doctors often achieve lower diagnostic accuracy. Our study proved that as their clinical experience increased, the diagnostic capability of ophthalmologists in recognizing retinal diseases on OCT images also improved (Figure 4B). Indeed, it is quite difficult for doctors to distinguish some diseases due to similar signs, such as DME and acute RVO, which usually present features like cystoid macular edema, subretinal fluid, and intraretinal hyperreflective foci. Surprisingly, our FMUE exhibited satisfying performance in identifying these two diseases, showing

Table 2. Performance comparison of our model, GPT-4V, and ophthalmologists on the HMC set (%)

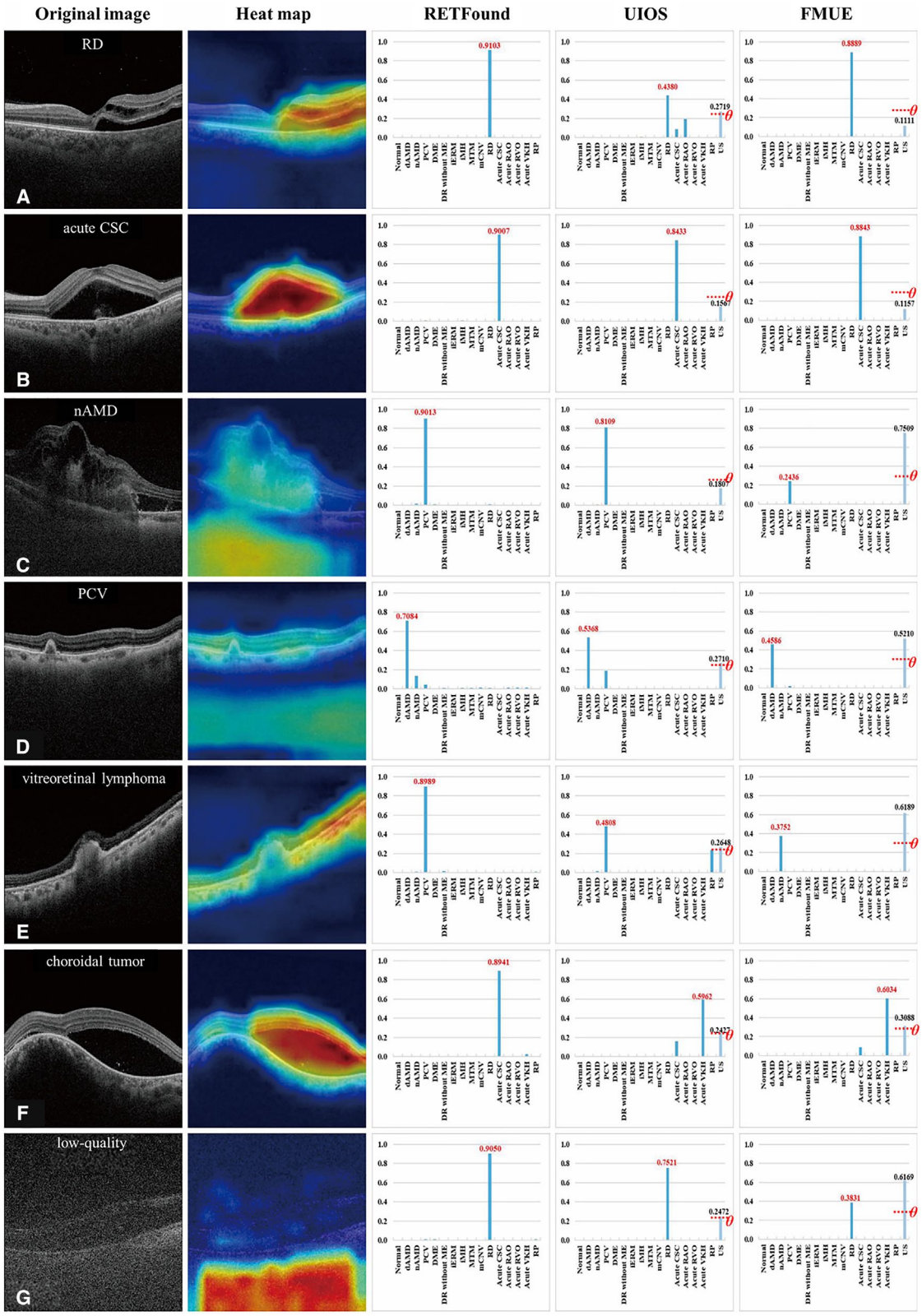
Model/doctors	F1 score	Sensitivity	Precision	Accuracy
FMUE	96.30	96.25	96.57	96.25
GPT-4V	32.39	29.38	20.92	30.00
Retinal experts				
E1	95.88	96.25	96.85	96.25
E2	90.26	90.00	91.39	90.00
E3	89.38	89.38	89.88	89.38
E4	88.89	89.38	92.22	89.38
E5	87.43	87.50	88.89	87.50
E6	86.61	86.88	88.42	86.88
E7	85.33	85.63	88.68	85.63
E8	83.28	83.75	84.07	83.75
E9	81.66	82.50	87.25	82.50
E10	80.77	81.88	84.86	81.88
Average	86.95	87.31	89.25	87.32
Senior doctors				
S1	89.55	90.00	91.98	90.00
S2	89.23	89.38	90.51	89.38
S3	87.48	86.88	90.04	86.88
S4	83.69	83.75	87.42	83.75
S5	82.73	83.13	85.50	83.13
S6	81.13	81.25	82.88	81.25
S7	80.14	81.88	86.00	81.88
S8	79.79	81.25	84.73	81.25
S9	79.75	80.63	81.33	80.63
S10	79.01	79.38	82.81	79.38
S11	77.32	78.75	82.22	78.75
Average	82.71	83.30	85.95	83.30
Junior doctors				
J1	79.60	80.00	84.32	80.00
J2	73.06	74.38	74.90	74.38
J3	72.68	73.75	77.68	73.75
J4	72.63	73.75	75.40	73.75
J5	70.26	71.25	72.55	71.25
J6	66.85	66.88	71.91	66.88
J7	57.74	60.00	62.35	60.00
J8	55.92	58.13	61.82	58.13
J9	50.17	54.38	49.27	54.38
Average	66.55	68.06	70.02	68.06

Abbreviations: HMC, human-model comparison; FMUE, foundation model with uncertainty estimation; GPT-4V, generative pretrained transformer 4 with vision; E, retinal expert; S, senior doctor; J, junior doctor.

relatively high F1 scores in both the internal and external test sets (Table 1). It is possible that our FMUE model is able to detect some distinguishable features that are often ignored by clinicians, such as the inner retinal hyperreflectivity, macrocystoid spaces, and perilesional hyperreflective foci that are more common in acute RVO cases.[13] Furthermore, our FMUE model achieved fast processing speed and provided visual interpretation of pathologic features (Figure 5). Despite the remarkable performance of our FMUE, it is recommended for use as a second reader due to the presence of samples with high uncertainty scores. As demonstrated in Table S6A, these samples are more prone to misclassification and therefore necessitate double-checking. To ensure diagnostic accuracy, clinicians may seek assistance from more experienced doctors or integrate other imaging modalities or clinical information.

There are a few studies investigating uncertainty in AI models for OCT images. Seebock et al. trained a model with health OCT images and applied it in anomaly detection based on epistemic uncertainty, but this model cannot differentiate different diseases.[14] Liu et al. used uncertainty strategy to enhance the accuracy of boundary localization of 7 retinal biomarkers but did not investigate image classification.[15] Some reports used uncertainty to enhance classification of 3, 4, and 5 categories of OCT images,[16–18] but they did not explore how to detect OOD samples based on uncertainty. Araújo et al. trained an efficient-Net V2-B0 using AMD staging images and used Dirichlet uncertainty estimation methods to detect near (DME, RVO, and Stargardt) and far OOD (CFP), but only achieved low AUC with this model.[19] In the current study, our model is capable of both classifications of 16 common conditions and detection of uncommon diseases unseen during training (NTC) as OOD data using an uncertainty thresholding strategy. To obtain the optimal threshold, Tran et al. gradually excluded uncertain samples with the highest entropy values on the test set and monitored changes in indicators such as AUC, accuracy, and rate of disease in the remaining subjects until these indicators decreased.[20] Unlike this method, our FMUE determines the threshold based on the validation set, which can effectively avoid data leakage in the test set, reduce the risk of overfitting, and only calculate the threshold once, without the need for repeated calculations on the test set, thereby simplifying the calculation process. Therefore, the setting of our study is more applicable to clinical scenarios. Furthermore, our study showed that the samples with uncertainty above the threshold had a 19.187-fold, 32.628-fold, and 9.211-fold higher risk of misclassification in the internal test set, the external-private set, and the external-public set, respectively, which have not been investigated before. These results suggest the effectiveness of uncertainty estimation in enhancing disease detection capabilities. In our method, the uncertainty score and prediction are optimized simultaneously,[9] which may explain the improvement of both the reliability and diagnostic performance with uncertainty estimation.

FMUE can identify low-quality OCT images that are difficult to diagnose due to indistinguishable features through uncertainty estimation. However, this concept of low-quality data is different from images with a low image quality score outputted by OCT devices, which is a quantitative indicator used to evaluate the physical quality of OCT images and is based on 3D images. In our study, we used 2D images for training and testing, and there may be inconsistencies between the 2D image quality and the 3D image quality score. Furthermore, in clinical practice, doctors mainly make diagnoses based on the characteristic features of OCT images, regardless of the image quality score. In some images with low-quality scores, doctors can still make correct



(legend on next page)

diagnoses based on critical features, as shown in Figure S4B (a). Conversely, some images with high-quality scores, such as in Figure S4B (b), may still be difficult for doctors to diagnose. Therefore, the decision to include an image for analysis is not solely based on the image quality score provided by the OCT device, but rather on whether the critical lesions are visible. This inclusion method is more in line with the clinical diagnosis process of doctors. Notably, we treat low-quality images as OOD data, demonstrating that our FMUE can detect abnormal data outside of the training categories and remind doctors to review images that are difficult to diagnose, thereby reducing the risk of misdiagnosis and missed diagnoses.

In conclusion, our FMUE combined with threshold strategy can not only provide reliable diagnostic results for 16 types of retinal diseases and conditions but also detect OOD samples that were not included during training, providing an automatic and trustworthy method for diagnosis of retinal diseases using OCT images in real-world clinical scenario.

Limitations of the study

We acknowledge several limitations in the current study and further studies are needed. Firstly, although our FMUE model can achieve relatively accurate predictions for various retinal diseases, there are still 29.65% of the samples in the external-public test sets exhibiting higher uncertainty than the threshold, requiring manual double-checking by experienced experts (Figure 3A, Table S5). It may be due to the instrument domain gap between the internal and external-public test set. In the next step, we will train more data obtained from different instruments to reduce the need for manual reconfirmation in these devices. In addition, for the data from Topcon devices, we believe that our model does not require additional training, while for data obtained from other devices, additional training is needed to ensure its generalization performance. Secondly, we only investigated the single-label classification, ignoring the issue of other coexisting diseases in the same OCT image. In the next stage, we will collect more multi-label classification data and explore uncertainty estimation methods to achieve reliable multi-label retinal disease detection. Thirdly, our FMUE only used single-mode OCT images and did not consider multi-modality imaging and valuable clinical text data, which will be explored in further investigation. Fourthly, the test datasets used in this study are sufficient in terms of diversity and representativeness, covering patients of different races, genders, and ages. However, our training dataset only includes OCT images of patients from China. In the future, we will continue to explore and incorporate datasets from more races to train our model, ensuring its generalization and robustness. Finally, this study lacked prospective multicenter studies to evaluate the effectiveness of FMUE in the real world. Additionally, although the test set contains a large number of OCT images, it is limited by the small number of eyes, which requires more data from different eyes to validate the performance of the model before clinical application. In the future, we will deploy it on local workstations or cloud platforms with GPUs for prospective clinical validation, involving data collected from more patients from multiple centers.

CONSORTIA

The members of the OCT reading group are Binwei Huang, Chaoxin Zheng, Chuang Jin, Dezhi Zheng, Dingguo Huang, Dongjie Li, Guihua Zhang, Hanfu Wu, Honghe Xia, Hongjie Lin, Huiyu Liang, Jingsheng Yi, Jinqu Huang, Juntao Liu, Man Chen, Qin Zeng, Taiping Li, Weiqi Chen, Xia Huang, Xiaolin Chen, Xixuan Ke, Xulong Liao, Yifan Wang, Yin Huang, Yinglin Cheng, Yinling Zhang, Yongqun Xiong, Yuqiang Huang, Zhenggen Wu, and Zijing Huang.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Haoyu Chen (drchenhaoyu@gmail.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Data from OCTDL are available at https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods.

Data from OCTID are available at https://dataverse.scholarsportal.info/dataverse/OCTID.

Data from Kaggle are available at https://doi.org/10.17632/rscbjbr9sj.3.

Data from ROCC are available at https://rocc.grand-challenge.org.

Data from RETORCH are available at https://retouch.grand-challenge.org.

The confidential medical records data reported in this study cannot be deposited in a public repository. To request access, please contact the lead contact.

Code is available at and data have been deposited at https://github.com/yuanyuanpeng0129/FMUE.

Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

This research was supported by the National Key R&D Program of China (2018 YFA0701700 to H.C. and X.C.), Agency for Science, Technology and Research (A*STAR) Career Development Fund (C222812010 to H.F.), Central Research Fund (“Robust and Trustworthy AI system for Multi-modality Healthcare” to H.F.), the National Nature Science Foundation of China (U20A20170 to X.C.), Shantou Science and Technology Program (190917085269835 to H.C.), Department of Education of Guangdong Province (2024ZDZX2024 to H.C.), and the University Natural Science Research Project of Anhui Province (2022AH040099 to Z.L. and 2023AH052070 to Y.P.).

Figure 5. The visualization results of FMUE by Grad-CAM and the detection results of seven samples of OCT images with RETFound, UIOS, and our FMUE

(A and B) Samples with typical features of target diseases.

(C and D) Samples with ambiguous features of target diseases.

(E–G) OOD samples that are not included in the training category. Unlike RETFound, UIOS and FMUE provide prediction results and the corresponding uncertainty score to reflect the reliability of the prediction results. θ, threshold theta.

AUTHOR CONTRIBUTIONS

Y.P.: conceptualization, methodology, data collection, experimental deployment, software, and writing – original draft; A.L.: data collection & annotation & curation and review & editing; M.W.: experimental deployment and review and editing, methodology, and writing – review and editing; T.L. and T.S.: data collection & annotation & curation; L.L., J.W., T.L., D.L., S.Y., D.S., J.L., and L.G.: data collection; K.Z.: experimental deployment and review and editing; L.F. and C.-Y.C.: manuscript revision and clinical support; Z.L.: experimental deployment and project administration; X.C.: supervision, project administration, methodology, and OCT reading group: online retinal OCT image reading; H.F.: supervision, project administration, methodology, and writing – review and editing; H.C.: supervision, data collection & annotation & curation, clinical support, and writing – review and editing.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

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

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Ethical approval
  - Target categories OCT datasets
  - OOD datasets
- METHOD DETAILS
  - Model development
  - Loss function
  - Implementation details
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Human-model comparison
  - Interpretation
  - Statistical analysis

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101876.

Received: July 29, 2024
Revised: October 5, 2024
Accepted: November 25, 2024
Published: December 19, 2024

REFERENCES

1. Pascolini, D., and Mariotti, S.P. (2012). Global estimates of visual impairment: 2010. Br. J. Ophthalmol. *96*, 614–618. https://doi.org/10.1136/bjophthalmol-2011-300539.
2. Lin, A., Mai, X., Lin, T., Jiang, Z., Wang, Z., Chen, L., and Chen, H. (2022). Research trends and hotspots of retinal optical coherence tomography: a 31-Year bibliometric analysis. J. Clin. Med. *11*, 5604. https://doi.org/10.3390/jcm11195604.
3. Lin, A., Xia, H., Zhang, A., Liu, X., and Chen, H. (2022). Vitreomacular interface disorders in proliferative diabetic retinopathy: an optical coherence tomography study. J. Clin. Med. *11*, 3266. https://doi.org/10.3390/jcm11123266.
4. Wu, Q., Zhang, B., Hu, Y., Liu, B., Cao, D., Yang, D., Peng, Q., Zhong, P., Zeng, X., Xiao, Y., et al. (2021). Detection of morphologic patterns of diabetic macular edema using a deep learning approach based on optical coherence tomography images. Retina *41*, 1110–1117. https://doi.org/10.1097/IAE.0000000000002992.
5. Shen, E., Wang, Z., Lin, T., Meng, Q., Zhu, W., Shi, F., Chen, X., Chen, H., and Xiang, D. (2024). DRFNet: a deep radiomic fusion network for nAMD/PCV differentiation in OCT images. Phys. Med. Biol. *69*, 075012. https://doi.org/10.1088/1361-6560/ad2ca0.
6. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C.S., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al. (2018). Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell *172*, 1122–1131.e9. https://doi.org/10.1016/j.cell.2018.02.010.
7. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al. (2018). Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat. Med. *24*, 1342–1350. https://doi.org/10.1038/s41591-018-0107-6.
8. Zhou, Y., Chia, M.A., Wagner, S.K., Ayhan, M.S., Williamson, D.J., Struyven, R.R., Liu, T., Xu, M., Lozano, M.G., Woodward-Court, P., et al. (2023). A foundation model for generalizable disease detection from retinal images. Nature *622*, 156–163. https://doi.org/10.1038/s41586-023-06555-x.
9. Wang, M., Lin, T., Wang, L., Lin, A., Zou, K., Xu, X., Zhou, Y., Peng, Y., Meng, Q., Qian, Y., et al. (2023). Uncertainty-inspired open set learning for retinal anomaly identification. Nat. Commun. *14*, 6757. https://doi.org/10.1038/s41467-023-42444-7.
10. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012–10022. https://doi.org/10.48550/arXiv.2103.14030.
11. Wenzel, F., Snoek, J., Tran, D., and Jenatton, R. (2020). Hyperparameter ensembles for robustness and uncertainty quantification. Adv. Neural Inf. Process. Syst. *33*, 6514–6527.
12. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., and Wang, L. (2023). The dawn of lmms: Preliminary explorations with gpt-4v (ision). Preprint at arXiv. https://doi.org/10.48550/arXiv.2309.17421.
13. Padilla-Pantoja, F.D., Sanchez, Y.D., Quijano-Nieto, B.A., Perdomo, O.J., and Gonzalez, F.A. (2022). Etiology of macular edema defined by deep learning in optical coherence tomography scans. Transl. Vis. Sci. Technol. *11*, 29. https://doi.org/10.1167/tvst.11.9.29.
14. Seeböck, P., Orlando, J.I., Schlegl, T., Waldstein, S.M., Bogunović, H., Klimscha, S., Langs, G., and Schmidt-Erfurth, U. (2019). Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. IEEE Trans. Med. Imag. *39*, 87–98. https://doi.org/10.1109/TMI.2019.2919951.
15. Liu, X., Zhou, K., Yao, J., Wang, M., and Zhang, Y. (2022). Contrastive uncertainty based biomarkers detection in retinal optical coherence tomography images. Phys. Med. Biol. *67*, 245012. https://doi.org/10.1088/1361-6560/aca376.
16. Wang, X., Tang, F., Chen, H., Luo, L., Tang, Z., Ran, A.R., Cheung, C.Y., and Heng, P.A. (2020). UD-MIL: uncertainty-driven deep multiple instance learning for OCT image classification. IEEE J. Biomed. Health Inform. *24*, 3431–3442. https://doi.org/10.1109/JBHI.2020.2983730.
17. Leingang, O., Riedl, S., Mai, J., Reiter, G.S., Faustmann, G., Fuchs, P., Scholl, H.P.N., Sivaprasad, S., Rueckert, D., Lotery, A., et al. (2023). Automated deep learning-based AMD detection and staging in real-world OCT datasets (PINNACLE study report 5). Sci. Rep. *13*, 19545. https://doi.org/10.1038/s41598-023-46626-7.
18. Singh, A., Sengupta, S., Rasheed, M. A., Jayakumar, V., and Lakshminarayanan, V. (2021). Uncertainty aware and explainable diagnosis of retinal disease. In Medical Imaging 2021: Imaging Informatics for Healthcare, Research, and Applications. 11601, 116-125. https://doi.org/10.1117/12.2581362.
19. Araújo, T., Aresta, G., Schmidt-Erfurth, U., and Bogunović, H. (2023). Few-shot out-of-distribution detection for automated screening in retinal OCT

images using deep learning. Sci. Rep. *13*, 16231. https://doi.org/10.1038/s41598-023-43018-9.

20. Tran, A.T., Zeevi, T., Haider, S.P., Abou Karam, G., Berson, E.R., Tharmaseelan, H., Qureshi, A.I., Sanelli, P.C., Werring, D.J., Malhotra, A., et al. (2024). Uncertainty-aware deep-learning model for prediction of supratentorial hematoma expansion from admission non-contrast head computed tomography scan. NPJ Digit. Med. *7*, 26. https://doi.org/10.1038/s41746-024-01007-w.
21. Kulyabin, M., Zhdanov, A., Nikiforova, A., Stepichev, A., Kuznetsova, A., Ronkin, M., Borisov, V., Bogachev, A., Korotkich, S., Constable, P.A., and Maier, A. (2024). Octdl: Optical coherence tomography dataset for image-based deep learning methods. Sci. Data *11*, 365. https://doi.org/10.1038/s41597-024-03182-7.
22. Gholami, P., Roy, P., Parthasarathy, M.K., and Lakshminarayanan, V. (2020). OCTID: Optical coherence tomography image database. Comput. Electr. Eng. *81*, 106532. https://doi.org/10.1016/j.compeleceng.2019.106532.
23. Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., et al. (2019). RETOUCH: The retinal OCT fluid detection and segmentation benchmark and challenge. IEEE Trans. Med. Imag. *38*, 1858–1874. https://doi.org/10.1109/TMI.2019.2901398.
24. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). Lora: Low-rank adaptation of large language models. Preprint at: arXiv. https://doi.org/10.48550/arXiv.2106.09685
25. Li, X., Shen, L., Shen, M., Tan, F., and Qiu, C.S. (2019). Deep learning based early stage diabetic retinopathy detection using optical coherence tomography. Neurocomputing *369*, 134–144. https://doi.org/10.1016/j.neucom.2019.08.079.
26. Russakoff, D.B., Mannil, S.S., Oakley, J.D., Ran, A.R., Cheung, C.Y., Dasari, S., Riyazzuddin, M., Nagaraj, S., Rao, H.L., Chang, D., and Chang, R.T. (2020). A 3D deep learning system for detecting referable glaucoma using full OCT macular cube scans. Transl. Vis. Sci. Technol. *9*, 12. https://doi.org/10.1167/tvst.9.2.12.
27. Sunija, A.P., Kar, S., Gayathri, S., Gopi, V.P., and Palanisamy, P. (2021). Octnet: A lightweight cnn for retinal disease classification from optical coherence tomography images. Comput. Methods Progr. Biomed. *200*, 105877. https://doi.org/10.1016/j.cmpb.2020.105877.
28. Gao, L., and Guan, L. (2023). Interpretability of Machine Learning: Recent Advances and Future Prospects. IEEE MultiMedia *30*, 105–118. https://doi.org/10.1109/MMUL.2023.3272513.
29. Martin, S.A., Townend, F.J., Barkhof, F., and Cole, J.H. (2023). Interpretable machine learning for dementia: a systematic review. Alzheimers Dement. *19*, 2135–2149. https://doi.org/10.1002/alz.12948.
30. 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. In Proceedings of the IEEE international conference on computer vision, 618–626.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
OCTDL	Kulyabin et al.[21]	https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods
OCTID	Gholami et al.[22]	https://dataverse.scholarsportal.info/dataverse/OCTID
Kaggle	Kermany et al.[6]	https://doi.org/10.17632/rscbjbr9sj.3
ROCC	N/A	https://rocc.grand-challenge.org
RETORCH	Bogunović et al.[23]	https://retouch.grand-challenge.org
Software and algorithms		
FMUE	This paper	https://github.com/yuanyuanpeng0129/FMUE
UIOS	Wang et al.[9]	https://github.com/LooKing9218/UIOS
RETFound	Zhou et al.[8]	https://github.com/rmaphoh/RETFound_MAE
Ensemble	Wenzel et al.[11]	https://github.com/google/edward2; https://github.com/google/uncertainty-baselines
Swin_T	Liu et al.[10]	https://github.com/microsoft/Swin-Transformer
Pytorch	N/A	https://pytorch.org/
SPSS	N/A	https://www.ibm.com/spss
Other		
NVIDIA Tesla K40 GPU	NVIDIA Corporation	https://www.nvidia.com/en-us/

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Ethical approval

This study adhered to the tenets of the Declaration of Helsinki and was approved by the Institutional Board of Joint Shantou International Eye Center of Shantou University and the Chinese University of Hong Kong (JSIEC). All the images were deidentified and encrypted to protect the security and privacy of personal information, and the informed consent from patients was waived. All the images were centered in the macula and collected from Joint Shantou International Eye Center using the electronic medical record. In addition, the images scanned at different times during the follow-up were also included.

Target categories OCT datasets

The retinal OCT images of target categories were obtained from two OCT devices: Triton DRI OCT (Topcon, Tokyo, Japan) and 3D OCT-2000 (Topcon, Tokyo, Japan). These datasets included 16 relatively common diseases and conditions: normal, dry AMD (dAMD), neovascular AMD (nAMD), polypoidal choroidal vasculopathy (PCV), DME, diabetic retinopathy without macular edema (DR without ME), idiopathic epiretinal membrane (iERM), idiopathic macular hole (iMH), myopic traction maculopathy (MTM), myopic choroidal neovascularization (mCNV), retinal detachment (RD), acute central serous chorioretinopathy (CSC), acute retinal artery occlusion (RAO), acute RVO, acute VKH, and retinitis pigmentosa (RP). The inclusion criteria of OCT images for these diseases/conditions are listed in Table S7.

Based on the classification criteria, all the retinal OCT images were annotated with 16 diseases or condition with the assistance of fundus photography. Three graders were trained to label the images and required to achieve a high agreement with an expert (Kappa $\geq$0.8) on a set of 100 images randomly selected from the dataset. After certification, the qualified graders performed two rounds of image labeling. In the first round, the masked images were sent to a junior grader to include those images with characteristic OCT features. Images with poor image quality that affected the image analysis and cases with uncertain diagnosis or comorbidity with

other retinal diseases were excluded. In the second round, the images that had passed the first round were independently labeled by two senior graders. In the event of disagreement, an experienced retinal specialist made the final decision (Figure S5).

After two rounds of annotation, a total of 102,468 OCT images from 1,376 eyes of 1,244 subjects were collected and our dataset contains multiple B-scan images from the same patient. Based on the patient-based split policy, the images of each disease/condition were randomly split into the training, validation and test set in the ratio of 6:2:2. The numbers of images in each category within each dataset and the demographic information of the dataset are listed in Table S8A and S8B, respectively.

To further evaluate the generalization ability of our FMUE in detecting retinal diseases, we also conducted experiments on a private dataset from other eye institutes (external-private set) acquired from Triton DRI OCT device and five public datasets obtained from various OCT instruments (external-public set).[6,21–23] For the public datasets, we only included target category samples with characteristic features from the original datasets named TC-OCTDL (OCTDL Data: https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods), TC-OCTID (OCTID Data: https://dataverse.scholarsportal.info/dataverse/OCTID), TC-Kaggle (Kaggle Data: https://doi.org/10.17632/rscbjbr9sj.3), TC-ROCC (ROCC Data: https://rocc.grand-challenge.org), and TC-RETOUCH (RETOUCH Data: https://retouch.grand-challenge.org), respectively. It must be noted that several types of OCT devices were used to obtain these images, including RTVue XR, Cirrus, Spetralis, 3D OCT-1000 and 3D OCT-2000. The numbers of images in each category within each dataset are listed in Table S8C.

OOD datasets

We used two NTC retinal disease datasets and a low-quality OCT image dataset to investigate the ability of FMUE in detecting retinal abnormalities outside the categories of the training set. The first was 3,598 OCT images collected from our clinic with retinal diseases outside the categories of the training set, which were obtained from Triton and 3D OCT-2000 OCT devices in our clinic and called the NTC-internal dataset. The second included 175 images of vitreoretinal lymphoma collected from three foreign institutes, and 31 images of epiretinal membrane with macular hole and vitreomacular traction from the OCTDL dataset (OCTDL Data: https://ieee-dataport.org/documents/octdl-optical-coherence-tomography-dataset-image-based-deep-learning-methods), which were called the NTC-external dataset. It should be noted that the OCT images in the NTC-external dataset were scanned using various types of OCT devices, including RTVue XR, Spectralis, and Cirrus. The low-quality OCT dataset was obtained from Triton OCT device in our clinic. It comprised 793 indistinguishable OCT images, primarily due to severe media opacity, image artifacts, or resolution reduction. The detailed information of NTC-internal and NTC-external is listed in Table S8D.

METHOD DETAILS

Model development

Overview of FMUE

Figure 1 shows the training and inference process of our proposed FMUE framework. In the training stage, we discarded the decoder of RETFound[8] and took its encoder as our backbone network to extract the high-level feature information contained in OCT images, followed by an uncertainty-based classifier to obtain the final prediction result with a corresponding uncertainty score, which was different from the standard AI model only assigning a probability value to each category of retinal disease included in the training set and taking the category with the highest probability value as the final prediction result without any information reflecting the reliability of the final decision. In addition, to effectively adapt the pretrained backbone network to our retinal disease classification task, we introduced a simple and effective adaption strategy, Low-Rank Adaption (LoRA), for model optimization.[24] After model training, we can use fine-tuned FMUE with threshold strategy for real clinical practice work, which takes a B-scan image as input and can generate the final prediction result with an uncertainty score as shown in the second step of Figure 1.

Uncertainty-based classifier

The standard AI model usually uses a Softmax classifier to produce the prediction results based on the features from the backbone network, which assigns a probability value to each category of retinal disease included in the training set, where the category with the highest probability value is used as the final prediction result, without any information reflecting the reliability of the final decision. However, if the standard AI model made incorrect predictions without any risk information prompts, it may bring serious consequences to clinical practice, especially in open-set clinical implementation. Uncertainty estimation can enable the capability of expressing the level of confidence and increase the credibility of AI model prediction results in open-set clinical implementation. Similar to our previous work,[9] an uncertainty-based classifier was achieved by evidential and Dirichlet distribution-based subjective logistic uncertainty theory, which mainly consisted of the following three steps.

(1) Softplus activation function was used to obtain the evidence feature $E = [e_1, e_2, ..., e_K]$ for different retinal diseases:

$$E = \text{Softplus}(F), \quad \text{Equation (1)}$$

where F was the feature from backbone network, K is the number of categories of retinal disease in this study.

(2) Evidence feature E was parameterized as a Dirichlet distribution:

$$\alpha = E + 1, \text{i.e.}, \alpha_k = e_k + 1, \quad \text{Equation (2)}$$

where e_k and α_k are evidence and Dirichlet distribution parameters of the k-th category, respectively. The Dirichlet distribution is an exponential family distribution, and its probability density is defined as follows:

$$Dir(P|\alpha) = \begin{cases} \frac{1}{\beta(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \; for \quad P \in S_k \\ 0 \qquad\qquad Otherwise \end{cases}, \quad \text{Equation (3)}$$

where $\alpha = [\alpha_1, \alpha_2, ..., \alpha_K] > 0$ is a dimensionless distribution parameter, $\beta(\alpha)$ represents the K-dimensional multinomial beta function. S_k is the K-dimensional unit simplex and is defined as follows:

$$S_k = \left\{ P \middle| \sum_{i=1}^{K} p_i = 1 \right\}, 0 \le p_i \le 1, \quad \text{Equation (4)}$$

(3) Calculating the belief masses and corresponding uncertainty score as follows:

$$b_k = \frac{e_k}{S} = \frac{a_k - 1}{S}, \mu = \frac{K}{S}, \quad \text{Equation (5)}$$

$$S = \sum_{k=1}^{K} (e_k + 1) = \sum_{k=1}^{K} \alpha_k, \quad \text{Equation (6)}$$

where S is the Dirichlet intensities.

LoRA-based optimization strategy

Different from the fully fine-tuning training strategy,[8,25–27] the LoRA-based adaption strategy kept the pretrained weights of the backbone network frozen and automatically adjusted the weights between layers in the backbone network to improve model performance and reduce memory consumption and training time. Due to the fact that the multi-head self-attention mechanism determines the region of interest based on cosine similarity, it is wise to apply LoRA to the projection layer of query (q), key (k), or value (v) to influence attention scores. Therefore, we applied the LoRA to query and value projection layers in multi-head attention block of RETFound. Figure S6 showed the LoRA-based optimization strategy in RETFound. Given the input token sequence $F_{in} \in \mathbb{R}^{B,N,C_{in}}$ and the output token sequence $F_{out} \in \mathbb{R}^{B,N,C_{out}}$ obtained from a projection layer $W \in \mathbb{R}^{C_{out},C_{in}}$, LoRA frozen the pretrained weights of RETFound to keep W fixed and injected the trainable rank decomposition matrices into each layer of multi-head attention block in the Transformer architecture by adding a bypass, greatly reducing the number of trainable parameters in our task. For convenience, we refer to this bypass as the LoRA layer, which consisted of two linear layers $A \in \mathbb{R}^{r,C_{in}}$ and $B \in \mathbb{R}^{C_{out},r}$, where $r \ll \min\{C_{in}, C_{out}\}$. It can be seen from Figure S6 that compared to multi-head attention block in RETFound, our FMUE added LoRA layer to the query and value projection layers. Based on LoRA layer, the processing of the updated layer W_{out} can be described as follows:

$$W_{out} = W + \Delta W = W + BA, \quad \text{Equation (7)}$$

$$F_{out} = W_{out} F_{in} \quad \text{Equation (8)}$$

Based on the above analysis, the LoRA-based optimization strategy for multi-head self-attention can be described as follows:

$$Att(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{C_{out}}} + B\right)V, \quad \text{Equation (9)}$$

$$Q = W_{out_q} F_{in} = W_q F_{in} + B_q A_q F_{in}, \quad \text{Equation (10)}$$

$$K = W_q F_{in}, \quad \text{Equation (11)}$$

$$V = W_{out_v} F_{in} = W_v F_{in} + B_v A_v F_{in}, \quad \text{Equation (12)}$$

where W_q, W_k and W_v are the frozen projection layers from RETFound, and A_q, B_q, A_v and B_v are trainable parameters in LoRA.

Uncertainty threshold strategy

In current study, we used the uncertainty distribution of uncertainty score in the validation dataset to obtain the threshold θ. Different from the standard AI classification model, our FMUE can not only provide the final diagnosis result but also obtain an uncertainty

score to indicate the reliability of the diagnosis result in inference stage. As shown in Figure 1, if the uncertainty score is higher than the threshold, a double-check by an experienced grader or ophthalmologist is required. In this scheme, to obtain the optimal threshold, we calculated the Precision-Recall (PR) curve in our validation dataset, all possible precision and recall for the wrong prediction based on ground truth $U' = [u'_1, ..., u'_k, ..., u'_n]$ and corresponding uncertainty score $U = [u_1, ..., u_k, ..., u_n]$ for each image of validation dataset, where n was the total number of samples in our validation dataset and the definition of U' can be described as follows:

$$u'_i = 1 - \mathrm{H}\{P_i, Y_i\}, \quad \text{Equation (13)}$$

where P_i and Y_i were the final prediction result and ground truth of i-th sample in validation dataset, and if $P_i = Y_i$, $\mathrm{H}\{P_i, Y_i\} = 1$, otherwise it is 0. Inspired by F1 score metric based on precision and recall, the threshold in validation dataset was calculated as:

$$T(\theta) = \frac{2 * precision(\theta) * recall(\theta)}{precision(\theta) + recall(\theta)}, \quad \text{Equation (14)}$$

Based on Equation 14, the final optimal threshold value can be obtained by Equation 15, and the optimal threshold θ was 0.2887 in this study.

$$\theta = argmax_\theta T(\theta), \quad \text{Equation (15)}$$

Loss function

As illustrated in Figure 1, our FMUE is an end-to-end deep learning (DL) classification network, which takes the OCT images as input and outputs the final diagnosis result with an uncertainty score to suggest the reliability of prediction result based on the uncertainty-based classifier. The uncertainty-based classifier was achieved by evidential and Dirichlet distribution based subjective logistic uncertainty theory. Based on the above analysis, FMUE combines uncertainty estimation loss with weighted cross-entropy classification loss to simultaneously optimize classification prediction and uncertainty score. Therefore, the loss of our FMUE is divided into two parts: uncertainty loss L_{UN} and weighted cross-entropy loss L_{WCE}. The total loss function combined is defined as follows:

$$L_{total} = L_{UN} + L_{WCE}, \quad \text{Equation (16)}$$

Similar to our previous study,[9] the uncertainty loss is defined as follows:

$$L_{UN} = L_{UN-CE} + \lambda * L_{KL}, \quad \text{Equation (17)}$$

where L_{UN-CE} and L_{KL} were used to ensure that the correct prediction for each sample yielded more evidence than other categories and the incorrect prediction would yield less evidence respectively, and λ was the balance factor that was gradually increased with the number of training epochs.

$$L_{UN-CE} = \sum_{k=1}^{K} y_k (\phi(S_k) - \phi(\alpha_k)), \quad \text{Equation (18)}$$

$$L_{KL} = \log\left(\frac{\Gamma\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)}{\Gamma(K)\prod_{k=1}^{K}\Gamma\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)}\right) + \sum_{k=1}^{K}(\hat{\alpha}_k - 1)\left[\phi(\hat{\alpha}_k) - \phi\left(\sum_{k=1}^{K} \hat{\alpha}_k\right)\right], \quad \text{Equation (19)}$$

where $\phi(\,)$ was the digamma function. $\hat{\alpha} = y + (1 - y) \odot \alpha$ is the adjusted parameter of Dirichlet distribution and $\Gamma(\,)$ is the gamma function.

As can be observed from Table S8A, there was a phenomenon of category imbalance on both datasets. To alleviate this problem, we use a weighted cross-entropy loss function in training process as follows:

$$L_{WCE} = -\sum_{k=1}^{K} w_k y_k \log(b_k), \quad \text{Equation (20)}$$

where b_k was the belief mass for k-th class. $w_k = \frac{\max(N)}{N(k)}$ was the weight of k-th class where $N(k)$ and $\max(N)$ represented the number of k-th class and the maximum number of samples for all classes in training set, respectively.

Implementation details

To reduce the computational cost and improve the computational efficiency of the model, all the images are resized to 256 × 256 by bilinear interpolation and normalized to [0,1]. In addition, online data augmentation, including random crop and resizing the cropped patches to 224 × 224, and random horizontal flipping, is adopted to prevent over-fitting and improve the robust ability of the model. We used LoRA to finetune the frozen query and value projection layers of the multi-head attention blocks, where the rank of LoRA was set to 4. The proposed FMUE was implemented based on the PyTorch platform. We use an NVIDIA Tesla K40 GPU with 12GB

memory to train the model with back-propagation algorithm by minimizing the loss function as shown in Equation 16. The batch size and epoch are set to 16 and 100, respectively. AdamW is used as the optimizer to minimize the loss functions and the first ten epochs are for learning warming up (from 0 to a learning rate of 6.25 × 10e−4), followed by a cosine annealing schedule (from 6.25 × 10e−4 to 3.125 × 10e−6 in the rest of the 90 epochs). During training, we save the model weights with highest AUC on the validation set for final evaluation.

QUANTIFICATION AND STATISTICAL ANALYSIS

Human-model comparison

To comprehensively evaluate the diagnostic ability of our proposed model, the human-model comparison was conducted between our FMUE and ophthalmologists. Thirty clinicians were invited to form the OCT reading group, comprising 10 retinal experts who have accumulated over a decade of experience, 11 senior doctors with experience ranging from 5 to 10 years, and 9 junior doctors with less than 5 years of experience. We created a new subset, the HMC set, comprising a total of 160 images by randomly selecting 10 samples per disease category from the internal test set. To reflect the genuine diagnostic abilities, the clinicians were not given any specific training or informed of the classification criteria of included diseases. During the annotation process, the doctors selected the diagnostic result based on the OCT image content via a customized online image reading system, where 16 disease options were provided as prompts on the webpage (Figure 4A).

Interpretation

DL models are often referred to as "black box" entities and lack interpretation of the results.[28,29] To improve transparency and interpretability, we applied the Grad-CAM technique to aid the interpretation of the results, which can capture the regions in the image that are relevant to the final classification result by calculating the gradient of a certain layer in the deep neural network.[30]

Statistical analysis

To comprehensively and fairly evaluate the classification performance of different methods, accuracy, precision, sensitivity and F1 score were used. In addition, AUC was calculated using the open-source package scikit-learn (version 1.0.2). The uncertainty rate was calculated as the proportion of images with high uncertainty scores among the total samples in a certain dataset. The factors associated with the model uncertainty were investigated using univariate logistic regression analysis. Associations were presented as OR with a 95% CI. Statistical analyses were performed using SPSS software version 19 (SPSS, Inc., Chicago, IL, USA).