

FedEye: Detecting Diabetic Eye Diseases using Federated Deep Learning and Handling Data Heterogeneity with Fedprox Aggregation

Seema Gulati

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India;
&

Department of Computer Science and Engineering,
Graphic Era Hill University, Dehradun, Uttarakhand, India.
E-mail: seemag.gulati@gmail.com, seemagulati@gehu.ac.in
ORCID: <https://orcid.org/0009-0001-3869-7240>

Kalpna Guleria

Chitkara University Institute of Engineering and Technology,
Chitkara University, Punjab, India.
Corresponding author: guleria.kalpna@gmail.com, kalpna@chitkara.edu.in
ORCID: <https://orcid.org/0000-0003-2359-8351>

Nitin Goyal

Department of Computer Science and Engineering, School of Engineering and Technology,
Central University of Haryana, Mahendergarh, Haryana, India.
E-mail: dr.nitingoyal30@gmail.com, dr.nitin@cuh.ac.in
ORCID: <https://orcid.org/0000-0001-7878-363X>

Sushil Kumar

Department of Computer Science and Engineering,
G. B. Pant DSEU Campus, Delhi, India.
E-mail: sushilyadav.thapar@gmail.com
ORCID: <https://orcid.org/0000-0003-1819-3712>

Sunita Tiwari

Department of Computer Science and Engineering,
G. B. Pant DSEU Campus, Delhi, India.
E-mail: sutiwari@gmail.com
ORCID: <https://orcid.org/0000-0003-1875-7928>

(Received on September 25, 2025; Revised on April 9, 2026 & July 9, 2026 & July 31, 2026;
Accepted on August 9, 2026)

Abstract

Diabetic Retinopathy (DR), Diabetic Macular Edema (DME), and glaucoma are significant ocular diseases that severely impair vision and require early automated detection for effective clinical management, but the traditionally adopted centralised deep learning based detection poses serious privacy problems for the participating organisations and also the data heterogeneity is a practical issue. The article proposes a privacy-preserving Federated Learning (FL) framework that enables collaborative model training across multiple decentralized datasets without sharing sensitive patient data for detecting DR, DME, and glaucoma from high-quality coloured fundus images, employing a Federated Deep Learning (FDL) approach coupled with the FedProx aggregation for effectively handling the non-IID data distributions in a multi-disease diagnostic setting. The framework evaluates four federated deep learning models based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures across multiple communication rounds to analyze their effectiveness in a decentralized learning environment. A curated dataset of 2,700 high-resolution fundus images (balanced across four classes) is constructed from multiple public sources and augmentation was applied to the training set, increasing the effective training samples while keeping the test set unchanged. The dataset is distributed across

three clients in a 30:35:35 ratio and trained over 20 communication rounds with consistent hyperparameters. The obtained results demonstrate that EfficientNetB0 and InceptionV3-based federated models consistently achieve more than 97% accuracy, with the EfficientNetB0-based model showing superior stability and emerging as the most suitable architecture with a prediction accuracy of 97.96% along with high precision, recall, and F1-score, indicating robust and stable multi-disease classification under heterogeneous federated settings.

Keywords- Federated learning, Automatic detection, Diabetic eye diseases, FedProx.

1. Introduction

Artificial Intelligence (AI), the revolutionary advancement of computer science creates systems that are built on mimicking human intelligence with decision-making capabilities, language translation, and visual perception. Many industries integrally use AI-enabled systems for not only providing capability but also productivity. Machine Learning (ML), a subset of AI, helps in building intelligent systems that are more adaptive, handle complex tasks effectively in diverse environments (Janiesch et al., 2021). Deep Learning (DL) models, considered to be a subset of ML, are data-hungry and handle complex problems using deep neural networks (Al-Fahdawi et al., 2024; Deepak & Bhat, 2024). The DL networks possess the capability to automatically learn the high-level features from the raw inputs and significantly enhance the performance of tasks such as image recognition, natural language processing, and speech synthesis (Qummar et al., 2020). DL algorithms are popular due to increased computational power, the availability of large datasets, and innovations in neural network architectures.

Progressive complications of diabetes, like Diabetic Macular Edema (DME), Diabetic Retinopathy (DR), and Glaucoma can lead to severe vision loss and blindness if not detected and treated early (Ding & Wong, 2012). The traditional detection methods are not only time consuming and tedious but also prone to human errors. Additionally, these detection methods are limited by the availability of ophthalmologists, especially in underserved regions. ML and DL techniques offer transformative solutions by automating retinal image analysis. These techniques efficiently process huge datasets of high-resolution fundus images, identifying early-stage biomarkers that may be difficult for human eyes to detect (Abramoff et al., 2010; Hughes-Cano et al., 2024). AI algorithms can be trained to distinguish between normal and diseased retinas with high precision (Nasajpour et al., 2022). The healthcare providers can ensure faster screening using these assistive techniques and thus provide timely treatment which is crucial for preventing irreversible damage. Furthermore, these systems can be deployed in telemedicine platforms, expanding access to quality eye care, especially in remote or low-resource settings. The scalability, speed, and accuracy of AI models make them essential tools in enhancing the diagnosis and management of retinal complications, improving medical results of patients, and reducing the burden on healthcare systems (Safi et al., 2018; Wilkinson et al., 2003).

Federated Learning which is a collaborative, privacy-preserved learning is also a subset of AI. It is a relatively more recent and innovative approach to privacy-protected learning and was introduced by Google company. FL is a decentralised form of ML where a data bank is not created at any central location. Unlike, the traditional AI systems where the data is pooled centrally, FL system allows data to remain at the host locations itself. The central server shoulders the responsibility for coordinating the learning process and incorporating the learned model into a combined global model (Zhang et al., 2021). This is done to address the privacy and data security problems resulting from data centralisation. FL enables model training across decentralised host devices that hold data samples locally, without sharing their samples to any central repository (Gulati et al., 2024; Yang et al., 2019). FL provides a collaborative and privacy-preserved learning framework, encouraging advancements in applications where data privacy is paramount, such as personalised healthcare and autonomous driving. FL also enables the development of robust disease

detection models without centralising sensitive patient data, thereby preserving privacy and complying with various privacy regulations such as GDPR and HIPAA (Chaddad et al., 2024). The primary advantage of an FL system, the data is held locally on devices within the organisation, reducing the risk of data breaches to a great extent and ensuring the confidentiality of the patients. Next, the FL architecture facilitates access to diverse data by enabling collaborative learning across multiple institutions. This diversity helps in developing more generalizable and robust disease detection models (Gulati et al., 2024; Li et al., 2020). Improved model performance is another key benefit of FL. By aggregating insights from multiple models trained on different datasets, FL enhances the overall robustness and reliability of disease detection models. Also, FL models can be continuously updated and adapted with new data, improving accuracy and reliability at runtime (Kairouz & McMahan, 2021; KhoKhar et al., 2022).

Therefore, FL has emerged as a promising paradigm for privacy-preserving model training across decentralized healthcare institutions, but the existing studies are largely limited to single-disease diagnosis (binary classification and detection) assuming IID data distributions and evaluate only a single model architecture. These limitations reduce their applicability in real-world clinical environments, where data are inherently heterogeneous (non-IID) and multiple co-existing retinal diseases must be detected simultaneously. To address these research gaps, the article proposes a multi-disease federated deep learning framework for the simultaneous detection of DR, DME, and glaucoma, incorporating the FedProx aggregation method to effectively handle data heterogeneity across clients. Additionally, a comparative evaluation of multiple CNN architectures is performed within a federated setting to identify the most stable and accurate model for decentralized retinal disease diagnosis.

The work presented in the article is an advancement of the federated learning framework proposed by Gulati et al. (2024), and Gulati et al. (2025) and by incorporating a systematic comparative assessment of multiple convolutional neural network architectures within a unified federated learning setting. The proposed work evaluates and identifies the most suitable architecture for privacy-preserving multi-disease retinal diagnosis under heterogeneous (non-IID) client data distributions.

The article presents a collaborative and privacy-preserved federated deep learning framework for the detection of diabetes induced eye diseases- Diabetic Macular Edema, Diabetic Retinopathy, and glaucoma. The key contributions and findings following the implementation of the framework are outlined below:

- The article proposes a unified federated deep learning framework for the simultaneous classification of multiple retinal diseases, including Diabetic Retinopathy, Diabetic Macular Edema, and Glaucoma. The proposed approach is a multi-disease modelling within a single decentralized architecture, unlike the previous works that focus on single disease detection with privacy preserved framework.
- The federated deep learning framework proposed in the article, for the detection of diabetic eye diseases enables collaborative model development across different institutions by integrating different open-source and harmonizing their labels into a common taxonomy. The harmonized dataset was distributed across different clients in an unequal split to represent heterogeneous, non-IID data. Thus, capturing realistic variations in data distributions, imbalanced classes and heterogeneous acquisitions of data, practically observed in decentralized healthcare systems.
- Thus, allowing for the development of robust models trained on a variety of datasets. Further, it increases the overall reliability and accuracy of the federated deep learning models.
- The proposed federated deep learning framework employs the FedProx aggregation method, which includes a proximal term to mitigate client drift and improve convergence stability in non-IID data distributions. the challenges posed by data heterogeneity across different clients. Thus, improving the stability and robustness of the global model highly suitable for privacy-preserving medical image analysis

in distributed environments.

- The proposed work systematically evaluates four federated deep learning models based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures. The federated deep learning model with EfficientNetB0 consistently demonstrates superior accuracy, stability, and reliability, achieving over 97.96% accuracy across communication rounds. The systematic evaluation provides insights into the convergence behaviour, stability, and cross-disease performance under heterogeneous data conditions.

The article introduces Federated Learning (FL) and its importance in preserving data privacy while enabling collaborative learning across decentralized datasets. Section II is the literature review that explores existing research on FL applications in healthcare, particularly in detecting diabetic eye diseases such as Diabetic Retinopathy, Diabetic Macular Edema, and Glaucoma. Section III of the materials and methods describes the federated learning framework implemented in this research, including the datasets used, deep learning models evaluated, and the FedProx aggregation method applied to address data heterogeneity. In Section IV, the results and discussion section, the performance of various models is analysed based on accuracy, precision, recall, and F1 scores. Finally, in section V, the conclusion section summarises the key findings, highlights the effectiveness of the proposed framework in ensuring privacy-preserving, accurate disease detection, and suggests potential avenues for future research.

2. Literature Review

The research work by Baptista et al. (2023) employed a horizontal cross-silo FL approach based on convolutional networks that have been used for glaucoma screening from retinal fundus images. The dataset used in the article was a curation of different publicly available retinal fundus image datasets: Origa, Refuge, Gamma, Drishti-GS, Eyepacs, and Airops. The results obtained range from 76.3% to 86.8% of prediction accuracy. The article focuses only on retinal fundus images, and domain generalization approaches are not covered in the article. The research article by Chetoui & Akhloufi (2023) proposed an approach to detecting DR (DR) using Federated Learning (FL) and Vision Transformer (ViT) models to preserve data privacy across multiple institutions. The research uses four publicly available datasets: APTOS 2019 (3,662 images), MESSIDOR-1 and MESSIDOR-2 (combined 2,948 images), IDRID (516 images), and a sample of 6,000 images from EyePACS. The results showed that the FLViT achieved 95% accuracy on the APTOS dataset (1% improvement), with similar improvements of 2-3% on other datasets.

Hossain et al. (2023) proposed a framework combining FL and Differential Privacy (DP) for predicting DR. The dataset used in the article is the Kaggle Diabetic Retinopathy Detection dataset. The article investigates four different federated deep learning models based on ResNet50, AlexNet, VGG16, and SqueezeNet1.1 for detecting DR. ResNet50 achieved the best performance, with an accuracy of 83.05% on the binary classification task (No-DR vs. DR) when no noise was added. The work primarily focused on IID data partitions and practically data is usually non-IID. The research by Singh & Singh (2023) aims to enhance the classification of DR using a DenseNet-based federated learning model. The DenseNet121 architecture was chosen for its efficiency and high accuracy in image classification tasks and improved. The dataset used in this article is the APTOS 2019 Blindness Detection dataset from Kaggle. The FedDDR model achieved an overall accuracy of 94.27% in classifying DR across all severity levels compared to existing methods.

Another research by Aljohani & Aburasain (2024) proposed a hybrid framework integrating VGG-16 and ResNet50 with Random Forest for the detection of glaucoma from retinal fundus images. This hybrid approach utilizes both image data and texture features for classification. The proposed framework was implemented using a pooled dataset from four publicly available sources: Acrima, G1020, Origa, and

Refuge. The proposed hybrid framework obtained an accuracy of 95.41%. The paper by Nam (2024) proposed a federated learning (FL) approach to improve the accuracy of the glaucoma disease detection model while maintaining patient data privacy. Publicly available fundus image datasets were used in the article for the proposed glaucoma detection: Origa, Refuge, and G1020. The segmentation of the optic cup and disc is completed using a U-net model. The approach obtained an overall detection accuracy of 88% for glaucoma detection. However, the work does not explicitly address the challenge of non-independent and identically distributed data across different clients.

Table 1. The summarization and comparison of the existing research work in the retinal disease detection.

Study	Task	Dataset	FL type	Disease scope	Method	Results	Limitation/Gap
Baptista et al. (2023)	Glaucoma screening	Origa, Refuge, Gamma, Drishti-GS, EyePACS, Airogs	Cross-silo FL	Single disease (Glaucoma)	CNN-based federated model (FedAvg Aggregation)	76.3% – 86.8% accuracy	Single disease
Chetoui & Akhloufi (2023)	DR detection	APTOS 2019, MESSIDOR-1, MESSIDOR-2, IDRiD, EyePACS	Cross-silo FL	Single disease (DR)	Vision Transformer (FLViT) (FedAvg Aggregation)	~95% accuracy on APTOS; 2–3% improvement on other datasets	Single disease
Hossain et al. (2023)	DR classification	Kaggle DR dataset	Cross-device FL	Single disease (DR)	ResNet50, AlexNet, VGG16, SqueezeNet + Differential Privacy (FedAvg Aggregation)	83.05% accuracy (ResNet50, binary DR classification)	IID partition
Singh & Singh (2023)	DR severity classification	APTOS 2019	Cross-device FL	Single disease (DR)	DenseNet121 (FedDDR) (FedAvg Aggregation)	94.27% classification accuracy	Single disease
Aljohani & Aburasain (2024)	Glaucoma detection	Acrima, G1020, Origa, Refuge	Centralized learning	Single disease (Glaucoma)	VGG16 + ResNet50 + Random Forest hybrid model	95.41% accuracy	Not federated
Nam (2024)	Glaucoma detection	Origa, Refuge, G1020	Cross-silo FL	Single disease (Glaucoma)	U-Net segmentation + federated model (FedAvg Aggregation)	88% detection accuracy	No non-IID evaluation
Gulati et al. (2024)	Diabetic eye disease detection	ODIR-5K, IDRiD, OCT & Eye Fundus Dataset, Robust Vessel Segmentation in Fundus Images	Cross-silo FL	Normal, DR, DME, Glaucoma	MobileNetV2 + FedProx Aggregation with $\mu = 0.01, 0.02, 0.03, 0.04, 0.05$.	Accuracy- 97.71% at $\mu = 0.03$.	Single backbone architecture.
Gulati et al. (2025)	Ocular disease detection	IDRiD and APTOS Blindness Detection	Cross-silo FL	Normal, DR, DME.	MobileNetV2 + FedAvg Aggregation	Accuracy: 96.53%, Precision: 96.52%, Recall: 96.53%, F1-score: 96.52% (best-performing client configuration)	Single backbone architecture.

Gulati et al. (2025) presented a collaborative federated learning framework using the MobileNetV2 architecture and FedAvg aggregation across different client configurations (2, 3, and 4 clients). The framework achieved an average prediction accuracy of 96%, with the 3-client configuration providing the

best balance between accuracy and stability. Although these studies demonstrated the feasibility of privacy-preserving federated learning for retinal disease diagnosis, they evaluated only a single deep learning architecture. Additionally, the authors Gulati et al. (2024) also proposed a privacy-preserving federated learning framework for detecting diabetic eye diseases using a MobileNetV2 backbone and the FedProx aggregation strategy to improve convergence under heterogeneous client data distributions. The framework achieved a maximum prediction accuracy of 98.70% ($\mu = 0.03$), while $\mu = 0.05$ demonstrated the most stable convergence.

Although there have been several studies on federated learning for retinal diseases diagnosis, most of the previous methods have centered around the single disease diagnosis either DR or glaucoma as summarized in **Table 1**. Many prior works have demonstrated the effectiveness of federated learning for retinal disease diagnosis; most existing approaches have been developed under centralized training settings or have evaluated only a single deep learning architecture within a federated framework. Moreover, many studies assume that data across participating clients are independently and identically distributed (IID), an assumption that rarely holds in real-world clinical environments where data collected from different healthcare institutions are inherently heterogeneous. Consequently, limited research focuses on the comparative robustness, convergence behaviour, and predictive performance of different convolutional neural network (CNN) architectures under realistic federated learning conditions. To address these limitations, the present research extends our previously proposed privacy-preserving federated learning framework for the detection of Diabetic Retinopathy (DR), Diabetic Macular Edema (DME), and glaucoma by incorporating the FedProx aggregation strategy to effectively handle heterogeneous non-IID client data distributions. Furthermore, a systematic comparative evaluation of four CNN architectures -EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 - is performed within an identical federated learning environment to identify the most robust and accurate architecture for privacy-preserving multi-disease retinal diagnosis.

3. Background

Federated Learning (FL) was launched by researchers at Google Inc., in the year 2016 for dealing with issues like the privacy of user-sensitive data and compliance with data regulations. Various issues, such as computational and storage requirements for enormous amounts of data are resolved using FL. Data privacy is of prime importance in critical fields like healthcare and finance. Therefore, the emergence of stringent data policies like the GDPR propelled the decentralization of data in order to maintain privacy (Hartmann, 2018). The concept of FL is based on the fact that the data is usually present on different servers, which are pooled onto a single server in a traditional machine learning system. But in an FL setup, the data remains in the originating host location and the data never leaves the host. Instead, machine learning is applied at the host location to form multiple local models (Gulati et al., 2025; Jin et al., 2023). These local models are used by a central coordinating server to form a global model. FL is essentially developed to hold data on local devices, enabling the training to be decentralized and distributed. Thus, the data never leaves the host site maintaining privacy and addressing bandwidth as well as storage issues. The computation is divided among multiple data holding sites. Thus, leveraging the computational power of the client devices (Zhang & Mao, 2022).

Federated learning is an iterative process with machine learning or deep learning as the backbone of the learning model. The process begins with the initialization of a global model at the central server. The server is responsible for the coordination of the federated learning process. The server initializes a global model that is either pre-trained or has random weights.

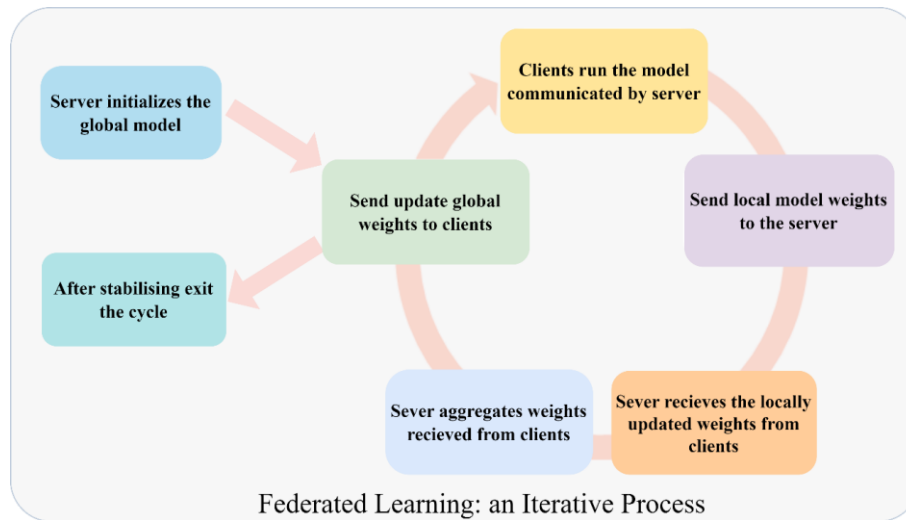


Figure 1. The cyclic process privacy-enhancing and collaborative federated learning framework.

The global model is then sent to all participating clients with their own dataset. Using the initial global model and their locally held dataset, the clients run a series of training epochs to build local models. All the participating clients upload their local model updates on the server. The central server then aggregates and incorporates the clients' local models into the global model; thus, forming an updated global model. The updated global model is transferred to the clients which again train using their local dataset and the process is iterated for a series of communication rounds until the model becomes stable, or attains a satisfactory prediction accuracy and converges. The aggregation at the server is usually an averaging function called the FedAvg. The cyclic nature of the privacy-preserved and collaborative federated learning architecture is shown in **Figure 1**. Once sufficiently trained, the final global model is deployed by distributing it back to the clients for local inference or broader centralized deployment. Additionally, the FL models can be adapted for continuous learning, allowing it to evolve as new data becomes available, all while maintaining privacy and efficiency.

The architectural representation of the privacy-preserved, collaborative federated learning framework with N clients is shown in **Figure 2**. It is a decentralized approach to machine learning where the data is not pooled and remains at the source location, ensuring the privacy of used sensitive data. Federated learning is an iterative process, as already discussed, with a single central server and multiple clients. The clients may be servers or edge devices as per the architecture.

The server is responsible for selecting participants in the current communication round. The server also selects and implements the aggregation strategy. The aggregation strategy is one of the crucial components in a federated learning scenario, as it decides how the client models will be combined into the global model. Federated Averaging, or FedAvg, is the most commonly used aggregation method. This method averages model updates, with contributions weighted by the number of data points per client, though it struggles with non-IID data. Federated Proximal or FedProx is an extension of the FedAvg aggregation method. FedProx adds a proximal term to the local objectives, which penalizes large deviations from the global model. Thus, with the help of the proximal term issues with non-IID data and slow clients are addressed. Another method is clustering-based aggregation, which groups clients based on similarities in their data or updates. Thereby improving performance in diverse data distributions.

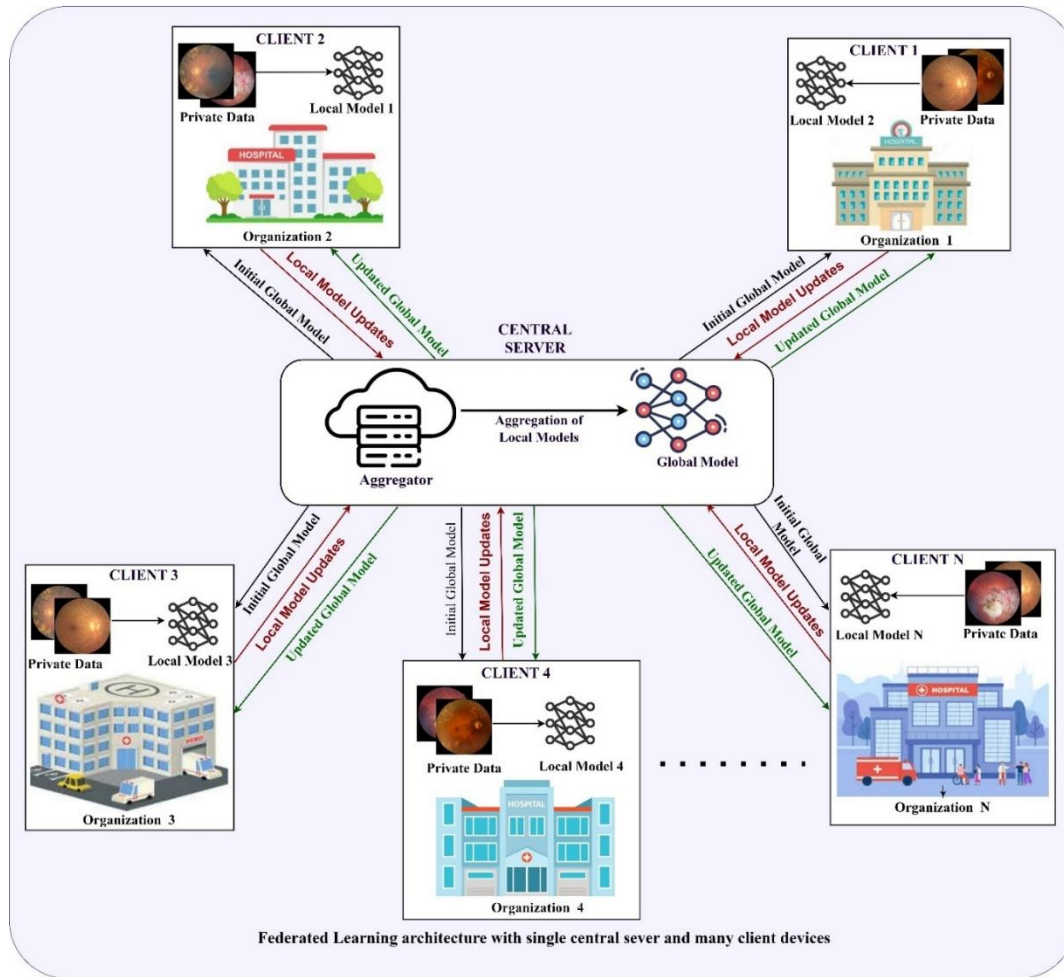


Figure 2. Architectural representation of a privacy-preserved and collaborative federated learning framework with N clients.

Federated Stochastic Gradient Descent or FedSGD, aggregates the gradients from each client rather than the model weights. The clients compute the gradients based on their local data which are then averaged at the server to update the global model. FedSGD is more efficient in scenarios where models are large and transferring gradients is less costly than transferring model weights.

4. Materials and Methods

The materials and methods section here, describes the materials in terms of the dataset and the methodology adopted in the implementation of the research for automatic detection of diabetic eye diseases, including Diabetic Retinopathy, glaucoma, and macular edema. The research employs a federated learning (FL) framework to ensure data privacy while enabling collaborative model training across multiple clients.

4.1 Dataset Description

The proposed Federated Deep Learning framework presented in the article uses a dataset of colored fundus images with three eye diseases- Diabetic Retinopathy, Diabetic Macular Edema, and Glaucoma, and the

fourth class is the Normal class fundus images. The image set used is a curation of four different datasets with high-quality fundus images available in the public domain for research purposes.

- i. The "OCT and Eye Fundus Dataset", is available on GitHub and uploaded by the Translational Visual Health Laboratory (2024). The dataset has labelled images from four Diabetes induced eye disease classes and is available on GitHub. The images belong to four main classes: Normal, DME, DR And Glaucoma.
- ii. The "Indian Diabetic Retinopathy Image Dataset (IDRiD)" (Porwal et al., 2018) another publicly open-source dataset, accessible through Kaggle and the also the IDRiD's website. The set comprises of labelled retinal fundus images for analysing and diagnosing DME and DR.
- iii. The Ocular Disease Intelligent Recognition (ODIR) dataset, is a structured ophthalmic database. It is assembled manually by Shanggong Medical Technology Co. Ltd. (2020) sourced from different medical organizations of China. This dataset is publicly accessible on Kaggle and represents real-life patient information.
- iv. Another dataset utilized in the research is the openly available "Robust Vessel Segmentation in Fundus Images," dataset created by Budai et al. (2013) in their research article published in the International Journal of Biomedical Imaging. The set comprises fundus images of persons with healthy eyes, eyes suffering from diabetic retinopathy and glaucomatous eyes.

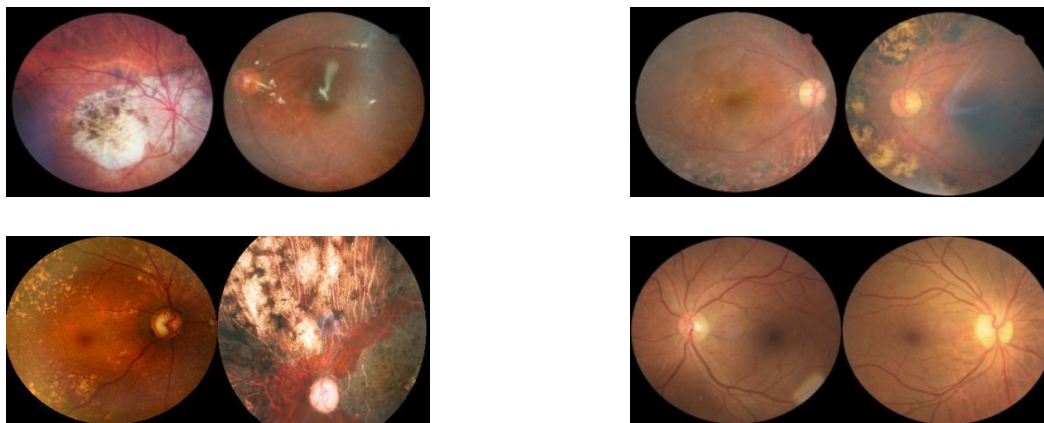


Figure 3. Data samples from: DME class, DR class, glaucoma class and normal class.

A multi-source data integration strategy was adopted since no single source was sufficient to provide all disease categories. Thereafter, a standardized processing pipeline was adopted in order to reduce the inter-dataset variability. The first step was the modality filtering for retaining only coloured fundus images and excluding all OCT images from the pooled dataset. Secondly, for datasets with multi-label annotations only single-disease instances were held. Next, images with poor illumination, blur or inadequate retinal visibility were removed to maintain quality of the dataset. Subsequently, all the images were transformed to a standard size for uniformity and normalised to ensure compatibility with deep learning models. The pooled dataset was obtained and due to differences in the dataset sizes and class distributions, the samples were balanced. From the pre-processes pool a total of 2700 images were selected, with 675 images from each class. The balanced representation supports robust model training in a federated learning setting with heterogeneous data distributions. **Figure 3** illustrates representative fundus images for each class: (a) DME, (b) DR, (c) glaucoma, and (d) normal, highlighting their key retinal characteristics.

Table 2. The source-wise image contribution from each dataset and the finally selected distribution.

Source dataset	Original dataset size	Available classes	Fundus images after filtering	Classes contributed	Images selected for final dataset
OCT & Eye Fundus Dataset	More than 5,000 (OCT + fundus)	Normal, DR, DME, Glaucoma	1,200	Normal, DR, DME, Glaucoma	800
IDRiD	516	DR, DME	516	DR, DME	400
ODIR-5K	Nearly 10,000 (left/right images)	Normal, DR, Glaucoma, others	6,000	Normal, DR, Glaucoma	1,480
Robust Vessel Segmentation in Fundus Images	40	Normal, DR	40	Normal, DR	20
Total Pooled Dataset	—	—	More than 7,700	—	—
Final Balanced Dataset	—	4 classes	—	Normal, DR, DME, Glaucoma	2700 (675 per class)

Table 2 summarizes the contribution of each dataset before and after preprocessing. The selective filtering of fundus only images followed by the label harmonization and quality filtering a balanced subset of 2700 images were constructed with 675 samples from each class. After that, augmentation and the total set of images used after augmentation was applied only to the training subset following the train–test split, increasing the training samples while keeping the test set unchanged.

4.2 Proposed Methodology

This subsection describes the detailed methodology proposed in the article for implementing the federated deep learning model used in the detection of diabetes induced eye diseases. The architecture used in the work is horizontal federated learning architecture with 3 clients and a dataset size of 2700 images. Augmentation is applied to the dataset to improve the prediction accuracy for detecting Glaucoma, DME, and DR. Enhancement of the dataset was done, and the number of images was increased three times, as shown in **Table 3**. Augmentation was applied only to the training set, increasing the number of training samples threefold, while maintaining the original test set unchanged.

The complete implementation process is diagrammatically depicted in **Figure 4**, which is given below, and the hyperparameters used in the execution are listed in **Table 3**, as shown below. The complete algorithm for the proposed methodology is given in Algorithm 1.

The process of setting up a federated deep learning system starts with the collection of data. The collection initially consists of the raw data from various sources (given in the previous subsection). The dataset is cleaned afterward, and only high-quality images are selected for the learning process. The augmentation was applied to the image set to improve the generalization and preventing the overfitting. The images were geometrically transformed after resizing, the transformations included were random rotations within 1 to 15 degrees. This introduces a minor variation in the original sample preserving the anatomical structure of the original image. Horizontal flipping was applied with a 0.5 probability, that reflects natural left- right variation and centre cropping was performed for retaining the original retinal region and ensuring that the central structures like optic disc and macula are retained in the frame. The augmentation pipeline was applied only to the training set after splitting the train and test data. The augmentation was applied in an on-the-fly manner during training where the transformed samples are generated dynamically at each epoch rather than being stored as separate images. Each of the original images contributes multiple variants across the iterations to increase data diversity without duplications. The transformations were carefully selected to be moderate for avoiding any distortions to the data. The augmentation was restricted to the train images the risk of data leakage is eliminated with improved model robustness and preventing compromising the evaluation integrity.

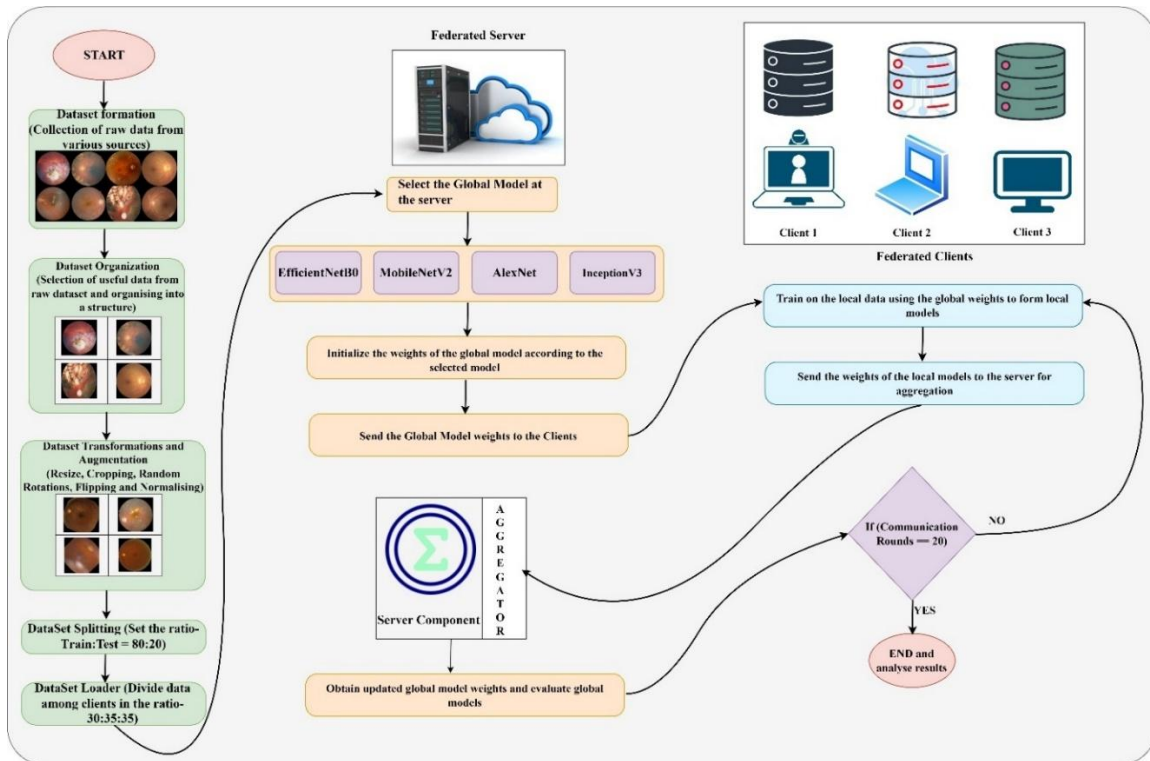


Figure 4. Flow diagram of the methodology for detecting diabetic eye diseases using federated deep learning models.

Furthermore, augmentation was implemented in an on-the-fly manner during training, where transformed samples are generated dynamically at each epoch rather than being stored as separate images. Each original training image contributes multiple augmented variants across training iterations, effectively increasing data diversity without physically duplicating the dataset. The applied transformations were deliberately moderate to avoid unrealistic distortions.

Although the effective training data is expanded, the transformations introduce meaningful variations in orientation and appearance rather than redundant copies. The use of center cropping further ensures structural consistency while reducing background noise. Since augmentation is restricted to the training set and performed dynamically, the risk of data leakage is eliminated, and the approach enhances model robustness without compromising evaluation integrity.

The fundus images were distributed across three clients in a 30:35:35 ratio after performing preprocessing of images and class balancing. Despite, the fact that the overall dataset maintains equal representation of all four classes, the allocation of images to each client was performed without enforcing identical class proportions. Therefore, each of the clients receive data originating from different source datasets with varying imaging conditions, such as differences in acquisition devices, illumination, and image quality. Therefore, this leads to variations in class distributions and underlying data characteristics across clients. Consequently, the local datasets exhibit non-identical (non-IID) distributions in terms of both class composition and feature space, thus, reflecting a realistic device-level heterogeneity practically observed in decentralized healthcare environments.

Table 3. The parameters used in implementing the federated deep learning model for detecting diabetic eye diseases.

Parameter	Value
Model Architecture	Horizontal Federated Learning
Deep Learning Models	EfficientNetB0, MobileNetV2, AlexNet and InceptionV3
Dataset Classes	Diabetic Macular Edema, Diabetic Retinopathy, Glaucoma and Normal
Number of Clients	3 clients
Data split among clients	30:35:35
Dataset Size	Original dataset-2,700 images Training set: 2160 originally and 6480 after augmenting Testing set: 540
Learning Rate	0.001
Optimizer	Adam
Batch Size	16
Epochs	5 (per round)
Communication Rounds	20
Aggregation Strategy	FedProx ($\mu=0.05$, μ is the proximal term)
Augmentation Techniques	Resize, Cropping, Random Rotations, Flipping and Normalizing
Train Test Ratio	80:20
Performance Metrics	Accuracy, Precision, Recall, and F1-score, Class-wise accuracy

The global model is selected, and the model weights and parameters are initialized accordingly. In this research work, 4 deep learning models are selected: EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3. These models were initialized with ImageNet weights and the last layer was adapted to classify into four classes. The training phase included training of higher layers meanwhile the convolutional layers were frozen, thus the generic feature representations were preserved and reduced overfitting.

The clients used same training configuration, that is the optimizer was Adam with a learning rate of 0.001, batch size was 16 with 5 local epochs per round. The training was conducted for 20 communication rounds without the use of early stopping and performance evaluation after every round was done.

After model selection, the initial model weights and the model parameters are communicated to all the participating clients that hold the data locally on their devices. The clients use the global model weights transferred by the central server for training on their local devices. The clients form local models after the training phase. These are then communicated to the server, and the aggregator at the server incorporates the client models into the global model.

The implementation uses a FedProx aggregation method in which a proximal term adds to the loss function of the training phase. The proximal term, denoted by μ , is a regularizer for limiting the deviation from the global model while training. The larger the deviation, the more the magnitude of the proximal term, and the loss increases. Therefore, the penalty increases for more deviation. The aim is to minimize the impact of heterogeneity by controlling the local updates, which leads to greater stability and consistency in the global model's performance.

The proximal term μ has been chosen carefully based on a prior evaluation conducted in an earlier work where the μ value was varied between 0.01 and 0.05 (Gulati et al., 2024). The results of the ablation demonstrated that the higher μ value of 0.05 performed better and converged earlier by restricting local model drift.

After the aggregation, the updated global model weights are again uploaded to the clients. The clients repeat the training process with these updated weights and build local models again using the FedProx loss

function. Clients again sent the newly built local models to the server for another round of aggregation. The server again receives all the client model weights, and the aggregator component aggregates or averages the client model weights to update the global model. The steps are iterated for a series of communication rounds until the model becomes stable. In the proposed work, the models are continuously evaluated for 20 communication rounds, and in each round, the clients run 5 epochs of training with batch size 16. After every communication round, performance metrics are calculated.

Algorithm 1. Federated learning with FedProx aggregation for detecting diabetic macular edema, diabetic retinopathy, and glaucoma.

Inputs

- N : Number of clients
- K : Total number of communication rounds
- T_i : Local data for client i
- μ : Proximal term (regularization parameter)
- η : Learning rate
- W^0 : Initial global model weights

Output

- Final global model weights W^K
-

1. Initialization:

- i. Server initializes the global model weights W^0
- ii. Set the number of communication rounds K .

2. For each round $k = 1$ to K :

- i. Server sends global model weights W^k each client i .
- ii. Each client i performs the following steps:
 - A. Receive global model weights W^k .
 - B. Set local model weights $W_i^k \leftarrow W^k$
 - C. Local Training on Client i :

For each local epoch (local iteration):

- a. Minimize the local objective function $f_i(W) = L(T_i, W) + \frac{\mu}{2} \|w - w^k\|^2$
- (1)

where

- $L(T_i, W)$ is the loss of local data T_i ,
- $\frac{\mu}{2} \|w - w^k\|^2$ is the proximal term.

- b. Update the local model using gradient descent: $w_i^{k+1} \leftarrow w_i^k - \eta f_i(w_i^k)$
- (2)

- D. Client sends updated local model weights w_i^{k+1} back to the server.

3. Server Aggregation:

- i. The server aggregates the updated local models from all clients: $W^{k+1} \leftarrow \frac{1}{N} \sum_{i=1}^N W_i^{k+1}$
- (3)

4. Repeat Steps 2-3 for K communication rounds.

5. Final Global Model:

- i. After completing K rounds, the server outputs the final global model W^K .

6. Exit

So, after all the models are executed for 20 communication rounds, the results are compared for each model, and inferences are drawn regarding the model's performance and stability.

5. Results and Discussions

In this section, the outcomes captured by following the methodology proposed are described in a graphical format for analysis. The section also discusses which deep learning model is apt for the problem of detecting the three diabetic eye diseases. The effectiveness of the model is assessed based on the performance metrics: prediction accuracy, precision, recall, F1-score, and class-wise accuracy. The four federated deep learning models simulated for the detection of diabetic eye diseases are based on the EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 models.

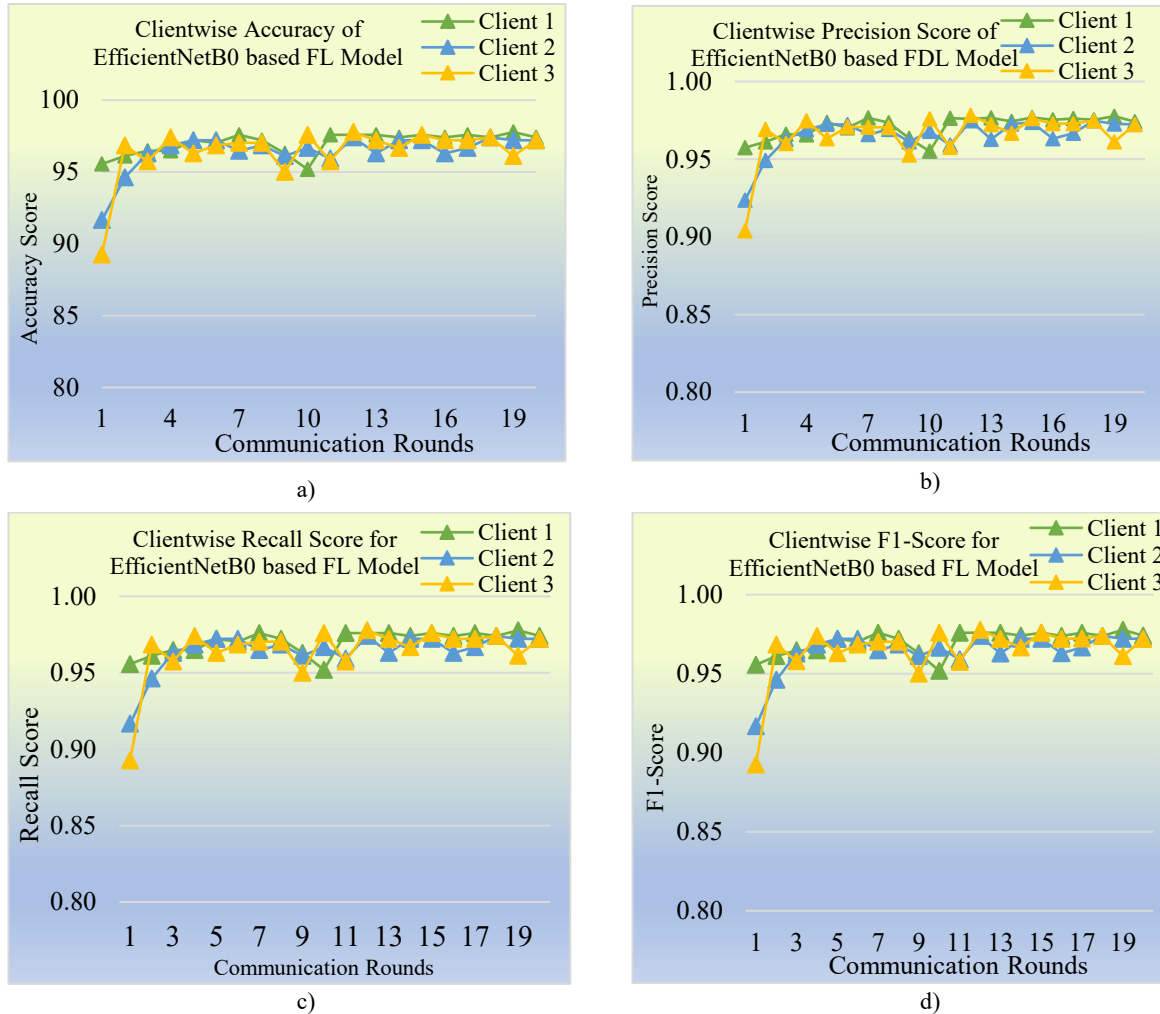


Figure 5. a) Clientwise Accuracy, b) Clientwise Precision, c) Clientwise Recall, and d) Clientwise F1-Score obtained by the efficientnetb0-based FL model.

Figures 5 to 8 present a comprehensive analysis of the performance of various FDL models, including EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3, across key evaluation metrics such as accuracy, precision, recall, and F1-score. These figures illustrate the clientwise results obtained for detecting diabetic eye diseases across multiple clients over successive communication rounds. **Figures 9 to 10** show the global metrics obtained by implementing the four FDL models based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures across the 20 communication rounds.

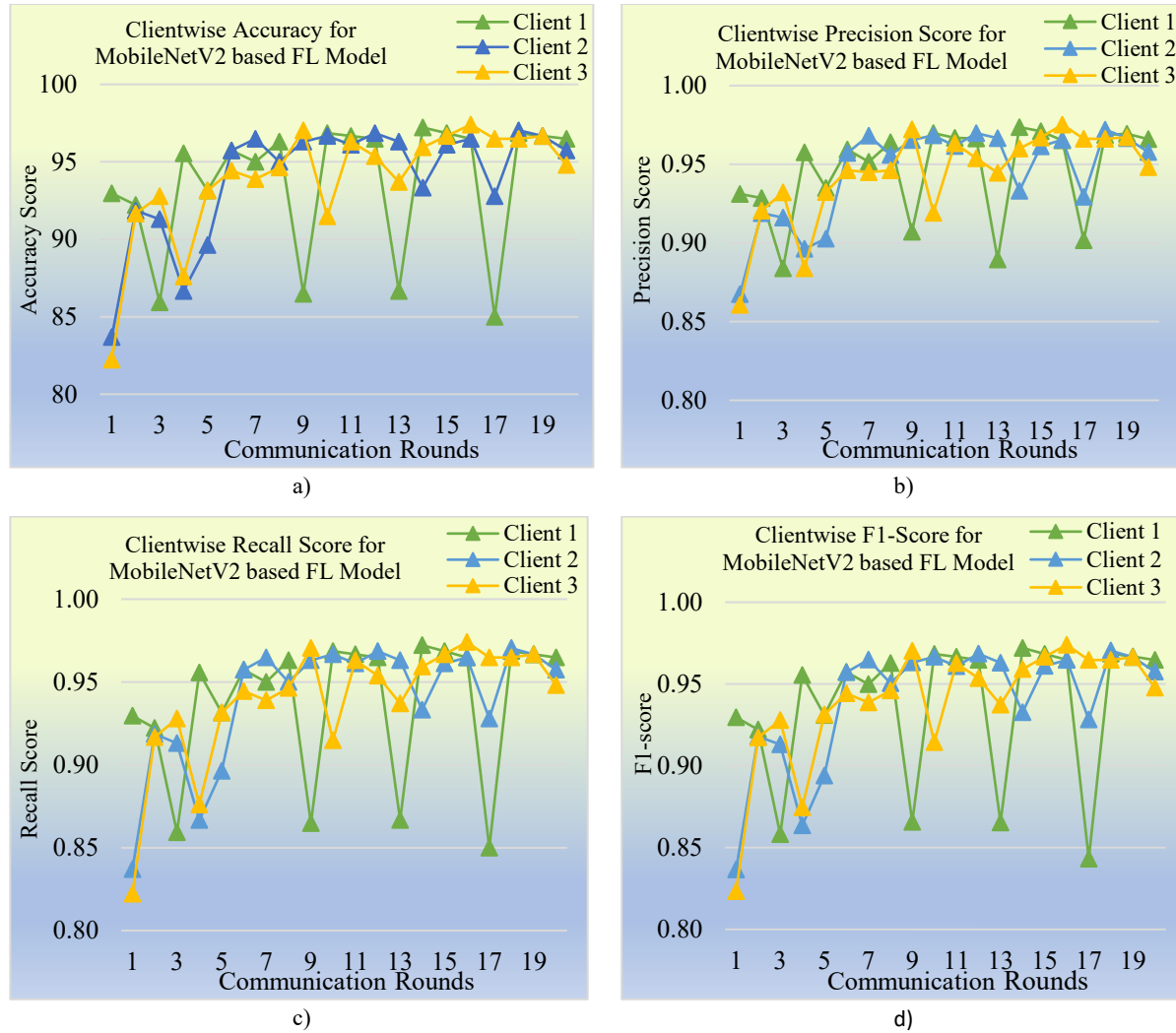


Figure 6. a) Clientwise accuracy, b) Clientwise precision, c) Clientwise recall, and d) Clientwise F1-score for mobileNetV2-based FL model.

Figure 5(a) below presents the accuracy scores obtained while training using the EfficientNetB0-based FDL model on various clients for detecting diabetic eye diseases. The performance of client 1 in terms of accuracy consistently increases and stabilizes by round 20. For client 2, a similar trend in accuracy performance is followed, indicating that the model performs well across rounds. The third client also achieves a high accuracy score by round 20. Thus, this suggests that the EfficientNetB0 model generalizes well. The client's precision score when the EfficientNetB0-based FDL model is utilized for client training is presented in **Figure 5(b)**. Precision is an important metric, especially in medical applications, as it measures the ability of the model to correctly identify positive cases without falsely classifying negatives as positives. For client 1, the precision score gradually increases before round 5, and then it remains high constantly. For client 2, the precision stabilizes from round 10 onwards, showing a consistent improvement in the model's performance over time. For client 3, the precision score steadily increases till round 10 and remains stable for the rest of the rounds. The precision scores for all three clients start strong and steadily improve over the first few communication rounds. By round 10, maintain this high precision level

consistently for the remainder of the rounds. **Figure 5(c)**, presents the client-wise recall score for the EfficientNetB0 based FDL model for detecting diabetic eye diseases. Recall, also known as sensitivity, is a critical metric in medical diagnostics because it measures the ability of the model to identify all actual positive cases correctly. All three clients demonstrate consistently high recall scores stabilising by around 20th communication round. The recall score gradually increases and stabilises for client 1, and client 2 and 3 follows a similar trend, with recall improving and becoming consistent after the 12th communication rounds. **Figure 5(d)**, shows the client-wise F1-score for the EfficientNetB0-based FDL model used to detect diabetic eye diseases. The F1-score is one of the important metrics in classification tasks, especially in medical diagnostics, since it measures the balance between precision and recall. Numerically, the F1-score is the harmonic mean of precision and recall, providing a balance between false positives and false negatives. It is observed that the F1-score for client 1 steadily increases in the initial rounds and stabilizes by round 5, indicating that the model has achieved a strong balance between precision and recall. However, client 2's F1-score shows a similar upward trend, where the F1-score improves significantly during the initial rounds and stabilizes by round 10. Whereas, client 3 starts with a lower F1-score compared to the other clients, but improves significantly by round 5. By round 10, client 3 achieves an F1-score comparable to the other clients and then stabilizes.

The client-wise accuracy for the MobileNetV2-based FDL model used for the detection of various diabetic eye diseases is diagrammatically shown in **Figure 6(a)**. It can be observed from **Figure 6(a)** that client 1 shows fluctuating accuracy in the early rounds, with an initial dip at round 3. Later, the accuracy improves and stabilises from round 8 onwards consistently achieving high accuracy in the later communication rounds. Client 2's accuracy shows a more inconsistent initial behaviour, with significant fluctuations across the first 10 rounds. From round 12 onwards, client 2's accuracy stabilizes, demonstrating that the MobileNetV2 model eventually converges to a high accuracy, though with some initial difficulties. Client 3 starts with the lowest initial accuracy among the three clients and shows a steady improvement in accuracy and after this point, the accuracy remains relatively high and stable. **Figure 6(b)** refers to the clientwise precision score for the MobileNetV2-based FDL model to detect diabetic eye diseases. For client 1, the precision score improves across rounds and stabilizes around round 8, maintaining a high precision score till the end of the 20 communication rounds. For client 2, initially the score shows fluctuations, but the model's performance improves over time, and it stabilizes by round 10. Similarly, for client 3, an overall steady increase in precision is demonstrated, which stabilizes after a few initial rounds. **Figure 6(c)** represents the client-wise recall scores for the MobileNetV2-based FDL model used to detect diabetic eye diseases. In client 1, the recall score initially fluctuates in the early rounds but shows a consistent improvement as the communication rounds progress. By round 10, the recall score stabilizes, maintaining a high value throughout the remaining rounds, which indicates that the model performs well in identifying positive cases in this client's data. In client 2 also shows some fluctuations in the initial rounds, but after round 10, the recall score stabilizes, indicating consistent performance in detecting positive cases. Client 3 follows a similar trend as client 2, where the recall score shows fluctuations early on, but the model's performance stabilizes in later rounds, reaching a high and steady recall value. **Figure 6(d)**, presents the client-wise F1-Score for the MobileNetV2-based Federated Deep Learning (FDL) model used to detect diabetic eye diseases. In the initial communication rounds, client 1 shows some fluctuation in F1-scores, reflecting inconsistency between precision and recall. However, after round 5, the F1-score stabilizes and remains high, indicating that the model performs well in balancing the correct identification of positive cases and minimizing false positives. Client 2 also shows an upward trend in the early rounds, around round 10 the F1-score stabilizes. Client 3 starts with a slightly lower F1-score compared to the other clients; by round 10, client 3 reaches a performance comparable to the other two clients, indicating that the MobileNetV2-based FDL model successfully balances precision and recall for this client as well.

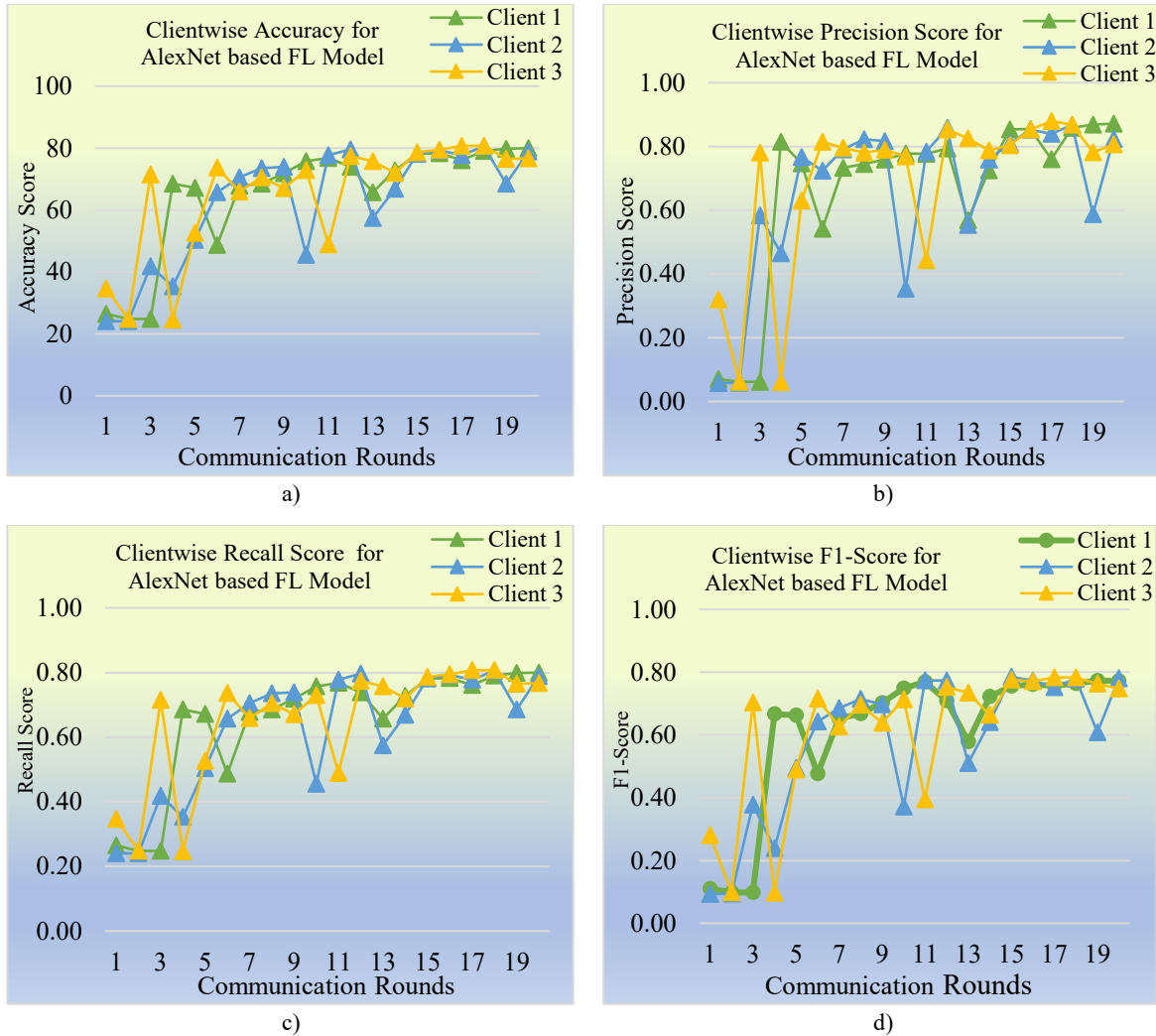


Figure 7. a) Clientwise accuracy, b) Clientwise precision, c) Clientwise recall, and d) Clientwise F1-score for AlexNet-based FL model.

Figure 7(a), presents the client-wise accuracy for the AlexNet-based FDL model used for detecting diabetic eye diseases such as Glaucoma, DME, and DR. For client 1, the accuracy in the initial rounds is very low and increases slowly. However, there is a sharp increase around round 4 and the accuracy keeps on increasing by round 7 and stabilizes by round 20. Despite this improvement, AlexNet struggles to maintain consistently high accuracy levels in the earlier rounds. In client 2, the accuracy scores in the starting rounds are lower than in client 1 and fluctuate significantly during the initial rounds. After round 5, the accuracy improves steadily, eventually reaching 78% by round 20. For client 3, initially better accuracy is shown than the other clients and fluctuates significantly during the first ten rounds. Thereafter, the model becomes stable. **Figure 7(b)**, represents the clientwise precision score for the AlexNet-based FDL model. Initially, in client 1, the precision score is extremely low, and from round 4 onwards, the precision score improves gradually. The precision peaks at 85% by round 15 and stabilizes by round 20. Similar to client 1, precision starts very low initially and keeps fluctuating till round 7, improves after round 7, and reaches a high by round 14. From 15 onwards, the precision score stabilizes, indicating a more consistent performance towards the later rounds. The precision score for client 3 starts very low as well, the precision score begins

to improve from round 4 and fluctuates around 80% between rounds 10 and 20, stabilizing by the final communication rounds. The client-wise recall score is presented in **Figure 7(c)** for the AlexNet-based FDL model. The recall score starts very low in the early communication rounds. Gradually, the recall improves and shows a steep increase starting from round 4 and increasing till round by round 7. The recall then stabilizes from round 10 to round 20. Similar to client 1, the recall for client 2 improves steadily after round 4 and reaches around 65-70% by round 10. By round 15, the recall stabilizes, indicating consistent performance in identifying positive cases in the later rounds. Client 3 begins with a higher recall score than the other two clients, by round 10, the recall stabilizes afterwards. The client-wise F1-score for the AlexNet-based FDL model is shown in **Figure 7(d)**. The F1-score for client 1 starts at a very low level of around in the first few communication rounds. From round 4 onwards, the F1-score improves significantly, by round 10. By round 15, the F1-score stabilizes around 75-77% and remains at this level until round 20. Client 2 also starts with a very low F1-score in the early rounds. The F1-score gradually improves after round 5, stabilizing by round 15. By round 20, client 2's F1-score reaches 77-78%, showing decent performance by the final rounds but still not ideal for medical use cases. Client 3 starts with a slightly higher F1-score in the early communication rounds. After round 5, the F1-score improves steadily till round 10 and stabilizes by 20th round.

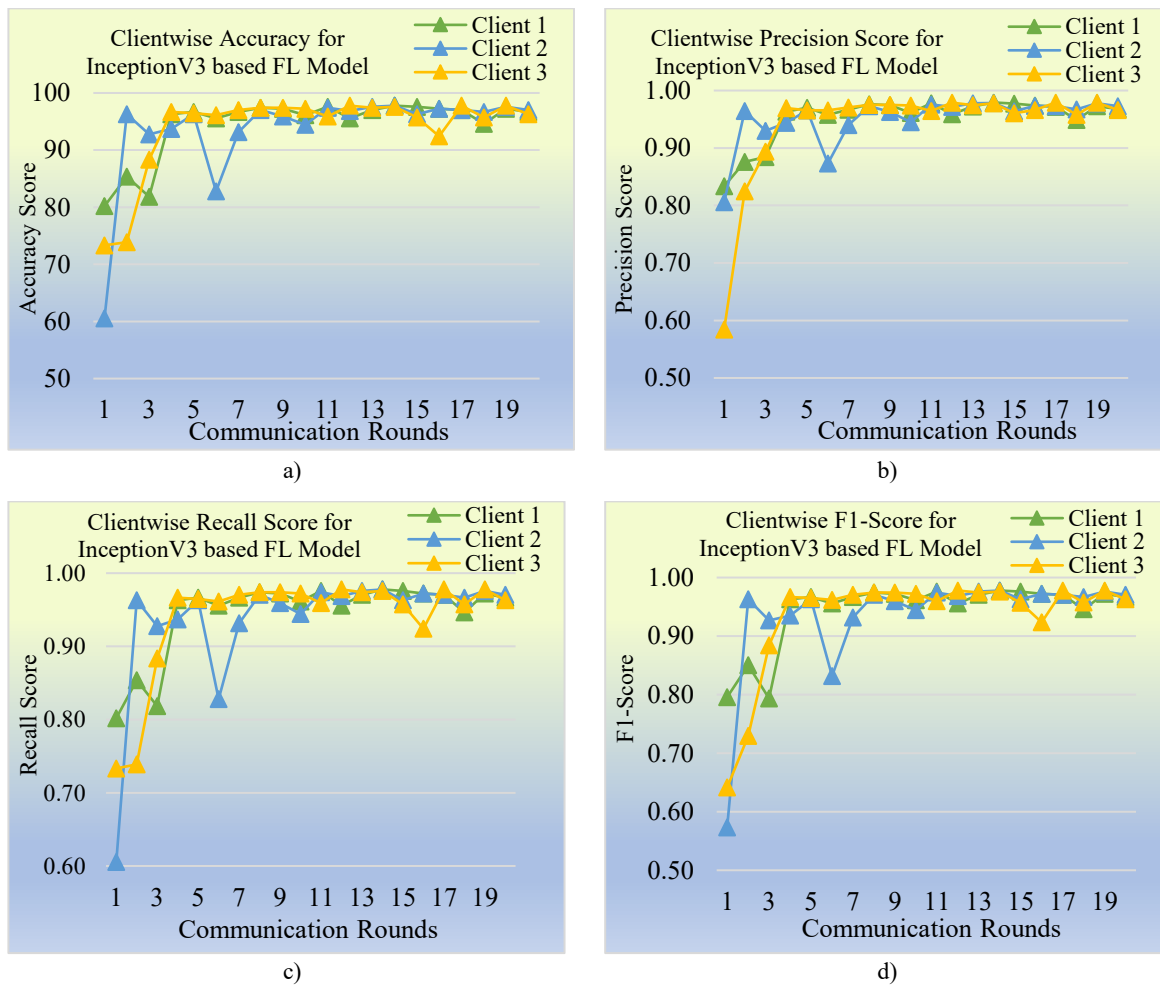


Figure 8. a) Clientwise accuracy, b) Clientwise precision, c) Clientwise recall, and d) Clientwise F1-score for inceptionV3-based FL model.

Figure 8(a) shows the client-wise accuracy for the InceptionV3-based FDL model used for detecting diabetic eye diseases. The accuracy for client 1 starts at around 80% in the first round, showing a strong initial performance. Accuracy consistently improves by round 4 and stabilizes around 97.22% between rounds 8 to 20. Client 2 starts with a significantly lower accuracy score compared to client 1 but improves dramatically after the 4th communication round; after that, the model's accuracy stabilizes. Client 3 starts at a medium accuracy level in the first round and improves steadily in the 4th to 8th round, stabilizes after round 8. The InceptionV3 architecture proves to be highly effective in this task, as evidenced by its ability to generalize well across all clients, even in the presence of decentralized and potentially heterogeneous data. The client-wise precision score for the InceptionV3-based FDL model used in detecting diabetic eye diseases is depicted in **Figure 8(b)**. The precision steadily increases for client 1 and stabilizes at a high value after the initial rounds (after round 5). Client 2 shows a similar pattern of increasing precision and stabilizes after round 8. Precision for client 3 improves gradually, starting lower than the other two clients but reaches stability around the 10th round at a similar level. The client-wise recall score for the InceptionV3-based FDL model used for detecting diabetic eye diseases is shown in **Figure 8(c)**. The recall score of clients 1 starts relatively high and remains stable throughout the communication rounds starting from 5th round. For client 2 the recall score shows similar patterns, starting around 60% in the initial rounds but quickly rising to stabilize around round 8. The recall score for client 3 starts at a medium level, improves steadily till round 4, and then stabilizes after round 4. The high recall values across all three clients demonstrate that the InceptionV3-based FL model is very effective at detecting true positive cases in the datasets used for diabetic eye disease detection. **Figure 8(d)** presents the client-wise F1-score for the InceptionV3-based FDL model used to detect diabetic eye diseases such as DR, DME, and Glaucoma. The F1-score for client 1 initially is high at around 79% and steadily increases to stabilize after the 5th round. Initially, client 2's F1 score starts lower compared to client 1, but it increases sharply within a few communication rounds and stabilizes after the 10th round. The F1-score for client 3 begins in the mid-range, but it quickly improves and reaches stability by the 8th round. The figure demonstrates that the F1-scores are consistent across all clients after a few communication rounds despite variations in their local datasets.

Figure 9(a) shows the prediction accuracy of FDL models with EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures in identifying DME, DR, Glaucoma, and healthy eyes. The FDL models with EfficientNetB0 and InceptionV3 showcase a higher accuracy performance with very small variations. However, the FDL model with the InceptionV3 model reaches the score after a few communication rounds. The FDL model with MobileNet V2 achieves a high prediction accuracy quite early, but it has more variations in comparison to the EfficientNetB0 and InceptionV3-based FDL models. In contrast, the AlexNet-based FDL model struggles with a lower peak accuracy and is considerably unstable in its performance over the various communication rounds. Therefore, based on prediction accuracy score comparison, the FDL model with EfficientNetB0 performs consistently well throughout with more stability and, hence, is suited for the privacy-enhanced diabetic eye disease detection, even though the data used is decentralised and heterogeneous. A comparative analysis of the precision scores achieved by the four FDL models based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures across different communication rounds is presented in **Figure 9(b)**. The EfficientNetB0 and InceptionV3-based FDL models achieve high precision scores while maintaining consistent performance throughout. This is indicative of the effectiveness in making accurate positive predictions. The FDL model with MobileNetV2-based architecture obtains a high precision score and has slight variations in comparison to the FDL models with EfficientNetB0 and MobileNetV2 but largely remains stable. Contrastively, the AlexNet-based FDL model achieves lower precision scores with a lot of variations across the communication rounds. Overall, the FDL model based on EfficientNetB0 architecture achieves a high precision from the beginning rounds and maintains high scores throughout the communication rounds. The recall scores achieved by four FDL

models with EfficientNetB0, MobileNetV2, AlexNet, and Inception architectures across various communication rounds are shown in **Figure 9(c)**. All three FDL models based on EfficientNetB0, MobileNetV2, and InceptionV3 perform well in terms of the recall scores achieved by these models. However, the InceptionV3 achieved a higher score a little later than the EfficientNetB0 and MobileNetV2-based architectures. Out of the three, the most stable is the EfficientNetB0, and MobileNetV2 has small deviations and is less stable when the standard deviation is considered. The AlexNet-based FDL model performance in terms of recall score is relatively much lower and suffers considerable variability. Thus, pointing to its inefficiency in correctly identifying true positive cases. Therefore, the best performance in accordance with the recall score is achieved by the EfficientNetB0 model. **Figure 9(d)** depicts the F1 scores for various FDL models—EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3—across different communication rounds within a federated learning framework. Again, the EfficientNetB0 and InceptionV3-based FDL models achieve high F1 scores, reflecting that they maintain an excellent balance between precision and recall, which is crucial for applications requiring accurate and consistent performance. The FDL model with MobileNetV2 architecture also achieves good F1 scores, showing effective balance with minor inconsistencies across rounds. The AlexNet-based FDL model displays significantly lower and more fluctuating F1 scores, indicating an inadequate balance between precision and recall.

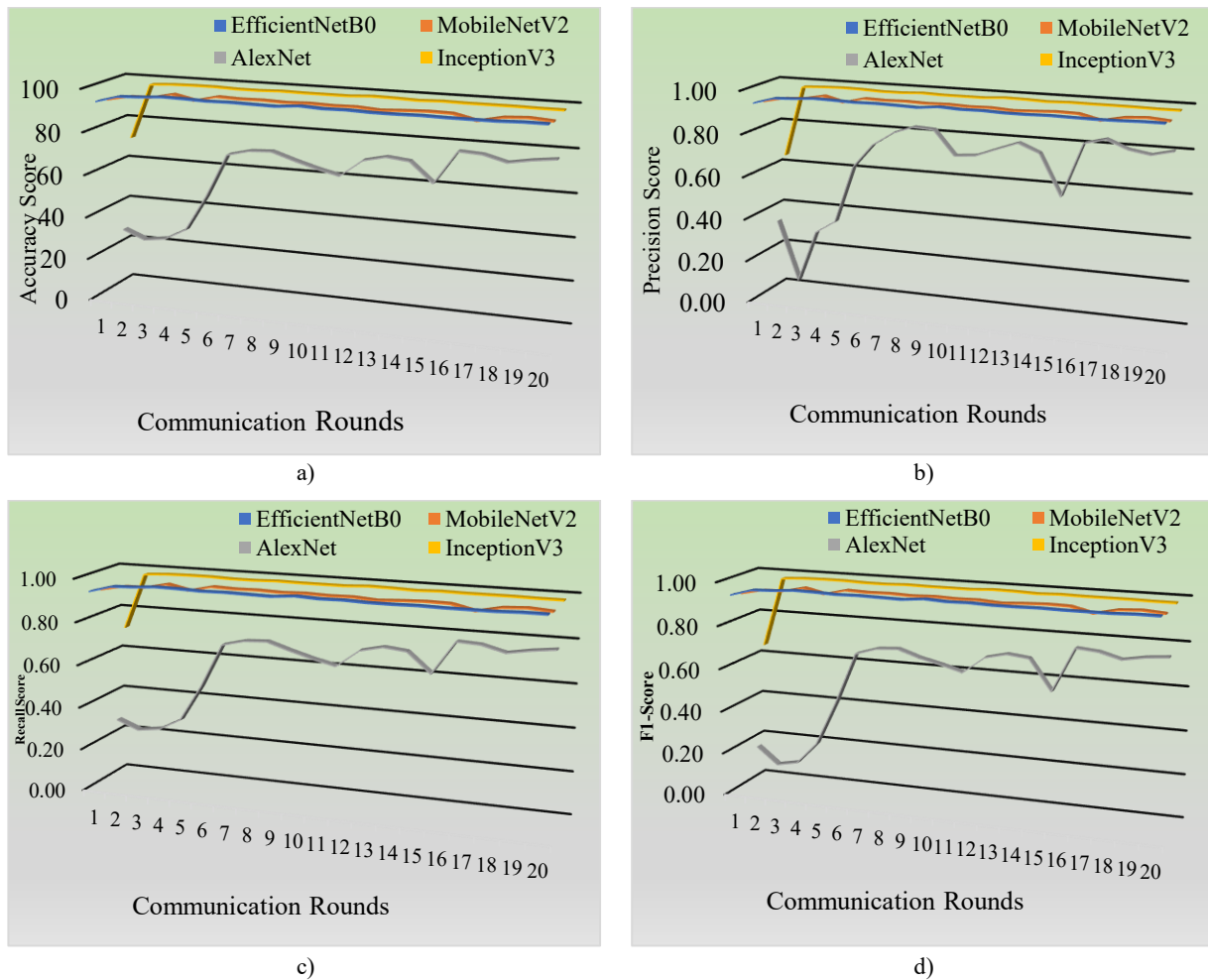


Figure 9. Comparison of the a) Global accuracy, b) Global precision, c) Global recall, and d) Global F1-score achieved by various models over communication rounds.

Figure 10(a) shows the prediction accuracy across the communication rounds of the four FDL models implemented for identifying DME. Both the EfficientNetB0 and MobileNetV2 architecture-based FDL models perform consistently well throughout the communication rounds, and they are also stable in their performance. Thus, demonstrating strong capabilities in identifying the DME disease in coloured fundus images. The FDL model is based on the InceptionV3 model is generally effective and maintains a commendable performance in the identification of DME, still it has some inconsistencies due to which it is unstable in between. The FDL model with AlexNet architecture displays a lower and inconsistent performance in comparison to the other models that achieve a high accuracy score in a few rounds. The performance comparison based on the accuracy of prediction of the FDL based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures for detecting DR is shown in **Figure 10(b)**. An exceptional performance in terms of accuracy score and model stability is shown by EfficientNetB0 and InceptionV3, proving to be highly capable of detecting DR. The MobileNetV2-based FDL model performed higher in terms of the accuracy achieved, yet it is not as consistent in comparison to the EfficientNetB0 and InceptionV3 models. The AlexNet-based FDL model does not perform as well when compared to the other FDL models and is not a stable model, showing high fluctuations across the communication rounds. Thus, the AlexNet-based FDL model is not reliable for identifying DR in coloured fundus images. The performance in terms of prediction accuracy of the four FDL models based on EfficientNetB0, AlexNet, InceptionV3, and MobileNetV2 architecture for the detection of Glaucoma is depicted in **Figure 10(c)** throughout the 20 communication rounds. A superior performance in terms of accuracy is showcased by the FDL models based on MobileNetV2 and EfficientNetB0 architectures. The FDL models based on MobileNetV2 and EfficientNetB0 are also more stable in their performance in comparison to the FDL models with InceptionV3 and AlexNet architectures. The FDL model with Inception V3 architecture exhibits a good performance but is not as stable and converges late when compared to the EfficientNetB0 and MobileNetV2 architectures in identifying Glaucoma. The InceptionV3 exhibits a good performance but is not as stable and converges late when compared to EfficientNetB0 and MobileNetV2 in identifying Glaucoma. However, the AlexNet-based FDL model's performance is not stable and not up to the mark when identifying Glaucoma. The results presented in **Figure 10(d)** are the comparison of accuracy scores for the Normal class obtained by FDL models based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 architectures across multiple communication rounds. The FDL models with EfficientNetB0 and InceptionV3 architectures demonstrate high performance after initial rounds and the accuracy becomes high consistently throughout. Therefore, both models are able to identify the normal class cases effectively. This means that models are robust and minimize false positive cases. The FDL model with MobileNetV2 architecture also performs well with little deviations. This shows that the MobileNetV2-based FDL model is also a reliable model for identifying healthy cases. In contrast, the FDL model with AlexNet architecture performs poorly and has a lot of fluctuations. Thus, indicating that the AlexNet-based model is not suitable for identifying the normal cases, this may lead to misclassification.

Figure 11 compares the proposed FDL model's performance with the state-of-the-art models for detecting diabetic eye diseases such as DR, DME, and glaucoma. The research works presented above differ in terms of learning scenarios, their dataset composition, and the task scope. Many of these are based on single-disease classification, while the proposed approach uses a federated multi-disease framework. Therefore, the comparison is made as a general performance reference rather than a strictly equivalent benchmark due to differences in datasets and experimental settings. The proposed work achieves a significantly higher accuracy score of 97.96%, while the other models, like the vision transformer-based model, achieve 95%, and an improved DenseNet model achieved 94.27% accuracy. This showcases the ability of the proposed model in handling federated learning environments while ensuring data privacy and achieving higher diagnostic accuracy for detecting multiple diabetic eye diseases.

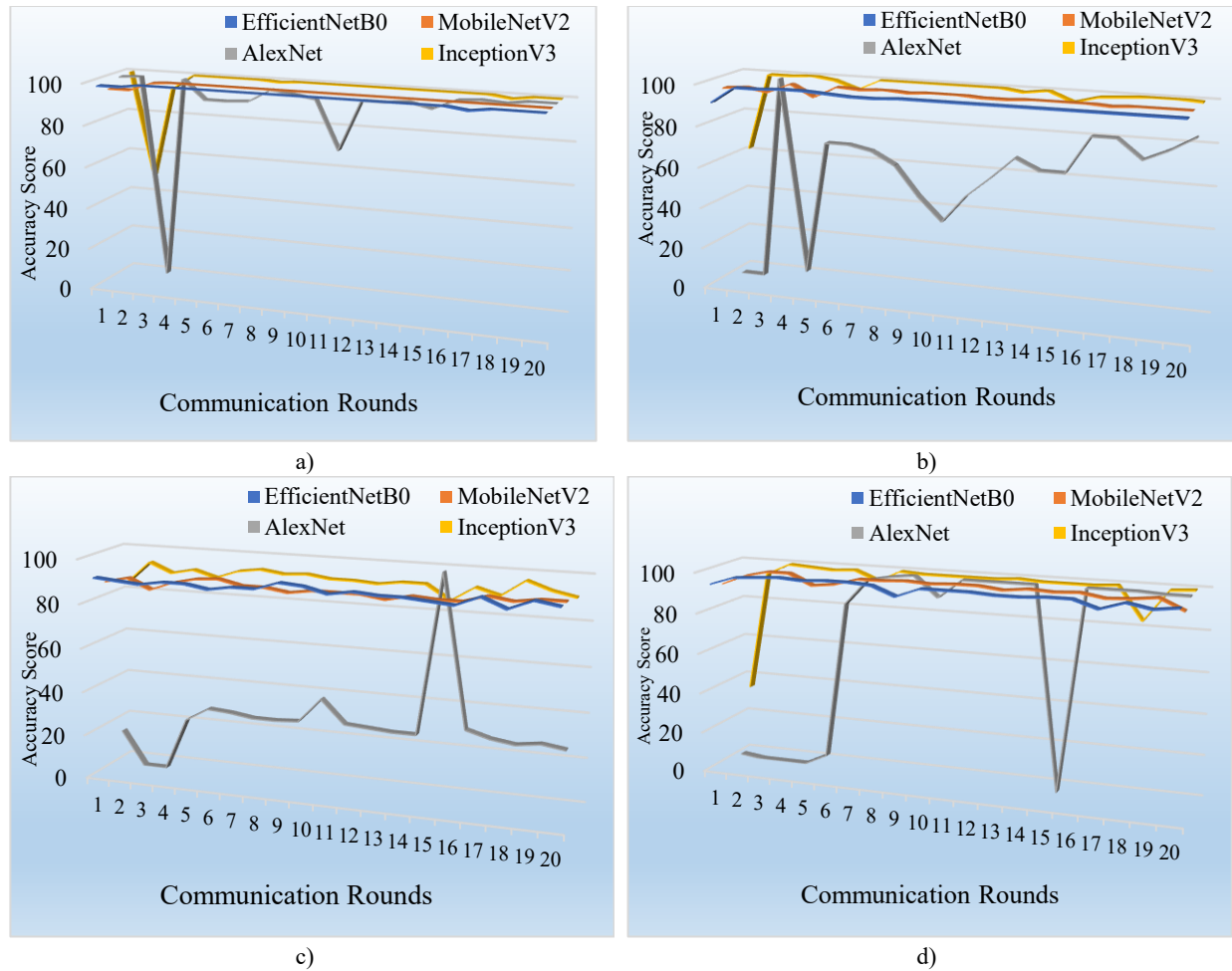


Figure 10. Presentation of the obtained accuracy scores for the a) DME, b) DR, c) Glaucoma, and d) Normal class with different models over communication rounds.

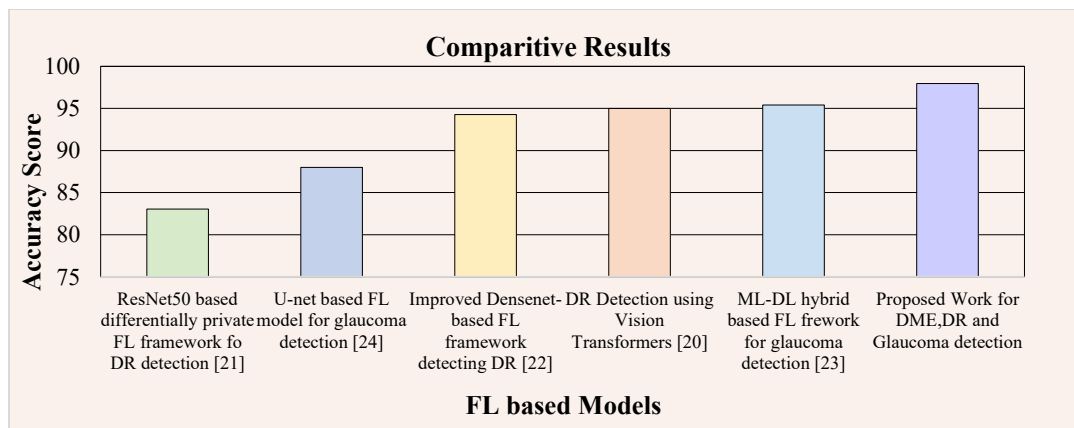


Figure 11. Comparison of the proposed work with the existing state-of-art models for the detection of diabetic eye diseases.

6. Conclusion and Future Work

The article presented a privacy-preserved and collaboratively trained federated deep learning framework for detecting diabetic eye diseases, including Diabetic Retinopathy (DR), Diabetic Macular Edema (DME), and Glaucoma. The proposed federated deep learning framework addresses the key challenge in centralized machine learning and data privacy. The use of federated learning ensures that sensitive patient information remains secure due to the training process being distributed across decentralised data sources. The utilisation of FedProx aggregation mechanism in the framework helps in dealing and mitigating the impact of data heterogeneity across clients. Thus, enhancing the stability and robustness of the global model.

The performance of multiple federated deep learning models based on EfficientNetB0, MobileNetV2, AlexNet, and InceptionV3 is evaluated and compared. The results demonstrated that the EfficientNetB0 and InceptionV3 based federated models outperformed others, achieving over 97.96% accuracy across communication rounds, making them highly reliable for detecting diabetic eye diseases. In terms of stability, the EfficientNetB0-based model is the highest, with the slightest deviation. In contrast, AlexNet exhibited lower accuracy and instability, highlighting the importance of selecting appropriate models for healthcare applications. Therefore, the EfficientNetB0-based federated deep learning model is the most suited for detecting diabetic eye diseases.

Future work may focus on further optimizing the framework for non-IID data, exploring additional models, and extending the application of federated learning to other critical healthcare domains. This research work contributes to the growing secure, decentralized machine learning field, paving the way for more advanced, privacy-conscious healthcare solutions. Additionally, cross-validation and multi-seed experiments might be added to further strengthen the statistical reliability of the results. Additionally, patient-wise data partitioning might also be explored where metadata is available to better simulate real-world federated learning scenarios.

Furthermore, the integration of advanced privacy-preserving techniques such as secure aggregation, encryption and differential privacy and analysis of communication overhead and communication-efficient strategies may be explored for improving practical deployment.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors would like to thank the editors and reviewers for their valuable comments and suggestions. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

AI Disclosure

The author(s) declare that no assistance is taken from generative AI to write this article.

Availability of Data and Material

The dataset used in this research is available at the following repositories.

<https://github.com/Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET>

<https://ieee-dataport.org/open-access/indian-diabetic-retinopathy-image-dataset-idrid>

<https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k>

<https://onlinelibrary.wiley.com/doi/epdf/10.1155/2013/154860>

Author Contributions

Conceptualization: Seema Gulati, Kalpna Guleria; Methodology: Seema Gulati, Kalpna Guleria; Formal Analysis: Seema Gulati, Kalpna Guleria, Nitin Goyal, Sushil Kumar, Sunita Tiwari; Validation: Seema Gulati, Kalpna Guleria, Nitin Goyal, Sushil Kumar, Sunita Tiwari; Investigation: Seema Gulati, Kalpna Guleria, Nitin Goyal, Sushil Kumar, Sunita Tiwari; Writing - original draft preparation: Seema Gulati, Kalpna Guleria; Writing - Review and Editing: Seema Gulati, Kalpna Guleria; Supervision: Kalpna Guleria, Nitin Goyal; All authors have read and agreed to this version of the manuscript.

References

- Abràmoff, M.D., Reinhardt, J.M., Russell, S.R., Folk, J.C., Mahajan, V.B., Niemeijer, M., & Quêlleç, G. (2010). Automated early detection of diabetic retinopathy. *Ophthalmology*, 117(6), 1147-1154. <https://doi.org/10.1016/j.ophtha.2010.03.046>
- Al-Fahdawi, S., Al-Waisy, A.S., Zeebaree, D.Q., Qahwaji, R., Natiq, H., Mohammed, M.A., Nedoma, J., Martinek, R., & Deveci, M. (2024). Fundus-DeepNet: multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. *Information Fusion*, 102, 102059. <https://doi.org/10.1016/j.inffus.2023.102059>
- Aljohani, A., & Aburasain, R.Y. (2024). A hybrid framework for glaucoma detection through federated machine learning and deep learning models. *BMC Medical Informatics and Decision Making*, 24(1), 115. <https://doi.org/10.1186/s12911-024-02518-y>
- Baptista, T., Soares, C., Oliveira, T., & Soares, F. (2023). Federated learning for computer-aided diagnosis of glaucoma using retinal fundus images. *Applied Sciences*, 13(21), 11622. <https://doi.org/10.3390/app132111620>
- Budai, A., Bock, R., Maier, A., Hornegger, J., & Michelson, G. (2013). Robust vessel segmentation in fundus images. *International Journal of Biomedical Imaging*, 2013, 154860. <https://doi.org/10.1155/2013/154860>
- Chaddad, A., Wu, Y., & Desrosiers, C. (2024). Federated learning for healthcare applications. *IEEE Internet of Things Journal*, 11(5), 7339-7358. <https://doi.org/10.1109/JIOT.2023.3325822>
- Chetoui, M., & Akhloufi, M.A. (2023). Federated learning for diabetic retinopathy detection using vision transformers. *BioMedInformatics*, 3(4), 948-961. <https://doi.org/10.3390/biomedinformatics3040058>
- Deepak, G.D., & Bhat, S.K. (2024). Deep learning-based CNN for multiclassification of ocular diseases using transfer learning. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 12(1), 2335959. <https://doi.org/10.1080/21681163.2024.2335959>
- Ding, J., & Wong, T.Y. (2012). Current epidemiology of diabetic retinopathy and diabetic macular edema. *Current Diabetes Reports*, 12(4), 346-354. <https://doi.org/10.1007/s11892-012-0283-6>
- Gulati, S., Guleria, K., & Goyal, N. (2024). Collaborative, privacy-preserving federated learning framework for the detection of diabetic eye diseases. *SN Computer Science*, 5(8), 1100. <https://doi.org/10.1007/s42979-024-03462-4>
- Gulati, S., Guleria, K., & Goyal, N. (2025). Privacy-preserving and collaborative federated learning model for the detection of ocular diseases. *International Journal of Mathematical, Engineering and Management Sciences*, 10(1), 218-248. <https://doi.org/10.33889/IJMEMS.2025.10.1.013>
- Hartmann, F. (2018). Federated learning [Bachelor's thesis, Freie Universität Berlin]. https://www.mi.fu-berlin.de/inf/groups/ag-ti/theses/download/Hartmann_F18.pdf
- Hossain, I., Puppala, S., & Talukder, S. (2023). Collaborative differentially private federated learning framework for the prediction of diabetic retinopathy. In *2023 IEEE 2nd International Conference on AI in Cybersecurity* (pp. 1-6). IEEE. Houston, TX, USA. <https://doi.org/10.1109/ICAIC57335.2023.10044122>

- Hughes-Cano, J.A., Quiroz-Mercado, H., Hernández-Zimbrón, L.F., García-Franco, R., Rubio Mijangos, J.F., López-Star, E., García-Roa, M., Lansingh, V.C., Olivares-Pinto, U., & Thébault, S.C. (2024). Improved predictive diagnosis of diabetic macular edema based on hybrid models: an observational study. *Computers in Biology and Medicine*, 170, 107979.
- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3), 685-695. <https://doi.org/10.1007/s12525-021-00475-2>
- Jin, Y., Zhu, H., Xu, J., & Chen, Y. (2023). Secure federated learning. In *Federated Learning: Fundamentals and Advances* (pp. 165-212). Springer Nature, Singapore. https://doi.org/10.1007/978-981-19-7083-2_4
- Kairouz, P., & McMahan, H.B. (2021). Advances and open problems in federated learning. *Foundations and Trends in Machine Learning*, 14(1-2), 1-210. <https://doi.org/10.1561/22000000083>
- KhoKhar, F.A., Shah, J.H., Khan, M.A., Sharif, M., Tariq, U., & Kadry, S. (2022). A review on federated learning towards image processing. *Computers and Electrical Engineering*, 99, 107818. <https://doi.org/10.1016/j.compeleceng.2022.107818>
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, 429-450. <https://doi.org/10.48550/arXiv.1812.06127>
- Nam, D.J. (2024). Distributed glaucoma detection method using privacy-preserving federated learning. *International Journal of High School Research*, 6(1), 20-24. https://terra-docs.s3.us-east-2.amazonaws.com/IJHSR/Articles/volume6-issue1/IJHSR_2024_61_20.pdf
- Nasajpour, M., Karakaya, M., Pouriyeh, S., & Parizi, R.M. (2022). Federated transfer learning for diabetic retinopathy detection using CNN architectures. In *SoutheastCon 2022* (pp. 655-660). IEEE. Mobile, AL, USA. <https://doi.org/10.1109/SOUTHEASTCON48659.2022.9764031>
- Porwal, P., Pachade, S., Kamble, R., Kokare, M., Deshmukh, G., Sahasrabuddhe, V., & Meriaudeau, F. (2018). Indian diabetic retinopathy image dataset. *IEEE Dataport*. <https://doi.org/10.21227/H25W98>
- Qummar, S., Khan, F.G., Shah, S., Khan, A., Din, A., & Gao, J. (2020). Deep learning techniques for diabetic retinopathy detection. *Current Medical Imaging*, 16(10), 1201-1213. <https://doi.org/10.2174/1573405616666200213114026>
- Safi, H., Safi, S., Hafezi-Moghadam, A., & Ahmadi, H. (2018). Early detection of diabetic retinopathy. *Survey of Ophthalmology*, 63(5), 601-608. <https://doi.org/10.1016/j.survophthal.2018.04.003>
- Shanggong Medical Technology Co., Ltd. (2019). *Ocular disease intelligent recognition (ODIR) [Data set]*. <https://odir2019.grand-challenge.org/dataset/>
- Singh, A., & Singh, K.K. (2023). FedDDR: a federated improved DenseNet for classification of diabetic retinopathy. In *Proceedings of the 6th International Conference on Informatics & Data-Driven Medicine* (pp. 73-81). CEUR Workshop Proceeding. Bratislava, Slovakia. <https://ceur-ws.org/Vol-3609/paper10.pdf>
- Translational Visual Health Laboratory. (2024). OCT and eye fundus dataset [Data set]. GitHub. <https://github.com/Traslational-Visual-Health-Laboratory/OCT-AND-EYE-FUNDUS-DATASET>
- Wilkinson, C.P., Ferris, F.L., Klein, R.E., Lee, P.P., Agardh, C.D., Davis, M., Dills, D., Kampik, A., Pararajasegaram, R., Verdaguer, J.T., & Global Diabetic Retinopathy Project Group. (2003). Proposed international clinical diabetic retinopathy and diabetic macular edema disease severity scales. *Ophthalmology*, 110(9), 1677-1682. [https://doi.org/10.1016/S0161-6420\(03\)00475-5](https://doi.org/10.1016/S0161-6420(03)00475-5)
- Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 1-19. <https://doi.org/10.1145/3298981>
- Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2021). A survey on federated learning. *Knowledge-Based Systems*, 216, 106775. <https://doi.org/10.1016/j.knosys.2021.106775>

Zhang, T., & Mao, S. (2022). An introduction to the federated learning standard. *GetMobile: Mobile Computing and Communications*, 25(3), 18-22. <https://doi.org/10.1145/3511285.3511291>

Original content of this work is copyright © Ram Arti Publishers. Uses under the Creative Commons Attribution 4.0 International (CC BY 4.0) license at <https://creativecommons.org/licenses/by/4.0/>

Publisher's Note- Ram Arti Publishers remains neutral regarding jurisdictional claims in published maps and institutional affiliations.