Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# Performance and Reliability Assessment of Deep Learning Models for OCT-Based Retinal Disease Classification

**Baohua Li**
Zhengzhou Seventh People's Hospital
**Xinfei Chen**
Henan Academy of Innovations in Medical Science
**Meng Li**
Beijing Anzhen Hospital Affiliated to Capital Medical University
**Kai Wang**
Tianjin Medical University
**Jingnan Li**
Zhengzhou Seventh People's Hospital
**Shu Li**
Zhengzhou Seventh People's Hospital
**Jie Du**
jiedubj@126.com

Henan Academy of Innovations in Medical Science

**Research Article**

**Keywords:** retinal diseases, class imbalance, deep learning, model reliability, image feature space

**Posted Date:** November 11th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11439143/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# Performance and Reliability Assessment of Deep Learning Models for OCT-Based Retinal Disease Classification

Baohua Li[2†], Xinfei Chen[1†], Meng Li[3], Kai Wang[4], Jingnan Li[2], Shu Li[2], Jie Du[1*]

[1] Institute of Biological Therapy, Henan Academy of Innovations in Medical Science, Zhengzhou, China

[2] Department of Ophthalmology, Zhengzhou Seventh People's Hospital, Zhengzhou, 450016, China

[3] Beijing Anzhen Hospital Affiliated to Capital Medical University, Beijing, China

[4] Department of Physiology and Pathophysiology, School of Basic Medical Sciences, Tianjin Medical University, Tianjin, China

[†] Baohua Li, Xinfei Chen contributed equally to this work.

[*] Correspondence: Jie Du jiedubj@126.com

# Performance and Reliability Assessment of Deep Learning Models for OCT-Based Retinal Disease Classification


## Abstract

**Background** Deep learning models have achieved high classification accuracy on OCT images for discriminating among retinal diseases. However, clinical OCT datasets often exhibit substantial class imbalance, and aggregate metrics may obscure unreliable predictions for underrepresented classes. The reliability of such models for clinical decision support therefore remains uncertain. This study aimed to comprehensively assess the reliability of a deep learning-based OCT model under class imbalance.

**Methods** Ten public OCT datasets were combined for a nine-class classification task of normal retina and eight retinal diseases. A Swin Transformer model was developed and evaluated on an internal test set, two independent public test sets (DUKE and OCTID), and a prospective clinical OCT cohort. Class-specific performance and bootstrap-derived 95% confidence intervals were analyzed alongside conventional metrics. Expected calibration error (ECE) assessed the reliability of predicted probabilities. LayerCAM was used to examine model attention to disease-related structures, and UMAP to visualized the deep feature space.

**Results** On the internal test set, the model achieved a macro-AUC of 0.9992 and macro-recall of 0.9886. Despite high recall for some underrepresented classes, bootstrap estimates of precision and F1-score varied more widely, underscoring the

need to consider sample size and uncertainty. Macro-F1 scores were 0.9884 on DUKE and 0.8944 on OCTID; recall for the underrepresented CSC class on OCTID was 0.7863. Pre-calibration ECE was 0.0048, indicating good agreement between predicted probabilities and observed outcomes. In the clinical cohort, macro-F1 was 0.8813 and macro-recall was 0.8365. LayerCAM activations corresponded spatially to disease-related retinal structures in many cases. UMAP showed class-specific clustering, with some overlap between classes.

**Conclusion** This study assessed the reliability of deep learning-based OCT models for retinal disease discrimination under class imbalance. Overall performance alone is insufficient to characterize model reliability in multi-disease clinical settings. Class-specific stability, uncertainty, probability calibration, and interpretability also warrant assessment. These findings support a risk-aware decision-support strategy in which model outputs are interpreted together with class representation and uncertainty, and uncertain or underrepresented cases are flagged for specialist review. Prospective human–AI studies are needed before clinical deployment.

**Keywords** retinal diseases; class imbalance; deep learning; model reliability; image feature space

## Introduction

Retinal diseases are major causes of visual impairment, and conditions such as diabetic retinopathy contribute substantially to the global disease burden [1, 2]. Optical coherence tomography (OCT) is widely used in ophthalmic practice because

it is noninvasive and provides high-resolution images of retinal microstructure [3, 4]. In clinical settings, OCT interpretation often requires distinguishing normal-appearing retinas from several diseases, rather than identifying a single condition. As examination volumes grow and the range of conditions encountered becomes more complex, artificial intelligence (AI)-assisted methods that can reliably distinguish among retinal diseases have become an important area of medical image research [5].

Deep learning models have shown strong performance in OCT-based classification of common retinal diseases, including age-related macular degeneration (AMD) and diabetic retinopathy (DR) [6-8]. Clinical datasets, however, are often class-imbalanced. Normal images and images of common diseases may dominate, while other categories have few samples because of differences in clinical frequency, data availability, or study conditions. This imbalance can affect class-specific performance, bias predictions, and impair probability calibration for underrepresented classes—disease categories with limited representation in the datasets [9].

The clinical value of an AI model depends not only on its classification performance but also on whether its confidence reflects its accuracy and whether it identifies uncertain or out-of-distribution cases [10, 11]. Estimates for underrepresented classes are especially sensitive to sampling variability, making any single metric an incomplete measure of model behavior. A fuller evaluation should therefore consider class-specific performance, uncertainty [12], probability calibration, and interpretability.

Despite advances in OCT-based AI, systematic assessments of model reliability

under class imbalance remain limited. Overall classification metrics can obscure differences between disease categories and uncertainty in class-specific estimates. The reliability of predicted probabilities and variation in performance across data sources also warrant closer examination [13]. Differences in OCT devices, scanning protocols, and patient populations can shift data distributions and reduce performance in external settings. Accordingly, evaluating model performance stability and predictive reliability across multiple dimensions under conditions that closely reflect real-world clinical practice is important for determining the application potential of AI-assisted diagnostic systems [14]. Equally important is defining the conditions under which such systems can be safely relied upon in clinical practice and when their outputs should be referred for human review.

In this study, we evaluated the performance and reliability of a deep learning OCT classifier under class imbalance. We combined ten public OCT datasets into a nine-class task comprising normal-appearing retinas and eight retinal diseases, then developed a model based on the Swin Transformer (Swin-T). Its hierarchical feature representations and shifted-window attention capture features at multiple scales while preserving local structure [15]. We evaluated the model on an internal test set, two independent external public test sets, and an independent clinical OCT cohort. Its stability and reliability across disease categories and data sources were assessed from multiple perspectives, including class-level classification performance and associated uncertainty, probability calibration, model interpretability, and distributions in the deep feature space. The overall study workflow is illustrated in Fig. 1.

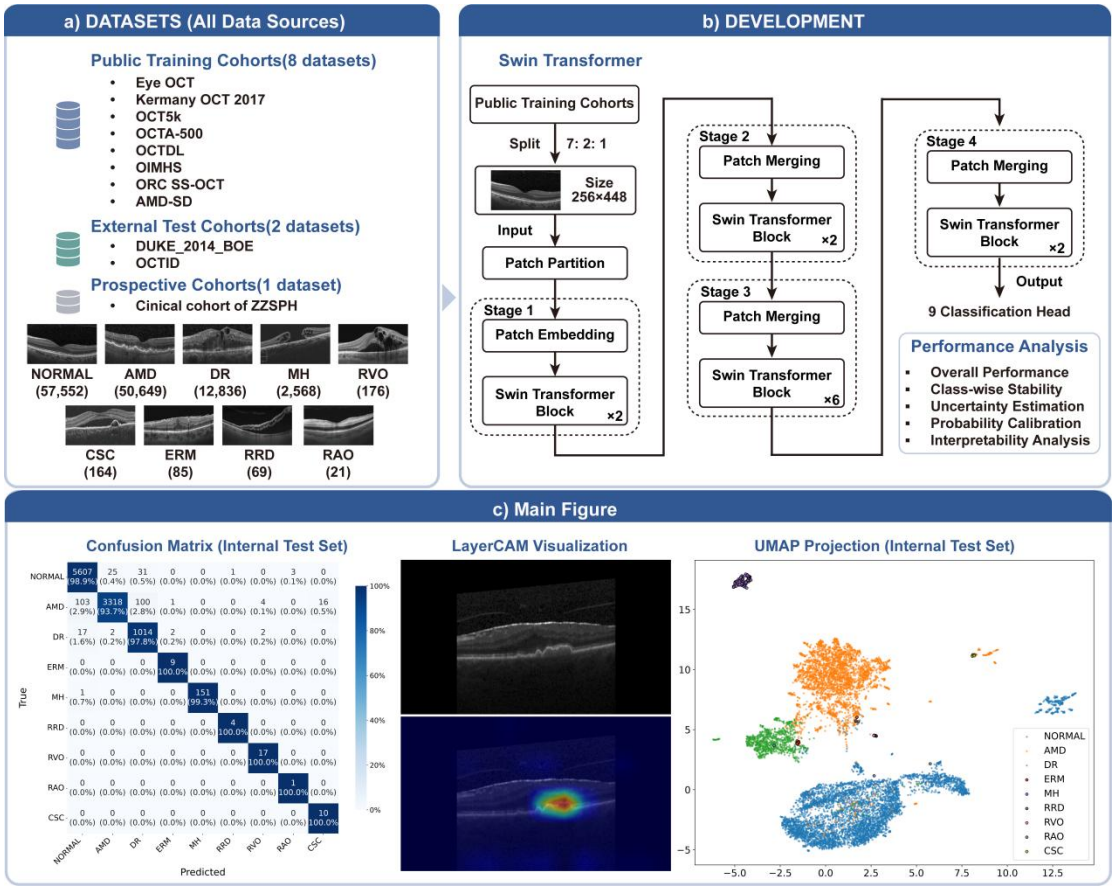



**Fig. 1** Overall flowchart. The flowchart illustrates the complete workflow of this study. a) multi-source OCT data integration, patient-level dataset partitioning; b) Swin-T development; c) Reliability evaluation, including the nine-class confusion matrix, LayerCAM interpretability analysis, and UMAP feature characterization.

## Materials and methods

We integrated multiple publicly available OCT datasets, mapped their labels to a unified disease classification scheme, and performed quality control to define a nine-class task comprising normal-appearing retinas and eight retinal disease categories. A Swin-T-based classifier was then developed, and class-weighted sampling was used to mitigate bias arising from class imbalance in the training data.

The model was evaluated on an internal test set, two independent external public test sets, and an independent clinical OCT cohort. In addition to conventional classification metrics, we assessed prediction reliability by examining class-specific

performance stability, uncertainty estimates, probability calibration, interpretability, and deep feature representations. These analyses characterized the model's performance and potential limitations across disease categories and data sources.

## Datasets

Publicly available OCT datasets were screened against predefined criteria. Eligible datasets contained clinically annotated retinal OCT images with clearly defined disease labels mappable to a common classification scheme and images of sufficient quality for model analysis. We excluded images or records with missing labels, ambiguous classifications, corrupted files, or clear duplicates. The included datasets spanned different institutions, geographic regions, and OCT device platforms, supporting model development and cross-source evaluation. Dataset details are provided in Supplementary Table S1.

Because diagnostic terminology varied across datasets, we established uniform label-mapping rules before model development. Based on the original diagnoses, CNV and DRUSEN were mapped to AMD, and diabetic macular edema (DME) was mapped to DR. Samples that could not be mapped confidently or had uncertain diagnostic definitions were excluded. The resulting nine classes were normal-appearing fundi (NORMAL), AMD, DR, macular hole (MH), retinal vein occlusion (RVO), central serous chorioretinopathy (CSC), epiretinal membrane (ERM), rhegmatogenous retinal detachment (RRD), and retinal artery occlusion (RAO).

The study included 124,120 images from ten public OCT datasets. OCTID and DUKE served as independent external public test sets; the other eight datasets were used for model development. Where patient identifiers were available, development data were split at the patient level and then divided into training, validation, and internal test sets at a 7:2:1 ratio, ensuring that images from the same patient remained in a single subset.

To assess generalizability in clinical practice, we also evaluated the model in an independent OCT cohort prospectively enrolled at Zhengzhou Seventh People's Hospital (ZZSPH). Consecutive eligible patients were recruited under a pre-approved protocol. These images were collected independently of model development and were not used for training, validation, or parameter optimization, thereby minimizing information leakage. All clinical images underwent standardized preprocessing and were screened for quality using clinically interpretable criteria. Written informed consent was obtained from all participants.

**Development**

All OCT images underwent standardized preprocessing, including correction of retinal structural orientation, cropping of the valid image region, proportional resizing followed by padding to 256 × 448 pixels, and normalization of pixel intensities to [0, 1]. To address class imbalance in the training set, we used class-weighted sampling to adjust each class's sampling probability during training; weights for the underrepresented ERM and RVO classes were set to 2.0. Random augmentation was

applied only during training and included geometric transformations, intensity perturbations, and noise simulation. Validation and test images underwent preprocessing without random augmentation.

We used Swin-T as the primary classifier. Its hierarchical feature extraction and shifted-window self-attention capture local retinal structures and broader contextual information. Preprocessed images were passed through the backbone to extract multiscale features, and the classification head produced probabilities for the nine disease categories. The model was initialized with ImageNet-pretrained weights and fine-tuned on the OCT datasets using transfer learning.

For architectural comparison, we included ResNet-50, DenseNet-121, ConvNeXt-Tiny, and Vision Transformer-Small (ViT) as baseline models. We also evaluated a Swin-T variant without the shifted-window mechanism to assess the contribution of window shifting to classification performance. All models were trained with multiclass cross-entropy loss. Detailed training parameters are provided in Supplementary Table S2.

**Assessment**

We first evaluated overall discriminative ability using classification metrics, then systematically assessed predictive reliability through class-specific stability, uncertainty, probability calibration, interpretability, and feature representation analyses.

## Classification performance assessment

We assessed performance using conventional classification metrics, including accuracy, macro-precision, macro-recall, macro-specificity, macro-F1 score, and macro-AUC. Because class sizes varied substantially, we emphasized macro-averaged metrics to reduce the influence of high-frequency classes. Multiclass AUC was calculated using a one-vs-rest approach. Confusion matrices were used to characterize misclassification patterns across disease categories.

## Stability assessment for underrepresented categories

Performance estimates may be unstable when a class has few test samples. To quantify this uncertainty, we generated 2,000 bootstrap samples by resampling the test set with replacement; each sample matched the original test set in size. We calculated 95% confidence intervals from the bootstrap distribution. We assessed class-specific stability in relation to the number of samples in each category, the range of metric variation, and confidence interval width, with particular attention to underrepresented classes.

## Probability calibration assessment

Expected calibration error (ECE) was calculated from the model output probabilities to assess the agreement between predicted probabilities and observed classification outcomes. The raw predicted probability for class $k$ was calculated using the model logits according to Equation 1.

$$\hat{p}_k = \frac{\exp(z_k)}{\sum_{i=1}^{C} \exp(z_i)} \quad \text{(Equation 1)}$$

The ECE was calculated according to Equation 2. Here, $B$ denotes the number of equal-width probability bins, $N_b$ denotes the number of samples in bin $B$, $conf(b)$ denotes the mean predicted confidence in bin $b$, and $acc(b)$ denotes the observed accuracy in bin $b$.

$$ECE = \sum_{b=1}^{B} \frac{n_b}{N} \left| acc(b) - conf(b) \right| \quad \text{(Equation 2)}$$

Three post hoc calibration methods were compared: temperature scaling, histogram binning, and isotonic regression. Temperature scaling introduced a global temperature parameter $T$ where $T>0$, to rescale the logits for all classes simultaneously. The calibrated probabilities were calculated according to the corresponding temperature-scaling equation.

$$\hat{p}_k^{TS} = \frac{\exp(z_k / T)}{\sum_{i=1}^{C} \exp(z_i / T)} \quad \text{(Equation 3)}$$

The optimal temperature parameter ($T^*$) was estimated by minimizing the multiclass negative log-likelihood on the validation set. Here, $\hat{P}_k^{TS}$ denotes the temperature-scaled probability for class $k$; $T$ is the temperature parameter shared by all classes; $N_{val}$ is the number of validation samples; and $y_n$ is the true class index of the $n$ th validation sample.

$$T^* = \arg\min_{T>0} \frac{1}{N_{val}} \sum_{n=1}^{N_{val}} \left[ -\log \hat{p}_{y_n}^{TS} \right] \quad \text{(Equation 4)}$$

Histogram binning divided the probability interval $[0,1]$ into $B$ contiguous bins of equal

width. The raw predicted confidence in each bin was mapped to the corresponding empirical event rate estimated from the validation set. The bin boundaries and mapping values were estimated using the validation set and then fixed before evaluation on the test set.

$$\hat{p}^{HB}(\hat{p}) = \sum_{b=1}^{B} \bar{y}_b \cdot \mathrm{II}(\hat{p} \in [a_{b-1}, a_b)) \quad \text{(Equation 5)}$$

where $\hat{p}^{HB}$ denotes the histogram binning calibration function; $\hat{p}$ is the raw predicted confidence to be calibrated; $B = 10$ is the number of bins; $a_0, a_1, ..., a_B$ are the bin boundaries, satisfying $a_0 = 0$ and $a_B = 1$, with all bins of equal width; $\bar{y}_b$ is the true positive rate of validation samples in the $b$ bin; $\mathrm{II}(\cdot)$ is the indicator function, which takes the value 1 when the condition inside the parentheses holds, and 0 otherwise. Once the bin boundaries and mapping values are determined on the validation set, they remain fixed.

Isotonic regression calibrated the predicted probabilities by learning a monotonically nondecreasing mapping function. A separate mapping function was fitted for each class. Each mapping function was estimated on the validation set by solving a monotonic constrained least-squares problem.

$$\hat{p}_k^{ISO} = f_k(\hat{p}_k) \quad \text{(Equation 6)}$$

The mapping function $f_k$ is obtained by solving the following monotonically constrained least-squares problem on the validation set:

$$f_k^* = \arg\min_{f_k} \sum_{i=1}^{N_{val}} (y_{k,i} - f_k(\hat{p}_{k,i}))^2 \quad \text{(Equation 7)}$$

Here, $\hat{p}_{k,i}$ denotes the isotonic-regression-calibrated probability for class $k$, $f_k$ is the

monotonically nondecreasing mapping function for classs $k$; $\hat{p}_{k,i}$ is the raw predicted probability for class $k$ in validation sample $i$, and $y_{k,i}$ is the corresponding one-hot class label.

All calibration methods were fitted using the validation set only, whereas the test set was reserved exclusively for final evaluation. Macro-average top-$k$ recall was also calculated to evaluate the model's ability to rank candidate classes and provide a candidate disease set for AI-assisted discrimination.

**Decision interpretability assessment**

LayerCAM was used to generate class-specific activation heatmaps and assess the spatial correspondence between the model's activated regions and disease-related OCT structures. For the Swin-T, the first normalization layer, norm1, of the last block in Stage 4 of the backbone was selected as the target feature layer for LayerCAM. The resulting feature map had a spatial size of 8 × 14. An ophthalmologist qualitatively reviewed the heatmaps to determine whether activations coincided with retinal abnormalities or instead focused on background regions, scale bars, or other disease-unrelated structures.

**Feature representation analysis**

To examine the distribution of image features learned for different retinal diseases, the activation vector immediately preceding the prediction layer was extracted as the feature embedding and projected into a two-dimensional space using Uniform

Manifold Approximation and Projection (UMAP). This analysis was used to examine feature clustering patterns and latent similarities among disease categories, as well as the distribution of underrepresented classes in the learned feature space. It provided supporting evidence for interpreting the model's internal representations of different retinal diseases. This analysis was not treated as an independent classification performance metric and was used only to visualize and interpret the learned feature space.

### Statistical analysis

Model performance evaluation and uncertainty analysis were based on the final predictions generated for the test set. Bootstrap resampling was performed for 2,000 iterations to calculate 95% confidence intervals for the primary evaluation metrics. No training-phase data were used in the test performance analyses, thereby maintaining the independence of the performance estimates.

## Results

### Baseline information on the datasets

A total of 10 publicly available OCT datasets comprising 124,120 OCT images were included in this study. The datasets covered normal-appearing fundi and eight categories of retinal disease (Table 1). Sample sizes differed substantially across disease categories. NORMAL, AMD, and DR accounted for most of the images, whereas RAO, RRD, ERM, CSC, and RVO had relatively few samples, indicating a

typical class-imbalanced distribution.

The internal test set comprised 10,439 OCT images obtained through patient-level splitting of the eight development datasets. It was not used for model training, parameter optimization, or model selection. The two independent external public test sets, DUKE and OCTID, comprised 441 and 428 OCT images, respectively. An independent clinical OCT test cohort comprising 312 images was also included to evaluate model performance in a real-world outpatient setting. The composition and class distribution of each dataset are summarized in Supplementary Table S3.

**Table 1** Sample Distribution and Sources by Disease Category

| Disease | Sample Size (Images) | Proportion (%) | Primary Dataset Source |
|---|---|---|---|
| NORMAL | 57,552 | 46.37 | Eye OCT, Kermany OCT2017, OCT5k, OCTA-500, OCTDL, Duke_2014_BOE, OCTID |
| AMD | 50,649 | 40.81 | AMD-SD, Kermany OCT2017, OCT5k, OCTA-500, OCTDL, Duke_2014_BOE, OCTID |
| DR | 12,836 | 10.34 | Kermany OCT2017, OCT5k, OCTA-500, OCTDL, Duke_2014_BOE, OCTID |
| MH | 2,568 | 2.07 | OCTDL, OIMHS, OCTID |
| RVO | 176 | 0.14 | OCTA-500, OCTDL |
| CSC | 164 | 0.13 | OCTA-500, OCTID |
| ERM | 85 | 0.07 | OCTDL |
| RRD | 69 | 0.06 | ORC SS-OCT |
| RAO | 21 | 0.02 | OCTDL |
| Total | 124,120 | 100 | - |

## Model classification performance and baseline comparison

To evaluate the baseline discriminative ability of different deep learning architectures in the nine-class OCT classification task, we compared the performance of Swin-T with that of ResNet-50, DenseNet-121, ConvNeXt-Tiny, and ViT (Table 2). All models were trained using the same data-splitting strategy and training procedure. All models demonstrated high overall classification performance on the internal test set. Swin-T achieved relatively higher macro-AUC and macro-Recall values, whereas some convolutional neural network models and ViT achieved slightly higher macro-F1 scores than Swin-T. The training curves for all models and the validation-set performance of Swin-T are shown in Supplementary Fig. S1. Considering the results across the different metrics, Swin-T was selected as the representative model for subsequent reliability assessment.

An ablation experiment was performed to further assess the effect of the shifted-window mechanism on model performance. On the OCTID external test set, the complete Swin-T model achieved a higher macro-Recall than the Swin-T-no-shift model, in which the shifted-window mechanism had been removed (0.8646 vs. 0.7381). A difference of comparable magnitude was not observed on the DUKE dataset (Supplementary Table S4). These findings suggest that window-based interactions may contribute to feature representation under some cross-source data conditions.

The confusion matrix for the internal test set showed high classification

accuracy, with misclassifications occurring mainly between disease categories with similar OCT imaging features (Fig. 2a). For example, some AMD samples were misclassified as NORMAL or DR. ROC curve analysis indicated good discrimination across all disease categories (Fig. 2b). The average inference time of Swin-T was 36.9 ms per OCT image. Comparisons of computational efficiency are provided in Supplementary Table S5 and Supplementary Fig. S2.

**Table 2** Performance comparison with deep learning models on internal test dataset

| Model | Macro-Precision | Macro-Recall | Macro-F1 | Macro-AUC |
|---|---|---|---|---|
| DenseNet-121 | 0.7155 [0.6797-0.7561] | 0.9692 [0.9651-0.9726] | 0.7791 [0.7483-0.8102] | 0.9949 [0.9942-0.9955] |
| ResNet-50 | 0.8075 [0.7630-0.8585] | 0.9375 [0.8999-0.9713] | 0.8527 [0.8132-0.8909] | 0.9983 [0.9979-0.9987] |
| ConvNeXt | 0.6640 [0.6125-0.7429] | 0.9831 [0.9799-0.9858] | 0.7474 [0.6989-0.8091] | 0.9984 [0.9980-0.9986] |
| ViT | 0.8031 [0.7771-0.8346] | 0.9704 [0.9446-0.9852] | 0.8509 [0.8249-0.8759] | 0.9987 [0.9984-0.9990] |
| Swin-T | 0.7684 [0.7142-0.8412] | 0.9886 [0.9865-0.9904] | 0.8343 [0.7903-0.8917] | 0.9992 [0.9990-0.9994] |

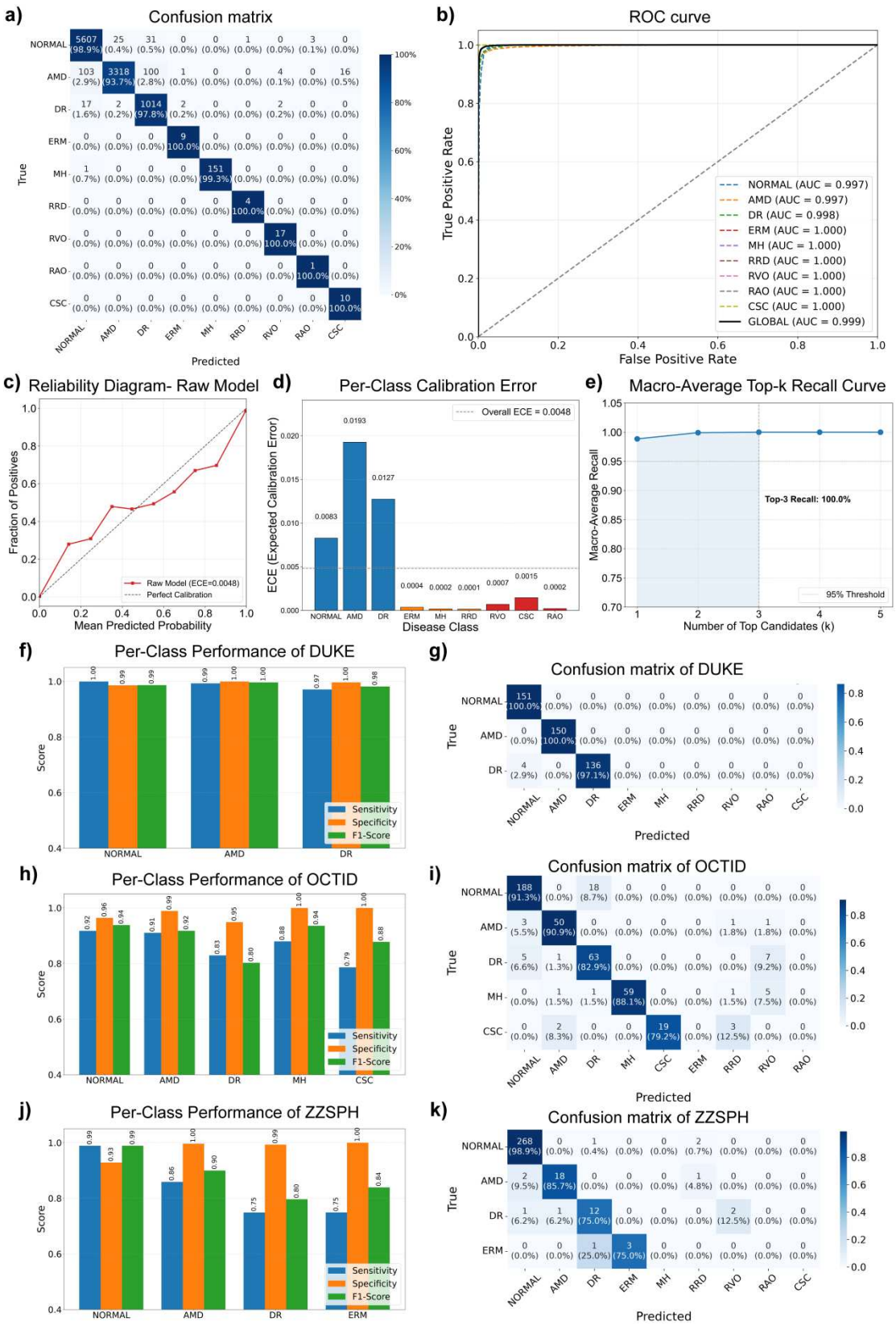



**Fig. 2** Performance and reliability of the Swin-T model on internal and independent cohorts. a) Nine-class confusion matrix on the internal test set. b) One-vs-rest ROC curves (global AUC = 0.999). c) Reliability diagram of the raw model (ECE = 0.0048). d) Per-class ECE across the nine categories. e) Macro-averaged top-k recall. f–g) External validation on DUKE: per-class

performance (f) and confusion matrix (g). h–i) External validation on OCTID: per-class performance (h) and confusion matrix (i). j–k) Prospective validation on ZZSPH: per-class performance (j) and confusion matrix (k).

## Performance stability under class imbalance

### Disease-Specific performance across categories

Although the model achieved high overall classification performance, class-level analysis revealed substantial differences among disease categories (Table 3). The high-sample categories, including NORMAL, AMD, and DR, showed relatively stable performance, whereas the performance metrics of underrepresented classes varied more substantially. In the internal test set, the numbers of samples were particularly limited for RAO (n = 1), RRD (n = 4), ERM (n = 9), CSC (n = 10), and RVO (n = 17). For example, the RAO category achieved a recall and AUC of 1.0000. However, because only one RAO image was included in the test set, this result cannot be interpreted as evidence of stable recognition performance for this category. Similarly, the lower F1 scores observed for some underrepresented classes were driven mainly by reduced precision, because a small number of false-positive predictions can substantially alter proportion-based metrics. These findings indicate that, under severe class imbalance, overall performance metrics may conceal instability in underrepresented classes. Class-level results should therefore be interpreted together with the corresponding sample sizes and estimation uncertainty.

**Table 3** Per-category diagnostic performance on internal test set

| Disease | Recall | Specificity | Precision | F1 Score | AUC |
|---|---|---|---|---|---|
| NORMAL | 0.9894 [0.9868-0.9919] | 0.9746 [0.9702-0.9788] | 0.9789 [0.9753-0.9823] | 0.9841 [0.9818-0.9864] | 0.9974 [0.9966-0.9981] |
| AMD | 0.9368 [0.9286-0.9449] | 0.9961 [0.9945-0.9975] | 0.9919 [0.9887-0.9949] | 0.9636 [0.9590-0.9680] | 0.9970 [0.9961-0.9979] |
| DR | 0.9779 [0.9691-0.9865] | 0.9861 [0.9836-0.9884] | 0.8859 [0.8683-0.9032] | 0.9296 [0.9189-0.9400] | 0.9982 [0.9975-0.9988] |
| ERM | 1.0000 [1.0000-1.0000] | 0.9997 [0.9993-1.0000] | 0.7684 [0.5625-1.0000] | 0.8650 [0.7200-1.0000] | 1.0000 [1.0000-1.0000] |
| MH | 0.9935 [0.9803-1.0000] | 1.0000 [1.0000-1.0000] | 1.0000 [1.0000-1.0000] | 0.9967 [0.9900-1.0000] | 1.0000 [1.0000-1.0000] |
| RRD | 1.0000 [1.0000-1.0000] | 0.9999 [0.9997-1.0000] | 0.8340 [0.5714-1.0000] | 0.9021 [0.7273-1.0000] | 1.0000 [1.0000-1.0000] |
| RVO | 1.0000 [1.0000-1.0000] | 0.9994 [0.9989-0.9998] | 0.7515 [0.6071-0.8947] | 0.8557 [0.7556-0.9444] | 1.0000 [1.0000-1.0000] |
| RAO | 1.0000 [1.0000-1.0000] | 0.9997 [0.9993-1.0000] | 0.3120 [0.1250-1.0000] | 0.4512 [0.2222-1.0000] | 1.0000 [1.0000-1.0000] |
| CSC | 1.0000 [1.0000-1.0000] | 0.9985 [0.9977-0.9991] | 0.3928 [0.2941-0.5263] | 0.5611 [0.4545-0.6897] | 1.0000 [1.0000-1.0000] |
| Global (Macro) | 0.9886 [0.9865-0.9904] | 0.9949 [0.9943-0.9955] | 0.7684 [0.7142-0.8412] | 0.8343 [0.7903-0.8917] | 0.9992 [0.9990-0.9994] |

## Uncertainty in performance estimates

To further quantify uncertainty in the performance estimates under class imbalance, the test-set metrics were evaluated using 2,000 bootstrap resampling iterations, and 95% confidence intervals were calculated (Table 3). Performance estimates for high-sample categories varied relatively little across bootstrap samples, whereas several metrics for underrepresented classes had wider confidence intervals. For example, the F1 score for RAO ranged from 0.2222 to 1.0000 within the 95% confidence interval, whereas the ERM F1 score was 0.8650, with a 95% confidence interval of 0.7200 to 1.0000. Some bootstrap intervals for recall or AUC were

degenerate or otherwise uninformative. For RAO, for example, the internal test set contained only one image, so repeated resampling could not provide meaningful statistical variation. These results should therefore not be interpreted as evidence of stable or perfect recognition performance. They must be considered in the context of the extremely small sample size. Taken together, these results indicate that a single point estimate may be insufficient to characterize model reliability in class-imbalanced OCT classification. Both confidence intervals and sample size are important when interpreting performance in underrepresented classes.

**Prediction probability reliability**

In addition to classification performance, the reliability of the model's predicted probabilities is an important component of overall model reliability. We therefore further evaluated the calibration of the model's predicted probabilities. The mean ECE was 0.0048 before calibration and 0.0047 after temperature scaling, indicating only a small difference between the two conditions. By contrast, histogram binning and isotonic regression did not further reduce the test-set ECE (Fig. 2c and Supplementary Fig. S3). These findings suggest that, under the current sample size and class distribution, nonparametric calibration methods may be limited by the number of validation samples. Consequently, their ability to improve the reliability of the predicted probabilities may be limited.

Class-level calibration analysis also showed differences among disease categories. For high-sample categories such as NORMAL, AMD, and DR, the

probability estimates were supported by more samples, and ECE provided a more stable description of the model's probabilistic behavior. Some underrepresented classes showed low ECE values, but the reliability of these estimates should still be interpreted in light of the limited number of samples. Within the high-confidence interval of 0.9 to 1.0, the observed accuracy for CSC was lower than the predicted confidence, indicating a degree of overconfidence (Fig. 2d). The macro-average top-3 recall was 99.9% (Fig. 2e), indicating that the true category was included among the three highest-ranked predictions for nearly all samples. This result suggests that the model may provide useful candidate information for clinical discrimination.

**External and clinical validation**

To evaluate generalizability across data sources and in a real-world clinical setting, the model was further evaluated on two independent external public test sets and an independent clinical OCT cohort.

**Validation on external public datasets**

The DUKE dataset comprised 441 OCT images from the NORMAL, AMD, and DR categories. The model maintained high and stable classification performance on this dataset (Table 4 and Fig. 2f). The confusion matrix showed that only a small proportion of DR cases were misclassified, accounting for 2.9% of the DR samples (Fig. 2g).

The OCTID dataset comprised 428 OCT images from the NORMAL, AMD, DR,

MH, and CSC categories. The model showed reduced diagnostic performance for DR and CSC on the OCTID dataset (Table 4 and Fig. 2h). Confusion matrix analysis showed that the main errors involved DR being misclassified as RVO and CSC being misclassified as RRD or AMD, which may reflect similarities in their OCT imaging phenotypes (Fig. 2i).

Overall, the external test results indicated that the model generalized to different data sources, although its performance was affected by inter-center differences in data distribution. The model's differential diagnostic performance was lower when disease categories had similar imaging phenotypes.

**Table 4** Performance on independent external validation datasets

| Dataset | Disease | Precision | Recall | F1-score | AUC |
|---|---|---|---|---|---|
| DUKE | NORMAL | 0.9741 [0.9497-0.9934] | 1.0000 [1.0000-1.0000] | 0.9869 [0.9742-0.9967] | 1.0000 [1.0000-1.0000] |
| | AMD | 1.0000 [1.0000-1.0000] | 0.9933 [0.9800-1.0000] | 0.9966 [0.9899-1.0000] | 1.0000 [1.0000-1.0000] |
| | DR | 0.9928 [0.9779-1.0000] | 0.9712 [0.9429-0.9929] | 0.9818 [0.9635-0.9964] | 0.9999 [0.9995-1.0000] |
| DUKE Macro-average | - | 0.9890 [0.9787-0.9978] | 0.9882 [0.9765-0.9976] | 0.9884 [0.9770-0.9977] | 0.9999 [0.9998-1.0000] |
| OCTID | NORMAL | 0.9598 [0.9314-0.9845] | 0.9175 [0.8786-0.9515] | 0.9380 [0.9128-0.9606] | 0.9845 [0.9720-0.9939] |
| | AMD | 0.9262 [0.8596-0.9815] | 0.9104 [0.8364-0.9818] | 0.9176 [0.8598-0.9643] | 0.9904 [0.9773-0.9989] |
| | DR | 0.7795 [0.7045-0.8533] | 0.8293 [0.7368-0.9079] | 0.8028 [0.7355-0.8645] | 0.9396 [0.9033-0.9688] |
| | MH | 1.0000 [1.0000-1.0000] | 0.8797 [0.7910-0.9552] | 0.9355 [0.8833-0.9771] | 0.9970 [0.9930-0.9995] |
| | CSC | 1.0000 [1.0000-1.0000] | 0.7863 [0.6250-0.9167] | 0.8779 [0.7692-0.9565] | 0.9747 [0.9301-1.0000] |
| OCTID Macro-average | - | 0.9331 [0.9105-0.9544] | 0.8646 [0.8189-0.9063] | 0.8944 [0.8598-0.9256] | 0.9772 [0.9644-0.9874] |

## Validation in the clinical cohort

To further evaluate model performance in routine clinical OCT examinations, we assessed the model on an independent clinical test cohort comprising 312 images. The cohort was derived from routine outpatient examinations and included NORMAL, AMD, DR, and ERM images. All OCT images were independently reviewed and annotated by two ophthalmologists. Cases with discrepant interpretations were adjudicated by a third senior ophthalmologist, and the consensus diagnosis was used as the reference standard. In the clinical cohort, the model achieved a macro-F1 score of 0.8813 and a macro-Recall of 0.8365 (Table 5 and Fig. 2j). Performance was most stable for the NORMAL category, whereas some misclassification occurred among AMD, DR, and ERM. The confusion matrix showed that errors in the clinical cohort occurred mainly between disease categories rather than as random classification errors (Fig. 2k). These findings suggest that the model retained disease-related classification ability in the real-world clinical setting, although its performance was affected by differences in disease phenotype and data distribution.

**Table 5** Diagnostic performance of Swin-T model on the prospective ZZSPH clinical cohort

| Disease | Precision | Recall | F1-score | AUC | Support |
|---|---|---|---|---|---|
| NORMAL | 0.9892 [0.9779-1.0000] | 0.9890 [0.9742-1.0000] | 0.9891 [0.9797-0.9963] | 0.9909 [0.9754-0.9996] | 271 |
| AMD | 0.9498 [0.8500-1.0000] | 0.8592 [0.7143-1.0000] | 0.9001 [0.8000-0.9767] | 0.9987 [0.9956-1.0000] | 21 |
| DR | 0.8625 [0.6923-1.0000] | 0.7490 [0.5000-0.9375] | 0.7966 [0.6207-0.9333] | 0.9745 [0.9335-0.9977] | 16 |
| ERM | 1.0000 [1.0000-1.0000] | 0.7490 [0.2500-1.0000] | 0.8393 [0.4000-1.0000] | 0.9493 [0.8417-1.0000] | 4 |
| Macro-average | 0.9503 [0.8985-0.9981] | 0.8365 [0.6996-0.9440] | 0.8813 [0.7571-0.9629] | 0.9783 [0.9500-0.9987] | 312 (Total) |

## Decision interpretability and feature representation

### Decision interpretability

LayerCAM was used to generate class-specific activation heatmaps and identify the image regions associated with the model's predictions. For representative categories, including NORMAL, AMD, DR, MH, CSC, and ERM, the regions with high activation were mainly located in areas associated with retinal structural abnormalities. These regions included macular morphological abnormalities, alterations in retinal layer structure, and areas associated with fluid accumulation (Fig. 3a). For the NORMAL class, activation was concentrated mainly in the foveal region. For AMD, activation was concentrated in areas corresponding to choroidal neovascularization or drusen. For DR, activation covered areas of retinal thickening associated with macular edema. For ERM, activation followed the highly reflective interface of the epiretinal membrane. For MH, activation corresponded to the central

defect of the macular hole. For CSC, activation was concentrated in areas of retinal neuroepithelial detachment. These spatial correspondences suggest that, for high-sample categories, the model learned image features associated with pathological structures rather than relying primarily on the image background or other nondiagnostic regions.

However, approximately 13% of samples in the underrepresented classes showed a spatial shift in the activation regions (Fig. 3b). In some CSC and ERM samples, the activation regions were not precisely aligned with the lesion core and instead extended into surrounding normal retinal layers or toward the image margins. In a small number of misclassified samples, the LayerCAM hotspots did not correspond to the typical anatomical locations of the predicted disease. For example, in an AMD sample misclassified as RVO, activation was concentrated within the right-sided elevated lesion rather than distributed across the overall lesion pattern. These findings indicate that LayerCAM can assist in identifying image regions associated with model predictions but cannot be regarded as direct evidence of the model's underlying decision mechanism. When model predictions are presented to clinicians, the corresponding interpretability information should also be provided.

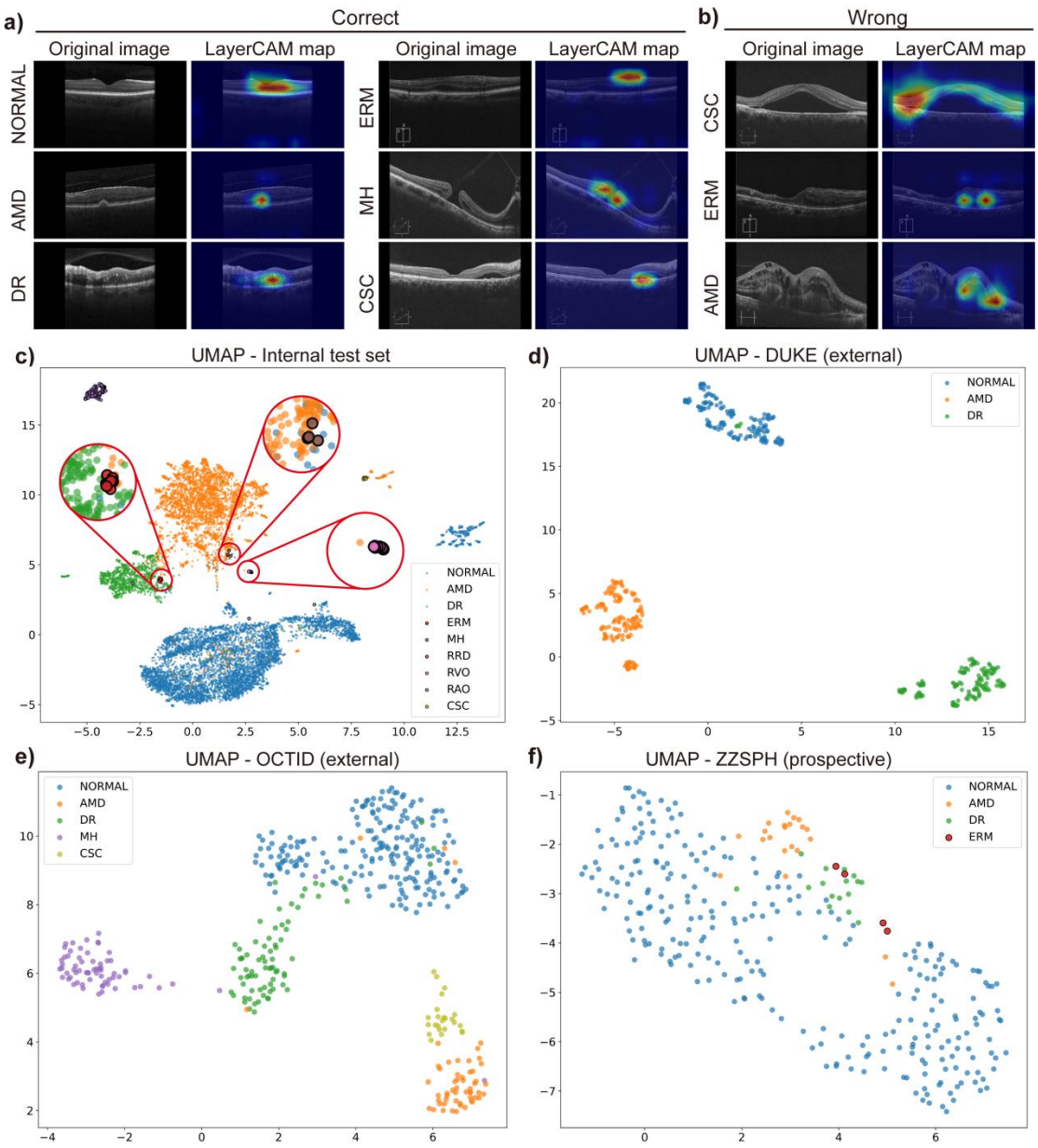



**Fig. 3** Model interpretability and feature-space visualization. a) Representative correctly classified cases with paired OCT B-scans and LayerCAM attention maps, showing that the model attends to clinically relevant retinal lesions. b) Representative misclassified cases, with LayerCAM maps highlighting the discriminatory regions driving the errors. c–f) UMAP projections of Swin-T embeddings on the internal test set (c), the DUKE external set (d), the OCTID external set (e), and the prospective ZZSPH cohort (f).

## Feature representation analysis

We further analyzed the distribution of image features learned by the model for different retinal diseases. The activation vector immediately preceding the model's

prediction layer was extracted as the feature embedding for each image, and UMAP was used to visualize these embeddings in two dimensions.

The UMAP projection of the internal test set showed that different disease categories formed partially distinct clusters in the feature space (Fig. 3c). This suggests that the model learned disease-related image representations with some degree of class separability. However, several underrepresented classes showed substantial feature overlap and spatial dispersion. This finding was consistent with the class-level performance and uncertainty analyses, suggesting that class imbalance may hinder the formation of stable feature distributions for some disease categories. For these categories, the observed feature dispersion may indicate a need for more detailed subtype classification or additional data.

The external public test sets showed different feature distributions (Fig. 3d and e). In the DUKE dataset, NORMAL, AMD, and DR formed relatively compact clusters, with only a small number of DR samples extending into the NORMAL cluster. In the OCTID dataset, the disease categories formed relatively distinct feature clusters. A continuous feature transition was observed between NORMAL and DR, whereas AMD and CSC showed localized feature overlap. These patterns suggest that macular diseases share some imaging features and that substantial lesion heterogeneity exists within disease categories.

In the clinical cohort, NORMAL samples showed a broad and continuous distribution (Fig. 3f). AMD formed a relatively compact subcluster that was generally separable from NORMAL, with only limited overlap at the cluster boundaries.

Because the numbers of DR and ERM samples were limited, their feature points were scattered around the boundaries between categories. Some of these samples overlapped with the AMD features, suggesting shared OCT imaging characteristics among DR, ERM, and AMD that may contribute to classification confusion.

Overall, analysis of the learned image feature space complemented the conventional classification metrics and helped clarify potential sources of performance differences under class imbalance. In the future, such representations may assist physicians in the discrimination of complex disease categories, although their clinical utility requires further validation.

## Discussion

We evaluated the reliability of a Swin-T-based OCT classifier in a nine-class task with substantial class imbalance, using an internal test set, two independent external public datasets, and a prospective clinical cohort. The model showed strong overall discrimination and high recall for some underrepresented classes. Nevertheless, class-specific results and bootstrap analyses showed that estimates for these classes were limited by small sample sizes; a single metric could therefore overstate reliability. Calibration, LayerCAM, and feature-space analyses offered further insight into model behavior, but each provided limited evidence on its own. Together, these findings show why OCT classifiers should be assessed beyond aggregate performance.

Class imbalance is common in medical imaging AI, particularly in clinical

datasets containing multiple disease categories, where the number of samples may differ substantially across classes [16]. We used class-weighted sampling and data augmentation to lessen its effects during training. Although recall was high for some underrepresented classes, high recall alone does not establish clinical reliability. False-positive predictions may still limit precision and increase the need for clinical review. Evaluation under class imbalance should therefore consider precision, recall, F1 score, and uncertainty in their estimates, rather than sensitivity or overall accuracy alone.

Performance estimates for underrepresented classes were also statistically uncertain. When test samples are scarce, class-specific AUC and F1 estimates may vary considerably, while confidence intervals may be wide or degenerate. High recall based on only a few cases, for example, may not reflect performance in a larger clinical population. Reports should present class-specific sample sizes and confidence intervals so that estimates from limited data are not overinterpreted. This is especially important in multi-disease OCT classification, where some diagnostically important conditions have few high-quality annotated examples.

Generalizability is an important consideration when assessing the potential clinical utility of medical AI [17]. We developed the model using multiple public OCT datasets and evaluated it on independent external data. Differences in image acquisition and data sources can affect performance outside the development setting [18]. External performance in our study differed from internal performance and was lower on OCTID, which had a more complex class composition and potentially

558 overlapping image features. Strong internal results alone are therefore insufficient to
559 establish clinical applicability; evaluation across independent data sources is also
560 needed.

561 In addition to classification performance, the reliability of predicted probabilities
562 is an important component of overall model reliability [19]. We used ECE to assess
563 agreement between confidence and observed outcomes and compared post hoc
564 calibration methods. Overall ECE was low, yet some high-confidence predictions for
565 underrepresented classes were incorrect. Aggregate calibration can thus conceal
566 class-specific problems, particularly when larger classes dominate the metric. ECE
567 should be considered alongside class sizes, prediction distributions, and
568 high-confidence errors.

569 Interpretability analyses provided additional, indirect evidence about model
570 behavior. LayerCAM activations corresponded spatially to retinal structures in many
571 cases, while UMAP showed clustering for some classes and overlap for others,
572 including underrepresented classes. Activation maps do not establish that a particular
573 structure caused a prediction, and a two-dimensional projection does not directly
574 measure representation quality. The cited saliency-map study supports caution in
575 interpreting activation maps but does not directly address UMAP [20]. These methods
576 can help examine model behavior, but they cannot replace clinical validation [21].
577 Classification performance, uncertainty, calibration, interpretability, and external
578 generalizability are best considered together.

579 The findings suggest a possible approach to clinical decision support. A system

could present each predicted label with its confidence, alongside relevant class-level reliability estimates, such as calibration error and bootstrap confidence intervals. Predictions for well-represented classes with stable performance, such as NORMAL and AMD, could potentially be handled with less routine review. Cases assigned to underrepresented classes, or classes with elevated calibration error or imprecise performance estimates, could instead be flagged for senior ophthalmologist review. Unusual activation patterns or feature representations might be identified by a dedicated detector or examined by a clinician using LayerCAM, with the physician retaining responsibility for the final decision. None of these workflow options was tested here. Prospective human-AI studies are needed before they can be adopted in practice.

We selected Swin-T as the representative model because its hierarchical features and shifted-window attention capture both local and broader spatial information [22]. It was competitive on some macro-averaged metrics, although precision, recall, and F1 score varied across architectures. Model selection should therefore reflect the needs of the clinical task. In ablation experiments, shifted windows were associated with better performance, most notably on OCTID. Cross-window information exchange may help distinguish conditions with similar local features but different broader structural patterns, although its effect varied across external datasets.

This study has several limitations. First, despite external and prospective clinical-cohort evaluations, model development relied mainly on public data. Further validation across devices, institutions, and patient populations is needed. Second,

some underrepresented classes had few test cases, leaving imprecise performance estimates that require confirmation in larger clinical datasets. Third, our analyses used primarily two-dimensional OCT images and did not fully use the information available in volumetric scans; three-dimensional or 2.5D approaches warrant investigation. Finally, we assessed model outputs rather than human-AI collaboration in routine care. Physician-involved studies are needed to determine effects on workflow efficiency, diagnostic agreement, and clinical decisions.

## Conclusions

The OCT-based Swin Transformer achieved high overall discrimination in a nine-class retinal disease task, but aggregate metrics did not fully characterize its reliability under class imbalance. Small sample sizes limited the precision of estimates for underrepresented classes, while external testing and calibration analyses revealed differences in performance and probability reliability across data sources and classes.

Evaluation of multi-disease OCT classifiers should therefore extend beyond conventional metrics to include class-specific stability and uncertainty, probability calibration, interpretability, and learned feature representations. The class-level reliability indicators examined here may help identify cases that warrant human review. Clinical value, however, remains to be established in larger multicenter cohorts and prospective studies of human–AI collaboration.

## Abbreviations

| Abbreviation | Full name |
|---|---|
| OCT | Optical Coherence Tomography |
| CNN | Convolutional Neural Network |
| Swin-T | Swin Transformer-Tiny |
| NORMAL | Normal Retinal |
| AMD | Age-related Macular Degeneration |
| DR | Diabetic Retinopathy |
| MH | Macular Hole |
| ERM | Epiretinal Membrane |
| RVO | Retinal Vein Occlusion |
| RAO | Retinal Artery Occlusion |
| RRD | Rhegmatogenous Retinal Detachment |
| CSC | Central Serous Chorioretinopathy |
| AUC | Area Under the Receiver-Operating-Characteristic Curve |
| Recall (Sensitivity) | True Positive Rate / Sensitivity |
| ECE | Expected Calibration Error |


## Acknowledgements

Not applicable.

### Authors' contributions

Baohua Li: Conceptualization, Data curation, Investigation, Formal analysis, Writing – original draft. Xinfei Chen: Conceptualization, Data curation, Methodology, Software, Writing – original draft. Meng Li: Methodology, Formal analysis, Investigation, Writing – review & editing. Kai Wang: Data curation, Investigation, Writing. Jingnan Li: Data curation, Investigation. Shu Li: Data curation, Investigation. Jie Du: Supervision, Project administration, Funding acquisition, Writing – review & editing. All authors have read and approved the final manuscript.

### Funding

This study was supported by the Zhengzhou Science and Technology Benefit-People Program (Grant No. 2024KJHM0029).

### Availability of data and materials

The datasets and code analysed during the current study are publicly available as follows. The prospective clinical OCT cohort from Zhengzhou Seventh People's Hospital (ZZSPH) is accessible at https://huggingface.co/spaces/dujie-lab/uncommon-retinal-disease-AI-OCT/tree/main. The ten public OCT datasets used for model development and external validation can be obtained from their respective original repositories, as detailed in Supplementary Table S1. The source code for model training and evaluation is available at https://github.com/dujie-lab/uncommon-retinal-disease-AI-OCT.

## Declarations

### Ethics approval and consent to participate

This prospective study was approved by the Medical Ethics Committee of Zhengzhou Seventh People's Hospital (approval number: 2026-KY-019). The study was conducted in accordance with the Declaration of Helsinki. All participants provided written informed consent prior to enrolment. Patient data used in this study were de-identified prior to analysis.

### Consent for publication

Not applicable. This manuscript does not contain any individual person's data in any form (e.g., identifiable clinical photographs or personal information).

## Competing interests

The authors declare no competing interests.

# References


1. Organization WH: **World report on vision**. In. Geneva: World Health Organization; 2019.
2. Teo ZL, Tham YC, Yu M, Chee ML, Rim TH, Cheung N, Bikbov MM, Wang YX, Tang Y, Lu Y *et al*: **Global Prevalence of Diabetic Retinopathy and Projection of Burden through 2045: Systematic Review and Meta-analysis**. *Ophthalmology* 2021, **128**(11):1580–1591.
3. Zeppieri M, Marsili S, Enaholo ES, Shuaibu AO, Uwagboe N, Salati C, Spadea L, Musa M: **Optical Coherence Tomography (OCT): A Brief Look at the Uses and Technological Evolution of Ophthalmology**. *Medicina (Kaunas)* 2023, **59**(12):2114.
4. Langlo CS, Amin A, Park SS: **Optical coherence tomography retinal imaging: narrative review of technological advancements and clinical applications**. *Ann Transl Med* 2025, **13**(2):17.
5. Zhang J, Zhong J, Lin L, Li J, Ma K, Li N, Li M, Pan Y, Meng Z, Zhou M *et al*: **Full end-to-end diagnostic workflow automation of 3D OCT via foundation model-driven AI for retinal diseases**. *npj Digital Medicine* 2026.
6. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, Askham H, Glorot X, O'Donoghue B, Visentin D *et al*: **Clinically applicable deep learning for diagnosis and referral in retinal disease**. *Nat Med* 2018, **24**(9):1342–1350.
7. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F *et al*: **Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning**. *Cell* 2018, **172**(5):1122–1131 e1129.
8. Laouarem A, Kara-Mohamed C, Bourennane EB, Hamdi-Cherif A: **HTC-retina: A hybrid retinal diseases classification model using transformer-Convolutional Neural Network from optical coherence tomography images**. *Comput Biol Med* 2024, **178**:108726.
9. Pan L, Zhang Y, Yang Q, Li T, Chen Z: **Long-tailed medical diagnosis with relation-aware representation learning and iterative classifier calibration**. *Comput Biol Med* 2025, **188**:109772.
10. Peng Y, Lin A, Wang M, Lin T, Liu L, Wu J, Zou K, Shi T, Feng L, Liang Z *et al*: **Enhancing AI reliability: A foundation model with uncertainty estimation for optical coherence tomography-based retinal disease diagnosis**. *Cell Rep Med* 2025, **6**(1):101876.
11. Penso C, Frenkel L, Goldberger J: **Confidence Calibration of a Medical Imaging Classification System That is Robust to Label Noise**. *IEEE Trans Med Imaging* 2024, **43**(6):2050–2060.
12. Zhao J, Lin T, Wang M, Sun L, Chen H, Zhang D, Fu H: **F2UE-NET: An Uncertainty-Driven Dynamic Expert Network for Long-Tail Retinal Disease Classification**. In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. 2025: 1–5.
13. Leyba-Mesa MV, Barkana BD: **Interpretable Deeply Supervised Networks for Class-Imbalanced OCT Classification**. *AI* 2026, **7**(5):164.

14. Zhu Y, Niu Y, Sun S, Liu W, Dou Y, Guo Y: **Performance of Deep Learning in Classifying Age-Related Macular Degeneration From Images: Systematic Review and Meta-Analysis**. *J Med Internet Res* 2026, **28**:e97174.
15. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B: **Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows**. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV).* Montreal, QC, Canada: IEEE; 2021: 10012–10022.
16. Wu Z, Guo K, Luo E, Wang T, Wang S, Yang Y, Zhu X, Ding R: **Medical long-tailed learning for imbalanced data: Bibliometric analysis**. *Comput Methods Programs Biomed* 2024, **247**:108106.
17. Matta S, Lamard M, Zhang P, Le Guilcher A, Borderie L, Cochener B, Quellec G: **A systematic review of generalization research in medical image classification**. *Comput Biol Med* 2024, **183**:109256.
18. Su Z, Guo J, Yang X, Wang Q, Coenen F, Hussain A, Huang K: **Navigating Distribution Shifts in Medical Image Analysis: A Survey**. *arXiv preprint arXiv:2411.05824,* 2024.
19. Vijayan M, Veerapu G, J SN: **A spatial adaptive multi-scale ConvNeXt framework for robust, calibrated, and statistically validated multi-class retinal OCT classification**. *Front Artif Intell* 2026, **9**:1898978.
20. Venkatesh K, Mutasa S, Moore F, Sulam J, Yi PH: **Gradient-Based Saliency Maps Are Not Trustworthy Visual Explanations of Automated AI Musculoskeletal Diagnoses**. *J Imaging Inform Med* 2024, **37**(5):2490–2499.
21. Carrick RT, Rowin EJ, Gasperetti A, Madias C, Maron M, Wu KC: **Enhancing explainability in clinical deep-learning models: Latent-space variable decoding is superior to gradient-weighted class activation mapping**. *Heart Rhythm O2* 2025, **6**(9):1248–1258.
22. Ait Hammou B, Antaki F, Boucher MC, Duval R: **MBT: Model-Based Transformer for retinal optical coherence tomography image and video multi-classification**. *Int J Med Inform* 2023, **178**:105178.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarymaterial.tar