International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

A Privacy-Preserving Federated Learning Framework for Multi-Class Diabetic Retinopathy Grading and Clinical Deployment

Hittarth Goyal , Pranjal Tomar, Prateek Jain

Abstract—Diabetic Retinopathy (DR) is a prominent microvascular complication of diabetes and a leading cause of preventable blindness worldwide. While Deep Learning, specifically Convolutional Neural Networks (CNNs), has established itself as the gold standard for automated retinal screening, the centralization of sensitive medical imagery for model training violates stringent data privacy regulations (e.g., HIPAA, GDPR, DPDP). Further-more, practical clinical deployment requires grading severity rather than simple binary detection. This paper presents a highly comprehensive, privacy-preserving distributed computing framework and an integrated Clinical Decision Support System (CDSS) designed for 5-class multi-label DR diagnosis. Leveraging Federated Learning (FL), the architecture empowers heteroge-neous healthcare institutions to collaboratively train a modified ResNet-50 architecture without transferring raw patient data. We address the critical bottleneck of communication overhead and non-IID (Independent and Identically Distributed) clinical data distributions by introducing a partial fine-tuning strat-egy—freezing all but the final 15 layers of the ResNet-50 back-bone. Evaluated on the APTOS 2019 Blindness Detection Dataset (3,662 retinal fundus images) distributed across 10 simulated clinical clients, the federated model's efficacy is tested under both IID and Non-IID (Quantity and Label Skew) conditions. The proposed model achieved an exceptional accuracy of ∼99.00% in IID settings and demonstrated high resilience by maintaining ∼96.00% accuracy under severe Non-IID quantity and label skew constraints. To bridge the gap between theoretical modeling and clinical application, the system incorporates a full-stack Flask web application featuring robust user authentication and real-time diagnostic analytics. This end-to-end system provides a scalable, privacy-compliant solution for comprehensive DR grading, effectively eliminating raw data leakage risks.

Index Terms—Diabetic Retinopathy, Federated Learning, Distributed Computing, ResNet-50, Non-IID Data, Partial Fine-Tuning, Privacy-Preserving AI.

I. Introduction

Diabetic Retinopathy (DR) is a serious and progressive form of microvascular complication that affects the blood vessels within the retina. According to WHO, diabetic retinopathy is still one of the major reasons for blindness worldwide. DR is divided into several levels of severity that include: no diabetic retinopathy, mild, moderate, severe non-proliferative diabetic retinopathy and proliferative diabetic retinopathy.

Early diagnosis of DR is essential in order to prevent the patient from experiencing serious complications that can ultimately cause visual impairment. Traditionally, an assessment of the retinal fundus images by specialized ophthalmologists is the key requirement of the current pipeline for diagnosis of DR. However, this process is extremely costly, time-consuming, and highly prone to inter-rater variability. In addition, the number of ophthalmologists per patient in developing countries is very low, causing further delays in diagnosing diabetic retinopathy.

The integration of Deep Learning (DL) into computer-aided diagnosis has revolutionized modern clinical workflows. Convolutional Neural Networks (CNNs) have repeatedly demonstrated expert-level accuracy in identifying pathological markers. However, training generalized, robust CNN models necessitates access to massive datasets. In the medical domain, consolidating raw patient data into a centralized data lake encounters severe legal and ethical bottlenecks. Global privacy frameworks, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the Digital Personal Data Protection Act (DPDP) in India, strictly regulate the movement and processing of Protected Health Information (PHI).

To overcome these centralized constraints, decentralized computing architectures have emerged as a highly secure paradigm. Federated Learning (FL) fundamentally reconfigures the traditional machine learning training pipeline. Instead of moving sensitive patient data to a central repository, FL brings the computational model directly to the localized data silos.

While FL inherently protects patient privacy, applying it to high-resolution ophthalmic imaging remains a complex open challenge due to the non-IID nature of clinical data. Different clinics see different volumes of patients and varying disease severity profiles, creating a phenomenon known as "Quantity and Label Skew." Furthermore, transmitting massive CNN architectures (like a 23-million parameter ResNet-50) over hospital networks iteratively introduces severe communication bottlenecks.

To address both the stringent privacy constraints of medical data and the engineering challenges of deployment, this paper introduces an optimized, end-to-end multi-class diagnostic system. The primary objectives and contributions are:

1) To create a distributed framework for training deep learning models that can effectively learn to classify diabetic retinopathy severity on the APTOS 2019 dataset, consisting of 3,662 images of patients and their corresponding severity classifications.
2) To minimize the amount of communication required to update the global model by using a partial fine-tuning strategy, specifically with ResNet-50 as the feature extraction backbone.
3) To evaluate and compare the performance of a global model through empirical means using standard fixed-data distribution metrics under both ideal IID conditions and realistic non-IID conditions involving extreme quantity and label skew.
4) To develop a full-stack clinical web application using the Flask framework that utilizes the finalized FL global model to enable medical practitioners to utilize the global model for patient diagnosis/management.

II. Background and Related Work

A. Deep Learning in Retinal Imaging

Historically, the method of performing automated screening of diabetic retinopathy (DR) has evolved from manual feature extraction techniques to using sophisticated deep learning approaches based on models such as convolutional neural networks (CNN) employing different architectures like ResNet, Inception, and VGG. The development of deep learning models with CNN architectures resulted in groundbreaking studies that demonstrated CNN models can achieve equivalent diagnostic accuracy to human ophthalmologists when trained on large, centralized datasets such as the Kaggle DR dataset, Messidor [3], and APTOS 2019 database.

Unlike traditional methods used in prior screening techniques, CNN-based architectures use unique hierarchical spatial feature extraction features to promote improved localization of features. The specific use of ResNet-50 [**?**] as a basis for developing CNN architectures in medicine is the introduction of residual skip connections to eliminate the effect of the vanishing gradient problem caused by increasing depth in complex architectures. However, grading NPDR into multiple classes remains a substantially more difficult task than simply detecting the presence or absence of NPDR simply due to the nature of the small morphological differences that occur between Mild, Moderate and Severe NPDR.

B. Limitations of Existing Centralized Systems

Centralized diagnostic systems today have severe structural limitations. For instance, the majority of commercial AI solutions require hospitals send the patient fundus scans to third-party cloud servers using REST APIs, significantly increasing the risk of data interception. Also, the assumption that one organization has access to a globally large dataset for training, which has often been demonstrated to be an inaccurate assumption for real-world heterogeneous health care systems.

There exists also a major gap in areas of Clinical Usability. Most published models developed in an academic setting, focus heavily on theoretical limit of accuracy and have limited/no usable Graphical User Interface (GUI) or longitudinal patient tracking, making them unusable for anyone except computer scientists.

C. Federated Learning in Healthcare

Federated Learning, originally introduced by McMahan et al. [4], directly addresses these vulnerabilities by decoupling model training from direct data access. FL has seen rapid adoption in edge computing and natural language processing [5].

Sheller et al. [6] established the feasibility of multi-institutional brain tumor segmentation using FL. Building upon these foundational works, this project applies decentralized federated aggregation specifically to multi-class retinal fundus imaging. We combine the mathematical rigor of FL with a robust client-server deployment model to address class imbalance and communication efficiency.

III. Mathematical Preliminaries and The Non-IID Challenge

A. The Federated Objective

In Federated Learning, the data is partitioned across K clients (clinics). Let D_k be the set of data points residing on client k, with $n_k = |D_k|$. The total number of data points is $N = \sum_{k=1}^{K} n_k$. The federated objective reformulates the global loss as a weighted average of local empirical risks:

$$\min_{w} F(w) = \sum_{k=1}^{K} \frac{n_k}{N} F_k(w) \tag{1}$$

where the local objective function for client k is $F_k(w) = \frac{1}{n_k} \sum_{i \in D_k} f_i(w)$.

B. Quantity and Label Skew

A fundamental challenge in federated medical imaging is that $F_k(w)$ could be an arbitrarily bad approximation of $F(w)$ if the data distribution $P_k(x, y)$ at clinic k is not representative of the global distribution $P(x, y)$. In real-world scenarios, clinics experience:

- **Quantity Skew:** Different clinics (k) have vastly different amounts of available training data (n_k). An urban hospital might have thousands of scans, while a rural clinic has dozens.
- **Label Skew:** The distribution of the target variable y (DR severity) varies across clinics. A specialized retina center will see far more ”Proliferative DR” cases than a general practitioner, who mostly sees ”No DR” cases.

Due to this combined ”Quantity and Label Skew”, the local model weights can diverge significantly during the local training epochs — also known as ”Client Drift” — making it much more difficult for the global model to converge.

IV. Proposed Federated Methodology

A. Dataset and Preprocessing Architecture

The model was trained on the APTOS 2019 Blindness Detection dataset, consisting of a total of 3,662 high-quality resolution retinal fundus images, divided into five separate severity classes. To ensure standardization across diverse clinical sensors, all images underwent a rigorous preprocessing pipeline.

Images were uniformally resized to a standard dimension of 224×224 pixels. To address illumination variances, a Gaussian blur filter was systematically applied. Pixel intensities across all RGB channels were scaled down to a $[0, 1]$ floating-point range to ensure numerical stability and accelerate gradient descent convergence.

B. Model Architecture and Partial Fine-Tuning

To perform multi-class grading locally at each participating edge node, we deployed a modified ResNet-50 Convolutional Neural Network. ResNet-50 consists of roughly 23 million parameters. Transmitting this massive weight matrix iteratively back and forth between 10 clients and a central server generates an unsustainable bandwidth overhead (roughly 100MB per client per round).

To mitigate this, we employed a **Partial Fine-Tuning** strategy. Pre-trained weights (initialized from large-scale image repositories) were loaded into the ResNet-50 backbone. We then deliberately ”froze” the weights of all layers except for the final 15 layers of the network. This strategic decision reduced the number of locally trainable parameters from approximately 23 Million down to approximately 5 Million. This allowed for much faster local convergence and significantly reduced the network communication payload, making the system viable for resource-constrained clinics.

The final classification layer was modified from a binary Sigmoid activation to a 5-node Softmax activation to output probability distributions across the 5 DR severity classes. Due to the inherent class imbalance in the APTOS dataset, the local models optimized a Weighted Cross-Entropy Loss function, penalizing misclassifications of rare, severe DR stages more heavily than the majority ”No DR” class.

C. Federated Averaging Orchestration

The system orchestrates training through a synchronous Federated Averaging (FedAvg) protocol. We defined a total network of $K = 10$ clients. In each communication round t, a subset consisting of 60% of the available clients was selected to participate.

The server broadcasts the global weights W_{t-1} to the selected clinics. Each clinic initializes its local weights and performs $E = 3$ local epochs of optimization using the Adam optimizer with a learning rate of $\eta = 1e-4$. The updated local weights w_t^k are securely transmitted back to the server. The central server aggregates these weights by computing a weighted average based on the volume of data samples n_k at each node.

To prevent overfitting, a central validation split of 10% was maintained, and Early Stopping was implemented with a patience of 10 communication rounds, terminating training globally if the Macro F1-score failed to improve. The global training was capped at a maximum of 100 communication rounds.

V. Data Distribution Strategies and Experimental Setup

To rigorously test the architecture, the 3,662 samples were distributed across the 10 clients using two distinct partitioning strategies to simulate both ideal and realistic clinical environments.

A. IID Distribution (Ideal Scenario)

For the IID setup, ”client datasets” were produced by randomly ”shuffling” the training data and splitting it up evenly between 10 different clients (”subsets”). As a result, each of the 10 clients received datasets that had equal proportions of all of the five class types across their total number.

B. Non-IID Distribution (Quantity and Label Skew)

In comparison, for the Non-IID setup, clients were given data that had both quantity and label distributions that were skewed. This non-uniform distribution (i.e., label/quantity skew across clients) is demonstrated by the use of heatmapping in Fig. 1 and the contrasting data provided by each of the 10 clients in comparison to all other clients as demonstrated by Fig. 2.

Fig. 1. Heatmap showing Non-IID Client Class Composition across ten simulated physical clinical nodes.

Fig. 1 provides a heatmap visualization of this Non-IID class composition across the 10 clients, while Fig. 2 demonstrates the sheer disparity in the quantity of data held by each individual clinic node.

International Journal of Creative and Open Research in Engineering and Management
ISSN: **3108-1754 (Online)**
Volume 02 Issue 05 May-2026 | Impact Factor: 3.5

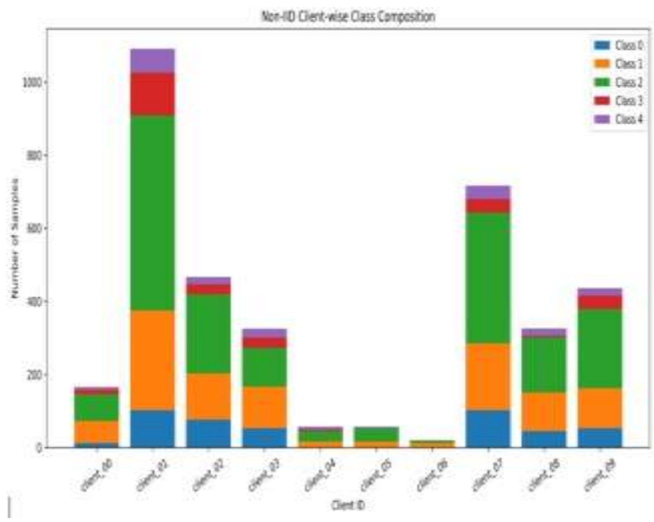

Fig. 2. Bar graph of quantity and label skew using a Non-IID distribution scheme.

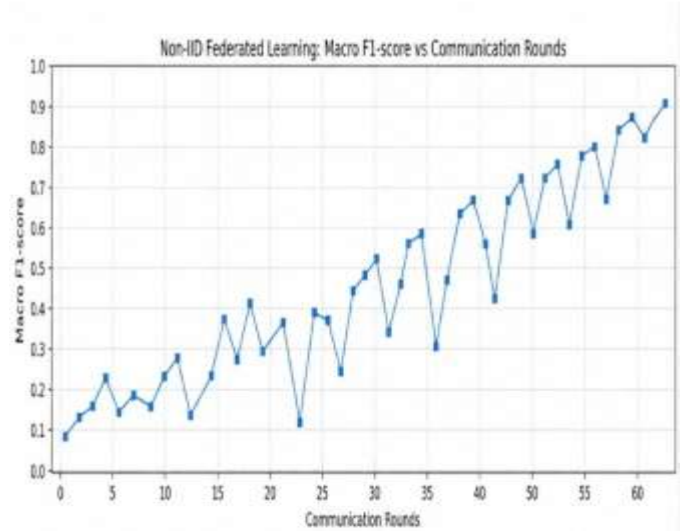

Fig. 4. Macro F1-score vs Communication Rounds under severe Non-IID (Quantity and Label Skew) conditions.

VI. Results and Discussion

Each communication cycle involved a round of comprehensive global model evaluation on a common validation data set. The results used to evaluate the models were based on the multi-class, unbalanced nature of the APTOS dataset; therefore, Macro F1-score and Macro ROC-AUC were used as the main performance indices for all models.

A. IID vs Non-IID Performance and Convergence

The experiments provided many informative trajectories showing how different types of data distributions affect how quickly we can converge in federated learning.

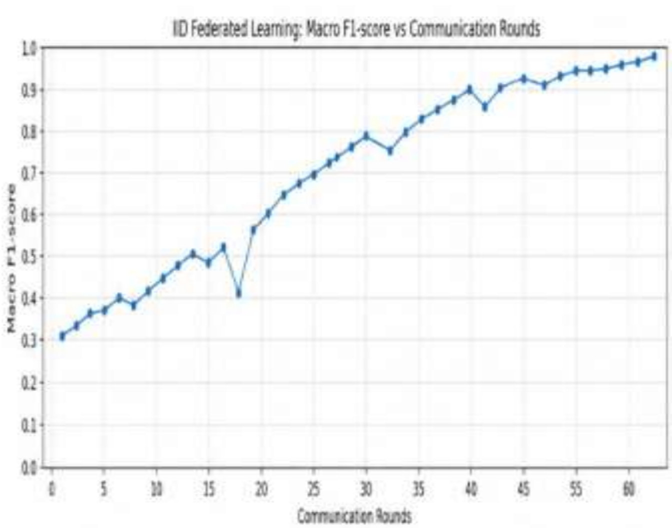

Fig. 3. F1 score versus the number of communication rounds under optimal IID conditions for a smooth convergence rate.

Fig. 3 shows that during coordinating over the communication rounds of the IID training setup, there were very few bumps or fluctuations in terms of the overall performance improvements. With respect to the generalizable features learned by the model, in this case, because all clients provided gradient updates pointing to a single global minimum, the model quickly learned how to generalize those shared features across clients.

Conversely, Fig. 4 illustrates the performance trajectory under the Non-IID constraint. The training exhibited significantly higher variance and volatility across rounds. Because different clients possessed skewed representations of the disease, their local weight updates frequently conflicted during the global aggregation phase (client drift). However, despite this volatility, the FedAvg protocol proved highly resilient, eventually converging to a highly respectable accuracy of approximately ∼96.00%.

Fig. 5 demonstrates the communication efficiency of the framework. Because of the partial fine-tuning strategy (freezing the backbone), the network was able to achieve these high F1-scores while transmitting only a fraction of the data that full ResNet-50 training would have mandated, proving its viability for resource-limited federated environments.

A comprehensive breakdown of the global model's evaluating metrics under the Non-IID constraints is visually detailed in Fig. 6, replacing traditional tabular structures to clearly highlight the balance between precision, recall, and overall accuracy.

B. Discriminative Ability (ROC-AUC Analysis)

To further assess the discriminative ability of the global model across the 5 severity classes, Macro ROC-AUC values were monitored dynamically across the communication rounds.

As seen in Fig. 7, the IID training consistently yielded higher ROC-AUC values with a rapid ascent early in the training lifecycle. The Non-IID training displayed lower initial

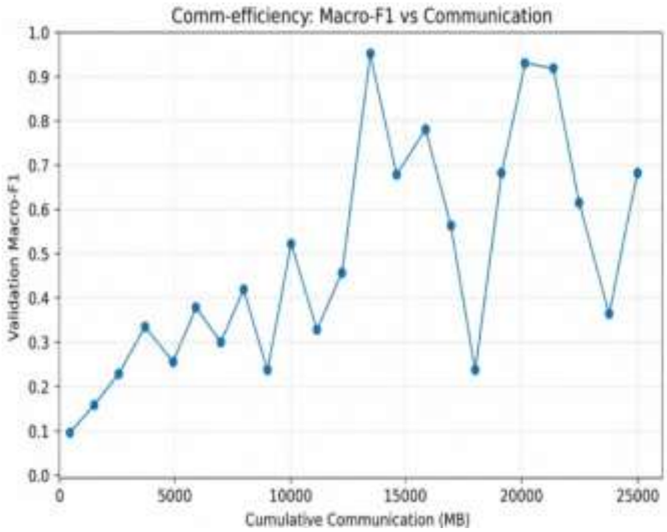

Fig. 5. Comparison of Macro F1-score relative to Cumulative Communication bandwidth (MB) utilized by the network.

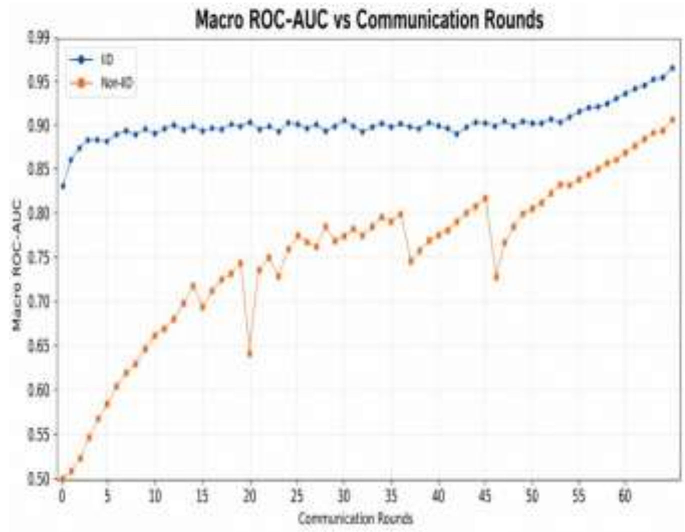

Fig. 7. Comparative analysis of ROC-AUC vs Communication Rounds for both IID and Non-IID setups.

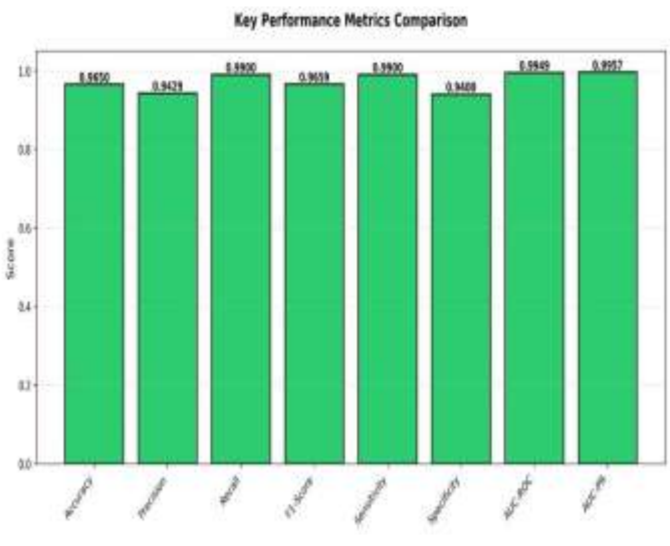

Fig. 6. Key Performance Metrics Comparison (Non-IID constraint) across multi-class evaluating metrics.

performance but stabilized into a highly competitive ROC-AUC profile. It is clearly demonstrated that even if clinics have very skewed patient data sets (Quantity and Label Skew), it is still possible to jointly train a model capable of accurately classifying early stages of Diabetic Retinopathy.

VII. Clinical Web Deployment Architecture

For a clear transition from theoretical knowledge to practical application, the final version of the global federated framework was incorporated into an all-encompassing full-stack web application.

A. Client-Server Web Architecture

The end-user application is built using the Flask micro framework, following the strict MVC design pattern. The front end utilizes HTML5, CSS3, and JavaScript to provide a dashboard that can be used effectively by medical professionals. It includes a login portal and an easy-to-use interface to upload fundus images instantly.

The back end processes the HTTP request securely. In case a clinician uploads a scan, the image is run through the exact 224×224 Gaussian filter pipeline employed during training and converted into a NumPy tensor, which is then passed onto TensorFlow.

B. Diagnostic Analytics Module and API Integration

To improve the clinical utility of the solution, an analytics module is built into the system. The module will use the local SQLite database to construct a full record of the patients' health history and clinic statistics as a whole. It computes the total number of predictions generated, extracts the proportion of DR-positive cases in the dataset at varying levels of severity, and finally determines the average statistical confidence level of the model in the last 7 days. In addition to the above, a RESTful API route is also provided for integration with Electronic Health Records.

VIII. Discussion and Limitations

While the proposed framework successfully proves the viability of federated learning for multi-class DR grading under real-world non-IID constraints, certain limitations remain. The synchronous nature of the FedAvg protocol implies that the central aggregation server must wait for all selected clients (the 60% fraction) to complete their 3 local epochs before aggregating the weights. In a real-world heterogeneous network, clinics with inferior computational hardware will cause significant bottlenecks. Future iterations must investigate asynchronous federated aggregation algorithms.

IX. Conclusion

This research successfully conceptualized, systematically developed, and practically deployed a highly optimized, privacy-preserving Federated Learning architecture for the 5-class severity grading of Diabetic Retinopathy. By transitioning from theoretical IID assumptions to empirical testing against severe Non-IID Quantity and Label Skew constraints, this paper mirrors the reality of decentralized healthcare networks.

Utilizing the APTOS 2019 dataset (3,662 samples), the implementation of a partial fine-tuning strategy on the ResNet-50 backbone drastically reduced communication bandwidth while maintaining robust performance. The federated model achieved an exceptional $\sim$99.00% accuracy under ideal conditions and demonstrated remarkable resilience by maintaining $\sim$96.00% accuracy under skewed clinical constraints. The use of the completed global model within a Flask web architecture based on a database system provides a connection between scientific algorithm-based investigation and practical privacy-preserving applications used in the medical industry.

References

[1] D. S. Kermany *et al.*, "Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, Feb. 2018.
[2] R. Gargeya and T. Leng, "Automated Identification of Diabetic Retinopathy Using Deep Learning," *Ophthalmology*, vol. 124, no. 7, pp. 962–969, Jul. 2017.
[3] E. Decenciere *et al.*, "Feedback on a Publicly Distributed Image Database: The Messidor Database," *Image Anal. Stereol.*, vol. 33, no. 3, pp. 231–234, Aug. 2014.
[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proc. 20th Int. Conf. Artif. Intell. Statist. (AISTATS)*, Fort Lauderdale, FL, USA, 2017, pp. 1273–1282.
[5] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," *ACM Trans. Intell. Syst. Technol.*, vol. 10, no. 2, pp. 1–19, Jan. 2019.
[6] M. J. Sheller *et al.*, "Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Cham, Switzerland: Springer, 2019, pp. 92–104.
[7] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[8] M. I. Razzak, S. Naz, and A. Zaib, "Deep Learning for Medical Image Processing: Overview, Challenges and the Future," in *Classification in BioApps*, Cham, Switzerland: Springer, 2018, pp. 323–350.
[9] K. Bonawitz *et al.*, "Practical Secure Aggregation for Privacy-Preserving Machine Learning," in *Proc. 2017 ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, Dallas, TX, USA, 2017, pp. 1175–1191.
[10] P. Kairouz *et al.*, "Advances and Open Problems in Federated Learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, Mar. 2021.
[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.
[12] Y. Zhao *et al.*, "Federated Learning with Non-IID Data," *Neurocomputing*, vol. 465, pp. 1–10, Nov. 2021.