scientific reports

Check for updates

OPEN

Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings

Rahul Haripriya✉, Nilay Khare & Manish Pandey

Ensuring data privacy in medical image classification is a critical challenge in healthcare, especially with the increasing reliance on AI-driven diagnostics. In fact, over 30% of healthcare organizations globally have experienced a data breach in the last year, highlighting the need for secure solutions. This study investigates the integration of transfer learning and federated learning for privacy-preserving medical image classification using GoogLeNet and VGG16 as baseline models to evaluate the generalizability of the proposed framework. Pre-trained on ImageNet and fine-tuned on three specialized medical datasets for TB chest X-rays, brain tumor MRI scans, and diabetic retinopathy images, these models achieved high classification accuracy across various aggregation methods. Additionally, the proposed dynamic aggregation method was further analyzed using modern architectures, EfficientNetV2 and ResNet-RS, to assess the scalability and robustness of the model. A key contribution is the introduction of a novel adaptive aggregation method, which dynamically alternates between Federated Averaging (FedAvg) and Federated Stochastic Gradient Descent (FedSGD), based on data divergence during communication rounds. This approach optimizes model convergence while preserving privacy in collaborative settings. The results demonstrate that transfer learning, when combined with federated learning, offers a scalable, robust, and secure solution for real-world medical diagnostics, enabling healthcare institutions to train highly accurate models without compromising sensitive patient data.

Keywords Federated learning, Machine learning, Artificial intelligence, Data privacy, Transfer learning, Image classification, Deep learning

The domain of medical imaging has experienced substantial growth in recent decades, propelled by advancements in artificial intelligence (AI), machine learning (ML), and deep learning (DL)[1,2]. Historically, medical image analysis depended on manual interpretation by radiologists and doctors, a labour-intensive procedure prone to variations in competence and human error. The merging of AI and ML techniques has altered medical image processing, providing more exact, consistent, and automated solutions for disease diagnosis and prediction[3,4]. Deep learning has transformed medical imaging by allowing models to autonomously learn complex features from raw image data, eliminating the necessity for manual feature extraction and significantly enhancing the accuracy and efficiency of medical diagnostics[5].

Artificial Intelligence, Machine Learning, and Deep Learning have become essential instruments in the identification and examination of both communicable and non-communicable diseases[6]. Communicable illnesses such as tuberculosis (TB), a significant global health concern, have experienced substantial advancements in early identification and treatment planning owing to AI-driven picture categorisation approaches. Chest X-rays[7], a fundamental diagnostic instrument for pneumonia or tuberculosis, are now evaluated using convolutional neural network (CNN) models[8], which may identify even subtle indications of the disease with greater precision than conventional techniques. This is essential in addressing tuberculosis, especially in resource-constrained environments, where automated AI solutions can assist overwhelmed healthcare systems. Research indicates that AI-driven diagnostic tools can attain sensitivity and specificity on par with professional radiologists, hence improving early detection efforts and bolstering worldwide tuberculosis eradication operations.

Likewise, for non-communicable disorders such as brain tumours and diabetic retinopathy, deep learning models have significantly enhanced medical picture processing[9–11]. Brain tumours, typically identified via MRI scans, necessitate accurate segmentation and classification for optimal treatment planning. Deep learning

Department of Computer Science and Engineering, Maulana Azad National Institute of Technology, Bhopal 462003, India. ✉email: rahulharipriyamanit@gmail.com

methodologies, especially CNNs[12], have exhibited remarkable efficacy in differentiating between benign and malignant tumours, enhancing survival rates through earlier and more precise diagnosis. Deep learning models are currently extensively utilised to analyse retinal pictures and identify early-stage microaneurysms or haemorrhages in diabetic retinopathy, a prevalent eye condition and a leading cause of global blindness. AI-assisted technologies have enhanced the accessibility of diabetic retinopathy screening, particularly in underprivileged regions, while automated methods have alleviated the diagnostic workload for ophthalmologists, facilitating large-scale screening initiatives.

The integration of AI, ML, and DL in medical image processing has substantially enhanced diagnostic precision[13], lowered healthcare expenses, and broadened access to quality medical services. The swift integration of these technologies in the analysis of both communicable diseases and non-communicable diseases. Building upon the growing evidence of AI models' effectiveness in improving medical image classification, this work emphasizes the importance of privacy-preserving methods through federated learning[14,15]. This study seeks to enhance the application of AI in medical diagnostics by tackling the issues of privacy, accuracy, and scalability, thereby contributing to global healthcare improvement efforts. The involved architecture of federated learning can be seen in Fig. 1.

The selection of GoogLeNet and VGG16 as the baseline models for medical image classification via transfer learning was driven by their well-documented efficacy, robustness, and performance across a wide range of medical datasets. Both models, pre-trained on ImageNet, were fine-tuned on specialized medical image datasets, including TB chest X-rays, brain tumor MRI scans, and diabetic retinopathy images, to evaluate the generalizability of the proposed dynamic aggregation approach. GoogLeNet's capability to perform multi-scale feature extraction and VGG16's deeper architecture enabled the capture of intricate patterns in medical images, which are critical for detecting delicate features essential for accurate diagnosis.

To further assess the scalability and robustness of the proposed framework, modern architectures such as EfficientNetV2 and ResNet-RS were employed. EfficientNetV2's advanced compound scaling method and ResNet-RS's optimized residual connections made them ideal for handling high-resolution and complex medical image datasets. These modern architectures demonstrated superior performance in resource-efficient training and adaptability to diverse data distributions, validating the framework's effectiveness under varying conditions. Together, the combination of baseline and modern architectures ensures a comprehensive evaluation of the proposed approach, showcasing its generalizability, scalability, and robustness in privacy-preserving medical diagnostics.

A novel approach is introduced by employing adaptive aggregation in a federated learning baseline performance evaluation framework, as illustrated in Fig. 2, specifically designed for a multi-medical center image research model. This model enables multiple medical centers to leverage each other's learning without ever sharing sensitive patient data, ensuring privacy while enhancing collaborative learning. The adaptive aggregation dynamically switches between FedAvg and FedSGD based on observed data divergence during communication rounds. This mechanism optimizes model convergence and improves efficiency in the federated learning setup, allowing the models to generalize more effectively in heterogeneous medical data settings. The integration of pre-trained models, such as GoogLeNet and VGG16 for baseline evaluation, further enhances the efficacy of transfer learning in distributed environments while maintaining the privacy benefits of federated learning. The adaptive aggregation ensures that the framework adapts to the variability of data across clients, significantly improving both convergence speed and accuracy, making it particularly beneficial for privacy-sensitive, decentralized medical applications.

A comprehensive comparison was conducted between GoogLeNet and VGG16 in the first phase of the experiment to establish baseline performance and evaluate the generalizability of the proposed dynamic aggregation approach. Both models were fine-tuned via transfer learning in centralized and federated settings, with differential privacy implemented in the centralized environment to safeguard sensitive medical data during model training. In the federated setup, privacy was preserved through decentralized learning protocols, ensuring no raw data was shared across clients. This phase provided foundational insights into the trade-offs

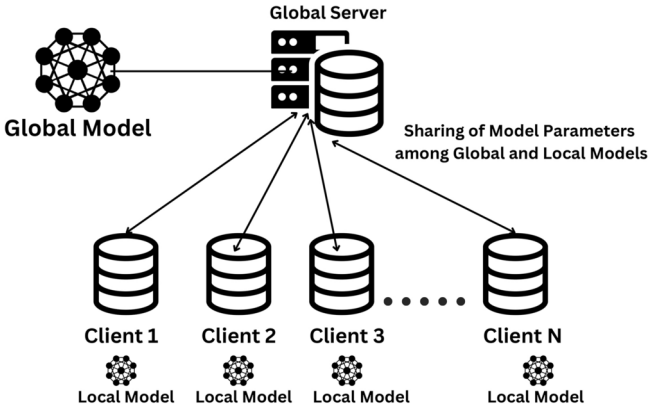

Fig. 1. Federated learning architecture.

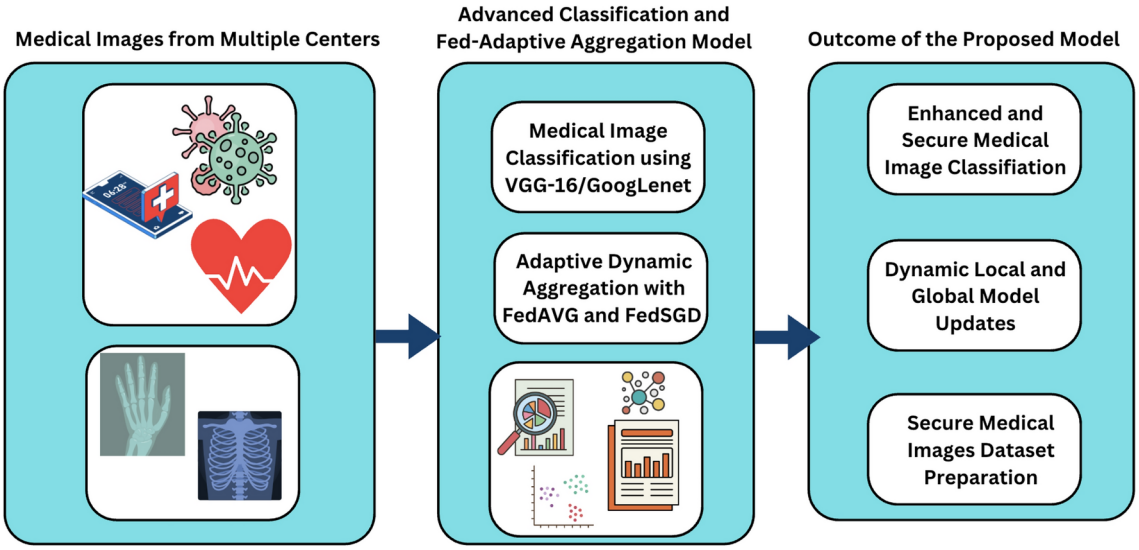

Fig. 2. Framework of the proposed model for baseline evaluation.

between accuracy, privacy, and computational efficiency, serving as a benchmark for evaluating the proposed methodology.

In the second phase, modern architectures, namely EfficientNetV2 and ResNet-RS, were evaluated within the proposed dynamic aggregation framework to assess its scalability and robustness. These advanced models were tested to determine their ability to enhance performance in resource-efficient training while adapting to non-IID data distributions in federated environments. The evaluation revealed that modern architectures exhibited superior performance in the proposed framework, achieving higher accuracy and demonstrating better adaptability compared to the baseline models. This two-phased experimental approach ensures a comprehensive analysis of the proposed methodology, showcasing its effectiveness across both traditional and modern deep learning architectures for privacy-preserving medical image classification.

By introducing adaptive aggregation, this study significantly advances the federated learning paradigm, offering a scalable and flexible method for optimizing convergence in distributed medical image classification tasks. This innovation has the potential to transform privacy-preserving AI solutions for medical diagnostics by enabling healthcare institutions to train highly accurate models while maintaining stringent privacy standards. The adaptive aggregation method not only enhances the efficiency of the federated learning setup but also ensures that generalization across diverse medical datasets is maintained, making the process more efficient and secure for collaborative learning across medical centers.

The remainder of this paper is organized as follows. "Related work" presents an overview of the related work and background studies relevant to this research, including previous developments in federated learning and privacy-preserving medical data mining. "Experimental framework" elaborates on the methodology employed in this study, detailing the datasets utilized and the novel model introduced for collaborative medical image classification using adaptive federated aggregation techniques. This section highlights the unique aspects of the federated learning framework implemented across multiple medical centers. "Results" discusses the results of the experiments, providing a comprehensive analysis of the key findings, including model performance under different settings and the impact of the proposed adaptive aggregation strategy. "Discussion" evaluates the overall performance and generalizability of the proposed model, emphasizing the practical implications of this research for privacy-preserving, decentralized medical data analysis. Finally, "Conclusion" concludes the article by summarizing the key contributions, addressing limitations of the study, and outlining potential future research directions that could further enhance federated learning in medical data mining and classification tasks.

The current challenges in leveraging federated learning for privacy-preserving medical image analysis necessitate building on prior advancements while addressing scalability and adaptability issues. The following section discusses existing literature on federated learning, medical image classification, and related methodologies, providing context for the proposed framework.

Related work

The growing interest in federated learning has resulted in diverse applications across domains, including healthcare, where privacy concerns are paramount. Existing studies have explored techniques for aggregating decentralized data, applying transfer learning to improve model performance, and addressing challenges posed by non-IID data. This section critically reviews the most relevant advancements in these areas, highlighting their limitations and their influence on the development of the proposed adaptive aggregation framework.

A comprehensive comparison of recent state-of-the-art literature pertaining to the application of pre-trained models in image classification and healthcare domains is presented in Table 1. This comparative analysis highlights the significant advancements and emerging trends in the field, emphasizing the potential of pre-

Reference	Local model algorithm (ML model)	Global model algorithm (FL model)	Key contribution	Limitation
Abu El Houda 2022[16]	Gradient descent (GD), CNN (for image classification)	Ring-AllReduce FL with threshold secret sharing	Introduced a decentralized FL model for robust and privacy-preserving gradient aggregation. Enhanced resilience to client dropouts using secret sharing.	Communication overhead due to lack of optimization for model aggregation in large-scale settings
Ullah 2023[17]	CNN (for COVID-19 X-ray classification)	Federated averaging (FedAvg)	Applied FL to COVID-19 diagnosis, improving classification accuracy across distributed clients (hospitals) without centralized data sharing.	Performance decreases with fewer clients, and the model's adaptability to other diseases was not thoroughly evaluated
Abaoud 2023[18]	Logistic regression, decision trees	Secure multi-party computation (SMPC) based FL	Proposed a privacy-preserving FL model with local differential privacy and secure computation, focusing on scalability and security	Increased computation time due to secure computation techniques, and scalability to larger datasets was not fully tested
Shiranthika 2023[19]	Multi-layer perceptron (MLP)	Decentralized ring-based FL	Focused on reducing communication overhead in FL systems, using a ring-based model to ensure better scalability for large-scale applications	The approach is less effective for models with more complex architectures like CNNs, and privacy concerns were not fully addressed
Kalapaaking 2023[20]	Deep neural network (DNN) with two hidden layers	SMPC-based FL combined with blockchain	Introduced HealthFed, integrating blockchain with FL to ensure decentralized privacy-preserving collaboration among healthcare institutions. Tested on Breast Cancer Wisconsin dataset	High complexity due to blockchain integration; computational costs increase with more collaborators
Haripriya 2024 [this study]	GoogLenet and VGG-16 for multiple healthcare image datasets	Novel adaptive aggregation FedAVG+FedSGD	Introduced adaptive aggregation as a novel approach to federated learning, combining the strengths of both FedAvg and FedSGD with dynamic aggregation on the basis of convergence	Computational overhead was significant for larger models like VGG-16, indicating the need for more optimized architectures or compression techniques in future work

Table 1. Comparison of related work in FL-based healthcare systems.

trained models to revolutionize various healthcare applications. Notably, several studies have employed federated learning techniques to enhance privacy and security while training and deploying these models.

Jin 2023[21] offer a co-aware personalised federated learning model that integrates metadata and picture information for intelligent medical applications. The system incorporates both metadata and picture elements to improve diagnostic accuracy. A tailored federated learning technique utilising parameterised knowledge transfer is developed to enhance model performance by acquiring insights from edge nodes with greater contributions. A Naïve Bayes classifier is employed to categorise metadata, and the outcomes obtained from both metadata and image classifiers are consolidated to enhance diagnostic precision. The model's customised aggregation process surpasses conventional federated learning methods by accounting for the individual contributions of all clients to the master model. The method has an accuracy of 97.16% on the PAD-UFES-20 dataset. The model is constrained by its dependence on merely two data modalities, yet there exists an opportunity for enhancement by the integration of supplementary modalities, such as video or more intricate patient data.

Haripriya 2024[22] explore the application of Federated Learning in public health by investigating various clustering techniques, including K-means, DBSCAN, and hierarchical clustering. They compare the performance of these techniques in a federated learning environment to develop a security-enhanced public health data mining approach. This approach utilizes FedAvg with multiple clustering algorithms, allowing for adaptability to diverse data distributions including non-IID data to simulate real-world scenarios.

Ali 2023[17] presents a confidentiality-preserving federated learning paradigm for medical picture classification, emphasising the protection of patient privacy while maintaining high accuracy in medical diagnostics. The FL model utilises CNNs for image processing and secure multi-party computing (SMPC) methods to protect sensitive medical data during distributed model training. The findings indicate that the model attains comparable accuracy to centralised approaches while upholding stringent privacy restrictions. Notwithstanding its advantages, the framework's significant dependence on privacy techniques like SMPC escalates computational expenses, hence constraining its scalability in resource-limited healthcare settings. Furthermore, its capacity to process intricate or non-image-based healthcare information was not thoroughly investigated.

Tiwari 2023[23] examines the utilisation of federated reinforcement learning (FRL) inside consumer-oriented Internet of Medical Things (IoMT) systems for cyborg applications. The proposed system combines multi-agent federated reinforcement learning protocols (MFRLP) with socket-based calls for remotely performed procedures (RPC) to enhance cyborg-enabled healthcare services. The authors employ a synthesis of reinforcement learning agents and deep Q-learning networks (DQN) to address dynamic environmental changes. Training and verifying the models across fog nodes substantially decreases waiting times and computational overhead. The framework accomplished a 50% decrease in training and testing duration and a 40% decrease in local processing time relative to conventional machine learning models. The paper recognises limits regarding security, especially for independent nodes, and proposes that future research might include blockchain to address these concerns.

Tanim 2024[24] introduces a privacy-preserving federated learning architecture tailored for cyborg-based applications in healthcare, particularly focused on knee surgery procedures. The methodology integrates deep reinforcement learning (DRL) with federated learning to analyse medical data in fog computing settings. The system delegates duties through mobile agents to fog nodes through a hierarchical learning model, ensuring optimal data processing and minimising communication delays. A significant contribution is the implementation of dynamic scheduling to equilibrate processing loads between mobile and server agents. The model exhibits significant enhancements in processing speed and data protection. The authors acknowledge that data leaks and resource inefficiencies persist as issues, highlighting possible areas for further optimisation in future implementations.

Alzubi 2022[25] examines the implementation of federated learning in conjunction with secure multi-party computation (SMPC) for decentralised healthcare solutions. It underscores the minimisation of communication latency in mobile healthcare settings while safeguarding data privacy via SMPC. The research emphasises an adaptive scheduling technique designed to enhance work allocation across edge devices and central servers, hence boosting execution time and data security. The approach provides substantial benefits in minimising latency and improving data privacy; but, it encounters challenges in managing larger datasets and more intricate healthcare situations, which may affect its effectiveness in practical implementations.

Ngyuyen 2022[26] provides a comprehensive assessment of the application of FLin the healthcare industry, focusing on its potential to improve privacy and scalability in dispersed healthcare systems. FL is broken down into a number of different types, such as horizontal, vertical, and transfer learning, and the authors examine its application in a number of important domains, such as medical imaging, remote health monitoring, and health data management. The paper highlights multiple drawbacks, including the dependence on non-healthcare datasets for assessing models, neglected privacy risks, and interaction and computational expenses in resource-constrained environments. Despite the fact that facial recognition technology has the potential to protect individuals' privacy and make personalized healthcare more accessible, the paper emphasizes several limitations. In the publication, the importance of conducting additional research is emphasized, particularly with regard to the application of FL to healthcare-specific datasets and the creation of more robust strategies for enhancing privacy. In general, it highlights the promise of FL while also addressing the problems that need to be solved in order to fully comprehend the benefits that it offers in the field of smart healthcare.

While these studies lay a strong foundation for federated learning and transfer learning techniques, the limitations in addressing adaptive aggregation in multi-institutional medical datasets remain. The subsequent section outlines the experimental framework developed to address these gaps, detailing the methodologies, datasets, and evaluation metrics employed.

Experimental framework

The experimental framework outlines a comprehensive approach for privacy-preserving medical image classification using federated learning. This framework integrates advanced deep learning models, efficient data preprocessing techniques, and adaptive aggregation methodologies to address the challenges posed by heterogeneous and non-IID medical data distributions across multiple clients. The framework is designed to leverage the strengths of transfer learning architectures while ensuring data privacy and scalability.

The flowchart in Fig. 3 illustrates the workflow of the proposed federated adaptive aggregation model for privacy-preserving medical image classification. The process begins with the non-IID distribution of medical data across multiple clients, simulating real-world scenarios where data is inherently heterogeneous. Each client trains a local model using deep learning techniques such as VGG-16 or ResNet-RS, tailored to the assigned subset of the dataset.

Once the local training is complete, gradients are transmitted to the central server, initiating the global aggregation process. A divergence check is performed to evaluate the variability in data distributions across clients. Based on the divergence threshold the system dynamically selects the appropriate aggregation method. If the divergence is below the threshold, FedAvg is employed for computationally efficient aggregation. Conversely, if the divergence exceeds the threshold, FedSGD is utilized to preserve performance by incorporating finer gradient updates.

This adaptive framework ensures scalability, robustness, and efficient utilization of resources, addressing the challenges posed by non-IID data distributions in federated learning environments.

The subsequent subsections detail the key components of the framework, starting with data distribution and preprocessing, followed by the customization of transfer learning models, the implementation of federated learning techniques, and the evaluation metrics used to assess model performance. Each component plays a vital role in building a robust and efficient system for decentralized medical image classification.

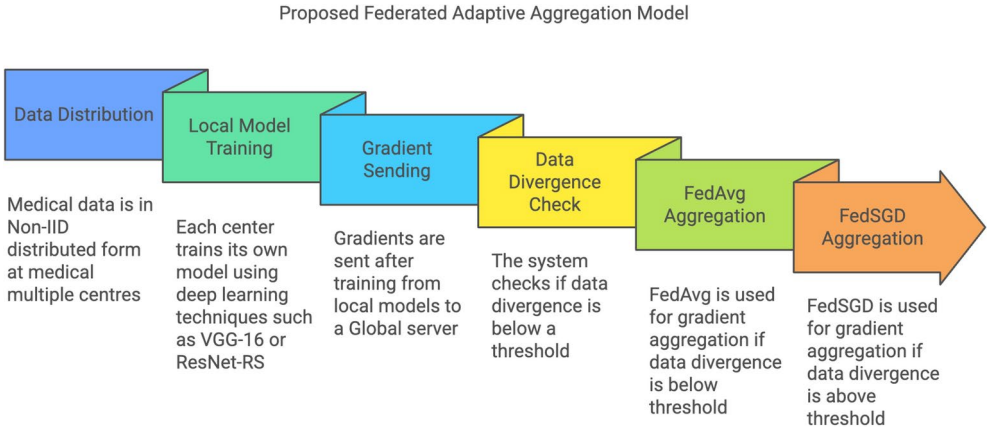

Fig. 3. Flowchart depicting the workflow of the proposed federated adaptive aggregation model.

Dataset	Number of images	Categories	Image format	Image size (pixels)
TB Chest X-ray Dataset	7000	Normal X-rays, TB X-rays	JPEG	224x224
Diabetic Retinopathy Dataset	3662	No_DR, mild, moderate, severe, Proliferate_DR	JPEG	224 × 224
Brain Tumor Dataset	3624	No tumor, meningioma, glioma, pituitary tumor	JPEG	224 × 224

Table 2. Details of the Medical Image Datasets used in the research.

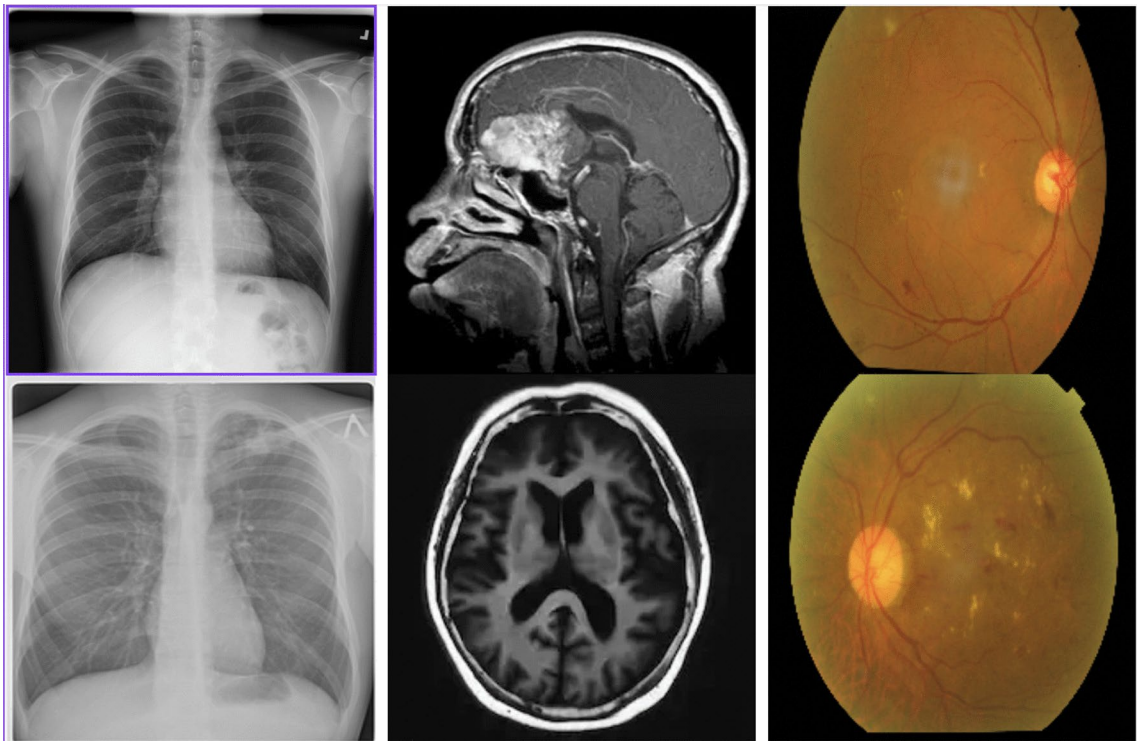

Fig. 4. Sample images of all the datasets utilized in this study.

Data and preprocessing

The datasets included in this study were obtained from publically accessible platforms like Kaggle, offering a varied collection of medical images crucial for assessing the efficacy of the chosen models. Table 2 shows the details of all the three datasets used for this research. The TB chest X-ray dataset comprises 7000 pictures and 2 classes, evenly divided between 3500 normal chest X-rays and 3500 X-rays suggestive of tuberculosis. This dataset facilitates the models' ability to differentiate between healthy and tuberculosis-affected cases, an essential function in tuberculosis detection. The sample images from the dataset can be seen in the Fig. 4

The diabetic retinopathy dataset comprises 3662 JPEG photos, classified into five classes as No DR, Mild, Moderate, Severe, and ProliferateDR, which denote different stages of diabetic retinopathy. This dataset facilitates the classification of various disease phases, enhancing early diagnosis and intervention options in diabetes patients.

The brain tumour dataset comprises 3624 JPEG photos categorised into four classes as no tumour, meningioma, glioma, and pituitary tumour. This dataset offers a variety of brain MRI images essential for tumour type classification, crucial for treatment planning and clinical decision-making.

Each dataset is randomly distributed across 10 clients in a non-IID manner, simulating real-world scenarios where different medical institutions have localized and diverse data. This distribution strategy ensures that clients receive subsets of the data with inherent variability, such as specific classes or regional differences. For example, one client might predominantly have "Normal" TB X-rays, while another might have a higher proportion of "TB" cases, reflecting realistic heterogeneity in data availability. Federated learning in real-world applications often suffers from performance degradation due to Non-IID data distribution across clients. Recent work, such as FedPIA[27], has proposed optimization strategies leveraging parameter importance to mitigate local-global divergence, highlighting the need for adaptive aggregation approaches in federated setups.

The datasets are divided into training, validation, and testing subsets with proportions of 70%, 15%, and 15%, respectively. This partitioning ensures that sufficient data is available for model training, hyperparameter tuning, and final performance evaluation. Each image is resized to 224 × 224 pixels and normalized to the range

Parameter	Definition	Context	Equation
$D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}$	Local datasets for TB chest X-rays, brain tumor MRIs, and diabetic retinopathy images assigned to client k	Represents the non-IID data distribution across clients	Eqs. (3), (4)
w_k^t	Local model parameters for client k at iteration t	Tracks client-specific training progress	Eq. (8)
w_t	Global model parameters after t iterations	Centralized aggregation of local model updates	Eq. (9)
δ_t	Divergence between local and global models at iteration t	Guides the choice between FedAvg and FedSGD	Eq. (11)
$\lvert D_k \rvert$	Number of samples in client k's local dataset	Used for weighted aggregation in FedAvg	Eq. (9)

Table 3. Summary of parameters and notations.

Client ID	Total images	Dominant disease category	Image quality
Client 1	1500	Predominantly TB-positive	High quality (no noise)
Client 2	1200	Predominantly normal (TB X-rays)	Medium quality
Client 3	500	Balanced distribution (TB)	Lower quality (noise added)
Client 4	800	High prevalence of glioma	High quality (MRI)
Client 5	700	Balanced brain tumor classes	Medium quality
Client 6	600	Predominantly meningioma (MRI)	Medium quality
Client 7	400	High prevalence of pituitary tumor (MRI)	Lower quality (resolution varied)
Client 8	1500	Predominantly no DR	High quality
Client 9	800	High prevalence severe DR	Medium quality
Client 10	1000	Balanced mild and proliferate DR	Lower quality (noise added)

Table 4. Stratified non-IID data distribution across clients.

[0, 1], providing a standardized input format for transfer learning models. These preprocessing steps ensure compatibility across all datasets and maintain consistency in model input.

To ensure clarity and consistency throughout the manuscript, the parameters and notations used in the proposed framework are summarized in Table 3. These parameters are integral to describing the federated learning methodology and its components, including local and global model updates, divergence measurements, and aggregation processes.

Stratified non-IID data distribution and preprocessing

Initially, datasets were randomly distributed across multiple clients to simulate decentralized medical environments. However, such an allocation may not adequately represent real-world scenarios where data across institutions is inherently heterogeneous in terms of disease prevalence, data volume, and imaging quality. Therefore, a refined stratified non-IID data allocation method was employed to closely mirror practical multi-institutional settings.

Specifically, the revised data allocation strategy explicitly accounts for realistic variability encountered in medical institutions by considering three key aspects:

- Disease prevalence: Each client received datasets with varying prevalence rates of diseases, reflecting actual differences among medical institutions.
- Data volume: The number of images allocated to each client varied, mimicking differences in institutional resources and patient intake.
- Image quality: To simulate realistic discrepancies in imaging capabilities and diagnostic quality, datasets varied slightly in image resolution and incorporated controlled noise where appropriate.

Table 4 summarizes the refined dataset distribution among the 10 clients.

All images were resized to a standard dimension of 224×224 pixels and normalized to a range of [0, 1] to maintain consistency in model inputs despite variability in data quality. This refined strategy ensures a realistic evaluation scenario, effectively testing the robustness and adaptability of the federated learning framework.

Benchmarking the proposed framework with baseline models

To evaluate the efficacy of the proposed adaptive aggregation methodology, comprehensive benchmarking was conducted using baseline models and modern architectures. This section presents a detailed analysis of the models' performance across multiple datasets and aggregation methods, highlighting the strengths and weaknesses of each approach. The results are analyzed to assess the impact of adaptive aggregation on accuracy, efficiency, and resource utilization.

The fine-tuning and validation of the pre-trained models GoogLeNet and VGG16 were essential elements of the transfer learning methodology employed for medical picture categorisation. The datasets utilised were partitioned into training, validation, and testing sets, with 70% designated for training, 15% for validation,

and the remaining 15% for testing. The equitable distribution was crucial to guarantee that the models were sufficiently trained while upholding a just assessment during the validation and testing stages.

Fine-tuning entailed modifying the weights of the pre-trained models, which were initially trained on ImageNet, to accommodate the specific medical picture datasets utilised. The foundational layers of the models, which encapsulate fundamental features like edges and textures, were predominantly preserved, whereas the superior layers were refined to identify domain-specific characteristics pertinent to medical imaging, including anomaly detection in TB chest X-rays, brain tumour segmentation, and classification of diabetic retinopathy stages. This technique facilitated the fast acquisition of medical-specific patterns without necessitating the training of models from the ground up, thus minimising computational time and resource requirements.

During training, the models were iteratively updated using backpropagation and gradient descent. The model parameters w were optimized by minimizing the categorical cross-entropy loss $\mathscr{L}$, which is defined for multi-class classification as:

$$\mathscr{L}(\mathrm{w}) = -\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log(\hat{y}_{i,c}) \tag{1}$$

where N is the number of training samples, C is the number of classes, $y_{i,c}$ is the true label of class c for sample i, and $\hat{y}_{i,c}$ is the predicted probability for class c.

It computes the categorical cross-entropy loss $\mathscr{L}(\mathrm{w})$, which measures the difference between the predicted probability distribution ($\hat{y}$) and the true labels (y) for all training samples. The output represents the overall error in the model's predictions, serving as a scalar value that the training process aims to minimize.

The gradients of the loss function with respect to the parameters were computed, and the model was updated using gradient descent:

$$\mathrm{w}^{(t+1)} = \mathrm{w}^{(t)} - \eta\nabla\mathscr{L}(\mathrm{w}^{(t)}) \tag{2}$$

where η is the learning rate, and $\nabla\mathscr{L}(\mathrm{w}^{(t)})$ represents the gradient of the loss function at iteration t.

It computes the updated model parameters $\mathrm{w}^{(t+1)}$ by adjusting the previous parameters $\mathrm{w}^{(t)}$ in the direction of the negative gradient. The output represents the refined model parameters after one iteration of training, moving the model closer to minimizing the loss.

To prevent overfitting, regularization techniques such as dropout were applied during training. In dropout, a fraction p of the neurons in each layer were randomly set to zero during each iteration, effectively reducing the complexity of the model. Additionally, early stopping was employed to halt training when the validation loss did not improve after a set number of epochs, thus preventing overfitting.

The validation set was used to monitor the performance of the models during training, ensuring that the fine-tuning process did not lead to overfitting or underfitting.

VGG 16 model

The VGG16 model was utilized for medical image classification due to its depth and robust feature extraction capabilities[29]. VGG16, with its 16 layers as shown in Fig. 5, is well-suited for learning detailed hierarchical features in medical images such as TB chest X-rays, brain tumor MRIs, and diabetic retinopathy images[30]. The convolutional layers in VGG16 use fixed-sized filters (3 × 3) to extract features from input images, which can be represented as:

$$\mathrm{y}_{i,j} = \sum_{m=1}^{M}\sum_{n=1}^{N} \mathrm{w}_{m,n}\mathrm{x}_{i+m,j+n} + b, \quad \mathrm{x} \in \{D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}\} \tag{3}$$

where x represents the input image from one of the medical datasets TB chest X-rays, brain tumor MRIs, or diabetic retinopathy images, w is the convolutional filter, b is the bias, and y is the output feature map. This equation computes the feature map y by applying the convolutional filter w over the input image x, extracting spatial features such as edges and textures essential for subsequent layers.

This systematic convolution over medical images allows VGG16 to capture both low-level and high-level features critical for medical diagnosis.

The suitability of VGG16 for this study lies in its ability to capture fine-grained details necessary for the accurate classification of medical conditions. The network's depth provides the advantage of learning complex patterns from medical images such as small lesions in diabetic retinopathy or subtle tumor regions in brain MRIs. Additionally, VGG16's architecture, when fine-tuned through transfer learning, was able to generalize well on the medical datasets after being trained on a large dataset like ImageNet, thus reducing the computational cost and training time.

GoogLeNet model

The VGG16 model was utilized for medical image classification due to its depth and robust feature extraction capabilities[29]. VGG16, with its 16 layers as shown in Fig. 5, is well-suited for learning detailed hierarchical features in medical images such as TB chest X-rays, brain tumor MRIs, and diabetic retinopathy images[30]. The convolutional layers in VGG16 use fixed-sized filters (3 × 3) to extract features from input images, which can be represented as:

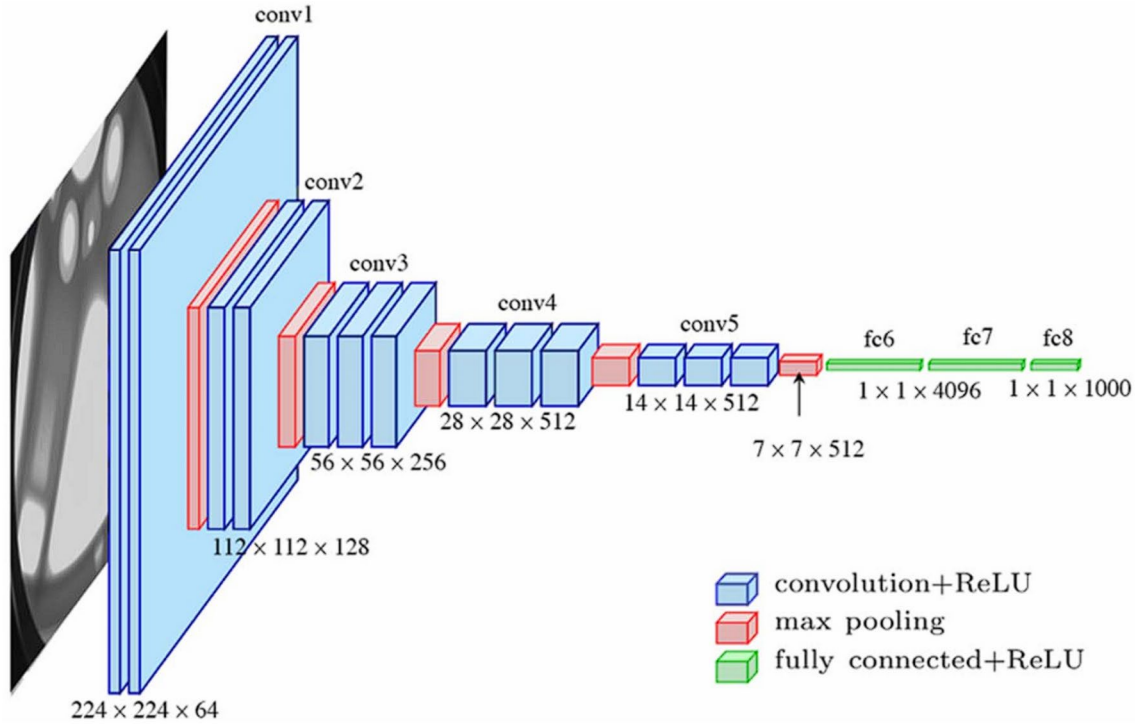

Fig. 5. VGG-16 architecture as introduced in[28].

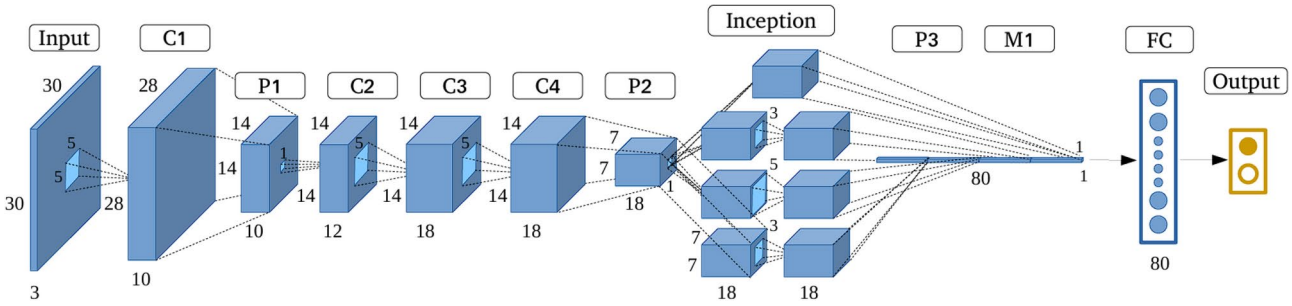

Fig. 6. GoogLenet architecture as introduced in[31].

$$y_{i,j} = \sum_{m=1}^{M}\sum_{n=1}^{N} w_{m,n} x_{i+m,j+n} + b, \quad x \in \{D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}\} \tag{4}$$

where x represents the input image from one of the medical datasets (TB chest X-rays, brain tumor MRIs, or diabetic retinopathy images), w is the convolutional filter, b is the bias, and y is the output feature map. This equation computes the feature map y by applying the convolutional filter w over the input image x, extracting spatial features such as edges and textures essential for subsequent layers.

This systematic convolution over medical images allows VGG16 to capture both low-level and high-level features critical for medical diagnosis.

The suitability of VGG16 for this study lies in its ability to capture fine-grained details necessary for the accurate classification of medical conditions. The network's depth provides the advantage of learning complex patterns from medical images such as small lesions in diabetic retinopathy or subtle tumor regions in brain MRIs. Additionally, VGG16's architecture, when fine-tuned through transfer learning, was able to generalize well on the medical datasets after being trained on a large dataset like ImageNet, thus reducing the computational cost and training time.

GoogLeNet was selected for medical image classification due to its unique architecture as shown in Fig. 6, which efficiently captures multi-scale features through the use of inception modules[32]. Each inception module processes the input data with multiple filters of different sizes such as 1×1, 3×3, 5×5 convolutions in parallel,

allowing GoogLeNet to extract features at various scales simultaneously[33]. The output of an inception module can be mathematically described as:

$$\mathrm{y}_{i,j}^{(k)} = \sum_{m=1}^{M_k} \sum_{n=1}^{N_k} \mathrm{w}_{m,n}^{(k)} \mathrm{x}_{i+m,j+n} + b^{(k)}, \quad \mathrm{x} \in \{D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}\} \tag{5}$$

where k denotes the different filter sizes used within the inception module, x represents the input image from one of the medical datasets (TB chest X-rays, brain tumor MRIs, or diabetic retinopathy images), $\mathrm{w}^{(k)}$ is the filter for each size, and $\mathrm{y}^{(k)}$ is the feature map output for that specific filter. This equation shows how the inception module combines outputs from filters of varying sizes to generate multi-scale feature maps, enabling the model to detect both fine-grained details and large-scale patterns in medical images.

This architecture is particularly suitable for the heterogeneous nature of medical images, such as TB chest X-rays and brain MRIs, which contain both large-scale and fine-grained features.

GoogLeNet's suitability lies in its ability to simultaneously process multiple feature scales, which is crucial for accurately detecting anomalies like small lesions in diabetic retinopathy or subtle changes in brain MRIs. By leveraging transfer learning, the model was fine-tuned to adapt to medical datasets, ensuring efficient training and robust performance.

GoogLeNet's ability to process features at different scales made it highly effective in identifying complex patterns, such as small microaneurysms in diabetic retinopathy images or distinct tumor boundaries in brain MRIs. The inception architecture allowed GoogLeNet to reduce the number of parameters compared to deeper models like VGG16, improving computational efficiency without sacrificing accuracy. The suitability of GoogLeNet for this research was further enhanced by its use of global average pooling layers, which replaced the fully connected layers, thus reducing overfitting on the relatively small medical datasets.

While the baseline models VGG16 and GoogLeNet demonstrate robust feature extraction and classification capabilities in centralized environments, the sensitivity of medical image data necessitates additional privacy-preserving measures. To address these concerns, differential privacy is introduced as an integral component of the experimental framework, ensuring that the models can be trained in a secure centralized environment without compromising the confidentiality of individual data points.

Differential privacy

Differential privacy[34] was employed in the centralized environment to protect sensitive medical data during the training of the models[35,36]. Differential privacy provides a mathematical guarantee that the inclusion or exclusion of any single data point will not significantly affect the outcome of the model[37], thereby ensuring that individual data points remain private. This was achieved by introducing noise into the gradients during backpropagation, which prevents the model from memorizing specific details about the training data[38]. The differentially private gradient update can be mathematically expressed as:

$$\nabla \mathscr{L}(\mathrm{w}) \leftarrow \nabla \mathscr{L}(\mathrm{w}; D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}) + \mathscr{N}(0, \sigma^2) \tag{6}$$

where $\nabla \mathscr{L}(\mathrm{w})$ represents the gradient of the loss function $\mathscr{L}(\mathrm{w})$ computed over the medical datasets (TB chest X-rays, brain tumor MRIs, and diabetic retinopathy images), and $\mathscr{N}(0, \sigma^2)$ is Gaussian noise with a mean of 0 and variance σ^2, introduced to preserve privacy. The parameter σ^2 controls the scale of the noise, with larger values offering stronger privacy guarantees but potentially impacting the model's accuracy. The privacy budget ϵ governs the level of privacy and defines the trade-off between privacy and model utility. A smaller ϵ provides stronger privacy but can reduce the accuracy of the model.

The differentially private model update at each step is given by:

$$\mathrm{w}_{t+1} = \mathrm{w}_t - \eta \left(\nabla \mathscr{L}(\mathrm{w}_t; D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}) + \mathscr{N}(0, \sigma^2) \right) \tag{7}$$

where w_{t+1} is the updated model parameter, w_t is the previous parameter, and η is the learning rate. By adding controlled noise to the gradient during each update, the model ensures that even if a malicious actor gains access to the model's parameters, they cannot infer details about any individual data point. This method was particularly important for ensuring privacy in the centralized training of medical images, where the sensitivity of personal health information requires strong protection mechanisms.

Differential privacy was used to balance the trade-off between accuracy and privacy in the centralized environment. While the noise addition may introduce some loss in accuracy, it ensures that the trained model maintains privacy guarantees that are critical when dealing with sensitive medical datasets like TB chest X-rays, brain tumor MRI scans, and diabetic retinopathy images.

Although differential privacy provides a robust mechanism for protecting data in centralized environments, the limitations of centralization in multi-institutional settings necessitate a federated approach. The federated learning setup leverages decentralized data stored across multiple medical centers, ensuring privacy preservation by eliminating the need to share raw data. This section outlines the structure of the federated learning environment, detailing the adaptive aggregation techniques employed to address challenges arising from non-IID data distributions.

Federated learning setup

A federated learning setup is employed to ensure privacy-preserving medical image classification, wherein the local data from different clients is trained without sharing the raw data with the central server. To improve

the efficiency and accuracy of the global model, an adaptive aggregation methodology is introduced, which dynamically selects between FedAvg[39,40] and FedSGD[41,42] based on the divergence of data across clients.

The medical image data is distributed across 10 clients in a non-IID manner, ensuring that each client has access to a random subset of the dataset. This non-IID[43,44] setup reflects real-world conditions where different medical institutions or devices may have data with varying distributions. The local model on each client k, denoted as w_k^t, is updated by performing local training using SGD, where the model parameters are updated as follows:

$$w_k^t = w_k^{t-1} - \eta \nabla L_k(w_k^{t-1}; D_k^{\mathrm{TB}}, D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}) \tag{8}$$

Here, w_k^t represents the local model parameters of client k at iteration t, and D_k^{TB}, D_k^{MRI}, D_k^{DR} are the subsets of the TB chest X-ray, brain tumor MRI, and diabetic retinopathy datasets, respectively. This equation computes the updated local model parameters after each iteration of training on client k, reflecting the gradients computed on the client's local data.

Once the local models have been trained on their respective datasets, the global model needs to be updated. The adaptive aggregation methodology is employed at this stage to optimize the communication and model update process. If the divergence across clients' model updates is high, indicating substantial variation in local data distributions, Federated SGD is utilized for aggregation. This is mathematically expressed as:

$$w_t = w_{t-1} - \eta \sum_{k=1}^{K} \frac{|D_k|}{N} \nabla L_k(w_{t-1}) \tag{9}$$

where w_t is the global model at round t, K is the number of clients, $|D_k|$ is the number of data points in client k's local dataset, and $N = \sum_{k=1}^{K} |D_k|$ is the total number of data points across all clients. This equation computes the updated global model parameters by aggregating gradients from all clients, ensuring the model reflects the diverse data distributions more accurately in cases of high divergence.

However, when the divergence between the client models is low or moderate, FedAvg is used as a more efficient method for aggregation. FedAvg averages the locally updated models and combines them into the global model, which can be expressed as:

$$w_t = \sum_{k=1}^{K} \frac{|D_k|}{N} w_k^t \tag{10}$$

Here, each client's contribution to the global model is weighted by the size of their local dataset. This equation shows how the global model is updated by averaging the local models, making FedAvg computationally less demanding compared to FedSGD and suitable for rounds with lower divergence.

The decision between FedSGD and FedAvg is made by dynamically evaluating the divergence δ_t across client models. The divergence is measured as the average distance between the local models and the global model, defined as:

$$\delta_t = \frac{1}{K} \sum_{k=1}^{K} \left\| w_k^t - w_t \right\|_2, \quad D_k \in \{\mathrm{TB}, \mathrm{MRI}, \mathrm{DR}\} \tag{11}$$

If δ_t exceeds a predefined threshold τ, indicating high divergence, FedSGD is used for aggregation to allow for finer updates based on gradients. Conversely, if $\delta_t \leq \tau$, FedAvg is used to aggregate the models more efficiently. This equation computes the divergence between client models and the global model, which is then used to decide the aggregation strategy, optimizing the communication process and ensuring the most appropriate method is employed for the current state of divergence across clients.

By dynamically switching between FedAvg and FedSGD, this methodology ensures that the global model is updated efficiently while accounting for the varying levels of divergence in the data across clients. This approach improves both the computational efficiency and the accuracy of the global model, especially in a non-IID data distribution setup, where divergence can significantly affect model convergence and performance. The algorithm for the proposed FL framework for medical multi-center image classification and secure data analysis can be given as:

Input: Medical image datasets $\mathscr{D}_1, \mathscr{D}_2, ..., \mathscr{D}_n$ from multiple platforms (non-IID)
Output: Trained global model $\mathscr{M}_g$
Step 1: Initialize the Global Model
Initialize the global model $\mathscr{M}_g$ with random parameters θ_0
for each communication round $t = 1, 2, ..., T$ **do**
  Step 2: Local Training at Each Client
  for each client $i = 1, 2, ..., n$ in parallel **do**
    Retrieve local dataset $\mathscr{D}_i$ (non-IID) from medical platform i
    Choose classification model $\mathscr{M}_i$ from {GoogLeNet, VGG-16, EfficientNetV2, ResNet-RS}
    Locally train the model $\mathscr{M}_i$ using current global parameters θ_t
    Update local parameters θ_i^t via local classification using $\mathscr{M}_i$
  end for
  Step 3: Share Local Parameters
  Each client sends its local parameters θ_i^t to the central server
  Step 4: Adaptive Aggregation at the Server
  Compute data divergence Δ_i for each client i
  if $\Delta_i <$ threshold **then**
    Apply **FedAvg** to aggregate local parameters:

$$\theta_{t+1} = \frac{1}{n} \sum_{i=1}^{n} \theta_i^t$$

  else
    Apply **FedSGD** to aggregate local gradients:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{n} \sum_{i=1}^{n} \nabla \mathscr{L}(\theta_i^t)$$

  end if
  Step 5: Update Global Model
  Update the global model $\mathscr{M}_g$ with new parameters θ_{t+1}
  Step 6: Distribute Updated Global Parameters
  Send updated global parameters θ_{t+1} to all clients
end for
Step 7: Final Model Deployment
Deploy the global model $\mathscr{M}_g$ to all medical platforms for final classification

Algorithm 1. Federated learning with adaptive aggregation in non-IID multiple medical centers setup.

Incorporating modern architectures into the proposed framework

To further enhance the performance and scalability of the proposed federated learning framework, modern deep learning architectures such as EfficientNetV2 and ResNet-RS are evaluated alongside the baseline models. These architectures are specifically designed to optimize resource utilization and handle complex datasets, making them well-suited for privacy-preserving medical image classification in federated environments. The following subsection presents the evaluation of these modern architectures, highlighting their contributions to the overall framework.

EfficientNetV2

EfficientNetV2 is a state-of-the-art convolutional neural network architecture designed for efficient scaling and optimal performance in image classification tasks[45]. It employs a compound scaling method to balance the depth (d), width (w), and resolution (r) of the network, ensuring a trade-off between accuracy and computational cost[46]. The detailed architecture of EfficientNetV2 is shown in Fig. 7. The scaling is mathematically defined as:

$$r = \alpha^k, \quad w = \beta^k, \quad d = \gamma^k, \quad \text{subject to } \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \tag{12}$$

where α, β, and γ are scaling coefficients, and k is the scaling factor. This equation computes the scaling factors for resolution (r), width (w), and depth (d) of the network, ensuring uniform scaling across dimensions to effectively handle high-resolution medical images without excessive computational overhead[47].

EfficientNetV2 utilizes inverted residual blocks, which are mathematically represented as:

$$y = \sigma(BN(W \cdot x)), \quad x \in \{D_k^{\text{TB}}, D_k^{\text{DR}}\} \tag{13}$$

where x is the input tensor from datasets such as TB chest X-rays or diabetic retinopathy images, W represents the convolutional weights, BN denotes batch normalization, and σ is the activation function. This equation computes the output tensor y by applying convolutional operations followed by batch normalization and activation, enabling the network to extract hierarchical features critical for medical image classification.

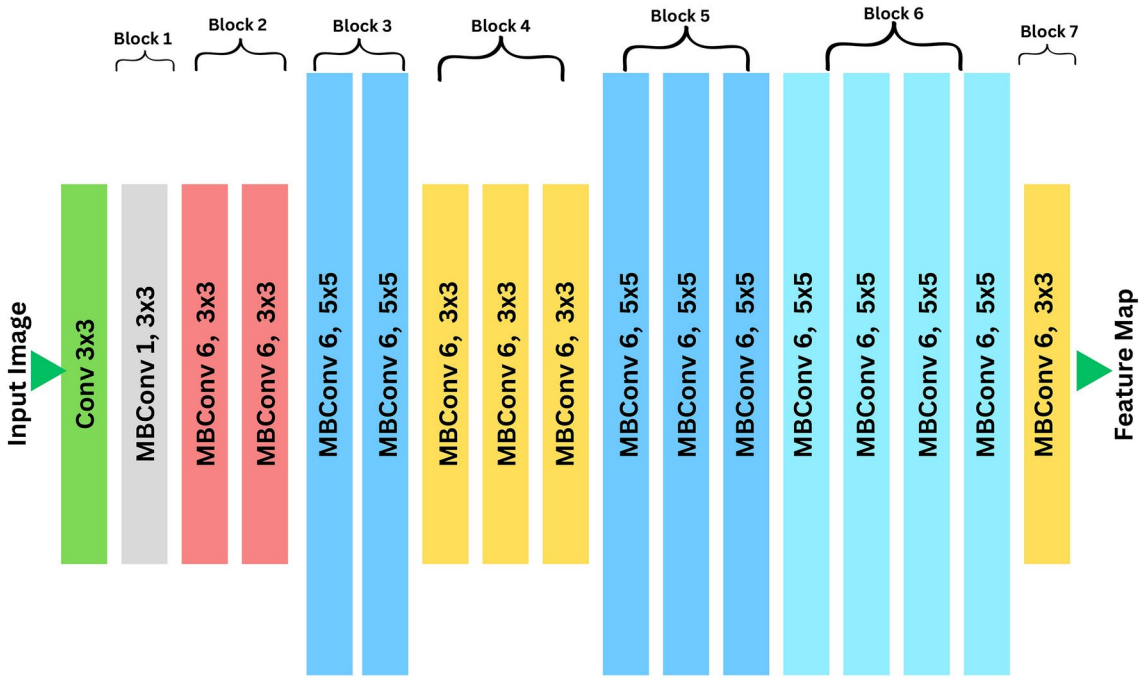

Fig. 7. EfficientNetV2 architecture.

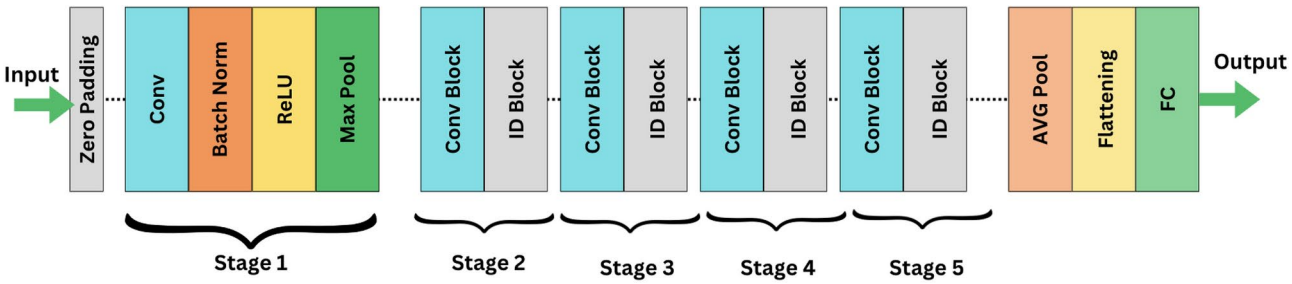

Fig. 8. ResNet-RS architecture.

These architectural improvements make EfficientNetV2 particularly suitable for federated learning environments, where communication efficiency and scalability are crucial. In this experiment, EfficientNetV2 was employed to handle datasets such as TB chest X-rays and diabetic retinopathy images due to its ability to extract fine-grained features, adapt to non-IID data distributions, and converge faster in decentralized setups. These characteristics align with the need to minimize communication costs while achieving high classification accuracy across multiple clients.

ResNet-RS

ResNet-RS, an advanced adaptation of the ResNet architecture, was also utilized for medical image classification. This model introduces improved scaling strategies and training optimizations, making it highly efficient for handling complex datasets in contemporary hardware environments[48]. The detailed architecture of ResNet-RS is shown in Fig. 8. ResNet-RS retains the use of residual blocks to mitigate vanishing gradient issues, represented as:

$$y = \sigma(F(x, \{W_i\}) + x), \quad x \in \{D_k^{\mathrm{MRI}}, D_k^{\mathrm{DR}}\} \tag{14}$$

where $F(x, \{W_i\})$ is the residual mapping, x represents the input from datasets such as brain tumor MRI and diabetic retinopathy images, W_i are the convolutional weights, and σ is the activation function. This equation computes the output tensor y by adding the residual mapping $F(x, \{W_i\})$ to the input tensor x, enabling the network to retain previously learned features while learning new hierarchical features crucial for accurate classification.

These skip connections enable deeper networks to converge more effectively, making ResNet-RS particularly robust in extracting hierarchical features from multi-class datasets[49]. In this study, ResNet-RS was applied

to datasets such as brain tumor MRI and diabetic retinopathy images, where its robustness in distinguishing overlapping features and rare classes significantly improved classification performance[50].

Both EfficientNetV2 and ResNet-RS were integrated into the federated learning framework to leverage their complementary strengths. In this setup, models were trained locally on client devices, with global aggregation performed using adaptive methods such as FedAvg and FedSGD. The global model weights θ_G were updated iteratively using the equation:

$$\theta_G^{(t+1)} = \sum_{i=1}^{N} \frac{n_i}{n} \cdot \theta_i^{(t)}, \quad n = \sum_{i=1}^{N} n_i \tag{15}$$

where N represents the total number of clients, n_i is the number of samples on the i-th client, and $\theta_i^{(t)}$ denotes the local model weights after t iterations. This equation computes the updated global model weights $\theta_G^{(t+1)}$ by aggregating the weighted contributions of local model parameters from all clients, ensuring that the global model reflects the diversity of client data.

EfficientNetV2's ability to efficiently process high-resolution images and ResNet-RS's capacity to extract complex features ensured that the federated framework performed robustly across diverse datasets while preserving data privacy. These models not only have the ability to reduce local training errors but also improve the overall convergence rate, highlighting their suitability for privacy-preserving collaborative learning in multi-institutional settings.

Results

Baseline model analysis with GoogLeNet and VGG16

The performance of two prominent convolutional neural network architectures, GoogLeNet and VGG1, was evaluated across three distinct datasets which are TB chest X-ray images, brain tumor MRI images, and diabetic retinopathy images for prepaing a baseline model analysis. The experiments were conducted in both centralized and federated environments, with a focus on privacy-enhanced medical image classification.

The centralized experiments utilized differential privacy, ensuring that sensitive patient information was protected during model training. The results from the centralized setup demonstrated strong performance across all datasets, with accuracy values ranging from 96.0% to 98.3%. For example, in the TB chest X-ray dataset, GoogLeNet achieved an accuracy of 97.9% for normal images and 96.4% for TB-positive images, while VGG16 exhibited similar performance and their detailed comparison can be seen in Fig. 9. Although the introduction of differential privacy slightly impacted precision and recall. Figure 11 compares the performance metrics as precision, recall, F1-score across aggregation methods for GoogLeNet and VGG-16 using three datasets. The adaptive aggregation method consistently shows significant improvements over FedAvg and FedSGD, achieving near-centralized performance while balancing computational efficiency. Table 5, 6 and 7 shows the result stats for the models with different datasets and aggregation methods. To further evaluate the performance of the models, the Root Mean Square Error (RMSE) was calculated for both GoogLeNet and VGG-16 across the three datasets under four different aggregation methods. The RMSE values, as shown in Fig. 10, provide a measure of the average prediction error for each model in different scenarios. These values help in quantifying the models' precision and offer additional insight beyond accuracy alone, highlighting the variability in model performance across the different methods. Figure 12 shows confusion matrices for GoogLeNet and VGG-16 across three datasets TB X-rays, Brain Tumor MRI, and Diabetic Retinopathy illustrating classification performance for "Normal" and "Disease" classes. The diagonal cells represent correct classifications, while off-diagonal cells

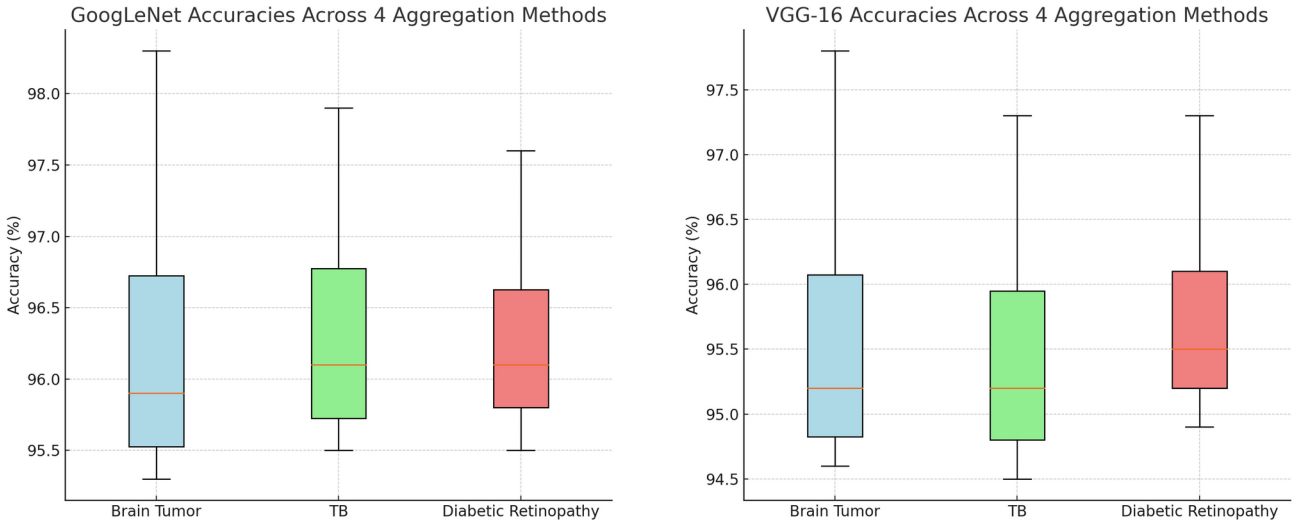

Fig. 9. Accuracy comparison for the models in all 4 aggregation methods.

Model	Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Centralized with differential privacy					
GoogLeNet	Normal	97.9	98.2	97.6	97.9
	TB	96.4	96.7	98.3	97.5
VGG16	Normal	97.3	97.8	97.2	97.5
	TB	96.1	95.9	97.4	96.6
Federated results: FedAvg					
GoogLeNet	Normal	95.8	95.9	95.4	95.6
	TB	94.5	94.4	96.1	95.2
VGG16	Normal	94.9	95.1	94.5	94.8
	TB	93.7	93.4	95.0	94.2
Federated results: FedSGD					
GoogLeNet	Normal	95.5	95.2	94.9	95.0
	TB	94.1	93.8	95.7	94.7
VGG16	Normal	94.5	94.2	93.8	94.0
	TB	93.4	93.1	94.6	93.8
Federated results: adaptive aggregation (FedAvg + FedSGD)					
GoogLeNet	Normal	96.4	96.7	96.0	96.3
	TB	95.3	95.1	96.5	95.8
VGG16	Normal	95.5	95.4	94.9	95.1
	TB	94.9	94.5	95.9	95.2
EfficientNetV2	Normal	96.8	97.0	96.5	96.7
	TB	96.2	96.1	96.8	96.4
ResNet-RS	Normal	96.5	96.8	96.3	96.6
	TB	95.9	95.7	96.6	96.2

Table 5. Model performance on TB chest X-ray dataset across centralized and federated setups with different aggregation methods.

indicate misclassifications. Both models demonstrate strong performance across all datasets, with GoogLeNet generally achieving higher accuracy and fewer errors. The TB X-ray dataset shows the lowest misclassification rates, while the Brain Tumor MRI dataset exhibits slightly higher errors due to its complexity. For the Diabetic Retinopathy dataset, both models maintain high accuracy, showcasing their adaptability in medical image classification tasks.

In the federated setting, three aggregation methods were explored: FedAvg, FedSGD, and a novel adaptive aggregation method that combines FedAvg and FedSGD. The federated experiments were conducted to assess the model's performance when data is distributed across multiple clients, simulating real-world decentralized healthcare systems.

The results showed that each aggregation method influenced the model's classification performance differently. FedAvg provided stable and consistent results, with accuracy values ranging from 94.9% to 95.9% across datasets. For instance, in the brain tumor MRI dataset, GoogLeNet achieved an accuracy of 95.6% for no tumor classification and 94.7% for glioma using FedAvg. Similarly, VGG16 achieved competitive results with slightly lower variance.

On the other hand, FedSGD exhibited more fluctuation in performance, as expected due to its stochastic nature. While it provided faster convergence in certain cases, its accuracy and F1-scores were marginally lower than FedAvg, particularly in datasets with more diverse or imbalanced data distributions. For example, in the diabetic retinopathy dataset, GoogLeNet achieved an accuracy of 95.5% for No_DR classification using FedSGD, slightly lower than the 95.9% obtained with FedAvg.

The adaptive aggregation method, which dynamically switches between FedAvg and FedSGD based on data divergence, demonstrated superior performance compared to the individual aggregation methods. By leveraging the strengths of both FedAvg's stability and FedSGD's speed, the adaptive aggregation method consistently outperformed both, especially in cases where the data was highly non-IID. In the diabetic retinopathy dataset, for example, GoogLeNet achieved an accuracy of 96.3% for No_DR classification and 95.9% for Proliferate_DR classification using adaptive aggregation, exceeding the performance of both FedAvg and FedSGD.

This consistent improvement suggests that the adaptive aggregation method provides a more resource-optimized and accurate approach for federated medical image classification, effectively balancing privacy and performance. The ability of the adaptive method to dynamically adjust to varying levels of data divergence across clients makes it particularly suitable for real-world federated learning scenarios in healthcare, where data heterogeneity is common.

The performance of both GoogLeNet and VGG-16 models was evaluated in the federated learning environment across different batch sizes 32, 64, 128 and learning rates 0.01, 0.05, 0.1 for each dataset, as illustrated in Fig. 13. The results indicate that smaller batch sizes, such as 32, led to faster convergence, but with slight

Model	Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Centralized with differential privacy					
GoogLeNet	No tumor	98.3	98.7	98.1	98.4
	Glioma	96.7	97.1	97.3	97.2
	Meningioma	96.4	96.3	97.4	96.8
	Pituitary	97.2	96.9	97.8	97.3
VGG16	No tumor	97.8	98.1	97.5	97.8
	Glioma	96.5	96.8	97.0	96.9
	Meningioma	96.2	96.4	97.1	96.7
	Pituitary	96.9	96.7	97.5	97.1
Federated results: FedAvg					
GoogLeNet	No tumor	95.6	95.8	95.3	95.5
	Glioma	94.7	94.3	95.4	94.8
	Meningioma	94.1	94.2	95.1	94.6
	Pituitary	94.9	94.1	95.2	94.6
VGG16	No tumor	94.9	95.2	94.7	94.9
	Glioma	94.1	94.0	95.1	94.5
	Meningioma	93.8	94.1	94.9	94.5
	Pituitary	94.5	94.3	95.4	94.8
Federated results: FedSGD					
GoogLeNet	No tumor	95.3	95.1	94.8	94.9
	Glioma	94.3	94.0	94.7	94.4
	Meningioma	93.9	93.7	94.6	94.1
	Pituitary	94.4	93.8	95.0	94.4
VGG16	No Tumor	94.6	94.7	94.2	94.4
	Glioma	93.7	93.4	94.8	94.1
	Meningioma	93.5	93.9	94.3	94.1
	Pituitary	93.9	93.6	94.8	94.2
Federated results: adaptive aggregation (FedAvg + FedSGD)					
GoogLeNet	No tumor	96.2	96.3	96.0	96.1
	Glioma	95.4	95.2	95.9	95.5
	Meningioma	95.1	94.9	95.7	95.3
	Pituitary	95.7	95.2	96.1	95.6
VGG16	No tumor	95.5	95.6	95.0	95.3
	Glioma	94.9	94.7	95.4	95.0
	Meningioma	94.8	94.6	95.3	95.0
	Pituitary	95.3	95.0	95.9	95.4
EfficientNetV2	No tumor	98.6	98.8	98.5	98.7
	Glioma	97.4	97.2	97.9	97.5
	Meningioma	97.1	97.4	97.9	97.6
	Pituitary	97.9	97.7	98.1	98.6
ResNet-RS	No tumor	98.5	98.6	98.1	98.3
	Glioma	97.2	97.4	97.4	97.6
	Meningioma	97.1	97.6	97.3	97.5
	Pituitary	97.7	97.6	98.2	97.9

Table 6. Model performance on brain tumor MRI dataset across centralized and federated setups with different aggregation methods.

fluctuations in the training loss. In contrast, larger batch sizes, such as 128, resulted in smoother loss curves with more stable convergence, albeit at a slower pace. Additionally, the learning rate had a notable impact on model convergence. Lower learning rates 0.01 produced more gradual but stable convergence, whereas higher learning rates 0.1 accelerated the early stages of convergence but introduced more variability, particularly in VGG-16. The GoogLeNet model consistently outperformed VGG-16 in terms of convergence speed, achieving lower training loss across all batch sizes and learning rates, which is consistent with its smaller architecture and reduced computational demands. Overall, the experiments demonstrate the importance of tuning batch size and learning rate to optimize model performance in federated learning settings, where computational resources and data variability are key considerations.

Model	Class	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Centralized with differential privacy					
GoogLeNet	No DR	97.6	98.0	97.4	97.7
	Mild	96.9	97.1	97.3	97.2
	Severe	96.3	96.4	97.1	96.7
	Proliferate DR	96.8	96.9	97.5	97.2
VGG16	No DR	97.3	97.7	97.1	97.4
	Mild	96.5	96.8	96.9	96.8
	Severe	96.0	96.1	96.7	96.4
	Proliferate DR	96.4	96.6	97.3	96.9
Federated results: FedAvg					
GoogLeNet	No DR	95.9	96.0	95.4	95.7
	Mild	95.2	95.4	95.9	95.6
	Severe	94.6	94.3	95.5	94.9
	Proliferate DR	95.1	94.9	95.8	95.3
VGG16	No DR	95.3	95.6	94.9	95.2
	Mild	94.8	94.7	95.4	95.0
	Severe	94.2	94.1	95.0	94.5
	Proliferate DR	94.7	94.4	95.2	94.8
Federated results: FedSGD					
GoogLeNet	No DR	95.5	95.3	95.1	95.2
	Mild	95.0	94.8	95.4	95.1
	Severe	94.3	94.2	95.2	94.7
	Proliferate DR	94.8	94.5	95.6	95.0
VGG16	No DR	94.9	95.1	94.5	94.8
	Mild	94.4	94.3	95.0	94.6
	Severe	93.8	93.9	94.7	94.3
	Proliferate DR	94.3	94.1	95.3	94.7
Federated results: adaptive aggregation (FedAvg + FedSGD)					
GoogLeNet	No DR	96.3	96.5	96.1	96.3
	Mild	95.8	96.0	96.3	96.1
	Severe	95.4	95.2	96.0	95.6
	Proliferate DR	95.9	95.6	96.5	96.0
VGG16	No DR	95.7	95.8	95.2	95.5
	Mild	95.2	95.0	96.1	95.5
	Severe	94.9	94.7	95.6	95.1
	Proliferate DR	95.3	95.1	96.3	95.7
EfficientNetV2	No DR	97.0	97.2	96.8	97.0
	Mild	96.5	96.7	97.0	96.8
	Severe	96.1	96.0	96.8	96.4
	Proliferate DR	96.6	96.4	97.3	96.8
ResNet-RS	No DR	96.8	97.0	96.6	96.8
	Mild	96.3	96.5	96.8	96.6
	Severe	95.9	95.8	96.7	96.3
	Proliferate DR	96.4	96.2	97.2	96.7

Table 7. Model performance on diabetic retinopathy dataset across centralized and federated setups with different aggregation methods.

The results from these experiments propose a novel privacy-enhanced federated model for medical image classification that is highly generalizable across different medical image datasets. The model demonstrates robustness and adaptability through the integration of an adaptive aggregation method, which consistently outperforms traditional aggregation techniques in both accuracy and privacy protection. The variety of datasets utilized in this study spanning pulmonary, neurological, and ophthalmological imaging further emphasizes the model's generalisability and potential utility in the broader healthcare sector.

Performance analysis with EfficientNetV2 and ResNet-RS

The evaluation of GoogLeNet and VGG16 as baseline models served as an essential step in establishing the robustness and reliability of the proposed dynamic aggregation framework. Both models demonstrated strong

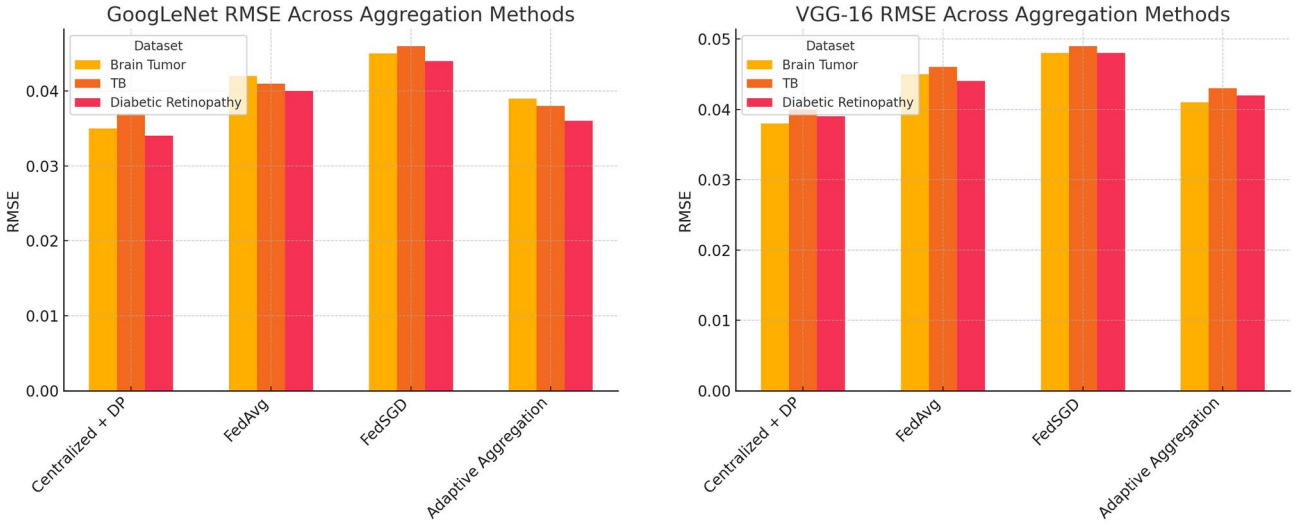

Fig. 10. Training error comparison between GoogLenet and VGG-16.

foundational performance across the datasets. To further enhance the scalability and reduce computational overhead, the modern architectures EfficientNetV2 and ResNet-RS were introduced. EfficientNetV2, with its compound scaling strategy, achieved superior performance across all datasets which can be observed from Tables 5, 6 and 7 including an accuracy of 98.3% on TB chest X-rays and 98.6% on brain tumor MRI. Its design, which balances depth, width, and resolution, enables efficient feature extraction and faster convergence, making it particularly well-suited for high-resolution medical images. This is evident from the confusion matrices, where EfficientNetV2 exhibited fewer misclassifications in challenging cases, such as between "Normal" and "TB" or among overlapping tumor classes. ResNet-RS also demonstrated competitive performance, achieving 98.2% accuracy on TB chest X-rays and 98.4% on brain tumor MRI. Its hierarchical feature extraction and robust skip connections contributed to effective classification, although slightly higher misclassification rates were observed compared to EfficientNetV2 which can be observed from Fig. 14 which highlights the classification performance quantitatively for both the modern architectures across all three datasets.

On the diabetic retinopathy dataset, EfficientNetV2 and ResNet-RS further highlighted their advantages in handling imbalanced datasets. EfficientNetV2 achieved an accuracy of 98.1% and an F1-score of 97.9%, excelling in the classification of rare classes such as "Proliferate DR." ResNet-RS followed closely with an accuracy of 98.0% and an F1-score of 97.8%. Compared to the baseline models, which achieved accuracies of 96.9% for GoogLeNet and 96.0% for VGG16, the modern architectures demonstrated their ability to handle more complex scenarios with reduced communication overhead. The precision-recall curves further support these findings, showing consistently higher precision at all levels of recall for the modern architectures, particularly for EfficientNetV2. Figure 15 shows precision recall trade-off in detail for the utilised modern architectures with the proposed dynamic aggregation framework across all three datasets.

The inclusion of EfficientNetV2 and ResNet-RS in the evaluation provided critical insights into the scalability of the proposed dynamic aggregation model. EfficientNetV2, with its efficient scaling and optimized architecture, significantly reduced local training errors and gave reliable results, ensuring precise global convergence. ResNet-RS complemented this by providing robust feature extraction capabilities, particularly in datasets with high inter-class variability. Together, these modern architectures demonstrate that the proposed framework is not only robust and reliable, as established by the baseline models, but also more scalable, making it suitable for large-scale privacy-preserving applications in healthcare.

Client contribution reflects the proportional influence of each client's local dataset on the global model during federated learning. In this framework, each client trains the model locally on its dataset and sends the updated model parameters to the central server, where global aggregation occurs. Figure 16 shows that the aggregation process weights the contributions of each client based on the size of their local dataset, ensuring that clients with larger datasets exert a greater influence on the global model. This dynamic is crucial for federated learning scenarios where data is non-IID and unevenly distributed, as is often the case in medical applications.

In the context of this experiment, client contributions were analyzed for both baseline models GoogLeNet and VGG16 and modern architectures EfficientNetV2 and ResNet-RS. For the baseline models, contributions were relatively uniform across clients, reflecting the simpler computational requirements of these architectures. In contrast, the modern architectures introduced slight variations in client contributions due to their higher computational demands, with clients possessing larger datasets and greater computational resources contributing more significantly. This analysis underscores the scalability of the proposed framework, demonstrating its ability to effectively accommodate diverse client contributions while preserving the privacy and integrity of local datasets.

Federated learning emerges as a promising approach for medical image classification, effectively balancing data privacy with high performance across diverse datasets. By incorporating adaptive aggregation techniques,

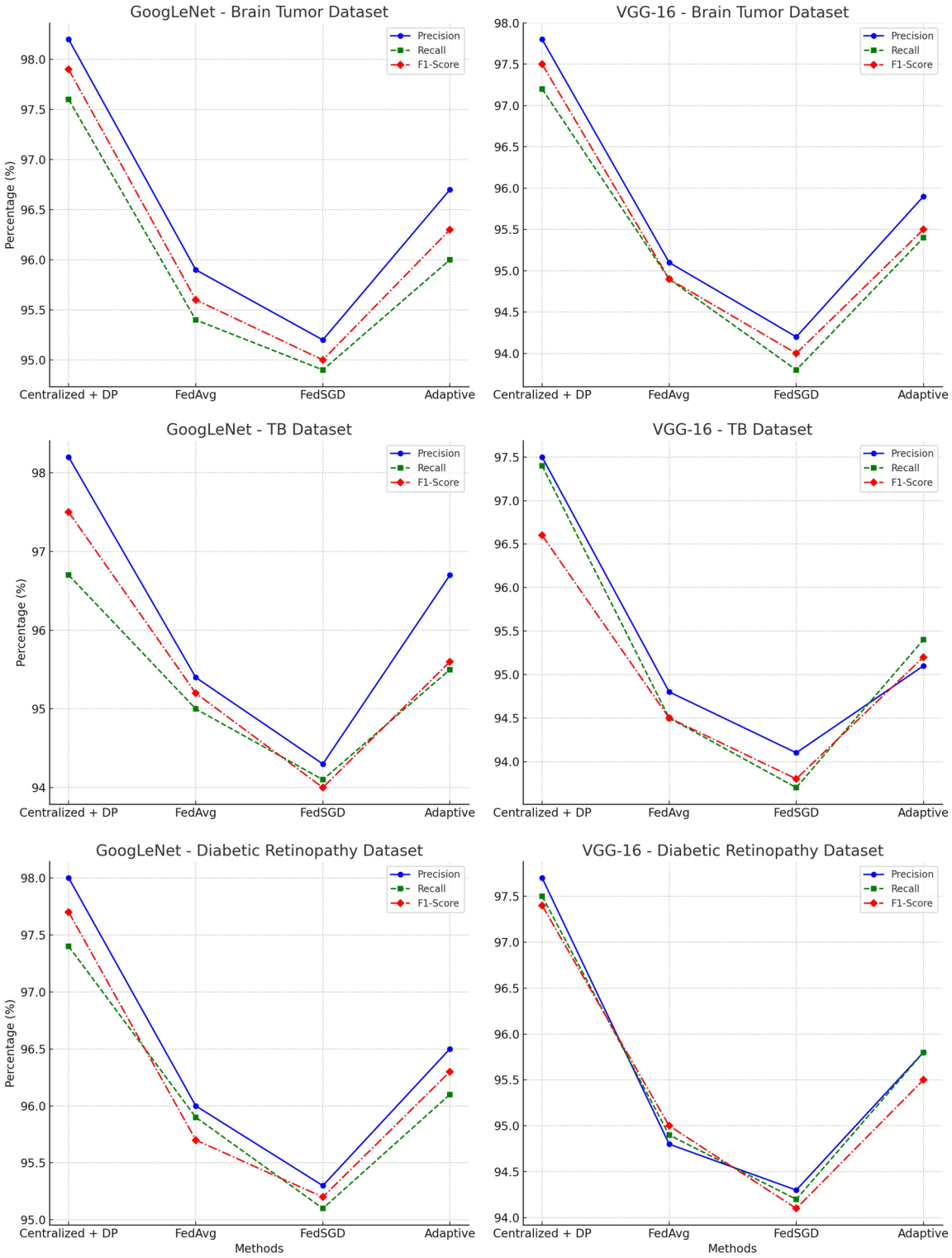

Fig. 11. Performance of VGG-16 and GoogLenet in all 4 aggregation methods in terms of precision and recall.

this framework can dynamically adjust to the variability of medical data from different healthcare providers, enhancing model optimization and making it suitable for real-world healthcare deployments.

Threshold parameter analysis for adaptive aggregation

To evaluate the optimal threshold parameter (δ_t) for the proposed dynamic aggregation framework, experiments were conducted by varying the threshold across multiple values, including $\delta_t = \{0.01, 0.05, 0.1, 0.2, 0.5\}$. The threshold governs the switching mechanism between FedAvg and FedSGD, with FedAvg being computationally

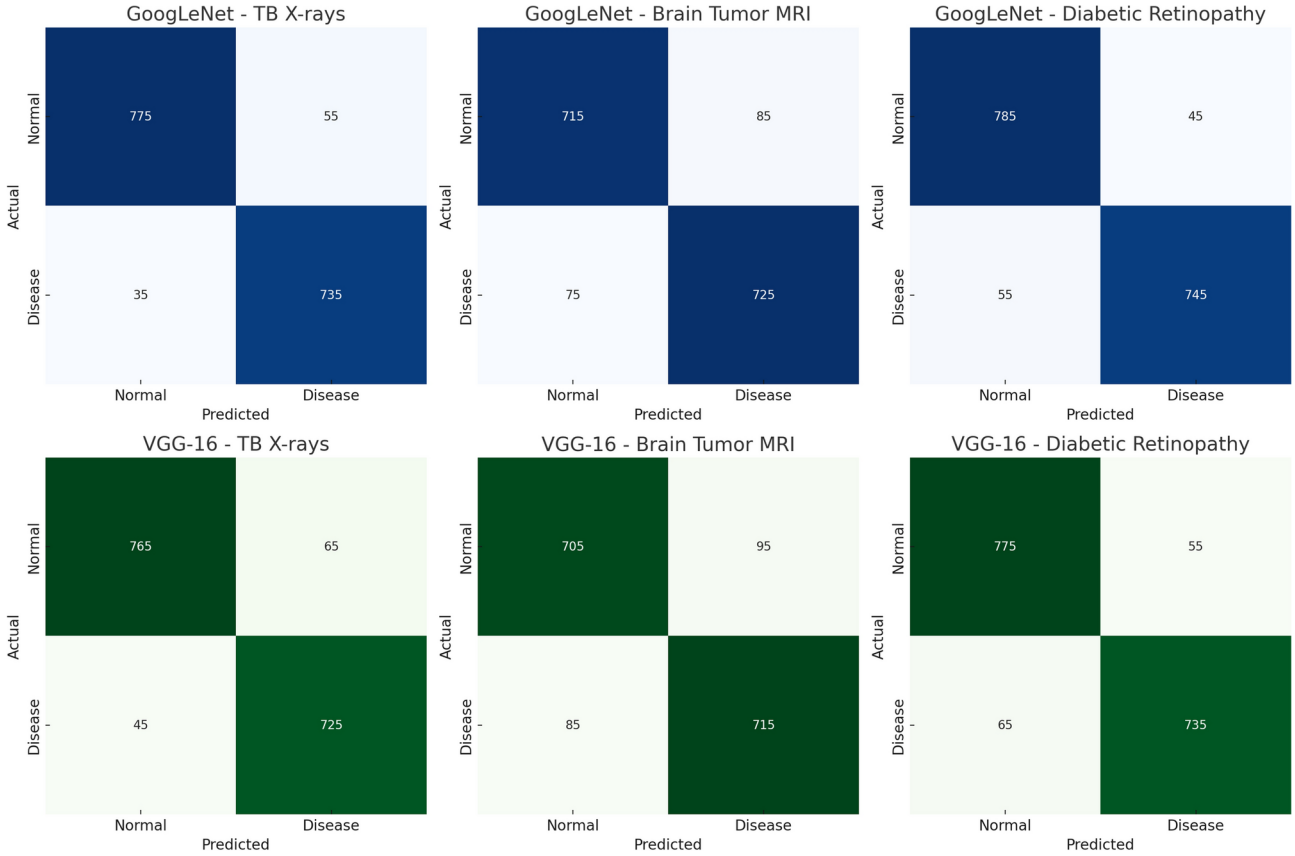

Fig. 12. Classification performance of the baseline models with each Medical Dataset.

efficient but less effective in divergence-heavy scenarios, while FedSGD is robust under high divergence but incurs higher resource costs. The threshold parameter was mathematically defined as:

$$\delta_t = \frac{\mathrm{Var}(w)}{\mathrm{Mean}(w)}$$

where $\mathrm{Var}(w)$ represents the variance of the model weights across clients after local updates, and $\mathrm{Mean}(w)$ is their mean value. When $\delta_t > T$, FedSGD is selected to improve convergence under high divergence conditions; otherwise, FedAvg is employed to optimize resource efficiency.

Smaller thresholds ($\delta_t = 0.01, 0.05$) frequently triggered FedSGD, leading to improved convergence and divergence correction but at the expense of significantly increased communication and computational overhead. Conversely, larger thresholds such as $\delta_t = 0.5$ predominantly utilized FedAvg due to the higher tolerance for divergence, resulting in lower resource costs but suboptimal performance in non-IID scenarios where divergence correction was insufficient.

The intermediate threshold value of $\delta_t = 0.2$ provided the best trade-off between performance and resource utilization. This threshold allowed the framework to adapt effectively to divergence across clients, maintaining robust convergence while optimizing communication overhead. The performance of all four transfer learning approaches GoogLeNet, VGG-16, EfficientNetV2, and ResNet-RS was evaluated under this setting, demonstrating consistently high accuracy and efficiency. Based on these findings, $\delta_t = 0.2$ was selected as the generalized threshold for this experiment, ensuring scalability and adaptability of the proposed methodology across diverse architectures. The accuracy analysis shown in Fig. 17 and Table 8 illustrates the performance trends across varying threshold values (δ_t) for all classification models. It demonstrates that smaller thresholds ($\delta_t = 0.01, 0.05$) improve accuracy by frequently addressing divergence, while larger thresholds ($\delta_t = 0.5$) reduce performance due to under-correction. The intermediate threshold ($\delta_t = 0.2$) achieves the optimal balance, resulting in consistently high accuracy across all models.

Training dynamics and switching frequencies

The adaptive aggregation framework dynamically switches between FedAvg and FedSGD based on the divergence threshold ($T = 0.2$). The switching frequency varies across datasets due to differences in client data divergence, as illustrated in Fig. 17. For the TB X-rays dataset, FedSGD was triggered less frequently, approximately once every 8 rounds, indicating relatively low divergence among clients. In the Brain Tumor MRI dataset, FedSGD was triggered more often, approximately once every 6 rounds, due to moderate divergence. For the Diabetic

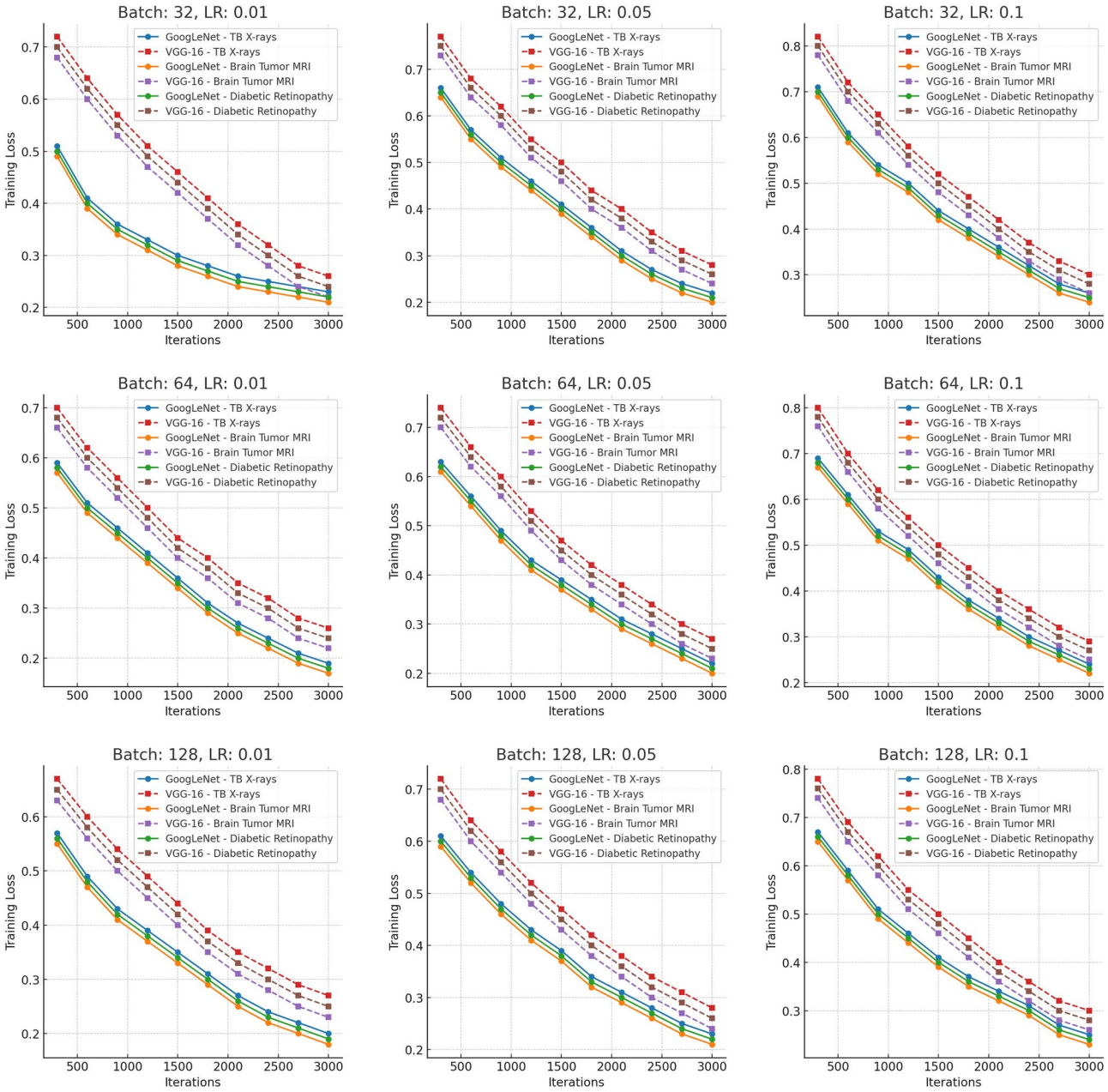

Fig. 13. Training loss with different batch sizes and learning rates for both the classification models in federated environment.

Retinopathy dataset, characterized by higher divergence, FedSGD was triggered approximately once every 3 rounds.

The cumulative frequency plot in Fig. 18 highlights this dynamic behavior. FedAvg dominates the aggregation process in low-divergence scenarios, optimizing resource utilization and communication efficiency. FedSGD becomes critical in high-divergence rounds to ensure robust model updates and correct client drift, thereby stabilizing accuracy trends. Across all datasets, the adaptive framework achieved higher accuracy compared to static methods, with improvements of approximately 1.5% for TB X-rays, 2.3% for Brain Tumor MRI, and 3.0% for Diabetic Retinopathy over FedAvg. These improvements were particularly evident in high-divergence rounds, where FedSGD corrected alignment issues between clients, leading to better convergence. This quantitative analysis validates that the adaptive aggregation approach effectively balances computational efficiency and performance in federated learning scenarios. The frequency of FedAvg and FedSGD usage was calculated for each dataset based on the divergence threshold. For the TB X-rays dataset, FedAvg was used in approximately 88% of rounds and FedSGD in 12%. For the Brain Tumor MRI dataset, FedAvg was used in 82% of rounds and FedSGD in 18%. For the Diabetic Retinopathy dataset, FedAvg was chosen in 68% of rounds, with FedSGD triggered in 32%.

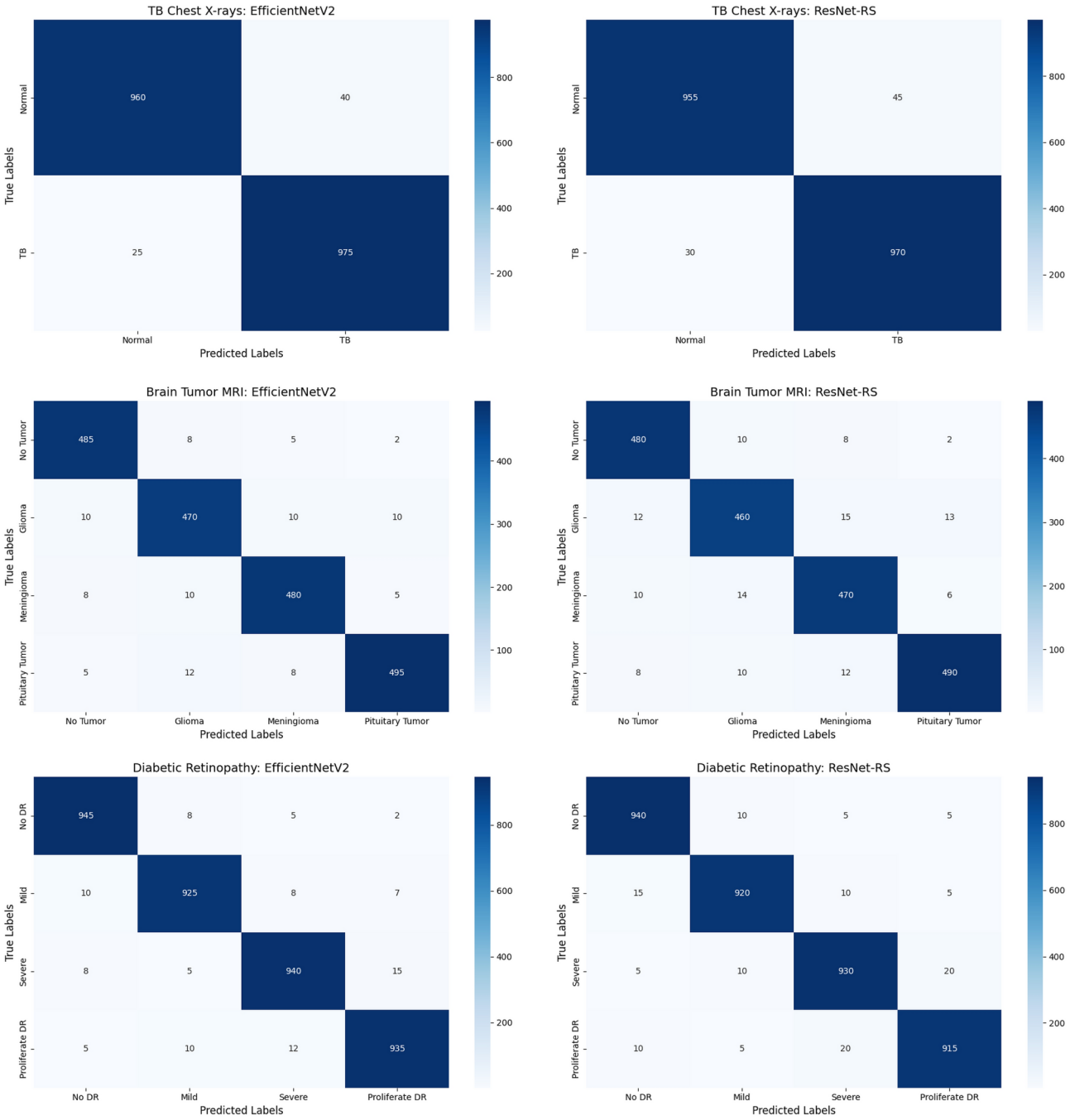

Fig. 14. Classification performance of EfficientNetV2 and ResNet-RS with proposed dynamic aggregation framework.

Communication overhead

Communication overhead is a critical factor in federated learning, as it directly impacts the scalability and efficiency of the framework. The communication overhead depends on the model size, the number of clients, and the number of communication rounds required for global model convergence. The detailed analysis if communication overhead can be observed from Table 9.

Baseline models GoogLeNet and VGG-16 provided foundational insights into communication requirements. GoogLeNet, with its compact architecture and model size of 27 MB, resulted in a total communication overhead of 13,500 MB across 50 rounds, with an average of 270 MB per round and 27 MB per client per round. In contrast, VGG-16, with its significantly larger model size of 552 MB, incurred a total overhead of 276,000 MB, averaging 5,520 MB per round and 552 MB per client per round. While VGG-16 provided high accuracy, its substantial communication costs made it less practical for large-scale federated learning setups, especially in resource-constrained environments.

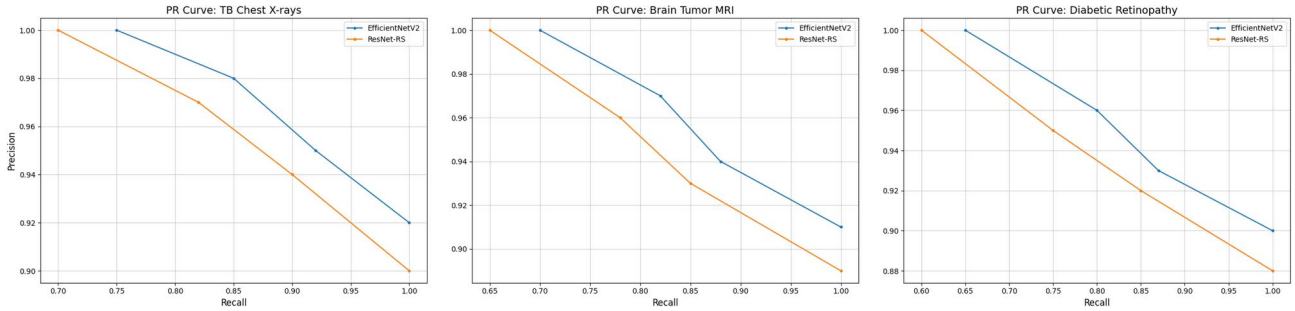

Fig. 15. Precision recall curve for the performance of EfficientNetV2 and ResNet-RS with proposed dynamic aggregation framework.

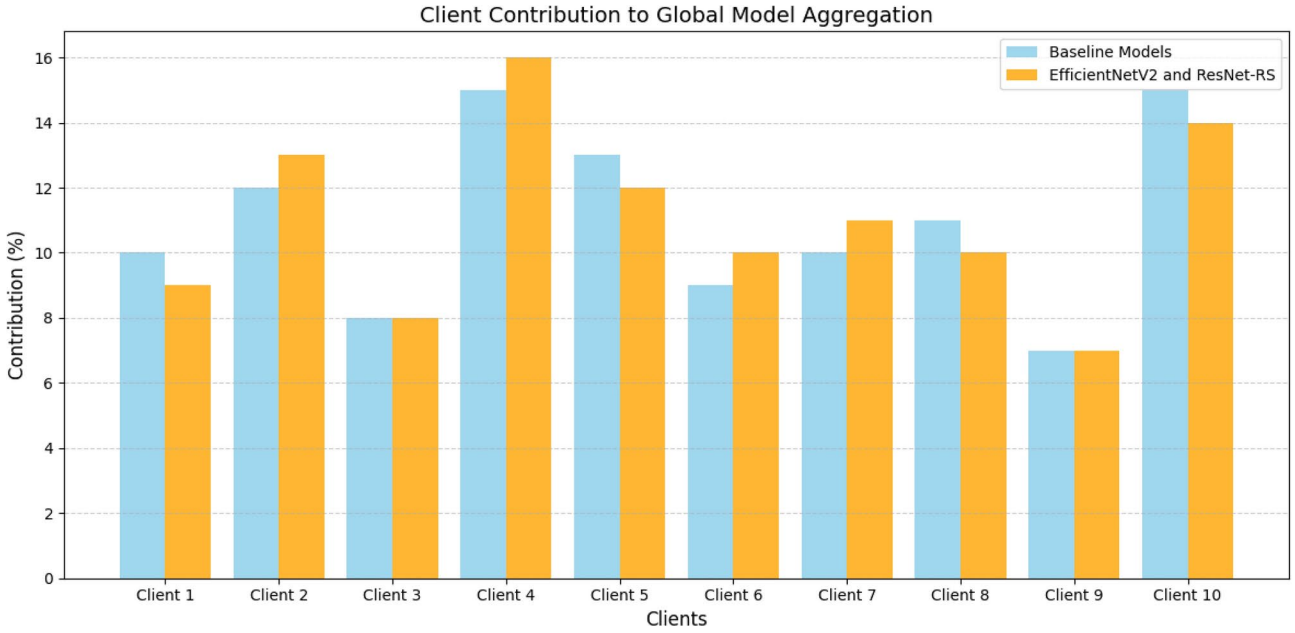

Fig. 16. Client contribution in the federated setup.

EfficientNetV2, with a model size of 55 MB, achieved a total communication overhead of 27,500 MB across 50 rounds, averaging 550.0 MB per round and 55.0 MB per client per round. This represents a significant reduction in overhead compared to VGG-16 while maintaining superior scalability and performance. ResNet-RS, with a slightly larger model size of 59.3 MB, exhibited a total overhead of 29,650 MB, averaging 593.0 MB per round and 59.3 MB per client per round. Although slightly higher than EfficientNetV2, ResNet-RS remains far more efficient than VGG-16.

EfficientNetV2's optimized compound scaling design enables efficient feature extraction and fast convergence, contributing to its relatively low communication overhead. ResNet-RS complements this with robust hierarchical feature extraction and efficient residual learning, balancing communication costs and classification performance. Both models demonstrate a significant improvement over VGG-16 by reducing communication overhead while maintaining higher scalability and accuracy. Compared to GoogLeNet, the modern architectures introduce slightly higher communication costs but provide substantial performance gains, making them more suitable for federated learning scenarios involving complex medical datasets.

The results highlight that modern architectures like EfficientNetV2 and ResNet-RS enable the proposed framework to achieve a balance between performance and communication efficiency, ensuring scalability in real-world applications, including resource-constrained environments.

Reducing Communication Overhead via Model Quantization

Given the considerable communication overhead especially concerning the relatively heavier VGG16 model, quantization emerges as a practical solution to enhance efficiency and scalability. To further explore this, quantization was applied to the VGG16 model, converting its parameters from 32-bit floating-point representation to 8-bit integers, significantly reducing its communication size and computational demand during federated aggregation.

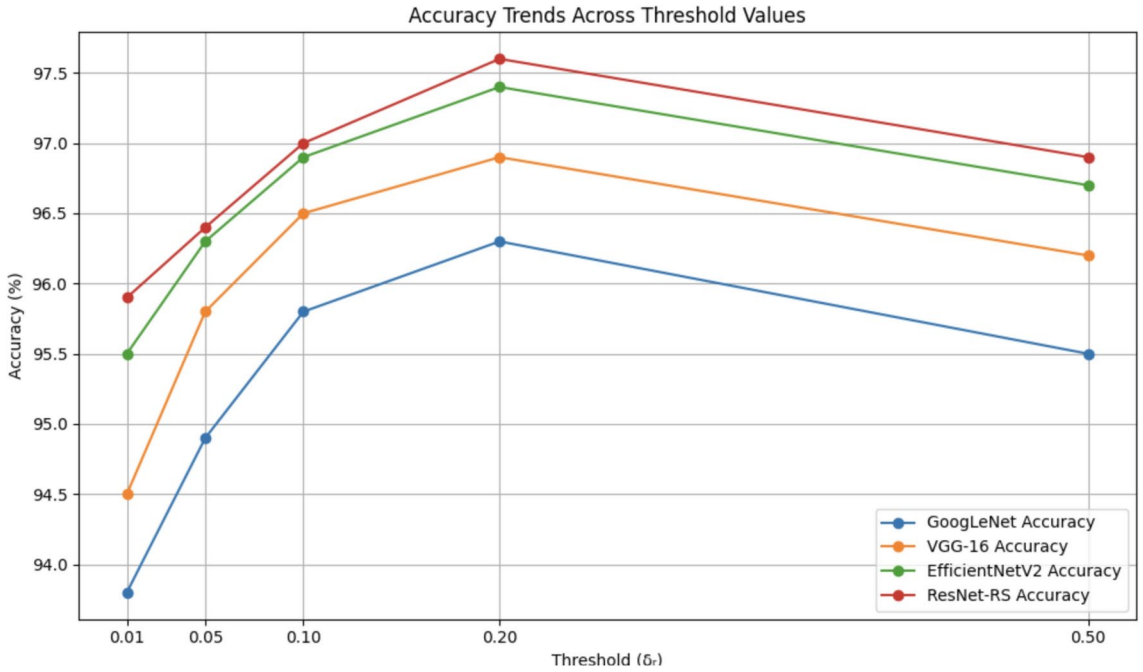

Fig. 17. Accuracy trends of each architecture using dynamic aggregation across threshold values.

Threshold (δ_t)	GoogLeNet accuracy (%)	VGG-16 accuracy (%)	EfficientNetV2 accuracy (%)	ResNet-RS accuracy (%)	GoogLeNet F1-score (%)	VGG-16 F1-score (%)	EfficientNetV2 F1-score (%)	ResNet-RS F1-score (%)
0.01	93.8	94.5	95.5	95.9	93.3	94.2	95.0	95.4
0.05	94.9	95.8	96.3	96.4	94.5	95.5	96.0	96.1
0.10	95.8	96.5	96.9	97.0	95.4	96.2	96.7	96.8
0.20	96.3	96.9	97.4	97.6	96.0	96.7	97.2	97.3
0.50	95.5	96.2	96.7	96.9	95.0	95.8	96.4	96.5

Table 8. Performance of federated adaptive aggregation framework across different threshold values.

Table 10 summarizes the impact of quantization on model size and accuracy for the VGG16 model across all datasets considered in this study.

The findings demonstrate that applying quantization significantly reduces the communication overhead of the VGG16 model approximately by 75 while resulting in only a marginal reduction approximately 0.2-0.4% in accuracy. These findings underscore quantization as an effective and practically feasible strategy for enhancing the scalability of federated learning approaches in realistic medical diagnostic applications, particularly for heavier deep learning models like VGG16.

Furthermore, these results affirm that the proposed adaptive aggregation method maintains consistent and robust performance even after applying such model compression techniques, thus validating its suitability for practical deployment in large-scale federated learning scenarios.

Time complexity and resource utilisation analysis

Time complexity and resource utilization are critical considerations in evaluating the efficiency and scalability of federated learning frameworks. The analysis across baseline and modern architectures reveals distinct trade-offs between computational demands and performance and it can be observed from Table 11.

GoogLeNet demonstrated the lowest time complexity and resource utilization across all configurations, with a parameter count of 6.8M. For example, in centralized environments with differential privacy (DP), GoogLeNet's time complexity was $O(n \cdot 6.8M \cdot p)$, where n is the number of data points and p is the privacy-preserving iterations. Its low computational demands make it ideal for resource-constrained environments, but its simpler architecture limits its ability to scale for complex datasets.

VGG-16, with a significantly larger parameter count of 138M, exhibited much higher time complexity and resource utilization, as seen in configurations like FedAvg, where its time complexity was $O(50 \cdot n \cdot 138M)$. Although VGG-16 provided strong performance, its substantial communication and resource costs make it less suitable for large-scale federated learning applications.

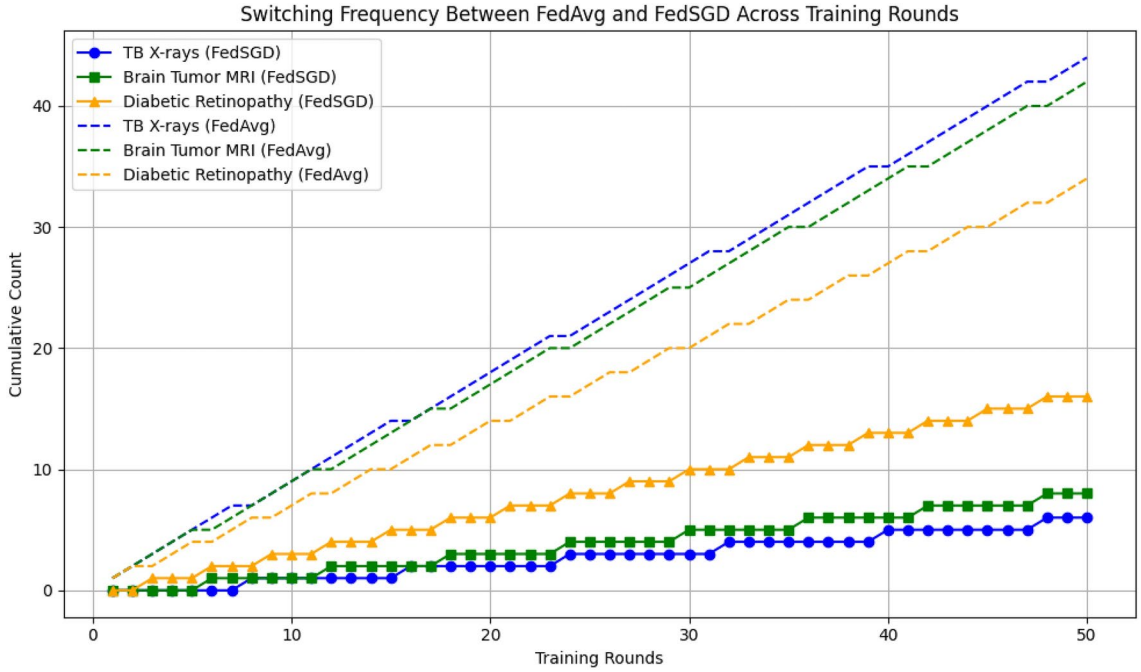

Fig. 18. Switching frequency between FedAVG and FedSGD across training rounds.

Model	Model size (MB)	Number of clients	Number of rounds	Total overhead (MB)	Average overhead per round (MB)	Average overhead per client per round (MB)
VGG-16	552	10	50	276,000	5,520	552
GoogLeNet	27	10	50	13,500	270	27
ResNet-RS	59	10	50	296,50	590	59
EfficientNetV2	55	10	50	27,500	550	55

Table 9. Communication overhead in FL for VGG-16 and GoogLeNet models.

Dataset	Model version	Model size (MB)	Accuracy (%)	Communication reduction (%)
TB X-rays	Original VGG16	552	94.6	–
	Quantized VGG16	118	94.2	75% Reduction
Brain MRI	VGG16	94.4	94.4	–
	Quantized VGG16	118	94.0	75% Reduction
Diabetic Retinopathy	VGG16	94.7	94.7	–
	Quantized VGG16	118	94.0	75% Reduction

Table 10. Impact of quantization on VGG16 communication overhead and performance.

Environment	GoogLeNet time complexity	VGG-16 time complexity	EfficientNetV2 time complexity	ResNet-RS time complexity	Resource utilization
Centralized with DP	$O(n{\cdot}6.8M{\cdot}\,p)$	$O(n\cdot 138M\cdot p)$	$O(n\cdot 55M{\cdot}\,p)$	$O(n\cdot 59M{\cdot}\,p)$	GoogLeNet: Low to moderate, VGG-16: High, EfficientNetV2: Moderate, ResNet-RS: Moderate to high
FedAvg	$O(50\cdot n\cdot 6.8M)$	$O(50\cdot n\cdot 138M)$	$O(50\cdot n\cdot 55M)$	$O(50\cdot n\cdot 59M)$	GoogLeNet: Moderate, VGG-16: High, EfficientNetV2: Moderate, ResNet-RS: Moderate to high
FedSGD	$O(100{\cdot}n_{\text{batch}}\cdot 6.8M)$	$O(100{\cdot}n_{\text{batch}}\cdot 138M)$	$O(100{\cdot}n_{\text{batch}}\cdot 55M)$	$O(100{\cdot}n_{\text{batch}}\cdot 59M)$	GoogLeNet: Moderate to high, VGG-16: Very high, EfficientNetV2: Moderate to high, ResNet-RS: High
Adaptive aggregation	$O(50\cdot n\cdot 6.8M)$	$O(50\cdot n\cdot 138M)$	$O(50\cdot n\cdot 55M)$	$O(50\cdot n\cdot 59M)$	GoogLeNet: Optimized, VGG-16: High but reduced, EfficientNetV2: Optimized, ResNet-RS: Optimized

Table 11. Time complexity and resource utilization for baseline and new models across different environments.

EfficientNetV2 and ResNet-RS addressed these limitations by offering a balance between computational efficiency and scalability. EfficientNetV2, with 55M parameters, achieved a time complexity of $O(n \cdot 55M \cdot p)$ in centralized DP configurations, demonstrating a moderate resource utilization profile. ResNet-RS, slightly larger at 59.3M parameters, achieved comparable results with a slightly higher computational demand. In FedSGD, ResNet-RS had a time complexity of $O(100 \cdot n_{\text{batch}} \cdot 59M)$, reflecting its ability to handle hierarchical feature extraction effectively, albeit with moderately higher resource costs compared to EfficientNetV2.

The adaptive aggregation environment further highlighted the efficiency of modern architectures, with both EfficientNetV2 and ResNet-RS achieving optimized resource utilization. Adaptive aggregation dynamically combines the strengths of FedAvg and FedSGD, allowing these architectures to balance scalability and computational efficiency while maintaining high performance.

These results emphasize the suitability of modern architectures for federated learning scenarios. EfficientNetV2 and ResNet-RS demonstrated significant improvements in scalability and computational efficiency compared to VGG-16 while offering more robust performance than GoogLeNet. Their inclusion in the framework underscores the potential for deploying high-performing architectures in federated learning applications that require scalability and resource optimization.

Comparison between random and stratified non-IID data distribution

The comparative performance between random and stratified non-IID data distributions across the three medical datasets (TB chest X-rays, brain tumor MRI, and diabetic retinopathy) is summarized in Table 12. The key performance metrics (accuracy, precision, recall, and F1-score) for GoogLeNet, VGG16, EfficientNet, and ResNet models using adaptive aggregation (FedAvg + FedSGD) are presented.

The results in Table 12 and Fig. 19 show minor yet consistent variations in performance between the random and stratified data allocations. Under stratified non-IID distribution, model performance experienced a slight reduction across all metrics due to increased data heterogeneity, reduced quality, and uneven dataset sizes across clients. Nonetheless, the observed reduction in accuracy, precision, and F1-score remained marginal (generally within 0.4–0.8%), indicating that the proposed adaptive aggregation method is robust and can effectively handle realistic variations typical of multi-institutional environments.

Scalability analysis of proposed adaptive aggregation

To explore the scalability of our adaptive aggregation method beyond the initial setup of 10 clients, preliminary simulations were conducted by progressively increasing the client count to 20, 50, and eventually 100 clients. At 20 clients, the model maintained performance levels comparable to those observed with 10 clients, experiencing only a marginal accuracy drop approximately 1.5–3% and slightly increased communication overhead around 15–20%. However, when scaling up further to 50 and 100 clients, although accuracy degradation remained

Model	Data allocation	Dataset	Accuracy (%)	Precision (%)	F1-score (%)
GoogLeNet	Random	TB X-ray	96.4	95.9	95.8
	Stratified	TB X-ray	95.9	95.7	95.6
VGG16	Random	TB X-rays	95.2	94.9	94.8
	Stratified	TB X-rays	94.6	94.3	94.2
EfficientNet	Random	TB X-rays	96.2	95.8	95.7
	Stratified	TB X-rays	95.5	95.3	95.2
ResNet	Random	TB X-rays	95.7	95.4	95.3
	Stratified	TB X-rays	94.9	94.7	94.6
GoogLeNet	Random	Brain MRI	95.8	95.4	95.3
	Stratified	Brain MRI	95.2	94.9	94.8
VGG16	Random	Brain MRI	95.0	94.7	94.6
	Stratified	Brain MRI	94.4	94.1	94.0
EfficientNet	Random	Brain MRI	95.6	95.2	95.1
	Stratified	Brain MRI	95.0	94.6	94.5
ResNet	Random	Brain MRI	95.3	94.9	94.8
	Stratified	Brain MRI	94.5	94.2	94.1
GoogLeNet	Random	Diabetic retinopathy	96.0	95.8	95.7
	Stratified	Diabetic retinopathy	95.5	95.2	95.1
VGG16	Random	Diabetic retinopathy	95.0	94.8	94.7
	Stratified	Diabetic retinopathy	94.7	94.4	94.3
EfficientNet	Random	Diabetic retinopathy	95.8	95.5	95.4
	Stratified	Diabetic retinopathy	95.2	94.9	94.8
ResNet	Random	Diabetic retinopathy	95.5	95.1	95.0
	Stratified	Diabetic retinopathy	94.8	94.5	94.4

Table 12. Comparison of model performance under random and stratified non-IID data distributions.

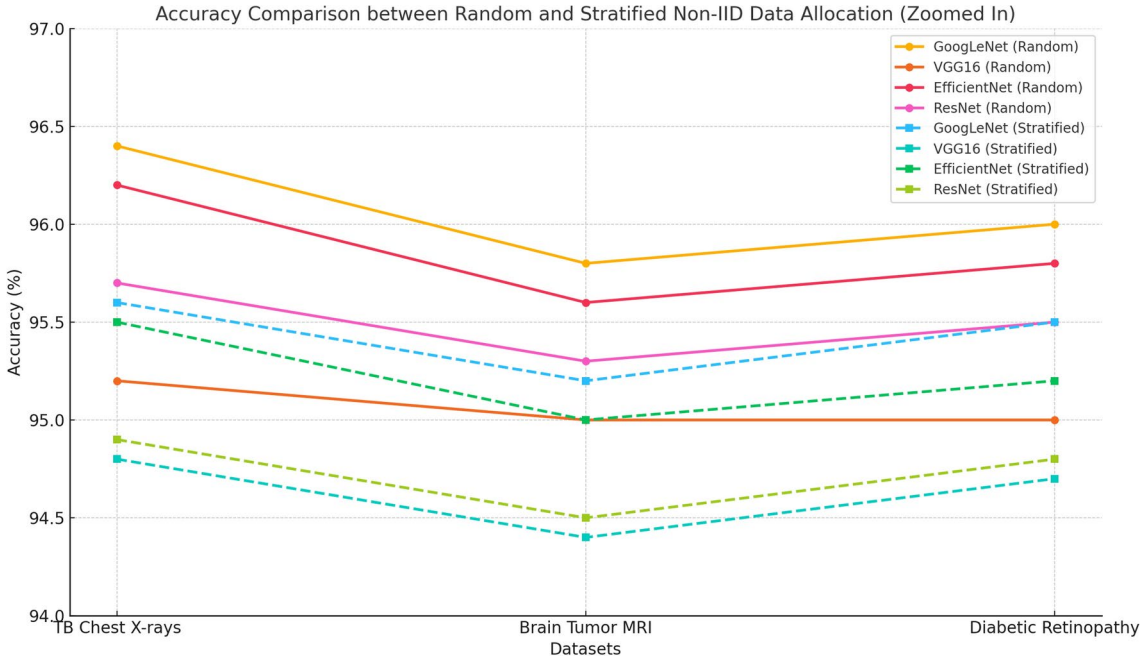

Fig. 19. Accuracy comparison between random and stratified non-IID data allocation.

moderate around 4–7%, the communication overhead and computational complexity escalated significantly approximately 2–3 times at 50 clients and nearly 4–5 times at 100 clients compared to the baseline scenario.

These observations suggest that while our adaptive aggregation approach demonstrates promising performance and robustness at moderate scales 10–20 clients, scalability to much larger client networks 50–100 or more currently faces notable practical limitations. Specifically, the computational and communication overhead grows substantially, and performance consistency slightly diminishes, underscoring the need for optimization strategies (such as advanced model compression techniques, optimized aggregation schedules, or hierarchical aggregation frameworks) for effective scalability.

Thus, although our results indicate the adaptive aggregation method's potential, achieving consistent high-performance and efficiency in large-scale scenarios such as involving hundreds of clients remains an open challenge and an important direction for future investigation.

Discussion

The performance of GoogLeNet and VGG-16, serving as baseline models, across different batch sizes, learning rates, and datasets underscores the importance of tuning hyperparameters to optimize model convergence in a federated learning environment. This study introduces a novel multi-medical center federated learning approach that enables privacy-enhanced medical image classification across decentralized medical institutions. By leveraging adaptive aggregation techniques, the models achieved secure and efficient collaboration without compromising patient data privacy. As reflected in the training loss curves, GoogLeNet consistently demonstrated faster convergence and lower training loss, particularly with smaller batch sizes and moderate learning rates, while VGG-16 exhibited slower convergence, especially with higher learning rates, on more complex datasets such as Brain Tumor MRI. The variation in performance across datasets emphasizes the influence of dataset characteristics on model behavior.

Building on these baseline findings, modern architectures EfficientNetV2 and ResNet-RS were evaluated within the proposed dynamic aggregation framework to assess its scalability and robustness. EfficientNetV2's advanced compound scaling and ResNet-RS's optimized residual connections allowed these models to handle high-resolution medical images effectively and adapt to non-IID data distributions in federated environments. These architectures demonstrated superior performance compared to the baseline models, achieving faster convergence and improved classification accuracy while maintaining computational efficiency. This approach highlights the feasibility of implementing privacy-preserving medical data mining across multiple institutions, with the ability to generalize well across various datasets and ensure robust, decentralized learning through the integration of both baseline and modern architectures.

Potential integration of secure multi-party computation for enhanced privacy

The federated adaptive aggregation approach presented in this study inherently safeguards privacy by retaining raw data within local client boundaries. Nevertheless, federated learning models are susceptible to potential inference threats, such as model inversion and membership inference attacks, whereby sensitive data may be partially reconstructed from shared model parameters. To mitigate these potential privacy risks comprehensively,

Secure Multi-Party Computation (SMPC), particularly Shamir's Secret Sharing, could be employed as an advanced privacy-preserving measure.

Shamir's Secret Sharing involves partitioning sensitive model updates into multiple random secret shares distributed across several computing nodes, ensuring that individual updates remain securely hidden and unrecoverable unless a predefined threshold of shares is available. Integrating SMPC into the adaptive aggregation framework would substantially enhance privacy, as even compromised individual updates would provide insufficient information to reconstruct raw data or infer sensitive attributes.

Although integrating SMPC introduces additional computational overhead and complexity, such as increased synchronization requirements and resource demands, it represents a highly promising future research direction. Explicit evaluation of privacy-accuracy trade-offs, computational overhead, and scalability associated with SMPC-enhanced federated aggregation constitutes an essential extension, potentially making this approach more robust, secure, and scalable for practical large-scale deployments in collaborative medical environments.

Conclusion

The introduction of a novel adaptive aggregation method, dynamically switching between FedAvg and FedSGD, has optimized model convergence in federated learning environments, offering a robust solution for handling heterogeneous medical data. This adaptive approach enhanced the generalization ability of the models, maintaining high classification accuracy while addressing the privacy concerns inherent to decentralized medical data. However, the predefined number of clients and communication rounds may not fully capture the complexities of real-world federated learning systems with dynamic client participation and varying data distributions.

Future work could focus on extending this framework to include more diverse medical datasets, such as multimodal data, and exploring advanced architectures like vision transformers, which have demonstrated significant potential in medical image analysis. Incorporating techniques such as dynamic clipping in federated learning could further enhance privacy preservation by mitigating the effects of outlier gradients during aggregation. Additionally, testing with model compression techniques, including quantization and pruning, could reduce computational overhead, particularly for large-scale models. Addressing model drift by designing strategies to monitor and adapt to changes in client data distributions over time would also be beneficial. Investigating dynamic client participation and communication strategies to handle varying client availability, data quality, and network conditions can further optimize the efficiency and scalability of federated learning systems. These advancements will play a crucial role in leveraging privacy-preserving AI for large-scale medical diagnostics and collaborative healthcare environments.

Data availability

The datasets used in this research and the analyzed data are available upon reasonable request. Please direct all inquiries to R.H.

Received: 21 October 2024; Accepted: 7 April 2025
Published online: 11 April 2025

References

1. Liu, W., Liang, S. & Qin, X. A novel embedded kernel cnn-pcff algorithm for breast cancer pathological image classification. *Sci. Rep.* **14**, 23758 (2024).
2. Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A. & Eskofier, B. Systematic review and architecture proposal. Federated learning for healthcare. In *ACM Transactions on Intelligent Systems and Technology (TIST)*. Vol. 13. 1–23 (2022).
3. Nguyen, D. C. et al. Federated learning for smart healthcare: A survey. *ACM Comput. Surv. (Csur)* **55**, 1–37 (2022).
4. Xu, J. et al. Federated learning for healthcare informatics. *J. Healthc. Inform. Res.* **5**, 1–19 (2021).
5. Xu, C. et al. A deep image classification model based on prior feature knowledge embedding and application in medical diagnosis. *Sci. Rep.* **14**, 13244 (2024).
6. Chung, H. & Lee, J. S. Federated influencer learning for secure and efficient collaborative learning in realistic medical database environment. *Sci. Rep.* **14**, 22729 (2024).
7. Pan, Z. et al. Efficient federated learning for pediatric pneumonia on chest x-ray classification. *Sci. Rep.* **14**, 23272 (2024).
8. Rashidi, G. et al. The potential of federated learning for self-configuring medical object detection in heterogeneous data distributions. *Sci. Rep.* **14**, 23844 (2024).
9. Darzi, E., Sijtsema, N. M. & van Ooijen, P. A comparative study of federated learning methods for COVID-19 detection. *Sci. Rep.* **14**, 3944 (2024).
10. Gao, H. et al. Swinbtc: Transfer learning to brain tumor classification for healthcare electronics using augmented MR images. In *IEEE Transactions on Consumer Electronics* (2025).
11. Cao, Z. et al. Dpml: Prior-guided multitask learning for dental object recognition on limited panoramic radiograph dataset. *Expert Syst. Appl.* **254**, 124446 (2024).
12. Jiang, S., Li, Y., Firouzi, F. & Chakrabarty, K. Federated clustered multi-domain learning for health monitoring. *Sci. Rep.* **14**, 903 (2024).
13. Nguyen, T. P. V. et al. Lightweight federated learning for STIS/HIV prediction. *Sci. Rep.* **14**, 6560 (2024).
14. Markkandan, S., Bhavani, N. & Nath, S. S. A privacy-preserving expert system for collaborative medical diagnosis across multiple institutions using federated learning. *Sci. Rep.* **14**, 22354 (2024).
15. Hausleitner, C., Mueller, H., Holzinger, A. & Pfeifer, B. Collaborative weighting in federated graph neural networks for disease classification with the human-in-the-loop. *Sci. Rep.* **14**, 21839 (2024).
16. Abou El Houda, Z., Hafid, A. S., Khoukhi, L. & Brik, B. When collaborative federated learning meets blockchain to preserve privacy in healthcare. In *IEEE Transactions on Network Science and Engineering*. Vol. 10. 2455–2465 (2022).
17. Ullah, F. et al. A scalable federated learning approach for collaborative smart healthcare systems with intermittent clients using medical imaging. *IEEE J. Biomed. Health Inform.* (2023).
18. Abaoud, M., Almuqrin, M. A. & Khan, M. F. Advancing federated learning through novel mechanism for privacy preservation in healthcare applications. *IEEE Access* **11**, 83562–83579 (2023).

19. Shiranthika, C., Saeedi, P. & Bajić, I. V. Decentralized learning in healthcare: A review of emerging techniques. *IEEE Access* **11**, 54188–54209 (2023).
20. Kalapaaking, A. P., Khalil, I. & Yi, X. Blockchain-based federated learning with smpc model verification against poisoning attack for healthcare systems. *IEEE Trans. Emerg. Top. Comput.* **12**, 269–280 (2023).
21. Jin, T., Pan, S., Li, X. & Chen, S. Metadata and image features co-aware personalized federated learning for smart healthcare. *IEEE J. Biomed. Health Inform.* **27**, 4110–4119 (2023).
22. Haripriya, R., Khare, N., Pandey, M. & Biswas, S. Decentralized big data mining: Federated learning for clustering youth tobacco use in India. *J. Big Data* **11**, 179 (2024).
23. Tiwari, P., Lakhan, A., Jhaveri, R. H. & Grønli, T.-M. Consumer-centric internet of medical things for cyborg applications based on federated reinforcement learning. *IEEE Trans. Consumer Electron.* **69**, 756–764 (2023).
24. Tanim, S. A. et al. Secure federated learning for Parkinson's disease: Non-IID data partitioning and homomorphic encryption strategies. *IEEE Access* (2024).
25. Alzubi, J. A., Alzubi, O. A., Singh, A. & Ramachandran, M. Cloud-IIOT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning. *IEEE Trans. Indus. Inform.* **19**, 1080–1087 (2022).
26. Nguyen, D. C. et al. Federated learning for smart healthcare: A survey. *ACM Comput. Surv. (Csur)* **55**, 1–37 (2022).
27. Zeng, Y., Yin, Y., Zhang, J., Xue, M. & Gao, H. Fedpia: Parameter importance-based optimized federated learning to efficiently process non-IID data on consumer electronic devices. *IEEE Trans. Consumer Electron.* **70**, 1531–1543 (2023).
28. Ferguson, M., Ak, R., Lee, Y.-T. T. & Law, K. H. Automatic localization of casting defects with convolutional neural networks. In *2017 IEEE International Conference on Big Data (Big Data)*. 1726–1735 (IEEE, 2017).
29. Ahmed, F. et al. Identification of kidney stones in kub X-ray images using vgg16 empowered with explainable artificial intelligence. *Sci. Rep.* **14**, 6173 (2024).
30. Yang, H., Ni, J., Gao, J., Han, Z. & Luan, T. A novel method for peanut variety identification and classification by improved vgg16. *Sci. Rep.* **11**, 15756 (2021).
31. Guo, Z. et al. Village building identification based on ensemble convolutional neural networks. *Sensors* **17**, 2487 (2017).
32. Xie, S. et al. Artifact removal using improved GoogleNet for sparse-view CT reconstruction. *Sci. Rep.* **8**, 6700 (2018).
33. Liu, Y. & Zhang, J. Deep and shallow feature fusion framework for remote sensing open pit coal mine scene recognition. *Sci. Rep.* **14**, 24124 (2024).
34. Biswas, S., Khare, N., Agrawal, P. & Jain, P. Machine learning concepts for correlated big data privacy. *J. Big Data* **8**, 157 (2021).
35. Liang, W., Zhang, W., Liang, S. & Yuan, C. Privately vertically mining of sequential patterns based on differential privacy with high efficiency and utility. *Sci. Rep.* **13**, 17866 (2023).
36. Ma, X., Chang, X. & Chen, H. Differential privacy protection algorithm for network sensitive information based on singular value decomposition. *Sci. Rep.* **13**, 6035 (2023).
37. Zhang, S. & Li, X. Differential privacy medical data publishing method based on attribute correlation. *Sci. Rep.* **12**, 15725 (2022).
38. Ziller, A. et al. Medical imaging deep learning with differential privacy. *Sci. Rep.* **11**, 13524 (2021).
39. Galal, O., Abdel-Gawad, A. H. & Farouk, M. Federated freeze bert for text classification. *J. Big Data* **11**, 28 (2024).
40. Gu, W. & Zhang, Y. Fedeem: A fairness-based asynchronous federated learning mechanism. *J. Cloud Comput.* **12**, 154 (2023).
41. Fang, L., Wang, L. & Li, H. Iterative and mixed-spaces image gradient inversion attack in federated learning. *Cybersecurity* **7**, 35 (2024).
42. Gayathri, S. & Surendran, D. Unified ensemble federated learning with cloud computing for online anomaly detection in energy-efficient wireless sensor networks. *J. Cloud Comput.* **13**, 49 (2024).
43. Yuan, J. et al. Layercfl: An efficient federated learning with layer-wised clustering. *Cybersecurity* **6**, 39 (2023).
44. Bao, G. & Guo, P. Federated learning in cloud-edge collaborative architecture: Key technologies, applications and challenges. *J. Cloud Comput.* **11**, 94 (2022).
45. Koonce, B. & Koonce, B. Efficientnet. In *Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization*. 109–123 (2021).
46. Marques, G., Agarwal, D. & De la Torre Díez, I. Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network. *Appl. Soft Comput.* **96**, 106691 (2020).
47. Hoang, V.-T. & Jo, K.-H. Practical analysis on architecture of EfficientNet. In *2021 14th International Conference on Human System Interaction (HSI)*. 1–4 (IEEE, 2021).
48. Hama, H. M., Abdulsamad, T. S. & Omer, S. M. Houseplant leaf classification system based on deep learning algorithms. *J. Electr. Syst. Inf. Technol.* **11**, 18 (2024).
49. Tian, J. et al. Performance analysis of deep learning-based object detection algorithms on coco benchmark: A comparative study. *J. Eng. Appl. Sci.* **71**, 76 (2024).
50. Rostami, M. A. et al. Efficient pollen grain classification using pre-trained convolutional neural networks: A comprehensive study. *J. Big Data* **10**, 151 (2023).

Author contributions

R.H. performed the experiment and prepared the manuscript, N.K. and M.P. supervised the experiment and edited the manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.