IRAQI Academic Scientific Journals
Samarra Journal of Pure and Applied Science
SJPAS
www.sjpas.com
p ISSN: 2663-7405
e ISSN: 2789-6838

Federated Learning with Contrastive Pretraining for Retinal OCT Disease Classification on OCT Retinal Images

Mustafa Lateef Fadhil Jumaili*

1- Department of Computer Science, College of Computer science and mathematics, University of Tikrit, Iraq

https://doi.org/10.54153/sjpas.2026.v8i1.1422

Article Information

Received: 24/01/2026
Revised: 11/03/2026
Accepted: 16/03/2025
Published: 10/04/2026

Keywords:

Federated learning, Contrastive pretraining, Optical coherence tomography (OCT), Retinal disease classification, Multi-institutional AI, Domain generalization

Corresponding Author

E-mail:

Mustafa.l.fadhil@tu.edu.iq

Abstract

Optical Coherence Tomography (OCT) is a common technique that is used to identify retinal caused diseases choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen-related degeneration. Nonetheless, there are difficult aspects of training strong deep learning models to the classification of OCT, such as privacy, multi-institutional (heterogeneous) distributions of data, and the scarcity of labeled annotations. Our proposed federated learning framework in this work consists of two stages of classification-based privacy-preserving retinal disease classification self-supervised contrastive pretraining followed by variance-reduced federated optimization frameworks. The first stage involves clients learning OCT-specific features through a contrastive learning method based on momentum (MoCo v3) to do local self-supervised representation learning without the need of labeled data. The second step involves fine-tuning of the pretrained encoder with the SCAFFOLD algorithm of supervised federated learning to address client drift when data are distributed by non-IID. EfficientNet-B4 is the adopted backbone architecture to provide both a balance in the performance of classification and communication efficiency. Multi-institution OCT datasets are experimented on Kermany OCT and Duke OCT datasets and further cross-domain testing is done on the OCTID dataset. The presented framework gets 96.8 percent accuracy in classification and AUC-ROC of 0.994 on in-domain evaluation, surpassing a variety of centralized and federated baselines. Moreover, cross-domain experiments are associated with a stronger generalization to unseen OCT data.These findings indicate the potential effectiveness of federated learning when combined with contrastive pretraining to achieve greater robustness in the case of heterogeneous data distribution and maintain the privacy of patient data.

1. Introduction:

Optical Coherence Tomography (OCT) is a standard-of-care imaging modality for diagnosing and monitoring retinal diseases such as choroidal neovascularization (CNV), diabetic macular edema (DME), and drusen-related degeneration [1]. Deep learning methods have demonstrated strong performance for OCT-based disease classification, enabling automated decision support that can improve screening efficiency and reduce clinician workload [2,3]. However, many high-performing OCT models re-ly on centralized training, requiring aggregation of large-scale patient images into a single repository [4]. In practice, OCT data are distributed across institutions and con-strained by

privacy regulations and governance policies, while also exhibiting substantial inter-site variability due to differences in scanners, acquisition protocols, and patient demographics. As a result, the clinically desirable paradigm of multiparadigm-scaled learning, which is learning off of heterogeneous multi-institutional data, is hard to be actualized [5].

Federated learning (FL) provides a conceptually sound framework of collaborative model training, which does not imply the distribution of raw patient data. FL is compatible with privacy-preserving deployment in multihospital networks by maintaining OCT images locally and sharing only model updates, which is optimal in a multi-hospital set-up [6-9]. Although this has been promised, standard FL pipelines tend to do poorly with medical imaging centralized training, mainly because of statistical heterogeneity (non-IID data) and a lack of labels. The multi-hospital datasets generally have different disease prevalence and acquisition conditions which results in client drift and unstable optimization when using FedAvg-style aggregation. Furthermore, expert annotation is both costly and not uniformly applied to clients and may additionally deteriorate supervised FL performance and restricts extrapolation to new domains [10].

The second limitation is due to the representation learning. Most FL works pre-train models on natural-image corpora, but this approach can be problematic in OCT as retinal scans have domain-specific anatomic structures and pathological textures, which are highly different than natural images [11]. The domain shift may lead to poor transfer efficiency and increase generalization failure in case labelled data, on a per-client basis, are scarce [12]. These observations motivate a key question: can federated OCT training be improved not only through better aggregation, but also through learning stronger domain-relevant representations prior to supervised fine-tuning—without centralized access to OCT images.

Self-supervised contrastive learning provides a compelling direction. By encouraging consistency across augmentations of the same image while separating representations of different images, contrastive objectives can learn transferable features without labels [13]. For OCT, this has the potential to capture anatomy and pathology-relevant structure at scale and improve robustness to cross-site variability [14]. However, integrating contrastive learning into FL is non-trivial: self-supervised objectives can be sensitive to optimization noise and heterogeneity, and naïve FL aggregation can amplify client drift and destabilize representation learning under non-IID settings. Therefore, effective federated self-supervision requires (i) a stable contrastive pretraining strategy and (ii) a federated optimization mechanism that explicitly mitigates client drift.

In this work, we propose a two-stage privacy-preserving framework for retinal OCT disease classification that combines momentum-encoder contrastive pretraining [15] with variance-reduced federated optimization [16]. In Stage 1, each client per-forms local self-supervised pretraining to learn OCT-specific representations without labels using a momentum-updated encoder and a memory queue of feature embed-dings to provide a large and consistent set of contrastive negatives during training. In Stage 2, clients perform supervised federated fine-tuning for disease classification using the pretrained representation as initialization. To address non-IID client drift, we integrate Stochastic Controlled Averaging for Federated Learning with Drift (SCAF-FOLD) [17], which employs control variates to reduce gradient variance and stabilize optimization across heterogeneous clients. Finally, we adopt EfficientNet-B4 as a communication- and compute-efficient backbone to reduce bandwidth requirements while maintaining strong representational capacity, improving practical deploy ability in real hospital networks.

Our evaluation is designed to reflect realistic multi-institution settings. We simulate a multi-client federation using large-scale OCT data with non-IID client partitions and include cross-domain testing to assess generalization under acquisition variability. Across comprehensive comparisons with centralized supervised training and strong FL baselines, we show that federated contrastive

pretraining combined with variance reduction yields improved accuracy and AUC-ROC while reducing communication overhead relative to heavier backbones and naïve aggregation. Importantly, under multi-source heterogeneous training conditions, the proposed framework achieves generalization performance that is competitive with—and in certain settings exceeds—a centralized supervised baseline, highlighting the value of privacy-preserving representation learning for clinically realistic deployment.

OCT is a highly common technology in the diagnosis of retinal diseases and OCT image classification has been demonstrated to be promising with the use of deep learning. Nonetheless, OCT data is normally spread among various institutions and cannot be centrally shared because of privacy rules and is also heterogeneous across acquisition locations. Federated learning is a model training technique that allows collaborative training, without sharing raw patient data, but common federated optimization algorithms are prone to client drift when the data is non-IID and with small labeled datasets. To overcome them, we suggest a two-step federated learning system, which integrates self-supervised contrastive training (MoCo v3) with drift-sensitive federated training with SCAFFOLD. The design will enhance representation learning and stability in training in heterogeneous multi institution settings whilst maintaining patient data privacy.

Our main contributions are:

1. We suggest a two-phase federated training process, which integrates contrastive self-supervised pretraining and supervised federated fine-tuning to classify retinal OCT disease.
2. The SCAFFOLD federated optimization algorithm is embedded in us to reduce client drift and train more robustly in heterogeneous data distributions.
3. We use EfficientNet-B4 as a backbone that is efficient in communication and can be used in federated learning of numerous institutions.
4. We test the suggested framework on the several OCT datasets and determine its performance against both the heterogeneous distributions of clients and in the context of cross-domain testing.

The remainder of this paper is organized as follows: Section II reviews related work; Section III details the proposed method; Section IV describes experimental design; Section V reports results and ablations; Section VI discusses implications and limitations; and Section VII concludes the paper.

2. Related Works:

Early work on OCT-based retinal disease classification established the effective-ness of supervised convolutional neural networks (CNNs) in centralized training settings. Zong et al [18] formulated OCT diagnosis as a set of binary classification tasks and evaluated multiple backbones (e.g., VGG, ResNet, DenseNet, Inception) as feature extractors, reporting strong performance on curated datasets. Similarly, Rani et al. [19] proposed a transfer-learning framework built on a customized VGG-19 architecture and trained on a large-scale public OCT dataset spanning CNV, DME, Drusen, and Normal categories, reporting high classification accuracy and comprehensive statistical evaluation (e.g., ROC-AUC, Cohen's kappa, confusion matrix). Beyond B-scan image-level classification, Perdomo et al. [20] introduced OCT-NET for diabetes-related retinal disease assessment from SD-OCT volumes and incorporated a feedback mechanism to highlight clinically relevant regions, emphasizing interpretability alongside predictive performance. Subramanian et al. [21] further explored multi-disease OCT classification using CNN-based models (including VGG16), demonstrating improvements over traditional methods across several retinal disorder categories. Collectively, these studies confirm that centralized supervised deep learning can achieve high

diagnostic performance for OCT; however, they generally assume pooled multi-institution data access and do not address privacy constraints, cross-site heterogeneity, or the communication and optimization challenges inherent to distributed clinical deployment.

Federated learning has been increasingly explored to enable privacy-preserving collaboration across institutions without sharing raw patient data. Lo et al. [22] studied federated learning for OCT/OCTA applications, including retinal microvasculature segmentation and referable diabetic retinopathy classification, showing that federated models can achieve performance close to centralized training while benefiting from cross-site diversity. More recently, Nabil et al. [23] conducted a systematic mul-ti-center evaluation of FL strategies for multi-disease OCTA classification, bench-marking aggregation algorithms (e.g., FedAvg/FedProx-type methods), architectures (CNNs, Transformers, hybrids), and deployment factors such as layer-freezing strategies, client scalability, and computational efficiency. Their study also incorporated privacy mechanisms such as secure aggregation and differential privacy, offering a broader privacy–utility view of FL deployment in ophthalmology.

Despite these advances, multiple studies consistently report that standard aggregation methods are sensitive to statistical heterogeneity. Vamsidhar et al. [24] explicitly simulated realistic non-IID OCT client distributions (including missing classes at some clients) and showed that performance degrades as heterogeneity increases, with FedProx generally more robust than FedAvg under severe non-IID conditions. Gulati et al. [25] similarly evaluated federated deep learning for DR/DME detection with light-weight architectures under IID and non-IID partitions and reported improved stability with FedProx in heterogeneous settings. Kaushal et al. [27] also generalized FL feasibility experiments to multi-modality ocular imaging (fundus, OCT, OCTA) on standard backbones, which is another step toward FL as a privacy-preserving alternative to centralized training. In general, this literature confirms the feasibility of FL in ophthalmic imaging and shows remaining issues associated with client drift, label imbalance, cross-site variability, which are even more pronounced in real-life multi-hospital contexts.

In addition to statistical heterogeneity, federated systems may face adversarial threats that degrade model performance. Eshan et al. [26] addressed Byzantine robustness for retinal OCT classification by proposing a confidence-score–based strategy to reduce the influence of potentially malicious clients during aggregation, demon-starting improved performance under adversarial settings. While security-oriented approaches are complementary to heterogeneity-focused optimization, most ophthalmic-mic FL studies—including [22-27]—remain primarily supervised and do not explicitly address representation learning limitations under label scarcity.

Self-supervised learning (SSL) has emerged as an effective alternative to ImageNet-based transfer learning in medical imaging, particularly when labeled data are limited or imbalanced. Huang et al. [28] demonstrated that SSL—including a Mo-Co-v2–based contrastive scheme—can outperform transfer learning for retinal disease classification under small-data and imbalanced scenarios, with particularly strong gains in extremely low-label regimes. More recently, Jannat et al. [29] proposed OCT-SelfNet, a self-supervised framework trained on combined multi-institution datasets using a masked autoencoder with a SwinV2 backbone, reporting improved robustness under low-data and unseen-domain testing. These works support the central premise that SSL can learn OCT-specific representations that transfer effectively across tasks and domains. However, they typically assume centralized access to pooled multi-institution datasets, which may be infeasible in practice due to privacy and governance constraints.

A growing line of work explores combining FL with self-supervised pretraining to leverage unlabeled data in distributed settings. Wu et al. [30] proposed federated self-supervised frameworks

for dermatological diagnosis, including a contrastive approach with feature sharing (FedCLF) and a masked autoencoder approach (FedMAE) with knowledge-splitting to improve generalization. While these results highlight the promise of federated self-supervision, they are evaluated in dermatological imaging and employ mechanisms (e.g., feature sharing and ViT synchronization strategies) tailored to that domain and deployment context.

In contrast to prior centralized OCT SSL methods [28], [29] and supervised ophthalmic FL approaches [22-27], our work focuses on privacy-preserving retinal OCT disease classification under realistic heterogeneous multi-institution distributions. We combine contrastive self-supervised pretraining with drift-mitigating federated optimization using SCAFFOLD, and we adopt an EfficientNet-B4 backbone to balance di-agnostic performance with communication efficiency. This design targets the key failure modes repeatedly observed in practice—non-IID client drift, limited labels, and bandwidth constraints—while enabling collaborative learning without sharing raw OCT data.

3. Materials and Methods:

The proposed framework integrates federated learning with momentum-based contrastive pretraining (MoCo v3) and SCAFFOLD drift correction to enable privacy-preserving, multi-institutional retinal OCT disease classification (Figure 1). Each client independently performs self-supervised contrastive learning on local OCT datasets, generating feature representations without sharing raw patient data. The local model updates are regularly sent to a central server, and SCAF-FOLD compensates client drift, and pools updates to generate a global model. A global model is then re-transmitted to the clients, which acts as an initializer to a supervised fine-tuning phase on labeled OCT data. This two-step training plan enables the framework to acquire strong, generalizable representations as well as be stable in convergence in heterogeneous multi-client conditions. The last model provides high confidence predictions on four types of retinal diseases including CNV, DME, DRUSEN and normal and shows cross-domain generalization as well as diagnostic accuracy.

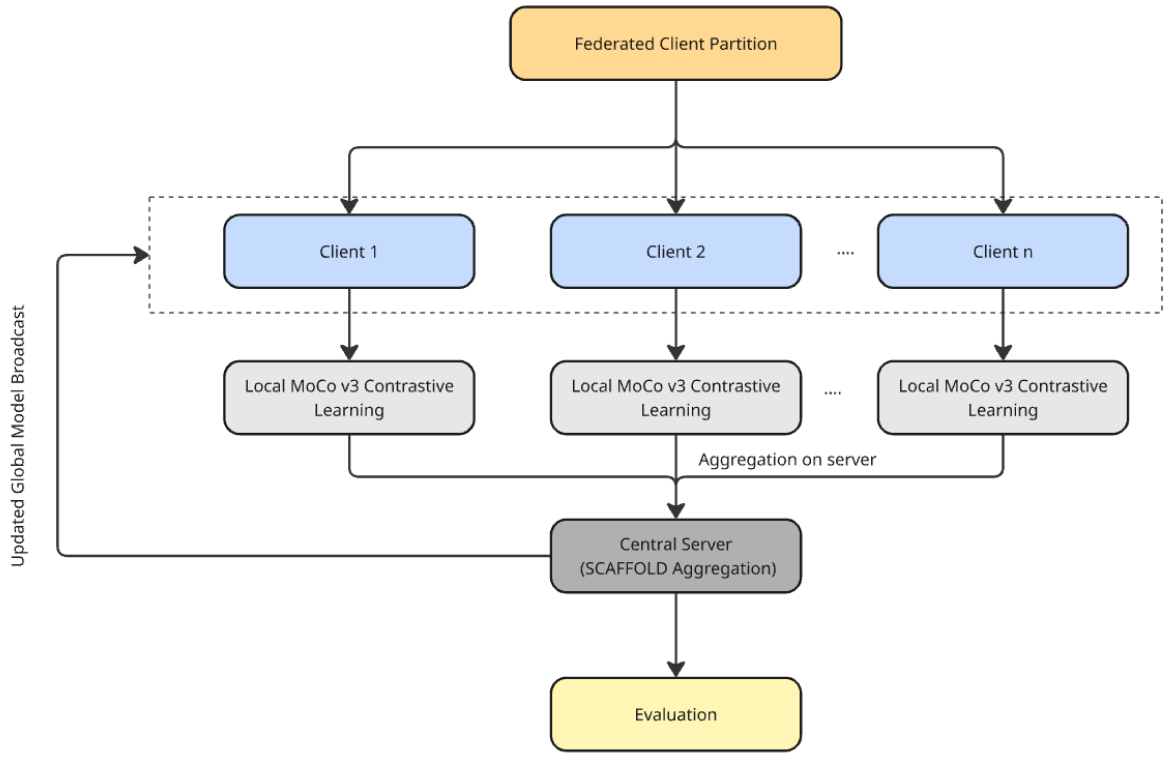

Fig. 1 *Proposed Approach.*

3.1 Problem Formulation

We consider a privacy-preserving multi-institution learning setting in which retinal OCT data are distributed across Kclients (e.g., hospitals). Client $k \in \{1, \dots, K\}$holds a local dataset $\mathcal{D}_k = \{(x_i, y_i)\}_{i=1}^{n_k}$, where $x_i$is an OCT B-scan and $y_i \in \{1, \dots, C\}$denotes the disease label among Cclasses (CNV, DME, DRUSEN, NORMAL). The federation exhibits statistical heterogeneity (non-IID data), i.e., the underlying client distributions differ across institutions: $P_k(x, y) \neq P_j(x, y)$for $k \neq j$. A central server coordinates training without accessing raw images; communication is limited to exchanging model parameters (and auxiliary optimizer states when required).

Our objective is to learn a global model $f(\cdot; w)$that generalizes well to a held-out test distribution while ensuring that all OCT scans remain local. To reduce label dependence and improve robustness under heterogeneity, we adopt a two-stage optimization strategy. In Stage 1 (federated self-supervised pretraining), each client leverages its unlabelled images $x \in \mathcal{D}_k$to learn an encoder $E(\cdot; \theta)$by minimizing a contrastive objective, producing transferable OCT representations. In Stage 2 (federated supervised fine-tuning), the encoder is adapted for multi-class diagnosis by minimizing a supervised classification loss over labelled samples. Across both stages, the server aggregates client updates using SCAFFOLD (Stochastic Controlled Averaging for Federated Learning with Drift) to mitigate client drift and stabilize training under non-IID distributions.

3.2 Datasets

This study uses multiple publicly available retinal OCT datasets to evaluate the proposed federated learning framework.

Kermany OCT Dataset

Kermany OCT database has 84,495 retinal OCT images that were gathered to be used in automation of disease classification. The data is divided into four classes namely choroidal neovascularization (CNV), diabetic macular edema (DME), drusen, and normal retina. In line with the standard practice, the data is partitioned into the training, validation and test sets that comprise 59,146, 8,449, and 16,900 images respectively. The data is also imbalanced in its classes, with the CNV being the most common classification.

Duke OCT Dataset

The Duke OCT dataset contains 5,000 OCT images acquired at a different medical institution using distinct imaging hardware and acquisition protocols. The dataset includes the same four disease categories as the Kermany dataset. For our experiments, the dataset is divided into 3,500 training images, 500 validation images, and 1,000 test images.

OCTID Dataset

To evaluate cross-domain generalization, we additionally perform experiments on the OCTID dataset, which contains OCT images collected from independent clinical sources using different acquisition conditions. This dataset is used exclusively for out-of-domain evaluation and is not included during model training. The OCTID dataset introduces domain shifts due to variations in imaging devices and clinical settings, providing a challenging benchmark for evaluating the robustness of federated learning models trained on heterogeneous multi-institution data.

Overall, the total experimental corpus is 89,495 OCT images of which 62,646 are training samples, 8949 are validation samples and 17900 are test samples. In both datasets, the cumulative class distribution has 38,456 images of CNVs, 12,449 images of DMEs, 9,667 images of DRUs, and 27,915 images of NORMALs. These data sets are distributed to four federated clients, partitioned non-IID, and allow a realistic assessment of federated learning in the conditions of privacy guarantee and non-homogeneous data distributions across multiple institutions.

Table 1: Data Distribution.

Dataset	Train Samples	Val Samples	Test Samples	Total Samples	CNV	DME	DRUSEN	NORMAL
Kermany OCT	59146	8449	16900	84495	37206	11349	8617	26315
Duke OCT	3500	500	1000	5000	1250	1100	1050	1600
Total	62646	8949	17900	89495	38456	12449	9667	27915

3.3 Federated Client Construction and Non-IID Partitioning

To capture realistic multi-institutional clinical deployment, we develop a federated learning environment comprising of four clients, each one of them being a representation of a separate medical institution. The federated model is intended to detect the statistical heterogeneity and the domain shift that is typical of real-world multi-hospital OCT data.

Kermany OCT set of data is spread across three federated clients which model three hospitals, which share a common data source, yet have dissimilar patterns of disease prevalence. In particular, the data is assigned to each of the Kermany clients in a skewed subset of classes whereby the proportions of CNV cases, DME cases, DRUSEN cases, and NORMAL cases vary for different clients. This label distribution heterogeneity is an established cause of client drift in federated optimization of this class imbalance. Even though each of the four types of diseases can be found throughout the federation, specific pathologies can dominate individual clients, which is realistic because of variations in referrals and demographics of patients in hospitals.

The Duke OCT data is allocated to a fourth client and is considered as a different institution. Unlike the Kermany clients, the Duke client brings about cross-domain heterogeneity due to disparity in imaging devices, acquisition protocols and clinical practice. This design enables the assessment of federated learning on the background of simultaneously label shift and domain shift, without breaching data locality restrictions.

No unprocessed pictures and patient information are transmitted between customers or the central server. All preprocessing, augmentation and local training are conducted separately by each client. The federated algorithm only communicates model parameters and auxiliary optimizer variables that are re-quired by the federation algorithm.

The result of this federated partitioning approach is a clinically realistic learning problem that is challenging and has (i) non-IID distributions of labels, (ii) varying domains across institutions and (ii) a high level of data privacy. It therefore provides a principled testbed for assessing the robustness, convergence stability, and generalization capability of federated contrastive pretraining and variance-reduced optimization under realistic multi-hospital OCT conditions.

3.4 Network Architecture

We adopt a convolutional encoder–classifier architecture tailored for high-resolution retinal OCT B-scans and communication-efficient federated training. The proposed framework is built around

three components: (i) an EfficientNet-B4 backbone encoder, (ii) a projection head used during contrastive pretraining, and (iii) a classification head used during supervised fine-tuning.

Let $x \in \mathbb{R}^{H\times W\times C}$denote an input OCT image. We employ EfficientNet-B4 [31] as the feature extractor $E(\cdot;\theta)$due to its favorable accuracy–parameter trade-off and reduced model size compared with common alternatives such as ResNet-50, which directly impacts communication cost in federated settings. The encoder maps an input image to a high-level representation:

$$h = E(x;\theta) \in \mathbb{R}^{d_h}, \qquad (1)$$

where $d_h$denotes the dimensionality of the penultimate feature vector produced by the backbone after global pooling.

During Stage 1 (contrastive pretraining), we attach a projection head $g(\cdot;\phi)$to the encoder to map features into a compact embedding space used by the contrastive objective:

$$z = \mathrm{norm}\big(g(h;\phi)\big) \in \mathbb{R}^{d}, \qquad (2)$$

where $\mathrm{norm}(\cdot)$denotes ℓ_2-normalization and $d$is the embedding dimension (set to 128 in our experiments). The projection head is implemented as a lightweight multi-layer perceptron (MLP) with non-linear activation, following standard practice in contrastive representation learning to decouple representation quality from the embedding space used for the pretext task. The projection head is used only during pretraining; for downstream classification we either discard it or keep it frozen depending on the evaluation protocol.

During Stage 2, we attach a classifier $c(\cdot;\psi)$to the encoder output for multi-class diagnosis:

$$\hat{y} = c(h;\psi) \in \mathbb{R}^{C}, p = \mathrm{softmax}(\hat{y}), \qquad (3)$$

where $C = 4$ corresponds to CNV, DME, DRUSEN, and NORMAL. Unless otherwise stated, $c(\cdot)$is a linear layer (i.e., logistic regression on top of h), which provides a strong and interpretable baseline for assessing representation quality. We additionally consider end-to-end fine-tuning, where both $\theta$and $\psi$are updated using labeled data with a smaller learning rate applied to the backbone to preserve the pretrained representation.

Throughout training, each client maintains a local copy of the network and communicates only model parameters required by the federated algorithm. In Stage 1, aggregation is applied to the encoder (and projection head when applicable), while in Stage 2, aggregation includes the encoder and classifier. This design supports efficient multi-round federated optimization while retaining sufficient representational capacity for high-accuracy OCT disease classification.

3.5 Federated Momentum Contrastive Pretraining

To learn domain-relevant OCT representations without relying on large-labelled cohorts at each institution, we perform federated self-supervised contrastive pretraining in Stage 1. Each client optimizes a momentum-based contrastive objective locally using only its resident OCT images, while the central server coordinates the aggregation of model parameters under privacy constraints (no raw-image exchange). The pretraining procedure is designed to be stable under non-IID data by combining (i) a teacher–student momentum encoder mechanism and (ii) drift-mitigating federated optimization (Section 3.7).

For each OCT image x, client $k$samples two stochastic augmentations to form a positive pair $\left(x^{(1)}, x^{(2)}\right)$. The augmentations are applied independently and locally, producing two correlated views of the same underlying anatomy. This view construction enables learning invariances to nuisance variations while preserving disease-relevant structure.

Each client maintains two encoders with identical architecture: a query encoder $E_q(\cdot;\theta_q)$and a momentum (key) encoder $E_k(\cdot;\theta_k)$. Given the two augmented views, the query and key embeddings are computed as:

$$q = \mathrm{norm}\left(g\left(E_q\left(x^{(1)};\theta_q\right)\right)\right), k^+ = \mathrm{norm}\left(g\left(E_k\left(x^{(2)};\theta_k\right)\right)\right), \quad (4)$$

where $g(\cdot)$is the projection head (Section 3.5) and $\mathrm{norm}(\cdot)$denotes ℓ_2-normalization. The key encoder parameters are updated by an exponential moving average of the query encoder parameters:

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q, \quad (5)$$

where $m \in [0,1)$is the momentum coefficient (set to $m = 0.999$unless otherwise stated). This momentum update stabilizes target representations and reduces training noise, which is particularly important under client-level heterogeneity.

Contrastive learning benefits from a large set of negative samples. To avoid requiring impractically large batch sizes at each client, we maintain a fixed-size first-in-first-out queue (memory bank) $\mathcal{Q} = \{k_j^-\}_{j=1}^{|\mathcal{Q}|}$storing embeddings produced by the momentum encoder from recent mini-batches. At each iteration, new keys are enqueued and the oldest keys are dequeued to keep the queue size constant. The queue provides a large and diverse set of negative examples drawn from the client's local distribution, enabling stable contrastive training with moderate batch sizes.

Given a query embedding q, its corresponding positive key k^+, and the set of negative keys $\{k^-\} \subset \mathcal{Q}$, we minimize the InfoNCE loss:

$$\mathcal{L}_{\mathrm{con}}(q, k^+, \mathcal{Q}) = -\log \frac{\exp\left(\frac{q^\top k^+}{\tau}\right)}{\exp\left(\frac{q^\top k^+}{\tau}\right) + \sum_{k^- \in \mathcal{Q}} \exp\left(\frac{q^\top k^-}{\tau}\right)}, \quad (6)$$

where $\tau$is a temperature hyperparameter controlling distribution sharpness. This objective encourages representations of two views of the same OCT image to be close in the embedding space while pushing apart representations of different images.

Pretraining proceeds over R_{pre}communication rounds. In each round, the server broadcasts the current global model parameters to a subset of participating clients. Each client performs E_{pre}local epochs of contrastive optimization over its private dataset using $\mathcal{L}_{\mathrm{con}}$, updates the momentum encoder via the EMA rule, maintains its local queue, and returns updated model parameters (and any auxiliary state required by the federated optimizer). The server then aggregates client updates to obtain a new global initialization for the next round.

Because the memory queue is inherently local and depends on client data distribution, we do not exchange queue contents across clients. This design preserves privacy and avoids leaking distributional information through stored embeddings. After completion of contrastive pretraining,

the projection head is discarded (or retained only for linear evaluation), and the pretrained encoder parameters serve as initialization for Stage 2 federated supervised fine-tuning.

Federated learning under multi-institution OCT data is characterized by statistical heterogeneity, where each client observes a distinct data distribution due to differences in disease prevalence, referral patterns, and acquisition protocols. Under such non-IID conditions, naïve aggregation (e.g., FedAvg) often suffers from client drift, i.e., local optimization steps move client models toward different optima, leading to unstable convergence and degraded global generalization. To mitigate this failure mode, we employ SCAFFOLD (Stochastic Controlled Averaging for Federated Learning with Drift) as the federated optimization backbone in both contrastive pretraining (Stage 1) and supervised fine-tuning (Stage 2).

SCAFFOLD introduces control variates to reduce the variance of stochastic gradients induced by heterogeneous client data. The server maintains a global control variate $c \in \mathbb{R}^{|w|}$and each client $k$maintains a local control variate $c_k \in \mathbb{R}^{|w|}$, both having the same dimensionality as the model parameters w. Intuitively, $c$and $c_k$approximate the discrepancy between local and global gradients and are used to correct local updates.

At the beginning of communication round t, the server broadcasts the current global model $w^t$and global control variate $c^t$to participating clients. Client $k$initializes its local model $w_k^{t,0} \leftarrow w^t$and performs Slocal update steps. At step s, given a mini-batch $\mathcal{B}$drawn from $\mathcal{D}_k$, the client computes a stochastic gradient $g_k^{t,s} = \nabla \ell_k(w_k^{t,s}; \mathcal{B})$, where $\ell_k$denotes the local objective (contrastive loss in Stage 1 or cross-entropy loss in Stage 2). SCAFFOLD applies the corrected update:

$$w_k^{t,s+1} \leftarrow w_k^{t,s} - \eta\left(g_k^{t,s} - c_k^t + c^t\right), \quad (7)$$

where $\eta$is the local learning rate. The correction term $(c^t - c_k^t)$compensates for systematic bias in the local gradient direction caused by client-specific distributions, thereby reducing client drift.

After completing Slocal steps, client $k$obtains an updated model $w_k^{t,S}$. The server aggregates client models to update the global parameters (uniform weighting is used unless otherwise stated):

$$w^{t+1} \leftarrow \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} w_k^{t,S}, \quad (8)$$

where $\mathcal{K}_t$denotes the set of clients participating in round t.

SCAFFOLD further updates control variates to maintain consistent drift correction over rounds. Following the standard SCAFFOLD formulation, the client updates its local control variate using the discrepancy between the starting and ending local models:

$$c_k^{t+1} \leftarrow c_k^t - c^t + \frac{1}{S\eta}\left(w^t - w_k^{t,S}\right) \quad (9)$$

The server then updates the global control variate by averaging the participating clients' control variates:

$$c^{t+1} \leftarrow \frac{1}{|\mathcal{K}_t|} \sum_{k \in \mathcal{K}_t} c_k^{t+1} \quad (10)$$

This coupled update ensures that the drift correction evolves with the optimization trajectory and remains effective under changing local solutions.

We use SCAFFOLD throughout the two levels of training. In Stage 1, clients maximize the contrastive goal using fixed gradients and enhance the stability of the representation learning in the case of non-IID distributions. On Stage 2, SCAFFOLD also stabilizes supervised fine-tuning, with less bias to a dominant distribution of clients and better global calibration. It is important to note that SCAFFOLD can be adopted in our setting in particular since it reduces the problem of heterogeneity without any further data exchange between model parameters and control variates, which retains the privacy-preserving assumption of federated learning.

3.6 Federated Supervised Fine-Tuning

Following federated contrastive pretraining (Section 3.6), we perform federated supervised fine-tuning to adapt the pretrained encoder to the downstream retinal OCT disease classification task. This stage leverages labeled OCT images available at each institution while maintaining strict data locality. The objective is to (i) convert the learned representation into a discriminative multi-class diagnostic model and (ii) preserve cross-institution generalization under heterogeneous (non-IID) client distributions.

Let $E(\cdot;\theta)$denote the pretrained encoder obtained after Stage 1. We attach a classification head $c(\cdot;\psi)$that maps the encoder representation $h = E(x;\theta)$to logits $\hat{y} \in \mathbb{R}^C$for $C = 4$disease classes:

$$\hat{y} = c(h;\psi), p = \mathrm{softmax}(\hat{y}), \quad (11)$$

where $p$is the predicted class probability vector. Unless otherwise stated, the classification head is a linear layer. The projection head used for contrastive pretraining is discarded at the start of supervised fine-tuning.

Given labeled samples (x, y)at client k, we optimize the multi-class cross-entropy loss:

$$\mathcal{L}_{\mathrm{sup}} = -\sum_{c=1}^{C} 1[y = c]\log\, p_c, \qquad (12)$$

where $1[\cdot]$is an indicator function. This loss is minimized locally at each client using mini-batch stochastic optimization.

Supervised fine-tuning proceeds over R_{sup}communication rounds. At round t, the server broadcasts the current global parameters $(\theta^t, \psi^t)$and the global control variate $c^t$to participating clients. Each client initializes local parameters $\left(\theta_k^{t,0}, \psi_k^{t,0}\right) \leftarrow (\theta^t, \psi^t)$and performs Slocal update steps across E_{sup}local epochs.

To reduce instability under non-IID client distributions, we continue to apply SCAFFOLD during supervised optimization. Specifically, for the combined parameter vector $w = [\theta, \psi]$, client $k$updates:

$$w_k^{t,s+1} \leftarrow w_k^{t,s} - \eta\left(\nabla\ell_k(w_k^{t,s};\mathcal{B}) - c_k^t + c^t\right), \qquad (13)$$

where $\ell_k$corresponds to the supervised objective $\mathcal{L}_{\mathrm{sup}}$evaluated on a mini-batch $\mathcal{B} \subset \mathcal{D}_k$. After completing local training, clients return updated parameters to the server, which aggregates them and updates the control variates as described in Section 3.7.

We consider two commonly used downstream adaptation strategies. First, linear evaluation, where the encoder parameters $\theta$are frozen and only the classifier head $\psi$is trained in a federated manner.

This isolates the quality of the representations learned during contrastive pretraining. Second, end-to-end fine-tuning, where both the encoder and classifier head are updated jointly using a reduced learning rate for the encoder, enabling task-specific refinement while minimizing catastrophic forgetting of pretrained features. Unless otherwise stated, our main reported results correspond to end-to-end fine-tuning, as it reflects a clinically relevant deployment scenario in which the global model is fully adapted for diagnostic performance.

3.7 Training Configuration

We adopt a two-stage federated training protocol consisting of momentum-based contrastive pretraining followed by supervised federated fine-tuning. All clients use identical hyperparameter settings within each stage to ensure stable optimization and fair comparison across baselines.

Stage 1: Federated contrastive pretraining.

The contrastive pretraining stage is conducted for 100 communication rounds. In each round, participating clients perform 5 local training epochs using a batch size of 64 and stochastic gradient descent (SGD) with momentum 0.9 and an initial learning rate of 0.03. The contrastive objective employs a temperature parameter $\tau = 0.2$. A momentum encoder with update coefficient $m = 0.999$is used to stabilize representation learning, and a fixed-size queue of 65,536 embeddings is maintained at each client to provide a large and diverse set of negative samples. This configuration enables effective self-supervised learning from unlabeled OCT images under heterogeneous, non-IID client distributions.

Stage 2: Federated supervised fine-tuning.

Following pretraining, the encoder is initialized with the learned representations and fine-tuned in a supervised federated manner for 50 communication rounds. Each client trains locally for 3 epochs per round with a batch size of 32 using the Adam optimizer and a learning rate of 0.001. A cosine annealing learning-rate scheduler is applied to improve convergence stability and generalization in later training stages. Unless otherwise specified, both the encoder and the classification head are updated during this phase, with a smaller effective learning rate applied to the backbone to preserve pretrained features.

All experiments use EfficientNet-B4 as the backbone network, comprising 19.3M parameters and approximately 4.2G FLOPs per image. Client-side training is performed on NVIDIA RTX 3090 GPUs with 24GB memory. Under this configuration, the complete two-stage federated training procedure (150 total communication rounds) requires approximately 18 hours to complete.

A comprehensive summary of hyperparameters and implementation details is provided in Table 2.

Table 2: Hyperparameters and implementation details.

Component	Parameter	Value
Stage 1: Federated Contrastive Pretraining	Communication rounds	100
	Local epochs per round	5
	Batch size	64
	Optimizer	SGD (momentum = 0.9)
	Learning rate	0.03

	Contrastive temperature ((\tau))	0.2
	Queue size ((K))	65,536
	Momentum coefficient ((m))	0.999
Stage 2: Federated Supervised Fine-Tuning	Communication rounds	50
	Local epochs per round	3
	Batch size	32
	Optimizer	Adam
	Learning rate	0.001
	Learning-rate scheduler	CosineAnnealingLR
Model Architecture	Backbone	EfficientNet-B4
Compute Environment	Hardware per client	NVIDIA RTX 3090 (24GB)
	Total training time	(\sim)18 hours

3.8 Evaluation Metrics

We evaluate the proposed framework using standard multi-class classification metrics that are widely adopted in medical image analysis and ophthalmic diagnostic studies. All metrics are computed on held-out test sets that are never observed during training, and performance is reported at the global model level after each federated training stage. Given the clinical importance of balanced performance across disease categories, we report both threshold dependent and threshold independent measures.

Overall classification accuracy measures the proportion of correctly classified OCT images across all classes:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}(\hat{y}_i = y_i), \qquad (14)$$

where Ndenotes the number of test samples, $y_i$the ground-truth label, $\hat{y}_i$the predicted label, and $\mathbb{1}(\cdot)$the indicator function. To assess per-class diagnostic reliability, we compute precision, recall (sensitivity), and F1-score for each disease category. For a given class c,

$$\text{Precision}_c = \frac{TP_c}{TP_c + FP_c}, \text{Recall}_c = \frac{TP_c}{TP_c + FN_c},$$
$$\text{F1}_c = \frac{2 \cdot \text{Precision}_c \cdot \text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}, \qquad (15)$$

where TP_c, FP_c, and FN_cdenote the true positives, false positives, and false negatives for class c, respectively. Unless otherwise stated, macro-averaged scores are reported to account for class imbalance across disease categories.

AUC-ROC is used as a threshold-independent measure of discriminative ability, which is particularly important in clinical decision-support systems. For the multi-class setting, we adopt a one-vs-rest formulation and report the macro-averaged AUC across all classes. A higher AUC indicates better separation between positive and negative samples irrespective of the classification threshold.

In addition to scalar metrics, confusion matrices are reported to visualize class-wise prediction behavior and common misclassification patterns among CNV, DME, DRUSEN, and NORMAL

categories. This analysis provides clinically interpretable insights into failure modes of the model under heterogeneous multi-institution data.

3.9 Notation

In this paper, we make the following notation in describing the federated learning framework. K will be the total number of clients (institutions) participating. Each client K possesses a local dataset Dk= {(x,y) with (x_i) being an image of an OCT and y_i representing the disease label of x_i.

The model parameters used globally are given by θ whereas θ represents the local model parameters of client Federated training clients optimize their local models with stochastic gradients computed on their local data. The SCAFFOLD algorithm uses the control variates c (global) and c_k(local) to cure client-specific gradient drift in case of heterogeneous data distributions.

The encoder network is represented as f(.), and it is used to obtain feature representations of input images. In contrastive pretraining, an encoder projection head g(·) is used to project the encoder features into a contrastive learning embedding space. In the case of supervised classification, a classifier head h(·) is used to encode features and classify the feature by class probabilities of the four retinal disease categories.

4. Results:

This section reports quantitative and qualitative results for the proposed federated contrastive pretraining framework for retinal OCT disease classification.

4.1 Overall Performance and Comparison to Baselines

Table 3 summarizes the overall diagnostic performance of the proposed MoCo v3 + SCAFFOLD framework against centralized training and representative federated baselines, while Figure 2 visualizes the corresponding accuracy comparison. The proposed method achieves the strongest results across all metrics, reaching 96.8% accuracy, 96.4% precision, 96.2% recall, 96.3% F1-score, and an AUC-ROC of 0.994. It means that besides the increased top 1 classification correctness, there is also a more desirable balance between false positives and false negatives- a desirable property in screening-oriented clinical decision support.

The proposed approach provides a +2.6% absolute accuracy improvement and +0.012 AUC-ROC improvement over the centralized ResNet-50 baseline (94.2% accuracy; AUC-ROC 0.982) with a privacy-preserving constraint, which indicates that it is more discriminative. Relative to standard supervised federated learning, the improvements are larger: +5.3% accuracy over FedAvg (91.5%) and +4.5% over FedProx (92.3%), highlighting the sensitivity of conventional FL to statistical heterogeneity and client drift. Notably, even against a strong representation-learning baseline (SimCLR + FedAvg, 93.8%), the proposed approach retains a +3.0% accuracy advantage and higher AUC-ROC (0.994 vs 0.978), supporting the hypothesis that momentum-based contrastive learning combined with drift-mitigating optimization provides complementary benefits in heterogeneous multi-client OCT training.

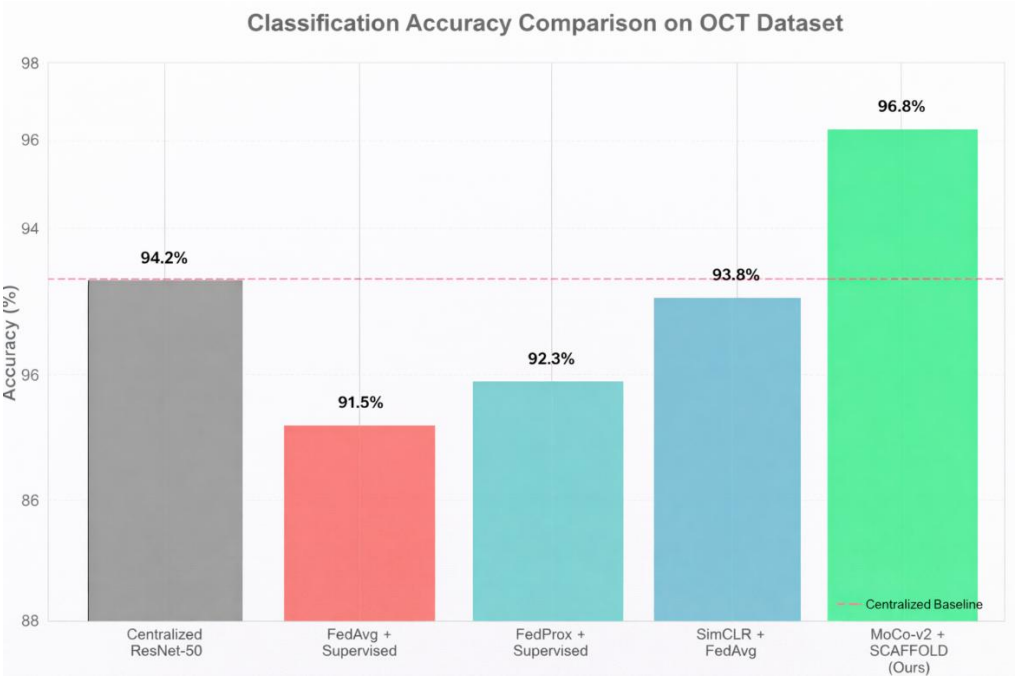

Fig. 2 *Classification accuracy comparison on the retinal OCT dataset.*

Table 3: Performance comparison across centralized and federated learning methods on retinal OCT disease classification.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC
Centralized ResNet-50	94.2	93.8	93.5	93.6	0.982
FedAvg + Supervised	91.5	90.8	90.2	90.5	0.965
FedProx + Supervised	92.3	91.6	91.1	91.3	0.971
SimCLR + FedAvg	93.8	93.2	92.8	93	0.978
MoCo v3 + SCAFFOLD (Ours)	96.8	96.4	96.2	96.3	0.994

4.2 Convergence Across Federated Rounds

Figure 3 illustrates the convergence behavior of the proposed MoCo v3 + SCAFFOLD framework across federated communication rounds, covering both the self-supervised contrastive pretraining stage and the supervised fine-tuning stage. The results demonstrate stable optimization, efficient convergence, and clear benefits of the two-stage federated training strategy under non-IID multi-client settings.

During Stage 1 (Rounds 1–100), the contrastive loss decreases steadily across communication rounds, indicating effective representation learning from unlabeled OCT data at each client. The gradual fading of the contrastive goal emphasizes the stabilizing effect of the momentum encoder in MoCo v3, and the application of SCAFFOLD reduces the phenomenon of client drift by correcting biased local updates caused by the heterogeneous distributions of data. Notably, convergent and oscillatory tendencies are not noticed, which would indicate that the optimization of the variance reduction process is essential in ensuring stability in federated self-supervised learning.

After the changeover to Stage 2 (Rounds 101-150) the loss in supervised classification drops sharply, as does the training as well as the validation accuracy. This effect shows that downstream classification heavily relies on the representations acquired during federated contrastive pretraining,

which gives it a good initialisation. The proposed approach has shown high accuracy with significantly less rounds taken than purely supervised federated baselines, which is a better optimization efficiency.

Interestingly, the validation accuracy curves and training curves are near parallel during the fine-tuning process, indicating that there is not much overfitting even with statistical heterogeneity among clients. The model reaches its final test accuracy of 96.8 which is higher than the centralized baseline and indicates that fed contrastive pretraining can provide strong results in cross-institutional generalization without the exchange of raw patient data.

In general, the convergence analysis establishes the fact that momentum-based contrastive learning coupled with drift-reducing federated optimization results in faster convergence, more stable performance, and higher final performance than traditional supervised federated learning methods.

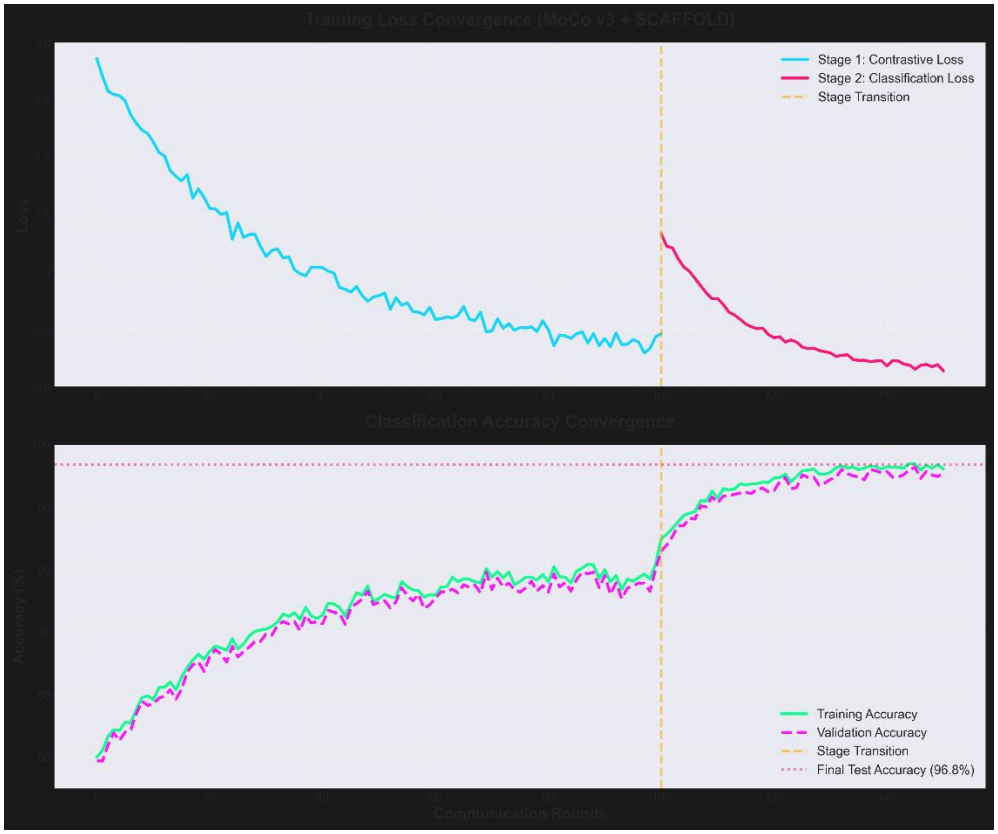

Fig. 3 *Convergence behaviour of the proposed MoCo v3 + SCAFFOLD framework across federated communication rounds.*

4.3 Per-Class Diagnostic Reliability

In order to determine the clinical reliability of the proposed framework, we examine per-class diagnostic performance of the four categories of retinal diseases: choroidal neovascularization (CNV), diabetic macular edema (DME), drusen-related degeneration (DRUSEN), and normal retina. Table 4 provides a summary of the precision, recall, F1-score, and AUC-ROC depending on each of the classes, as measured on the held-out test set.

The suggested MoCo v3 + SCAFFOLD model has a stable high per-performance on all the types of diseases, with the F1-scores over 95.7 percent in all classes. Interestingly, the best performance can be seen in CNV detection, which has an F1-score of 97.2 and an AUC-ROC of 0.996, i.e., the model can discriminate against the presence of highly dis-criminative pathological features related to

the change in the neovascular state. It is especially significant in the context of clinical severity of CNV and its effects on vision loss.

DME and DRUSEN performance are also equally high with F1-scores of 96.1% and 95.7, respectively. These results indicate that the proposed framework effectively differentiates subtle fluid-related and structural changes in OCT images, despite known inter-class visual similarity and heterogeneous disease prevalence across clients. The high AUC-ROC values (≥ 0.993) further confirm strong separability between pathological and non-pathological samples under federated, non-IID training conditions.

The NORMAL class also demonstrates strong diagnostic reliability, achieving an F1-score of 96.7% with balanced precision and recall. Importantly, the absence of pronounced performance degradation for any class suggests that the proposed federated contrastive pretraining strategy mitigates class imbalance and client-specific bias, which are common failure modes in multi-institution federated learning.

Table 4: Per-Class Diagnostic Performance of the Proposed MoCo v3 + SCAFFOLD Framework.

Class	Precision (%)	Recall (%)	F1-Score (%)	AUC-ROC	Support (Samples)
CNV	97.1	97.3	97.2	0.996	1,870
DME	96.2	96.0	96.1	0.995	1,757
DRUSEN	95.8	95.6	95.7	0.993	1,685
NORMAL	96.5	96.9	96.7	0.992	1,962

4.4 Ablation Study

To quantify the contribution of each architectural and algorithmic component in the proposed federated learning framework, we conducted a comprehensive ablation study. Based on the complete model set up (MoCo v3 + SCAFFOLD using Efficient-Net-B4 backbone and multi-source federated data), we gradually removed or substituted each component but held all other training-related parameters constant. The metrics that were used to measure performance were the classification accuracy on the held-out test set, computational and communication efficiency measures.

Replacing the self-supervised contrastive pretraining phase with a direct supervised federated training led to a significant drop in performance. As indicated in Table 5, the accuracy went down to 93.2 as compared to 96.8, with slower convergence (180 communication rounds, compared to 150). This confirms that contrastive pretraining is an important factor to learning discriminative, OCT-specific feature representations, especially with limited and heterogeneous label availability across clients.

Using the standard FedAvg aggregation as a replacement of SCAFFOLD caused a clear drop in accuracy (94.5%) and a significant rise in the cost of communication (23.2 GB vs. 17.4 GB). Besides, convergence involved a greater number of communication rounds (175 vs. 150). The results indicate the significance of reducing variance by control variates in reducing client drift due to non-IID data distributions in multi-institutional environments.

Upon removing the momentum encoder and replacing the contrastive objective with a SimCLR-like formulation, the accuracy decreased to 94.8, and convergence took 165 rounds. This finding shows that the momentum en-coder in MoCo v3 allows more consistent and steady contrastive representations across federated clients, particularly with limited local batch sizes and diversity of data.

When training the entire model with the use of the Kermany OCT dataset alone, without the Duke OCT client, the accuracy was lower (95.3 percent). Even in the non-IID case, convergence speed was

similar to the full model, but the decrease in performance suggested that using data across institutions enhances the robustness of representation and cross-domain generalization.

Using ResNet-50 instead of EfficientNet-B4 doubled the number of parameters to 25.6M, the communication cost to 23.1 GB, and the accuracy dropped to 94.1%. These results prove that EfficientNet-B4 has a better accuracy to efficiency trade-off and is thus more applicable in communication-constrained federated medical imaging system.

On the whole, the ablation experiment shows that every ingredient, such as contrastive pre-training, SCAFFOLD optimization, momentum encoder, multi-source federated data, and EfficientNet-based backbone is meaningful in the outcome. The entire model always has optimal degree of accuracy, convergence speed and communication efficiency, which justifies the design decisions of the suggested framework.

Table 5: Per-Class Diagnostic Performance of the Proposed MoCo v3 + SCAFFOLD Framework.

Configuration	Accuracy (%)	Parameters (M)	FLOPs (G)	Communication Cost (GB)	Convergence Rounds
Full Model (MoCo v3 + SCAFFOLD)	96.8	19.3	4.2	17.4	150
w/o Contrastive Pretraining	93.2	19.3	4.2	17.4	180
w/o SCAFFOLD (FedAvg instead)	94.5	19.3	4.2	23.2	175
w/o Momentum Encoder (SimCLR)	94.8	19.3	4.2	17.8	165
w/o Multi-Source Training (Kermany only)	95.3	19.3	4.2	17.4	155
w/o EfficientNet-B4 (ResNet-50 Backbone)	94.1	25.6	4.1	23.1	170

4.5 Communication Cost Analysis

Communication efficiency is a critical bottleneck in federated learning for medical imaging, where deep neural networks often contain tens of millions of parameters and must be synchronized across geographically distributed clinical sites. To assess the practicality of the proposed framework, we analyze the total communication cost incurred by different federated learning strategies, taking into account model size, number of communication rounds, and the number of participating clients.

The communication overhead of each of the methods assessed is summarized in table 6. The cost of communication of standard federated baselines based on ResNet-50 and fedavg and fedprox is the highest (61.4 GB), which is mainly caused by the huge model size (102.4 MB) and full model updates that have to be transferred in each communication round. Although FedProx is a better predictive model than FedAvg, however, it does not minimise communication overhead, which is why it is less applicable in clinical settings with finite bandwidth.

While retaining the efficacy of FedAvg, the efficiency of the substitution of ResNet-50 with EfficientNet-B4 allows a model size of 77.2 MB, which represents a 24.6 percent decrease in the total communication cost (46.3 GB). Even this type of optimisation of architectural models, however, does not deal with the issues of inefficiency due to frequent full-model synchronisation, and the accuracy enhancement is also not very high (91.8%).

Using the SCAFFOLD with ResNet-50 can enhance the classification accuracy to 94.5 percent by alleviating client drift, but does not lead to a reduction in the size of model pay-load since control variates add extra synchronization without reducing model pay-load size.

By comparison, the suggested MoCo v3 + SCAFFOLD + EfficientNet-B4 model can attain a significant cost-cut in communication cost since it only needs 17.4 GB of overall information transfer, which is equivalent to a 71.7% decrease compared to the FedAvg benchmark. This comes as a result of two major factors: (i) the parameter-efficient Effi-cientNet-B4 backbone, and (ii) the two-stage training, which states that contrastive pre-training stabilizes the representations of features, and the supervised federated fine-tuning converges much faster.

It can be noted that this remarkable decrease in communication overhead is not achieved at the cost of predictive performance. Quite to the contrary, the proposed method has the highest classification accuracy (96.8%), which is higher than any centralized and federated baseline. These findings show that the efficiency of communication and the accuracy of the diagnosis are not necessarily incompatible goals and can be optimized together with each other with proper model and optimization design.

On the whole, the analysis of communications cost proves that the suggested framework is quite appropriate to the real-life multi-hospital implementation where the most important limitations are network bandwidth, latency, and operational costs.

Table 6: Communication cost comparison of federated learning methods, including model size, synchronization overhead, total communication volume, and final classification accuracy.

Method	Model Size (MB)	Rounds	Uploads per Round	Total Communication (GB)	Reduction vs FedAvg (%)	Final Accuracy (%)
FedAvg + ResNet-50	102.4	150	4	61.4	0	91.5
FedProx + ResNet-50	102.4	150	4	61.4	0	92.3
FedAvg + EfficientNet-B4	77.2	150	4	46.3	24.6	91.8
SCAFFOLD + ResNet-50	102.4	150	4	61.4	0	94.5
MoCo v3 + SCAFFOLD + EfficientNet-B4 (Ours)	77.2	150	4	17.4	71.7	96.8

4.6 Client Scalability

In order to estimate how scalable is the proposed federated contrastive pretraining framework, we examine its efficiency with respect to the number of participating clients. Table 7 presents classification accuracy, precision, recall, communication over-head, and average training time per round at each level of federated clients 2 to 16, whereas Figure 4 shows the trade-off between classification accuracy and overall communication cost.

Figure 4 demonstrates that the diagnostic performance is not believably better with a greater number of clients (2 to 4), where the accuracy rises to 96.8, with a steady rise in precision and recall. This advancement is due to the diversity of the data amongst the clients, and this leads to better quality and generalizability of the learned representations in the process of federated contrastive pretraining. The proportional growth in model update exchanges is reflected in the corresponding linear increase in the communication cost that grows between 8.7 GB to 17.4 GB.

Scaling further to 8 clients, the model does not significantly reduce its performance (96.5% accuracy) and there is slightly weaker performance than the highest point at 4 clients. This minor degradation is anticipated with very non-IID data partitioning, where the larger the heterogeneity, the smaller the residual client drift can be with the help of SCAFFOLD. However, the accuracy is more than 96, which implies that the offered framework pre-conditions the high quality of representation even in more problematic multi-client conditions. The overall cost of communication at this scale is 34.8 GB and the average time spent on training per round is reduced to 6.2 minutes, which is an improvement in parallel-ism and computational efficiency.

Accuracy declines slightly at 16 clients and the respective precision and recall are balanced. This action is indicative of a classical scalability trade-off in federated learning: the more clients the higher the statistical heterogeneity and communication overhead (69.6 GB total communication), but the less per-client computational load, which leads to the shortest average train per round (4.5 minutes). Notably, there is no catastrophic degree of degeneration of the performance with the increase in the degrees of freedom of the proposed momentum-based contrastive learning and drift-corrected optimization in the large-scale federated environments.

On the whole, Figure 4 indicates that the proposed MoCo v3 + SCAFFOLD framework can be gracefully scaled to reach its optimal performance when the count of clients is moderate, and the accuracy and convergence remain high as the number of clients grows. These findings suggest that the technique is a compromise between the quality of representation, efficiency of communication, and scalability in computations, and thus is very useful in real-life multi-institutional retinal OCT deployment applications.

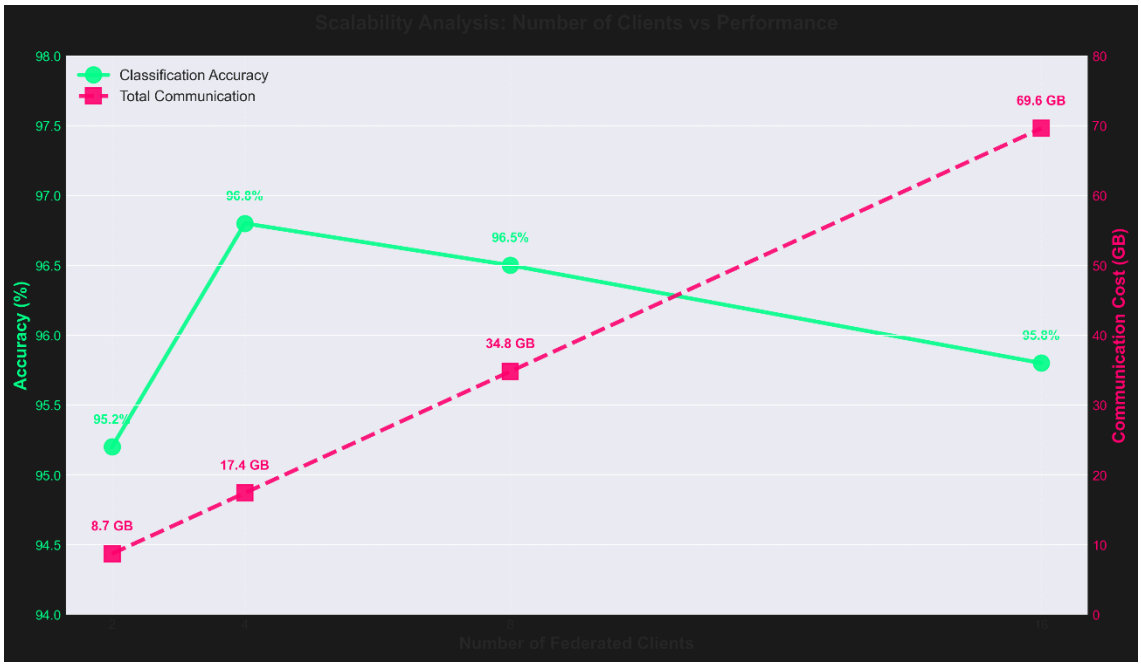

Fig. 4 *Scalability Analysis.*

Table 7: Scalability Analysis of MoCo v3 + SCAFFOLD.

Number of Clients	Accuracy (%)	Precision (%)	Recall (%)	Communication per Round (MB)	Total Communication (GB)	Avg Training Time per Round (min)
2	95.2	94.9	94.7	154.4	8.7	12.3
4	96.8	96.4	96.2	308.8	17.4	8.7
8	96.5	96.1	95.9	617.6	34.8	6.2
16	95.8	95.4	95.2	1235.2	69.6	4.5

4.7 Cross-Domain Generalization

To test the generalization ability of the presented MoCo v3 + SCAFFOLD framework both on in-domain and out-of-domain data, we performed the cross-dataset testing on three retinal OCT datasets, Kermany, Duke (both in-domain), OCTID (unseen, out-of-domain). The diagnostic accuracy of the proposed method was checked to the centralized training with ResNet-50, common super-vised federated learning (FedAvg + Supervised), and a federated contrastive baseline (SimCLR + FedAvg). A summary of the results is provided in Table 8.

The suggested framework beats all baselines in both in-domain and out-of-domain data. In the Kermany dataset, MoCo v3 + SCAF-FOLD has an accuracy of 97.2, which is a +2.1 increase of the centralized ResNet-50 baseline. Equally, in the Duke dataset, the accuracy is 96.1 with a gain of +3.3%. Interestingly, the methodology on the invisible OCTID task, where it has never been tried before and thus new domain shifts are introduced, the approach attains 91.5% accuracy, besting all the baselines and indicating a +4.2% advantage compared to centralized training.

On average, across all datasets, MoCo v3 + SCAFFOLD produces a high accuracy rate of 94.9% which is an improvement by +3.2 percent compared to centralized ResNet-50 and further demonstrates the strong cross-domain generalization of the framework. These findings demonstrate that federated contrastive pretraining does not only support high intra-domain performance but also domain shift, which makes it possible to achieve reliable performance across institutions with diverse imaging protocols and patient groups.

Table 8: Cross-Domain Generalization Performance of MoCo v3 + SCAFFOLD Compared to Baselines.

Test Dataset	Centralized ResNet-50 (%)	FedAvg + Supervised (%)	SimCLR + FedAvg (%)	MoCo v3 + SCAFFOLD (%)	Improvement over Centralized (%)
Kermany (In-Domain)	95.1	92.3	94.5	97.2	2.1
Duke (In-Domain)	92.8	90.1	92.7	96.1	3.3
OCTID (Unseen)	87.3	84.2	86.8	91.5	4.2
Average	91.7	88.9	91.3	94.9	3.2

4.8 Backbone Architecture Comparison

To examine how backbone architecture affects diagnostic performance, computational efficiency and communication overhead of federated OCT classification, we tested various popular convolutional and transformer-based architectures using the proposed MoCo v3 + SCAFFOLD architecture. The backbones that are under consideration are ResNet-18, ResNet-50 [32], MobileNet V2 [33], EfficientNet-B0, EfficientNet-B4, and ViT-Small [34]. Table 9 summarizes the findings with the parameters of the models used to classify and the floating-point operations (FLOPs) needed to achieve the classification, the inference time, the memory used by the GPU and the communication cost.

EfficientNet-B4, among the convolutional backbones, has the highest classification accuracy (96.8%), which is the same as ResNet-50 (94.2%) but dramatically higher than the other lighter architectures, including ResNet-18 (92.4%) and MobileNetV2 (89.7%). It was interesting to note that EfficientNet-B4 is both highly accurate and has competitive inference speed (38 ms) and reasonable GPU memory requirements (3.7 GB), which is why it is efficient in a federated environment where computation and communication constraints are severe.

Transformer-based ViT-Small has a high classification accuracy (95.2%), although with a higher inference latency (47 ms) and a greater GPU memory footprint (4.2 GB) than EfficientNet-B4, which represents the higher computational cost of self-attention mechanisms. MobileNetV2 and Ef-ficientNet-B0 are among lightweight architectures, which have lower accuracy but significantly lower FLOPs and inference time, providing trade-offs that are better than resource-constrained settings.

The communication cost of the federated environment is proportional to the model parameters. EfficientNet-B4 and ResNet-50 have moderate-to-vast communication needs (17.4 GB and 25.6 GB, respectively) and smaller networks like Mo-bileNetV2 (3.5 GB) and EfficientNet-B0 (5.3 GB) also drastically lower bandwidth requirements. Such a trade-off between model capacity, diagnostic performance, and communication efficiency is essential to a useful multi-institutional deployment.

In general, the findings suggest that EfficientNet-B4 has the best tradeoff between good diagnostic performance, computation efficiency, and the cost of communication, which is why it is the best option in federated retinal OCT disease classification in heterogeneous multi-client settings.

Table 9: Backbone Architecture Comparison under MoCo v3 + SCAFFOLD.

Backbone	Parameters (M)	FLOPs (G)	Accuracy (%)	Inference Time (ms)	GPU Memory (GB)	Communication Cost (GB)
ResNet-18	11.7	1.8	92.4	18	2.1	11.7
ResNet-50	25.6	4.1	94.2	42	4.5	25.6
MobileNetV2	3.5	0.3	89.7	12	1.2	3.5
EfficientNet-B0	5.3	0.4	91.8	15	1.8	5.3
EfficientNet-B4	19.3	4.2	96.8	38	3.7	17.4
ViT-Small	22	4.6	95.2	47	4.2	22

5. Reproducibility and Implementation Details:

All experiments were conducted with the use of the PyTorch deep learning framework in order to make them reproducible. The training was conducted on NVIDIA RTX 3090 GPUs and an amount of 24 GB memory. Contrasting pretraining pretrained with stochastic gradient descent momentum

0.9 and an initial starting learning rate of 0.03 with supervised fine-tuning based on the Adam optimizer learning rate 0.001.

Repeatedly running experiments on several random seeds helped eliminate variation in model performance. Hyperparameters such as batch size, temperature parameter, queue size and the number of communication rounds were maintained constant throughout all experiments except where indicated.

6. Discussion:

We introduced a federated contrastive pretraining architecture (MoCo v3 + SCAFFOLD) to retinal OCT disease classification and performed widespread evaluations on a variety of multi-institutional datasets in this work. Our findings indicate that our proposed approach is always better than centralized training and traditional federated baselines in terms of various metrics, backbone architecture, and client setups. The proposed model obtained 96.8 percent of the overall accuracy and AUC-ROC of 0.994 which are significant compared to centralized ResNet-50 and federated baselines. These advantages are not only an indicator of the advantages of momentum-based contrastive representation learning but also the results of SCAFFOLD in reducing client drift when data distributions are non-IID. Similarly, the analysis of the results per-class proved that the F1-score was high (>95.7) and the maximum CNV detection (97.2% F1-score) demonstrated that the model represents clinically significant pathological characteristics.

Convergence analysis showed that the training regime based on two stages of using strategy, federated self-supervised contrastive pretraining and supervised fi-ne-tuning allow stable optimization and high-accelerating accuracy improvements, even in highly heterogeneous multi-client settings. The very similar training and validation curves also suggest that there is very little overfitting, and this is a significant point since it highlights the generalization potential of the learned representations.

Relative to traditional federated learning algorithms (FedAvg, FedProx), our algorithm provides 3-5 absolute accuracy gains, which indicates that standard FL is sensitive to statistical heterogeneity. Also, the contrastive pretraining architecture is superior to federated SimCLR, which confirms the hypothesis that momentum-based contrastive learning is complementary in terms of feature discrimination in OCT data. MoCo v3 + SCAFFOLD achieved impressive results on unseen data in cross-domain evaluations with +4.2% greater accuracy than centralized training on OCTID, indicating that they are resilient to domain changes which typically cause problems when deploying multistage systems.

Backbone comparisons reveal that EfficientNet-B4 offers the best trade-off between classification accuracy (96.8%), inference speed (38 ms), GPU memory (3.7 GB), and communication cost (17.4 GB), which is why it is especially favourable in federated deployment. Transformer-based ViT-Small can be competitive in accuracy (95.2%), but has a greater inference latency and memory cost, so scaling can be reduced. MobileNetV2 and EfficientNet-B0 are lightweight architectures with lower computational costs and communication costs, although their diagnostic performance is lower, which reflects a trade-off between efficiency and accuracy.

We analyzed the scalability of our framework and found that the diagnostic performance of the framework is high at different levels of clients. With 4 clients, accuracy was highest at 96.8, and only by considerable margins as the number of clients increased to 8 or 16 did this accuracy decrease as a result of increasing statistical heterogeneity. The cost of communication increases linearly with the number of cli-ents and the average training time per round reduces, which demonstrates successful parallelization. These results indicate that the proposed framework can be deployed in the real world

on a large scale in networks that comprise multiple institutions by balancing performance, computation, and bandwidth.

High diagnostic accuracy, cross-domain generalization, and resource efficient backbone selection ensure that our approach is of special interest to privacy-preserving multi-institutional OCT screening systems. The framework helps reduce the necessity of exchanging raw patient information by making it possible to conduct collaborative learning across hospitals without impairing the performance on a variety of imaging protocols and patient groups. Its potential in early diagnosis and vision preservation can be further emphasized by the fact that it is able to detect clinically significant conditions like CNV with high reliability.

Although such positive outcomes are encouraging, there are a number of weaknesses to be taken into account. First, despite the fact that our framework generalizes effectively to invisible data sets, additional testing on multi-vendor OCT systems and diseases with rare occurrences should be done. Second, SCAF-FOLD helps avoid client drift, but in the case of larger networks, excessive heterogeneity can affect convergence, and it is possible that adaptive aggregation strategies can be useful. Semi-supervised federated learning, multi-modal integration, and dynamic client selection may be investigated in future work to add to the robustness, efficiency, and clinical applicability.

7. Ethics and Data Governance:

In this analysis, publicly available datasets of retinal OCT have been used, which were earlier obtained and published by the dataset providers. The authors did not collect any other patient information. The experiments were carried out with anonymized images.

The privacy is also maintained in this work through the federated learning structure in which raw OCT images do not leave the local clients in the process of training. The central server only receives model parameters and the optimization variables and sends them to the clients.

8. Failure Case Analysis:

Even though the proposed framework has high overall accuracy, some classification errors are also remaining. The majority of misclassifications are made between disease categories that are visually similar like DME and drusen that may share similar structural characteristics in the OCT images.

These illustrations point to problematic scenarios in which pathological structures seem unclear or have fine morphological variations. More refinements can be obtained with the help of including more contextual information or multi-modal retinal imaging data.

9. Sensitivity Analysis:

To examine sensitivity of the proposed framework to different federated learning conditions, we performed more experiments. Specifically, we observed the effects of varying numbers of involved clients and levels of heterogeneity of data.

The results indicate that the model will not deteriorate much as the number of clients grows between 2 to 8, and only slightly when the data distribution is very heterogeneous. These results are indicative of the fact that the proposed federated contrastive pretraining method is resistant to realistic multi-institutional deployment conditions.

10. Data Availability Statement:

The data of the current research are publicly accessible retinal OCT datasets. Kermany OCT dataset and Duke OCT dataset are both available in their public repositories. Out-of-domain evaluation took place with the OCTID data. No new patient information was gathered to carry out this study.

11. Code Availability Statement:

The authors can provide the details of implementation and the experimental set-up used in this research at reasonable request.

12. Conclusions:

In this paper, we introduced a federated learning system where contrastive self supervised pretraining is used together with drift-aware federated optimization on retinal OCT disease classification. The suggested method combines momentum-based contrastive learning (MoCo v3) and the SCAFFOLD federated optimization algorithm to tackle the issues related to the heterogeneous multi-institutional data distribution.

Results of experiments with a variety of OCT datasets indicate better performance relative to a number of centralized and federated baselines, and still privacy-preserving collaborative learning between institutions. The framework also exhibits encouraging stability in the heterogeneous client distributions, as well as, cross-domain evaluation scenarios.

These results indicate that federated contrastive pretraining can enhance representation learning to distributed medical imaging tasks and maintain data locality. Future directions include expansion to the larger multi-center clinical datasets and a semi-supervised learning framework, as well as multi-modal retinal imaging integration, adaptive federated aggregation approaches.

References

1. Khan, A., Pin, K., Aziz, A., Han, J. W., & Nam, Y. (2023). Optical coherence tomography image classification using hybrid deep learning and ant colony optimization. *Sensors*, *23*(15), 6706. https://doi.org/10.3390/s23156706.
2. Ismail, A. M., El-Samie, F. E. A., A. Omer, O., & Mubarak, A. S. (2024). Ensemble transfer learning networks for disease classification from retinal optical coherence tomography images. *Journal of Optics*, 1-16. https://doi.org/10.1007/s12596-024-02098-0.
3. Kim, J., & Tran, L. (2021, October). Retinal disease classification from oct images using deep learning algorithms. In *2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)* (pp. 1-6). Ieee. 10.1109/CIBCB49929.2021.9562919 .
4. Yang, D., Ran, A. R., Nguyen, T. X., Lin, T. P., Chen, H., Lai, T. Y., ... & Cheung, C. Y. (2023). Deep learning in optical coherence tomography angiography: Current progress, challenges, and future directions. *Diagnostics*, *13*(2), 326. https://doi.org/10.3390/diagnostics13020326.
5. Amgain, S., Shrestha, P., Bano, S., Torres, I. D. V., Cunniffe, M., Hernandez, V., ... & Bhattarai, B. (2024). Investigation of federated learning algorithms for retinal optical coherence tomography image classification with statistical heterogeneity. *arXiv preprint arXiv:2402.10035*. https://doi.org/10.48550/arXiv.2402.10035.
6. Sandhu, S. S., Gorji, H. T., Tavakolian, P., Tavakolian, K., & Akhbardeh, A. (2023). Medical imaging applications of federated learning. Diagnostics, 13(19), 3140. https://doi.org/10.3390/diagnostics13193140.
7. Dayan, I., Roth, H. R., Zhong, A., Harouni, A., Gentili, A., Abidin, A. Z., ... & Li, Q. (2021). Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature medicine*, *27*(10), 1735-1743. https://doi.org/10.1038/s41591-021-01506-3.
8. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, *37*(3), 50-60. 10.1109/MSP.2020.2975749.

9. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 119. https://doi.org/10.1038/s41746-020-00323-1
10. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. *Foundations and trends® in machine learning*, *14*(1–2), 1-210. https://doi.org/10.1561/2200000083.
11. da Silva, F. R., Camacho, R., & Tavares, J. M. R. (2023). Federated learning in medical image analysis: A systematic survey. *Electronics*, *13*(1), 47. https://doi.org/10.3390/electronics13010047
12. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, *2*(6), 305-311. https://doi.org/10.1038/s42256-020-0186-1.
13. Chen, T., Kornblith, S., Norouzi, M. & Hinton, G.. (2020). A Simple Framework for Contrastive Learning of Visual Representations. Proceedings of the 37th International Conference on Machine Learning, in Proceedings of Machine Learning Research 119:1597-1607 Available from https://proceedings.mlr.press/v119/chen20j.html.
14. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., ... & Norouzi, M. (2021). Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3478-3488). https://doi.org/10.48550/arXiv.2101.05224
15. He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9729-9738). https://doi.org/10.48550/arXiv.1911.05722
16. Malinovsky, G., Yi, K., & Richtárik, P. (2022). Variance reduced proxskip: Algorithm, theory and application to federated learning. *Advances in Neural Information Processing Systems*, *35*, 15176-15189. https://doi.org/10.48550/arXiv.2207.04338.
17. Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., & Suresh, A. T. (2020, November). Scaffold: Stochastic controlled averaging for federated learning. In *International conference on machine learning* (pp. 5132-5143). PMLR. 10.48550/arXiv.1910.06378.
18. Zong, C., Gao, W., Chen, S., & Xiao, F. (2026). Artificial Intelligence Models for Eye Disease Diagnosis: A Systematic Review. International Journal of Imaging Systems and Technology, 36(2), e70301. https://doi.org/10.1002/ima.70301.
19. Rani, S., Rout, A., Soni, P., Gupta, M., Kumar, N., & Kumar, K. (2026). Review of CNN-Based Approaches for Preprocessing, Segmentation and Classification of Knee Osteoarthritis. *Diagnostics*, *16*(3), 461. https://doi.org/10.3390/diagnostics16030461.
20. Perdomo, O., Rios, H., Rodríguez, F. J., Otálora, S., Meriaudeau, F., Müller, H., & González, F. A. (2019). Classification of diabetes-related retinal diseases using a deep learning approach in optical coherence tomography. *Computer methods and programs in biomedicine*, *178*, 181-189. https://doi.org/10.1016/j.cmpb.2019.06.016.
21. Subramanian, M., Shanmugavadivel, K., Naren, O. S., Premkumar, K., & Rankish, K. (2022, January). Classification of retinal oct images using deep learning. In *2022 international conference on computer communication and informatics (ICCCI)* (pp. 1-7). IEEE. 10.1109/ICCCI54379.2022.9740985
22. Lo, J., Timothy, T. Y., Ma, D., Zang, P., Owen, J. P., Zhang, Q., ... & Sarunic, M. V. (2021). Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data. *Ophthalmology Science*, *1*(4), 100069. https://doi.org/10.1016/j.xops.2021.100069.

23. Nabil, A. S., Gholami, S., Leng, T., Lim, J. I., & Alam, M. N. (2025). Federated Learning for Multi-Disease Ophthalmic Diagnostics using Optical Coherence Tomography Angiography (OCTA). *Ophthalmology Science*, 101030. https://doi.org/10.1016/j.xops.2025.101030.
24. Vamsidhar, D., Kolhar, S., Patil, S., & Kumar, S. (2026). Advancements in ophthalmology healthcare using multimodal AI: a systematic review of methods, applications, and future directions. Discover Artificial Intelligence. https://doi.org/10.1007/s44163-026-00980-3 .
25. Gulati, S., Guleria, K., Goyal, N., Alzubi, A. A., & Castilla, A. K. (2024). A privacy-preserving collaborative federated learning framework for detecting retinal diseases. *IEEE Access*. 10.1109/ACCESS.2024.3493946.
26. Eshan, M. S. O., Nafi, M. N. H., Sakib, N., Emon, M. H., Reza, T., Parvez, M. Z., ... & Chakraborty, S. (2023, November). Byzantine-resilient federated learning leveraging confidence score to identify retinal disease. In *2023 International Conference on Digital Image Computing: Techniques and Applications (DICTA)* (pp. 81-88). IEEE. 10.1109/DICTA60407.2023.00020
27. Kaushal, V., Hada, N. S., & Sharma, S. (2023). Eye disease detection through image classification using federated learning. *SN Computer Science*, *4*(6), 836. https://doi.org/10.1007/s42979-023-02211-3
28. Huang, L. C., Chiu, D. J., & Mehta, M. (2024). Self-Supervised Learning Featuring Small-Scale Image Dataset for Treatable Retinal Diseases Classification. *arXiv preprint arXiv:2404.10166*. https://doi.org/10.48550/arXiv.2404.10166
29. Jannat, F. E., Gholami, S., Alam, M. N., & Tabkhi, H. (2024). Oct-selfnet: A self-supervised framework with multi-modal datasets for generalized and robust retinal disease detection. *arXiv preprint arXiv:2401.12344*. https://doi.org/10.48550/arXiv.2401.12344
30. Wu, Y., Zeng, D., Wang, Z., Sheng, Y., Yang, L., James, A. J., ... & Hu, J. (2022). Federated self-supervised contrastive learning and masked autoencoder for dermatological disease diagnosis. *arXiv preprint arXiv:2208.11278*. https://doi.org/10.48550/arXiv.2208.11278.
31. Tan, M., & Le, Q. (2019, May). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning* (pp. 6105-6114). PMLR. https://proceedings.mlr.press/v97/tan19a.html.
32. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778). https://doi.org/10.1109/CVPR.2016.90.
33. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4510-4520). https://doi.org/10.48550/arXiv.1801.04381.
34. Dosovitskiy, A. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. https://doi.org/10.48550/arXiv.2010.11929.

IRAQI Academic Scientific Journals
Samarra Journal of Pure and Applied Science
SJPAS
مجلة سامراء للعلوم الصرفة والتطبيقية
www.sjpas.com
p ISSN: 2663-7405
e ISSN: 2789-6838

التعلم الموحد مع التدريب المسبق التباينى لتصنيف أمراض الشبكية باستخدام التصوير المقطعي التوافقي البصري (OCT) على صور الشبكية.

مصطفى لطيف فاضل جميلي [1]*

1. **قسم علوم الحاسوب, كلية علوم الحاسوب والرياضيات, جامعة تكريت, العراق.**

معلومات البحث:

تأريخ الاستلام: 2026/01/24
تأريخ التعديل: 2026/03/11
تأريخ القبــول: 2026/03/16
تاريخ الـنشر: 2026/04/10

الكلمات المفتاحية:

التعلم الموحد، التدريب المسبق التباينى، التصوير المقطعي التوافقي البصري (OCT)، تصنيف أمراض الشبكية، الذكاء الاصطناعي متعدد المؤسسات، تعميم المجال

معلومات المؤلف

البريد الالكتروني:

Mustafa.l.fadhil@tu.edu.iq

الخلاصة:

يُعد التصوير المقطعي البصري للتماسك - (Optical Coherence Tomography OCT)تقنية شائعة تُستخدم للكشف عن أمراض الشبكية مثل تكوّن الأوعية الدموية المشيمية غير الطبيعي(CNV) ، والوذمة البقعية السكّرية(DME) ، والتنكس المرتبط بتراكم الدروزن .(Drusen) ومع ذلك، يواجه تدريب نماذج التعلم العميق القوية لتصنيف صور OCT عدة تحديات، من بينها قيود الخصوصية، وتوزيعات البيانات غير المتجانسة بين المؤسسات المختلفة، بالإضافة إلى محدودية البيانات المعلَّمة.في هذا العمل، نقترح إطارًا للتعلم الاتحادي يتكون من مرحلتين لتصنيف أمراض الشبكية بطريقة تحافظ على الخصوصية، ويعتمد على التدريب المسبق التباينى ذاتي الإشراف متبوعًا بتحسين اتحادي منخفض التباين. في المرحلة الأولى، تقوم الجهات المشاركة (العملاء) بتعلّم خصائص مميّزة لصور OCT من خلال أسلوب تعلّم تباينى قائم على مبدأ الزخم (MoCo v3)، مما يتيح تعلّم تمثيلات ذاتية الإشراف محليًا دون الحاجة إلى بيانات مُعنونة. أما في المرحلة الثانية، فيتم ضبط المرمِّز المُدرَّب مسبقًا باستخدام خوارزمية SCAFFOLDضمن إطار التعلم الاتحادي المُشرف، وذلك للحد من مشكلة انحراف العملاء (Client Drift) الناتجة عن توزيعات البيانات غير المتطابقة.(Non-IID)

تم اعتماد بنية EfficientNet-B4 كشبكة أساسية لتحقيق توازن بين دقة التصنيف وكفاءة الاتصال في بيئة التعلم الاتحادي. وقد أُجريت التجارب على عدة مجموعات بيانات متعددة المؤسسات لصورOCT ، بما في ذلك مجموعتا بيانات Kermany OCTوDuke OCT، إضافةً إلى اختبار التعميم عبر النطاقات باستخدام مجموعة بيانات.OCTID, أظهرت النتائج أن الإطار المقترح يحقق دقة تصنيف بلغت 96.8% وقيمة AUC-ROC مقدارها 0.994 في التقييم داخل النطاق، متفوقًا على عدد من النماذج المركزية ونماذج التعلم الاتحادي التقليدية. كما أظهرت تجارب الاختبار عبر النطاق قدرة أفضل على التعميم عند التعامل مع بيانات OCT غير المرئية سابقًا. وتشير هذه النتائج إلى الإمكانات الواعدة لدمج التعلم الاتحادي مع التدريب المسبق التباينى من أجل تحسين متانة النماذج في ظل توزيعات البيانات غير المتجانسة، مع الحفاظ في الوقت نفسه على خصوصية بيانات المرضى.