Journal of the American Medical Informatics Association, 2025, **32(5)**, 800–810
https://doi.org/10.1093/jamia/ocaf021
Advance access publication 4 March 2025
Research and Applications

Research and Applications

Development and validation of a multi-stage self-supervised learning model for optical coherence tomography image classification

Sungho Shim, BS[1], Min-Soo Kim, PhD[2], Che Gyem Yae, MD[3], Yong Koo Kang, MD[3], Jae Rock Do, MD[3], Hong Kyun Kim, MD, PhD*[,3], Hyun-Lim Yang, PhD*[,4,5,6]

[1]Department of Electrical Engineering and Computer Science, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu 42988, Republic of Korea, [2]School of Computing, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Republic of Korea, [3]Department of Ophthalmology, School of Medicine, Kyungpook National University, Daegu 41944, Republic of Korea, [4]Office of Hospital Information, Seoul National University Hospital, Seoul 03080, Republic of Korea, [5]Innovative Medical Technology Research Institute, Seoul National University Hospital, Seoul 03080, Republic of Korea, [6]Department of Medicine, College of Medicine, Seoul National University, Seoul 03080, Republic of Korea

*Corresponding authors: Hyun-Lim Yang, PhD, Office of Hospital Information, Seoul National University Hospital, 101, Daehak-ro, Jongno-gu, Seoul 03080, Republic of Korea (hlyang@snu.ac.kr) and Hong Kyun Kim, MD, PhD, Department of Ophthalmology, School of Medicine, Kyungpook National University, 130, Dongdeok-ro, Jung-gu, Daegu 41944, Republic of Korea (okeye@daum.net)

Abstract

Objective: This study aimed to develop a novel multi-stage self-supervised learning model tailored for the accurate classification of optical coherence tomography (OCT) images in ophthalmology reducing reliance on costly labeled datasets while maintaining high diagnostic accuracy.

Materials and Methods: A private dataset of 2719 OCT images from 493 patients was employed, along with 3 public datasets comprising 84 484 images from 4686 patients, 3231 images from 45 patients, and 572 images. Extensive internal, external, and clinical validation were performed to assess model performance. Grad-CAM was employed for qualitative analysis to interpret the model's decisions by highlighting relevant areas. Subsampling analyses evaluated the model's robustness with varying labeled data availability.

Results: The proposed model outperformed conventional supervised or self-supervised learning-based models, achieving state-of-the-art results across 3 public datasets. In a clinical validation, the model exhibited up to 17.50% higher accuracy and 17.53% higher macro F-1 score than a supervised learning-based model under limited training data.

Discussion: The model's robustness in OCT image classification underscores the potential of the multi-stage self-supervised learning to address challenges associated with limited labeled data. The availability of source codes and pre-trained models promotes the use of this model in a variety of clinical settings, facilitating broader adoption.

Conclusion: This model offers a promising solution for advancing OCT image classification, achieving high accuracy while reducing the cost of extensive expert annotation and potentially streamlining clinical workflows, thereby supporting more efficient patient management.

Key words: optical coherence tomography; deep learning; self-supervised learning; pre-trained model.

Introduction

Optical coherence tomography (OCT) plays a pivotal role in medical diagnostics, particularly in ophthalmology, for obtaining high-resolution, cross-sectional images of eye tissues. This noninvasive imaging technique enables ophthalmologists to visualize and assess the retina and optic nerve,[1] thus aiding in the early detection and accurate diagnosis of ocular diseases, such as age-related macular degeneration (AMD), drusen, and diabetic macular edema (DME). Early detection is of paramount importance for the prevention of vision loss and the preservation of visual acuity.[2,3] As of 2020, approximately 196 million individuals globally were affected by AMD, with a projected increase to 288 million by 2040. This highlights the significant global burden of disease and the necessity for early detection.[4–6]

OCT enables clinicians to identify subtle disease progression signs. However, the intricate nature of OCT images necessitates manual assessment by skilled ophthalmologists, which can result in significant levels of fatigue and subjectivity. Consequently, developing an automated, precise, and efficient OCT image classification system is imperative. The integration of artificial intelligence in medical imaging has enhanced efficiency, as evidenced by reductions in diagnostic time, clinician workload, and healthcare costs.[7–9]

Deep learning models have demonstrated efficacy in image classification problems, autonomously extracting features and analyzing intricate relationships from high-resolution images.[10–14] Numerous deep learning models for OCT image classification have been introduced to seize its outstanding performance.[15–17] Lee et al[15] achieved 87.63% accuracy by

Received: March 21, 2024; **Revised:** November 3, 2024; **Editorial Decision:** January 13, 2025; **Accepted:** January 23, 2025

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association. This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

training the model based on the VGG16 using 101 002 OCT images with 2 labels (normal and AMD). Li et al[16] achieved 97.30% accuracy by training the model based on the ResNet-50 using 21 357 OCT images with 4 labels (normal, choroidal neovascularization [CNV], DME, and drusen). Kamran et al[17] achieved 99.34% accuracy by training the model based on the ResNet-50 using 84 484 OCT images. However, these methods necessitate a substantial number of OCT images, incurring high labeling costs. Consequently, the effectiveness of these models depends on the data quality and quantity.

Self-supervised learning techniques have been proposed for the effective training of deep learning models with limited data.[18–21] Among the various self-supervised learning methods, the contrastive self-supervised learning approaches, such as SimCLR,[21] MOCO,[20] and SwAV[22] are widely utilized. This approach involves learning representations by contrasting positive data pairs against negative data pairs, which drives the model to differentiate between them in the embedding space. In numerous studies, self-supervised contrastive learning has consistently demonstrated superior model performance, even in medical fields with limited labeling.[23–26] For instance, Soni et al[23] employed contrastive learning to diagnose heart and lung diseases from digital stethoscope data. Similarly, Zhang et al,[24] Han et al,[25] and Azizi et al[26] utilized contrastive learning to diagnose diseases from chest radiography data. Most relevantly, Fang et al[27] achieved superior performance in OCT image classification compared with conventional deep learning models by utilizing a self-supervised learning scheme. However, the model's generalizability is limited because it was only trained and evaluated on a single specific OCT dataset.

This study proposes a novel method that employs a self-supervised learning approach, augmented by multi-staging techniques, to achieve high performance that can be widely applicable in practical clinical environments, even when using a limited amount of annotated data. Our model incorporates a multi-staging approach comprising 2 phases: a self-supervised learning phase derived from pre-trained datasets, and a fine-tuning phase for use with a private or target dataset. The model was trained on 2 large public datasets and subsequently evaluated on another public dataset or a private dataset collected from a tertiary general university hospital. Furthermore, the pre-trained model and associated code are made available to the public as a user-accessible training framework.

Our major contributions are summarized as follows:

- We achieved state-of-the-art performance in OCT image classification task by leveraging our proposed multi-stage self-supervised learning scheme.
- The proposed method can decrease the costs of annotating OCT images while retaining the superior performance of the deep learning model.
- We validated the effectiveness of the model which was trained using our proposed method, on both publicly available datasets and a dataset collected from a real clinical setting.
- To the best of our knowledge, this is the first publicly available pre-trained OCT image classification model that provides a training framework that allows the use of personal clinical data, thus making it accessible to everyone.

Methods

This study was planned in accordance with the "Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research" statements.[28]

Data collection, preprocessing, and ethical approval

Two publicly available OCT image datasets were utilized to train our model: OCT2017 and Srinivasan2014.[29,30] The publicly available datasets were employed for self-supervised learning, with each dataset corresponding to a specific stage of the process. It is notable that the training dataset labels were excluded during self-supervised learning training, given that only the image data itself was being considered. In addition, sets of predefined test images were reserved for internal validation and comparison of performance with other studies (see Figure S1).

Furthermore, 2 additional OCT image datasets were employed to validate our approach: a public dataset and a private dataset. For external validation of the trained model, this study utilizes another small open dataset[31] known as OCTID. To validate the clinical utility, we constructed a private OCT dataset, abbreviated as KNUH-OCT in this study, from a clinical tertiary hospital. The OCT images were acquired using a Heidelberg Spectralis OCT machine at Kyungpook National University Hospital (KNUH). Three skilled ophthalmologists (7, 13, and 15 years of experience) performed OCT image labeling. Two independent ophthalmologists (C.G.Y. and J.R.D.) labeled the OCT images into 4 categories; in case of inconsistencies, a senior masked retinal specialist (Y.K.K.) independently re-evaluated all images and adjudicated the discrepancies. For clinical validation, 200 OCT images were randomly selected, while the remaining 2519 images were used for training. The acquisition and analysis of the private KNUH-OCT dataset have been approved by the Institutional Review Board of KNUH (IRB-2019-07-022). Table 1 and Method S1 present detailed information about the class labels and statistics included in each dataset.

The OCT images from the public datasets (OCT2017, Srinivasan2014, and OCTID) were resized to 512 × 512 pixels and normalized by their respective means and standard deviations. Conversely, images from the private KNUH-OCT dataset were cropped from 1520 × 596 pixels to 1024 × 496 pixels and normalized in a manner analogous to that described above. To ensure robust training and enhance the generalizability of our model, we employed image augmentation techniques such as random resized cropping, random horizontal flipping, and Gaussian blurring. A detailed description of the image preprocessing is provided in Method S1.

Multi-stage self-supervised learning and model training

We propose a novel model-building strategy that considers the consistency of OCT image characteristics and the variability of OCT image quality resulting from different measurement instruments. Since OCT images consistently present the patient's condition in a similar way, we anticipate that a self-supervised learning scheme can enhance the model's performance by retrieving inherent feature information from multiple sources of data and learning from the images themselves. During deep learning model training, subtle differences or

Table 1. OCT dataset statistics.

Dataset	Class type	OCT2017	Srinivasan-2014	OCTID	KNUH-OCT
Train	Normal	26 315	1302	186	525
	CNV[a]	37 205	–	–	1028
	DME[b]	11 348	996	–	458
	AMD[c]	–	618	35	–
	Drusen	8616	–	–	508
	MH[d]	–	–	82	–
	CSR[e]	–	–	82	–
	DR[f]	–	–	87	–
Test	Normal	250	105	20	50
	CNV[a]	250	–	–	50
	DME[b]	250	105	–	50
	AMD[c]	–	105	20	–
	Drusen	250	–	–	50
	MH[d]	–	–	20	–
	CSR[e]	–	–	20	–
	DR[f]	–	–	20	–
Total images		84 484	3231	572	2719

[a] CNV: choroidal neovascularization
[b] DME: diabetic macular edema
[c] AMD: age-related macular degeneration
[d] MH: macular hole
[e] CSR: central serous retinopathy
[f] DR: diabetic retinopathy

variances in equipment from various manufacturers and settings could lead to training inefficiencies. In consideration of these issues, we developed a multi-stage self-supervised learning approach and trained a novel deep learning model.

In each stage of our self-supervised learning phase, we employed SimCLR according to our previous experiments.[32] The stages were divided according to the dataset used. SimCLR learns similarities by maximizing the similarity between positive pairs, generated from different views of the same image, and the dissimilarity between negative pairs, generated from different images. It was selected due to its demonstrated efficacy in addressing the distinctive attributes of OCT images, including the low variance in lesion distribution and fixed image orientation, by employing a contrastive learning scheme. SimCLR has demonstrated superior performance in capturing discriminative features essential for differentiating between diseased and non-diseased images. The ResNet-50[14] architecture was employed as the baseline convolutional neural network (CNN) model in this study.

Figure 1 depicts the proposed modeling framework. In Stage 1, self-supervised learning was conducted on the OCT2017[29] open dataset, which constituted a large-scale OCT image dataset. Subsequently, the CNN model that had been trained in Stage 1 was transferred to the CNN model that was to be trained in Stage 2. In Stage 2, self-supervised learning occurred on the Srinivasan2014[30] open dataset using a CNN model that was initialized with weights transferred from Stage 1. Then, the CNN model trained in Stage 2 was transferred to the CNN model in Stage 3. Stage 3 is designed to be optional, providing users with the flexibility to train the CNN model with their own downstream OCT image dataset if they so desire. This customization allows the model to be tailored to specific clinical environments, thus enhancing its suitability for specific clinical applications. Stage 3-1 follows, during which the CNN model was fine-tuned using a supervised learning scheme with labels. A small subset (20%) of labeled images from the training datasets was used for fine-tuning in Stage 3-1. Further details regarding the data usage in our method are provided in Figure S1.

In all experiments, training was performed on the entire training images, and evaluation was performed using isolated testing images from each dataset (see Figure S1). The model training for all experiments consisted of 100 epochs. When transferring the CNN model, we implemented full-weight transfer without freezing any layers. Further details about the model training can be found in Method S2.

Performance evaluation

To ensure a comprehensive assessment of the models, several performance metrics were employed for evaluation. These include precision, recall, and F-1 score for each class, in addition to accuracy, macro precision, macro recall, and macro F-1 score, which are employed to evaluate the models' performance across all classes. Moreover, confusion matrices were constructed to assess the model's strengths and weaknesses in each class.

To assess the effects of the multi-stage self-supervised learning, we compared model performance through an ablation study involving 4 datasets: 2 open datasets for internal validation (OCT2017, Srinivasan2014), 1 open dataset for external validation (OCTID), and a private dataset for clinical validation (KNUH-OCT). First, the model was trained through a supervised learning scheme. Second, the model was trained via self-supervised learning without staging and only trained on a single dataset. Finally, the model was trained using our proposed multi-stage self-supervised learning method. For internal validation, Stage 3 was omitted and proceeded directly to Stage 3-1 (fine-tuning) with only 20% of the labeled images before evaluation. To comprehensively assess the effectiveness of our method, we conducted 2 evaluations. A predefined dataset evaluation was conducted on discrete test sets that had been previously partitioned for evaluation purposes to facilitate comparison with existing literature. In order to ensure the robust assessment of model stability, a 5-fold cross-validation (5-CV) was performed. In addition, to evaluate the statistical difference between models trained on different methods, the 5 × 2 cv test[33] was employed. A P-value of $<.001$ was deemed to be statistically significant.

To evaluate the efficacy of multi-staging, we compared the model's performance with that of baseline and single-stage models trained by combining all the data used for self-supervised learning into one, without stage distinction. The training data were shuffled in this experiment for baseline and single-stage model training to ensure robust training.

Furthermore, the performance of the model was also compared based on the order of the datasets used in multi-stage training, where each dataset was trained on separate stages in ascending and descending order by data size.

Subsampling analyses were conducted to evaluate the efficiency of our framework in a small dataset. For the OCTID dataset, we trained our models by reducing the number of randomly sampled images from the data from 472 to 120. Similarly, for the KNUH-OCT dataset, we trained the models by decreasing the number of randomly selected images from 2519 to 500. These experiments were designed to assess how well our proposed model would perform in a small hospital with limited data.

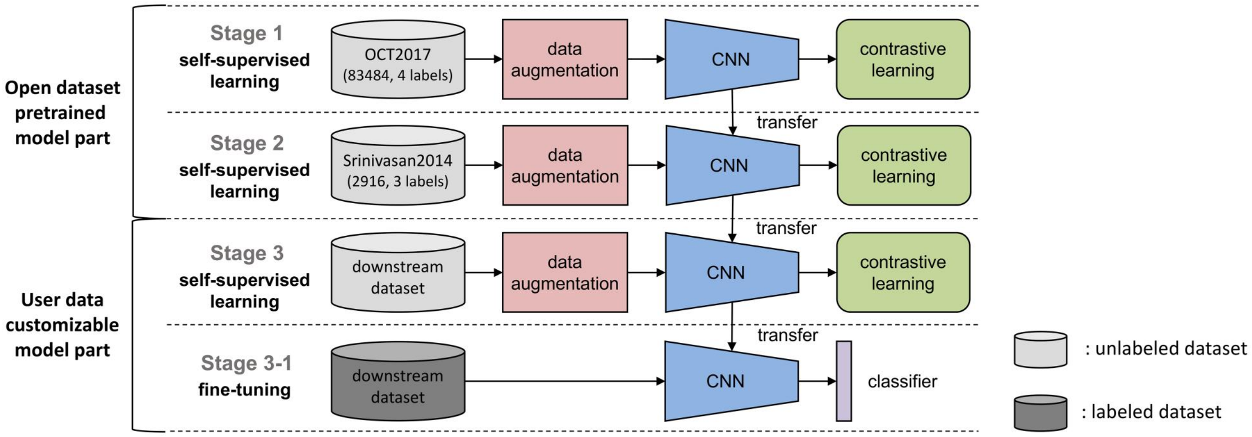


Figure 1. Multi-stage self-supervised learning model. The model consists of 3 self-supervised learning stages and 1 fine-tuning stage. In the self-supervised learning stage, the model is trained with only the training set images themselves without labels to learn robust feature representations. In the fine-tuning stage, the model is fine-tuned with a small subset (20%) of the training dataset with labels to adapt the model to specific tasks. The CNN encoder trained in the previous stage is transferred to the CNN encoder of the next stage for continuous learning. Abbreviations: CNN = convolutional neural network.

Grad-CAM analysis

The qualitative analysis of Gradient-weighted Class Activation Mapping (Grad-CAM)[34] visually examines how effectively our model, trained using multi-stage self-supervised learning, focuses on disease-relevant areas compared to the supervised learning-based baseline model. We analyzed and compared the Grad-CAM results from both the baseline model and our model using the diagnosis results of ophthalmologists. To facilitate comparison with the Grad-CAM heatmaps, 3 ophthalmologists delineated the disease regions in red circles. The annotation of the disease regions was subjected to a second review and confirmation by 3 ophthalmology specialists, who applied the same labeling rules utilized in annotating our private dataset. This clinical analysis enabled us to assess whether the regions highlighted by our model correspond to those that are clinically important for diagnosing conditions like CNV and DME.

Data and code availability

All data, code, and the pre-trained model are publicly available except for the KNUH-OCT dataset. Further details regarding data availability can be found in the "Data Availability" section below.

Results

OCT dataset statistics

Table 1 presents the number of images in each class for the training set and test set of the 4 OCT datasets. The OCT2017[29] dataset comprises 84 484 OCT images labeled as normal, CNV, DME, and drusen. The Srinivasan2014[30] dataset comprises 3231 OCT images labeled as normal, DME, and AMD, obtained from 45 patients. The OCTID[31] dataset consists of 572 OCT images labeled as normal, AMD, macular hole (MH), central serous retinopathy (CSR), and diabetic retinopathy (DR). All the aforementioned publicly available datasets include each predefined testing dataset. This testing dataset comprises 1000 images from 633 patients, 315 images from 3 patients, and 100 images for OCT2017, Srinivasan2014, and OCTID, respectively. The KNUH-OCT dataset comprises 2719 OCT images labeled as normal, CNV, DME, and drusen, from 493 patients. The training set comprises 2519 images from 457 patients and the test set comprises 200 images from 36 patients. The demographic information of the KNUH-OCT dataset can be found in Table S1.

Performance evaluation

Table 2 presents the results of the predefined dataset evaluation comparing the performance of the proposed model with that of previous works, the baseline model, and the single-stage self-supervised learning model for each dataset. These results demonstrate that our method outperforms the baseline model across all OCT datasets, achieving higher accuracy, macro precision, macro recall, and macro F-1 score. Furthermore, our approach demonstrates superior performance to single-stage self-supervised learning, except for the OCT2017 dataset. Notable performance enhancements were identified, particularly for small-scale data such as the OCTID with an improvement of 22.00% from supervised learning. Moreover, our model achieved state-of-the-art performance compared to the base performance of previous works.

Figure 2 shows each model's confusion matrices for the KNUH-OCT dataset classification results from predefined dataset evaluation. The matrices highlight the superior performance of our model compared to the baseline and single-stage self-supervised learning models. Confusion matrices for the other dataset classification results are provided in Figure S2.

Table 3 provides the performance from the 5-CV evaluation. Our method consistently achieved superior performance over both the baseline and single-stage self-supervised learning models, excelling in all performance metrics. *P*-values for comparisons with both the baseline and single-stage self-supervised learning models were below .001, indicating statistically significant improvements, except for the OCT2017 dataset, where our model showed a statistically superior performance to the baseline model. These results highlight the robust performance improvement of our method, especially

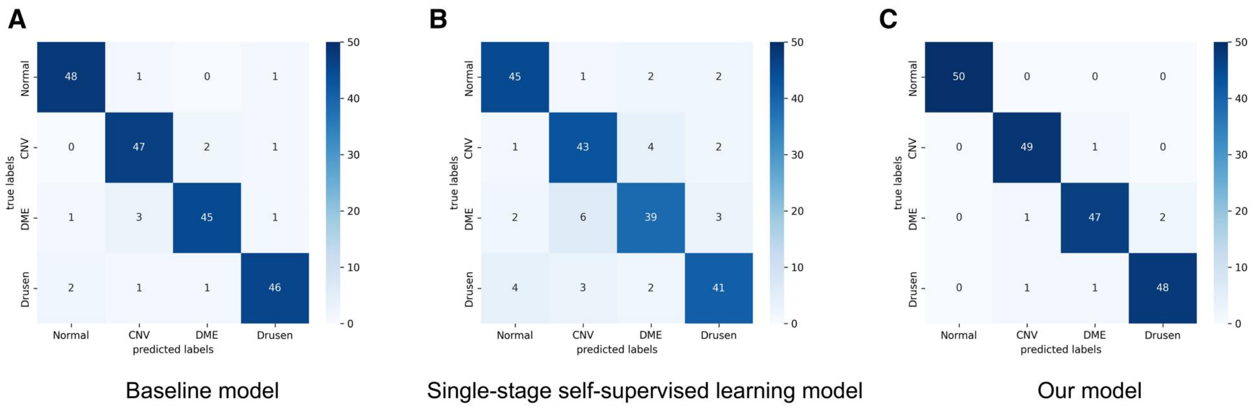
Table 2. Performance of the multi-stage self-supervised learning (predefined dataset evaluation).

Dataset (# classes)	Models	Classes	Precision	Recall	F-1 score	Accuracy	Macro precision	Macro recall	Macro F-1 score
OCT2017 (4)	Best performance of previous works[a][11]	Normal	100.00	99.60	99.80	99.30	99.30	99.30	99.30
		CNV	99.20	98.80	99.00				
		DME	98.81	99.60	99.20				
		Drusen	99.20	99.80	99.20				
	Baseline model (Supervised learning)	Normal	100.00	99.20	99.60	99.10	99.10	99.10	99.10
		CNV	98.80	98.40	98.80				
		DME	98.41	99.20	98.80				
		Drusen	98.81	99.60	99.20				
	Single-stage self-supervised learning model	Normal	100.00	100.00	100.00	99.90	99.90	99.90	99.90
		CNV	100.00	99.60	99.80				
		DME	99.60	100.00	99.80				
		Drusen	100.00	100.00	100.00				
	Multi-stage self-supervised learning model (our model)	Normal	100.00	100.00	100.00	99.80	99.80	99.80	99.80
		CNV	99.60	99.60	99.60				
		DME	99.60	99.60	99.60				
		Drusen	100.00	100.00	100.00				
Srinivasan2014 (3)	Best performance of previous works[a][11]	Normal	95.33	97.14	96.23	95.24	95.24	95.24	95.23
		DME	96.17	94.29	95.19				
		AMD	94.29	94.29	94.29				
	Baseline model (Supervised learning)	Normal	95.33	97.14	96.23	96.19	96.20	96.19	96.19
		DME	97.12	96.20	96.65				
		AMD	96.15	95.24	95.69				
	Single-stage self-supervised learning model	Normal	90.74	93.33	92.02	91.11	91.12	91.11	91.11
		DME	92.23	90.48	91.35				
		AMD	90.38	89.52	89.95				
	Multi-stage self-supervised learning model (our model)	Normal	100.00	100.00	100.00	100.00	100.00	100.00	100.00
		DME	100.00	100.00	100.00				
		AMD	100.00	100.00	100.00				
OCTID (5)	Baseline model (Supervised learning)	Normal	65.38	85.00	73.91	69.00	69.45	69.00	68.87
		AMD	57.89	55.00	56.41				
		MH	77.78	70.00	73.68				
		CSR	68.42	65.00	66.67				
		DR	77.78	70.00	73.68				
	Single-stage self-supervised learning model	Normal	58.33	70.00	63.64	56.00	56.47	56.00	56.00
		AMD	40.91	45.00	42.86				
		MH	61.11	55.00	57.89				
		CSR	58.82	50.00	54.05				
		DR	63.16	60.00	61.54				
	Multi-stage self-supervised learning model (our model)	Normal	90.91	100.00	95.24	91.00	91.01	91.00	90.92
		AMD	84.21	80.00	82.05				
		MH	94.74	90.00	92.31				
		CSR	90.48	95.00	92.68				
		DR	94.74	90.00	92.31				
KNUH-OCT (4)	Baseline model (Supervised learning)	Normal	94.12	96.00	95.05	93.00	93.03	93.00	92.99
		CNV	90.38	94.00	92.16				
		DME	93.75	90.00	91.84				
		Drusen	93.88	92.00	92.93				
	Single-stage self-supervised learning model	Normal	86.54	90.00	88.24	84.00	84.02	84.00	83.95
		CNV	81.13	86.00	83.50				
		DME	82.98	78.00	80.41				
		Drusen	85.42	82.00	83.67				
	Multi-stage self-supervised learning model (our model)	Normal	100.00	100.00	100.00	97.00	97.00	97.00	96.99
		CNV	96.08	98.00	97.03				
		DME	95.92	94.00	94.95				
		Drusen	96.00	96.00	96.00				

[a] The model was retrained using the author's source code to prevent dataset leakage concerns. See the "Discussion" section for the details.
Abbreviations: CNV = choroidal neovascularization; DME = diabetic macular edema; AMD = age-related macular degeneration; MH = macular hole; CSR = central serous retinopathy; DR = diabetic retinopathy.

on smaller datasets. Detailed results of the statistical test can be found in Table S4.

Table S2 illustrates the effect of stage distinction in our training process using 5-CV evaluation. Note that the baseline and single-stage self-supervised learning models were trained on a merged dataset comprising all available datasets, our multi-stage self-supervised learning model utilized distinct datasets at different stages of the training process. The multi-stage model consistently achieved the highest performance across all datasets, demonstrating its effectiveness. Table S3 compares the performance of the models based on the order of the datasets used at each stage in the multi-stage



Baseline model

Single-stage self-supervised learning model

Our model

Figure 2. Confusion matrices for the KNUH-OCT dataset classification results. The matrices show the classification performance across 4 classes: Normal, choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen. (A) Confusion matrix of the baseline model (conventional supervised learning model) on the KNUH-OCT dataset. (B) Confusion matrix of the single-stage self-supervised learning model on the KNUH-OCT dataset. (C) Confusion matrix of the multi-stage self-supervised learning model on the KNUH-OCT dataset. Abbreviations: OCT = optical coherence tomography.

training. The results show that the multi-stage training with descending order by data size outperforms that of the ascending order across all OCT datasets.

Subsampling analysis

Table 4 presents the results of the subsampling analysis conducted on the OCTID dataset and the KNUH-OCT dataset. The accuracy of the model increased steadily as the data size grew, from 84.00% to 91.00% for OCTID and 92.00% to 97.00% for KNUH-OCT. Our model outperformed the supervised learning-based baseline model at all training data sizes. In the analysis of 120 images from the OCTID dataset, our model demonstrated an accuracy of 84.00%, whereas the supervised learning-based model exhibited an accuracy of only 50.00%. In the analysis of 500 images from the KNUH-OCT dataset, our model achieved an accuracy of 92.00%, whereas the supervised learning-based model achieved only 74.50%. Subsampling analysis for OCT2017 and Srinivasan2014 datasets are provided in Table S5.

Grad-CAM analysis

Figure 3 displays the Grad-CAM results obtained from the supervised learning-based baseline model and our multi-stage self-supervised learning model for 4 different types of OCT images. The first column shows the original OCT images. The second and third columns present the Grad-CAM results obtained from the baseline and our proposed models, respectively. In these results, the red-marked heatmap area indicates the model's focus.

In Figure 3(2), the baseline model was unable to focus on CNV lesions. Conversely, our model effectively identified CNV lesions in both the outer retina and the retinal pigment epithelium, as well as the thickened choroidal layer. This figure illustrates the characteristics of polypoidal choroidal vasculopathy, a type of neovascular AMD associated with choroidal thickening.[35] In Figure 3(3), our model distinctly targets the outer retina and choroidal layer in DME lesions. DR is marked by choroidal thinning due to ischemia,[36] and our model accurately reflects this feature, whereas the baseline model focused on completely different regions. Furthermore, Figure 3(4) confirms that the baseline model focused on other regions other than drusen lesions. In contrast, our model identified the exact location of drusen bodies, which are well-defined hyperreflective deposits located between the basal membrane of the retinal pigment epithelium and Bruch's membrane. Grad-CAM results of OCT2017, Srinivasan2014, and OCTID datasets are provided in Figure S3.

Discussion

In this study, we propose a multi-stage self-supervised learning approach that accurately classifies OCT images with limited data. We trained a model using the proposed method and compared its performance with both external and clinical validation. Our model outperforms traditional supervised learning-based models and the best-performing models in the literature. Across the 4 datasets used for evaluation, our model demonstrated an improvement in accuracy ranging from 0.7% to 22% compared with the conventional supervised learning-based model. Our model still showed superior performance in the 5-CV evaluation, highlighting its robustness. Remarkably, our models exhibited acceptable performance even with extremely small-scale OCT datasets. When we trained our model using only 20% of the data collected in a clinical setting, the supervised learning-based model underperformed by 18.5% compared to the full data training, whereas our model underperformed by only 5%. These findings support the efficacy of the proposed multi-stage learning approach. Moreover, our proposed model demonstrated superior performance compared to models trained on pooled datasets across all datasets. This underscores the efficacy of our multi-stage self-supervised learning strategy, which leverages the distinctive attributes of each dataset at different stages, rather than relying on a single merged dataset. The findings highlight that the enhanced performance of our model is a direct consequence of the stage distinction, rather than merely the quantity of data utilized for training. We also found that the performance is influenced by the order in which data are used during multi-stage training. When initially trained on smaller datasets, the performance decreased up to 3%. This finding suggests that transferring the latent features of OCT images from large OCT datasets to the next stage improves accuracy.

Table 3. Performance of the multi-stage self-supervised learning (5-CV evaluation).

Dataset (# classes)	Models	Classes	Precision	Recall	F-1 score	Accuracy	Macro precision	Macro recall	Macro F-1 score	*P*-value[a]
OCT2017 (4)	Baseline model (Supervised learning)	Normal	98.67±0.42	98.86±0.40	98.76±0.41	98.44±0.56	97.84±0.74	97.68±0.92	97.76±0.83	0.0012
		CNV	99.14±0.39	99.24±0.16	99.19±0.26					
		DME	97.39±0.86	97.13±1.38	97.26±1.11					
		Drusen	96.16±1.31	95.49±1.81	95.82±1.55					
	Single-stage self-supervised learning model	Normal	99.35±0.21	99.32±0.38	99.34±0.28	99.18±0.38	98.85±0.59	98.84±0.39	98.84±0.49	0.5725
		CNV	99.58±0.18	99.54±0.35	99.56±0.27					
		DME	98.39±0.89	98.95±0.83	98.67±0.83					
		Drusen	98.07±1.09	97.55±0.33	97.81±0.60					
	Multi-stage self-supervised learning model (our model)	Normal	99.32±0.25	99.40±0.28	99.36±0.26	99.21±0.29	98.94±0.38	99.00±0.31	98.97±0.35	N/A
		CNV	99.57±0.17	99.40±0.28	99.48±0.22					
		DME	98.51±0.66	98.80±0.52	98.66±0.58					
		Drusen	98.34±0.53	98.41±0.29	98.38±0.36					
Srinivasan 2014 (3)	Baseline model (Supervised learning)	Normal	95.54±1.18	95.81±1.19	95.67±1.14	94.62±1.30	94.28±1.35	94.19±1.33	94.23±1.33	<0.001
		DME	94.83±1.38	94.92±1.51	94.87±1.44					
		AMD	92.49±1.75	91.84±1.57	92.16±1.44					
	Single-stage self-supervised learning model	Normal	93.50±0.61	90.97±1.11	92.22±0.78	90.56±0.93	89.76±1.01	90.26±0.92	89.99±0.97	<0.001
		DME	90.49±1.13	91.55±0.88	91.02±0.95					
		AMD	85.30±1.53	88.24±0.97	86.75±1.24					
	Multi-stage self-supervised learning model (our model)	Normal	99.29±0.50	99.50±0.60	99.40±0.51	99.32±0.56	99.34±0.56	99.21±0.63	99.27±0.59	N/A
		DME	99.28±0.75	99.64±0.38	99.46±0.52					
		AMD	99.44±0.59	98.48±1.13	98.96±0.82					
OCTID (5)	Baseline model (Supervised learning)	Normal	82.23±2.49	74.75±7.26	78.23±5.15	68.52±4.82	64.57±4.84	65.29±4.61	64.67±4.83	<0.001
		AMD	36.82±9.44	47.27±7.61	41.29±8.66					
		MH	67.42±5.70	67.71±3.55	67.39±2.93					
		CSR	66.65±6.55	65.71±2.95	66.13±4.58					
		DR	69.75±4.83	71.00±6.27	70.30±4.95					
	Single-stage self-supervised learning model	Normal	79.30±1.75	57.79±4.03	66.82±3.18	55.43±3.20	53.66±2.87	53.34±3.05	52.52±3.05	<0.001
		AMD	19.50±2.51	38.18±4.07	25.80±3.10					
		MH	54.97±5.51	56.90±3.29	55.87±4.17					
		CSR	54.33±3.12	55.90±2.84	55.09±2.77					
		DR	60.20±3.68	57.92±3.53	59.02±3.88					
	Multi-stage self-supervised learning model (our model)	Normal	99.01±1.35	95.63±2.67	97.28±1.61	90.38±2.80	86.54±3.49	87.32±3.36	86.81±3.45	N/A
		AMD	59.62±8.21	69.09±8.13	63.95±8.03					
		MH	92.78±2.80	88.19±2.92	90.40±2.34					
		CSR	90.29±5.84	89.29±3.85	89.75±4.55					
		DR	91.03±3.01	94.42±1.96	92.66±1.88					
KNUH-OCT (4)	Baseline model (Supervised learning)	Normal	88.44±1.84	89.04±1.91	88.73±1.74	90.58±1.48	89.40±1.58	90.08±1.48	89.72±1.54	<0.001
		CNV	95.96±1.01	92.67±1.70	94.28±1.23					
		DME	85.49±2.73	88.98±1.58	87.19±2.13					
		Drusen	87.73±1.44	89.61±1.34	88.66±1.29					
	Single-stage self-supervised learning model	Normal	77.83±1.89	79.30±1.56	78.56±1.62	82.79±1.26	80.85±1.43	81.42±1.43	81.11±1.42	<0.001
		CNV	91.73±0.57	88.50±1.29	90.08±0.74					
		DME	75.61±1.92	79.93±2.09	77.71±1.95					
		Drusen	78.23±1.67	77.95±1.80	78.09±1.71					

(continued)

Table 3. (continued)

Dataset (# classes)	Models	Classes	Precision	Recall	F-1 score	Accuracy	Macro precision	Macro recall	Macro F-1 score	*P*-value[a]
	Multi-stage self-supervised learning model (our model)	Normal	94.31±1.13	95.13±0.78	94.72±0.93	96.06±0.88	95.59±0.94	95.56±0.81	95.57±0.88	N/A
		CNV	98.05±0.62	98.14±1.18	98.10±0.90					
		DME	94.51±1.59	94.69±1.11	94.60±1.32					
		Drusen	95.47±0.61	94.27±0.47	94.86±0.49					

[a] *P*-value was calculated using 5 × 2 cv[33] test between each model and multi-stage self-supervised learning (our) model.

Abbreviations: CNV = choroidal neovascularization; DME = diabetic macular edema; AMD = age-related macular degeneration; MH = macular hole; CSR = central serous retinopathy; DR = diabetic retinopathy.

Table 4. Subsampling analysis.

Dataset (Training data size)	Models	Accuracy	Macro precision	Macro recall	Macro F-1 score
OCTID (472)	Baseline model (Supervised learning)	69.00	69.45	69.00	68.87
	Single-stage self-supervised learning model	56.00	56.47	56.00	56.00
	Multi-stage self-supervised learning model (Our model)	91.00	91.01	91.00	90.92
OCTID (237)	Baseline model (Supervised learning)	55.00	53.87	55.00	54.60
	Single-stage self-supervised learning model	51.00	51.48	51.00	51.10
	Multi-stage self-supervised learning model (Our model)	87.00	87.22	87.00	87.07
OCTID (120)	Baseline model (Supervised learning)	50.00	49.88	50.00	49.82
	Single-stage self-supervised learning model	48.00	48.49	48.00	48.06
	Multi-stage self-supervised learning model (Our model)	84.00	84.12	84.00	84.02
KNUH-OCT (2519)	Baseline model (Supervised learning)	93.00	93.03	93.00	92.99
	Single-stage self-supervised learning model	84.00	84.02	84.00	83.95
	Multi-stage self-supervised learning model (Our model)	97.00	97.00	97.00	96.99
KNUH-OCT (1500)	Baseline model (Supervised learning)	90.00	90.11	90.00	89.99
	Single-stage self-supervised learning model	82.00	81.96	82.00	81.96
	Multi-stage self-supervised learning model (Our model)	95.50	95.62	95.50	95.53
KNUH-OCT (500)	Baseline model (Supervised learning)	74.50	74.50	74.50	74.48
	Single-stage self-supervised learning model	78.50	78.47	78.50	78.48
	Multi-stage self-supervised learning model (Our model)	92.00	92.07	92.00	92.01

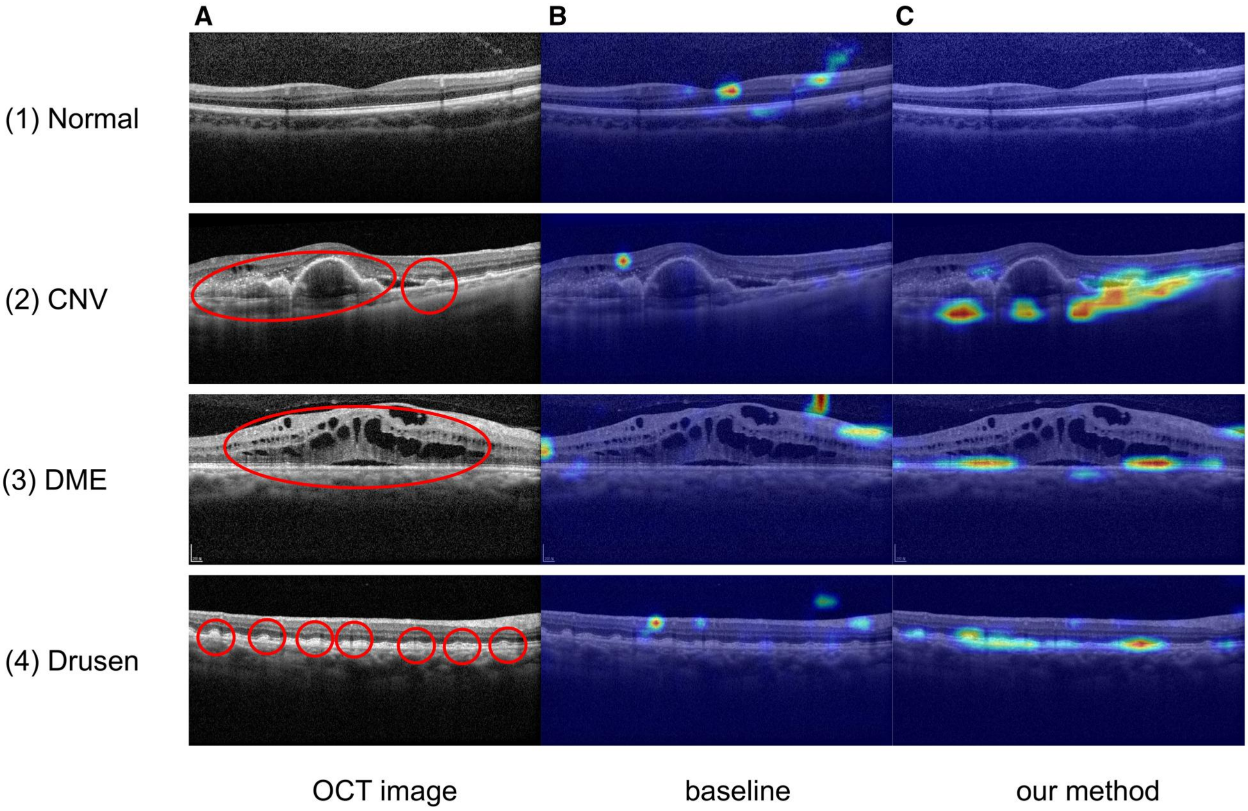


Figure 3. An example of Grad-CAM results in the KNUH-OCT test set. (A) Original OCT images from the test set, showing different ophthalmic conditions. For comparison with the Grad-CAM heatmaps, 3 ophthalmologists have circled the disease areas in red. (B) Grad-CAM heatmaps generated by the baseline model, highlighting the regions used for predictions. (C) Grad-CAM heatmaps generated by our model, showing the regions of interest that the model focused on for making predictions. Abbreviations: OCT = optical coherence tomography; Grad-CAM = gradient-weighted class activation mapping.

Clinical image data, such as OCT, possess unique characteristics that differ from those of general images. First, labeling incurs high costs. As clinical experts must manually inspect and label each image, obtaining and maintaining the large amounts of labeled data required for high-performance deep-learning model training are challenging. Second, variability exists in the clinical environment, such as regarding type of machine, settings, and measurer. Finally, owing to privacy concerns, hospitals seldom release sensitive clinical data, impeding the retention of large data. Consequently, most hospitals or users interested in building an OCT classifier typically have access to only small-scale clinical datasets, making it difficult to generate a robust deep-learning model.

We developed a method that considers relevant clinical data characteristics. Initially, we decreased labeling costs and the demand for a large dataset using self-supervised learning.

The model acquires knowledge of the OCT images through contrastive learning by utilizing 2 open OCT datasets of significant sizes. Subsequently, we address the inefficiency of training a model owing to OCT image quality variability by adopting multi-stage learning. By ensuring that the model learns only from a single dataset in a single stage, learning inefficiencies that may arise from mixing images of varying qualities are avoided. Additionally, we released a pre-trained model with large-scale public data and provided a framework for users to fine-tune the model according to our methodology. This enables each hospital or user to own a superior OCT classifier that uses only their small data with minimal privacy concerns.

Most previous studies on classifying OCT images used supervised learning models that require abundant data and labels such as over 10 000 OCT images.[15,16,37] Additionally, most current studies that trained and assessed deep learning models used an undisclosed OCT dataset created by the researchers themselves. To ensure a reliable performance evaluation of deep learning models, it is essential to test their performance using various datasets, such as external or clinical validation datasets.

When evaluating model performance, the testing data should be entirely separate from other data, such as training or validation data. Unfortunately, we found that the OpticNet,[17] the current best-performing model for the OCT2017 and Srinivasan2014 datasets, had data leakage concerns. Specifically, the source code they provided utilized the same dataset for both validation and testing. Consequently, although the model may perform well in a given test dataset, its real-world clinical utility can be called into question. An accuracy of 99.80% in the OCT2017 dataset and a perfect score of 100.00% in the Srinivasan2014 dataset were reported; however, after re-training with separate testing data, our reproduced results showed a decline in performance to 99.30% and 95.24% for each respective dataset.

Our model's performance was evaluated both reliably and objectively. We analyzed the performance internally, externally, and clinically and confirmed the superiority of the model trained using our proposed method. Furthermore, we publicly released the codes for transparent replication of our results.

The model has significant clinical potential. It enhances diagnostic accuracy even with limited data, which is critical for precision in resource-constrained environments. This automated, reliable classification model facilitates faster diagnostics and timely medical interventions, providing valuable second opinions and assisting clinicians with challenging conditions. By making our pre-trained models publicly available, we enable hospitals to enhance their diagnostic capabilities with minimal additional data and training, democratizing advanced tools and ensuring broader access.

Our study has some limitations. First, the CNN model performance can be enhanced by optimizing the model architecture and hyperparameters, which were not rigorously explored in this study. However, we made all the training and testing source codes of our proposed method available, enabling users to update the model architecture as desired. Second, the scope of our experiments was confined to OCT image classification. However, the proposed method has the potential to lay the groundwork for expanding its application to various modalities and tasks. Third, although we evaluated our model externally and clinically, this was a retrospective study that requires prospective validation for real-world clinical usage.

Conclusion

We developed and validated a multi-stage self-supervised learning model to classify OCT images. This approach allows the model to learn inherent features without high labeling costs and train efficiently without confusion over image quality. We believe that our contributions will facilitate the adoption of high-performing OCT classifiers in clinical practice, improving patient outcomes.

Author contributions

Sungho Shim, Min-Soo Kim, Hyun-Lim Yang (Conceptualization), Sungho Shim, Che Gyem Yae, Yong Koo Kang, Jae Rock Do, Hong Kyun Kim (Data curation), Sungho Shim, Hyun-Lim Yang (Formal analysis), Hong Kyun Kim (Funding acquisition), Sungho Shim, Min-Soo Kim, Hyun-Lim Yang (Investigation), Sungho Shim, Hyun-Lim Yang (Methodology), Min-Soo Kim, Che Gyem Yae, Yong Koo Kang, Jae Rock Do, Hong Kyun Kim (Resources), Sungho Shim (Software), Min-Soo Kim, Hyun-Lim Yang (Supervision), Sungho Shim, Hyun-Lim Yang (Validation), Sungho Shim (Visualization), Sungho Shim, Hyun-Lim Yang (Writing—original draft), Sungho Shim, Min-Soo Kim, Hong Kyun Kim, Hyun-Lim Yang (Writing—review and editing).

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

This work was supported by a Biomedical Research Institute Grant, Kyungpook National University Hospital (2018); and the National Research Foundation of Korea (NRF) grant funded by the Ministry of Science and ICT (MSIT), Korea [grant numbers NRF-2018R1A5A1060031 and NRF-2021R1A2C2013939].

Conflicts of interest

The authors have declared that no competing interest exists.

Data availability

The data used to develop and validate this study are publicly available, except for the KNUH-OCT dataset. Due to the ethical restrictions imposed by the IRB, the private dataset is only available upon request. The source code, including data preprocessing, model building, model training, performance evaluation, weights of the pre-trained CNN model, and a user-customizable training framework, is available for use with personal clinical data through a public repository (https://github.com/Mercury22/multi-stage-self-supervised-learning-model-for-OCT-classification). The pre-trained CNN model has undergone training until Stage 2. The framework provided can execute both Stage 3 and Stage 3-1 consecutively when the user follows the prescribed folder structure to store their own OCT images.

References

1. Huang D, Swanson EA, Lin CP, et al. Optical coherence tomography. *Science*. 1991;254:1178-1181.
2. Coleman HR, Chan C-C, Ferris FL, et al. Age-related macular degeneration. *Lancet*. 2008;372:1835-1845.
3. Lee R, Wong TY, Sabanayagam C. Epidemiology of diabetic retinopathy, diabetic macular edema and related vision loss. *Eye Vis*. 2015;2:1-25.
4. Resnikoff S, Pascolini D, Etya'Ale D, et al. Global data on visual impairment in the year 2002. *Bull World Health Organ*. 2004;82:844-851.
5. Song P, Du Y, Chan KY, et al. The national and subnational prevalence and burden of age–related macular degeneration in China. *J Glob Health*. 2017;7:020703.
6. Colijn JM, Buitendijk GH, Prokofyeva E, et al.; European Eye Epidemiology (E3) consortium. Prevalence of age-related macular degeneration in Europe: the past and the future. *Ophthalmology*. 2017;124:1753-1763.
7. Van Leeuwen KG, de Rooij M, Schalekamp S, et al. How does artificial intelligence in radiology improve efficiency and health outcomes? *Pediatr Radiol*. 2022;52:2087-2093.
8. Dembrower K, Wåhlin E, Liu Y, et al. Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. *Lancet Digit Health*. 2020;2:e468-e474.
9. Lång K, Dustler M, Dahlblom V, et al. Identifying normal mammograms in a large screening population using artificial intelligence. *Eur Radiol*. 2021;31:1687-1692.
10. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst*. 2012;25:1097-1105.
11. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I. Springer International Publishing; 2014:818-833.
12. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. Published Online First: September 4, 2014. Accessed March 20, 2024. https://arxiv.org/abs/1409.1556
13. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2015:1-9.
14. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2016:770-778.
15. Lee CS, Baughman DM, Lee AY. Deep learning is effective for classifying normal versus age-related macular degeneration OCT images. *Ophthalmol Retina*. 2017;1:322-327.
16. Li F, Chen H, Liu Z, et al. Deep learning-based automated detection of retinal diseases using optical coherence tomography images. *Biomed Opt Express*. 2019;10:6204-6226.
17. Kamran SA, Saha S, Sabbir AS, et al. Optic-net: a novel convolutional neural network for diagnosis of retinal diseases from optical tomography images. In: 2019 *18th IEEE International Conference On Machine Learning And Applications*. IEEE Computer Society; 2019:964-971.
18. Noroozi M, Favaro P. Unsupervised learning of visual representations by solving jigsaw puzzles. In: *European Conference on Computer Vision*. Springer International Publishing; 2016:69-84.
19. Gidaris S, Singh P, Komodakis N. Unsupervised representation learning by predicting image rotations. Published Online First: March 21, 2018. Accessed March 20, 2024. https://arxiv.org/abs/1803.07728
20. He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2020:9729-9738.
21. Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations. In: *International Conference on Machine Learning*. PMLR; 2020:1597-1607.
22. Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments. *Adv Neural Inf Process Syst*. 2020;33:9912-9924.
23. Soni PN, Shi S, Sriram PR, et al. Contrastive learning of heart and lung sounds for label-efficient diagnosis. *Patterns*. 2022;3:100400.
24. Zhang Y, Jiang H, Miura Y, et al. Contrastive learning of medical visual representations from paired images and text. In: *Machine Learning for Healthcare Conference*. PMLR; 2022:2-25.
25. Han Y, Chen C, Tewfik A, et al. Pneumonia detection on chest x-ray using radiomic features and contrastive learning. *Proc IEEE Int Symp Biomed Imaging*. 2021;2021:247-251.
26. Azizi S, Mustafa B, Ryan F, et al. Big self-supervised models advance medical image classification. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. IEEE Computer Society; 2021:3478-3488.
27. Fang L, Guo J, He X, et al. Self-supervised patient-specific features learning for OCT image classification. *Med Biol Eng Comput*. 2022;60:2851-2863.
28. Luo W, Phung D, Tran T, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. 2016;18:e323.
29. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*. 2018;172:1122-1131.e9.
30. Srinivasan PP, Kim LA, Mettu PS, et al. Fully automated detection of diabetic macular edema and dry age-related macular degeneration from optical coherence tomography images. *Biomed Opt Express*. 2014;5:3568-3577.
31. Gholami P, Roy P, Parthasarathy MK, et al. OCTID: optical coherence tomography image database. *Comput Electr Eng*. 2020;81:106532.
32. Shim S, Yang H-L, Kim M-S. Comparison of self-supervised learning methods for optical coherence tomography image classification. In: *Proceedings of 2024 8th International Conference on Medical and Health Informatics*. ACM; 2024:41-46.
33. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Comput*. 1998;10:1895-1923.
34. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE Computer Society; 2017:618-626.
35. Ryu G, Moon C, van Hemert J, et al. Quantitative analysis of choroidal vasculature in polypoidal choroidal vasculopathy using ultra-widefield indocyanine green angiography. *Sci Rep*. 2020;10:18272.
36. Li Z, Yang F, Deng X, et al. Association between choroidal thickness and diabetic macular edema: a meta-analysis. *Acta Diabetol*. 2024;61:951-961.
37. Yang H-L, Kim JJ, Kim JH, et al. Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images. *PLoS One*. 2019;14:e0215076.

© The Author(s) 2025. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Journal of the American Medical Informatics Association, 2025, 32, 800–810
https://doi.org/10.1093/jamia/ocaf021
Research and Applications