information

MDPI

Article

Federated Learning for Breast Cancer Classification: A Comparative Study of Aggregation Methods

Nadjat Saàdia Lachemi [1,2,*], Medjeded Merati [2,3] and Saïd Mahmoudi [4,*]

[1] Laboratoire de Génie Energétique et Génie Informatique (L2GEGI), Faculty of Mathematics and Computer Science, University of Tiaret, BP P 78 Zaäroura, Tiaret 14000, Algeria
[2] Department of Computer Science, Faculty of Mathematics and Computer Science, University of Tiaret, Tiaret 14000, Algeria; medjeded.merati@univ-tiaret.dz
[3] Laboratoire d'Informatique et Mathematique (LIM), Faculty of Mathematics and Computer Science, University of Tiaret, BP P 78 Zaäroura, Tiaret 14000, Algeria
[4] ILIA Department, Faculty of Engineering, University of Mons (UMONS), 20 Place du Parc, 7000 Mons, Belgium
* Correspondence: nadjetsaadia.lachemi@univ-tiaret.dz or nadjat.laachemi@gmail.com (N.S.L.); said.mahmoudi@umons.ac.be (S.M.); Tel.: +213-54-297-1543 (N.S.L.)

Abstract

Federated Learning (FL) allows healthcare institutions to collaboratively develop machine learning models while safeguarding patient data, making it ideal for privacy-sensitive medical imaging. This study explores the effects of data heterogeneity on federated breast cancer classification using MobileNetV2 across five simulated clients. Five aggregation methods—FedAvg, FedProx, FedNova, FedDyn, and SCAFFOLD—were assessed under various data distributions, including balanced, imbalanced, non-homogeneous, and non-IID. Results indicate that aggregation performance is significantly affected by data distribution; FedAvg excels in balanced settings but falters in heterogeneity, whereas FedProx shows robustness in extreme non-IID cases, achieving up to 98.466% accuracy. FedDyn and SCAFFOLD also demonstrate adaptability but are less consistent in severe imbalance scenarios. Beyond accuracy, recall and robustness under extreme non-IID conditions were analyzed to assess clinical reliability in cancer detection. These results underscore the necessity of choosing suitable aggregation methods for effective medical federated learning.

Keywords: federated learning; breast cancer classification; MobileNetV2; aggregation methods; non-IID data; cross-silo learning

Academic Editor: Heming Jia

Received: 29 April 2026
Revised: 24 May 2026
Accepted: 26 May 2026
Published: 2 June 2026

1. Introduction

Breast cancer remains one of the most prevalent cancers among women worldwide, accounting for approximately 24.5% of all new cancer cases and 15.5% of cancer-related deaths in women [1]. According to the World Health Organization (WHO), early detection and accurate diagnosis significantly improve survival rates, making imaging-based classification crucial for timely intervention [1]. Mammography is the most widely used screening technique, assisting radiologists in detecting abnormalities such as benign and malignant lesions. However, manual interpretation is subjective and prone to inter-observer variability, necessitating the integration of artificial intelligence (AI) to enhance diagnostic accuracy [2].

Deep learning, a subset of AI, has shown remarkable performance in medical image classification, particularly in detecting and classifying breast cancer lesions [3]. Convolutional Neural Networks (CNNs) have demonstrated superiority in extracting relevant

features from mammograms, outperforming traditional machine learning techniques [4,5]. Among various CNN architectures, MobileNetV2 has gained attention due to its lightweight design and efficiency, making it suitable for resource-constrained environments [6].

Despite these advances, the use of deep learning methods often requires centralized data collection, where large-scale datasets are aggregated into a single repository for model development [7]. In medical imaging, however, this strategy poses several challenges, primarily concerning data accessibility and privacy:

1. Privacy and Security Concerns: Patient data is highly sensitive and subject to strict privacy regulations, such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) [8,9]. Hospitals and medical institutions are often reluctant to share data due to ethical concerns and potential legal restrictions, limiting access to diverse and high-quality datasets.
2. Data Heterogeneity: Medical imaging datasets are often distributed across multiple hospitals and research centers, each with distinct imaging protocols, devices, and demographic variations. Aggregating such data into a centralized repository introduces bias and data heterogeneity issues [10,11].
3. Computational and Storage Constraints: Transferring large-scale medical image datasets incurs high computational and storage costs, further hindering the feasibility of centralized learning approaches [12].

Given these constraints, there is a growing need for decentralized learning frameworks that enable collaborative training without compromising data privacy and security. This has led to the emergence of Federated Learning (FL) as a promising alternative.

FL is a decentralized machine learning paradigm that enables multiple institutions to collaboratively train a shared model while keeping their data localized [13]. Unlike traditional centralized learning, where data is collected and stored in a single location, FL allows hospitals to train models locally and share only model updates (gradients) with a central server [14]. This approach ensures compliance with privacy regulations while leveraging diverse medical data to improve model generalization.

In healthcare, the application of FL has gained momentum due to its potential to address data-sharing barriers while preserving model performance [7,15]. Recent studies have increasingly explored federated learning frameworks for breast cancer prediction and pathology image classification to enable collaborative medical AI training while preserving patient privacy [16,17]. In breast cancer classification, FL enables multiple medical centers to participate in collaborative learning, ensuring that models benefit from diverse patient populations without exposing sensitive patient data [18]. However, FL introduces new challenges, such as dealing with non-IID (non-independent and identically distributed) data, communication inefficiencies, and heterogeneous computational resources across participating institutions [13,18].

To enhance FL's robustness in medical applications, several aggregation strategies have been proposed. Among them, FedAvg is widely used due to its simplicity and efficiency [13]. However, it struggles with non-IID data distributions. To mitigate this, FedProx introduces a proximal term to stabilize training in heterogeneous settings [19]. FedNova improves performance by normalizing update contributions across clients [20], while FedDyn dynamically adjusts optimization to handle complex data distributions effectively [21]. Additionally, SCAFFOLD addresses drift in local updates [22].

Despite the growing adoption of federated learning in medical imaging, several limitations remain in existing studies. Many previous works evaluate federated learning models under relatively simplified experimental settings, often relying on homogeneous or mildly heterogeneous client distributions and focusing on only a limited subset of aggregation strategies. Furthermore, most studies primarily emphasize overall classification accuracy

while providing limited analysis of aggregation behavior under severe class imbalance, non-IID distributions, multi-source client variability, and distorted imaging conditions that commonly occur in real-world healthcare environments. As a result, the robustness and reliability of aggregation methods in clinically realistic federated learning scenarios remain insufficiently explored. To address this gap, this study presents a systematic comparative analysis of five widely used federated aggregation methods under five progressively challenging data distribution configurations designed to emulate realistic cross-silo medical imaging environments.

Unlike previous studies that investigate federated learning under limited or idealized settings, this study aims to systematically evaluate the effectiveness, robustness, stability, and clinical reliability of multiple federated aggregation methods for breast cancer classification using MobileNetV2 under progressively heterogeneous and clinically realistic data distributions. The research focuses on addressing the following questions:

- How does FL perform in a binary breast cancer classification task?
- How do different FL aggregation methods (FedAvg, FedProx, FedNova, FedDyn, and SCAFFOLD) impact classification accuracy and convergence?
- How does data heterogeneity (balanced, imbalanced, non-IID, and non-homogeneous distributions) affect FL model performance?

The primary objective of this work is to assess the feasibility of federated learning for breast cancer classification by examining widely used aggregation strategies under progressively realistic and heterogeneous medical data configurations. Unlike studies that focus primarily on proposing new optimization algorithms, this work emphasizes a clinically motivated evaluation framework designed to analyze the stability, robustness, and reliability of federated aggregation methods under real-world challenges such as class imbalance, non-IID distributions, client drift, and image heterogeneity. The main contributions of this study are summarized as follows:

1. We conduct a comprehensive comparative evaluation of five federated learning aggregation strategies (FedAvg, FedProx, FedNova, FedDyn, and SCAFFOLD) for breast cancer classification using a unified MobileNetV2 architecture.
2. We design five progressively challenging federated configurations combining IID, non-IID, class imbalance, multi-source heterogeneity, and client-specific image distortions to better simulate realistic cross-silo hospital environments.
3. We provide an in-depth analysis of aggregation method behavior under heterogeneous federated conditions, including client drift, distribution shift, and minority-class sensitivity, with particular emphasis on clinically relevant metrics such as recall and false-negative detection.
4. We demonstrate that aggregation strategies specifically designed to mitigate client heterogeneity provide improved robustness and stability compared to conventional averaging-based methods in decentralized medical imaging applications.
5. We highlight the importance of evaluating clinical reliability beyond overall accuracy in federated healthcare systems, showing that high accuracy may still correspond to poor malignant-case detection under highly imbalanced distributions.

By addressing these challenges, this work contributes toward the development of more reliable and privacy-preserving federated learning frameworks for real-world breast cancer diagnosis.

The remainder of this paper is structured as follows: Section 2 reviews related work, including previous studies on breast cancer classification using deep learning, the limitations of centralized medical imaging approaches, and recent advancements in federated learning for healthcare applications. Section 3 outlines the methodology, detailing the class

distributions, selected datasets, preprocessing steps, data configuration scenarios and the aggregation methods (FedAvg, FedProx, FedNova, FedDyn, and SCAFFOLD). Section 4 introduces the federated learning setup with the MobileNetV2 model architecture, training strategies and the evaluation metrics, and computational resources. Section 5 presents and discusses the experimental results, comparing the aggregation methods across five different data configurations. Finally, Section 6 concludes the paper by summarizing the key findings and suggesting future research directions.

2. Related Work

2.1. Deep Learning and the Limitations of Centralized Medical Imaging Approaches

Over the past decade, deep learning, especially Convolutional Neural Networks (CNNs), has excelled in breast cancer classification. Initial work by Spanhol et al. [23] achieved 80–85% accuracy with histopathological images. Subsequent studies, such as Ragab et al. [24] enhanced performance using multiple CNN architectures and deep feature fusion, while Arevalo et al. [25] developed a framework for classifying breast mass lesions in mammography. Their approach, which utilized a dataset of 736 biopsy-proven images, achieved an AUC of 0.822 and improved to 0.826 by combining learned and hand-crafted features.

Traditional centralized learning in medical imaging aggregates data from multiple sources, enhancing dataset diversity but presenting privacy and logistical challenges. Health institutions face significant challenges in sharing patient data due to strict regulations such as HIPAA (Health Insurance Portability and Accountability Act) [8] and GDPR (General Data Protection Regulation) [9] which prioritize patient confidentiality and data security. The centralization of data creates vulnerabilities, making institutions susceptible to data breaches. Furthermore, variations in imaging protocols, annotation standards, and patient demographics among different institutions lead to domain shifts, compromising the generalizability of AI models and hindering their effective application in real-world healthcare settings [26].

2.2. Federated Learning in Breast Cancer Prediction and Medical Imaging

Federated Learning (FL), first introduced by McMahan et al. [13], offers a promising alternative by enabling collaborative model training in medical imaging without sharing raw patient data, allowing clients to train models locally and share only updated parameters for aggregation.

In medical imaging, Sheller et al. [27] pioneered federated learning (FL) for brain tumor segmentation, achieving a Dice score of 0.852, which is near the centralized training benchmark of 0.862- while outperforming other collaborative methods and maintaining data privacy. Li et al. [28] developed a privacy-preserving FL method with domain adaptation for multi-site fMRI classification, showing that reliable biomarkers can be identified without centralizing data. Their work highlighted the importance of addressing domain shifts across institutions; In contrast, these studies focused mainly on domain adaptation and did not provide a comparative evaluation of multiple federated aggregation methods under highly imbalanced conditions. Li et al. [29] successfully applied FL for histopathological breast cancer classification, demonstrating robust performance across non-IID datasets. Although their results confirmed the potential of FL in breast cancer analysis, the evaluation was limited to a narrower range of heterogeneity scenarios and aggregation behaviors compared to real-world multi-institutional environments. Almufareh et al. [16] implemented a federated framework across multiple hospitals and reported 97.54% accuracy, 96.5% precision, and 98.0% recall for breast cancer prediction. Similarly, Gupta et al. [17] employed YOLOv6 in a federated ensemble setup, achieving 98% and 97% accuracy on the

BreakHis and BUSI datasets, respectively. Their model outperformed ResNet-50, VGG-19, and InceptionV3, demonstrating the potential of FedAvg with privacy-preserving techniques. Despite the strong predictive performance, their studies focused mainly on overall classification metrics without extensively analyzing model robustness under severe non-IID distributions or client-specific variability.

Recent studies have further explored advanced federated learning architectures for breast cancer prediction. Al-Hejri et al. [30] proposed a hybrid explainable FL framework combining Vision Transformers and CNNs with LIME-based interpretability, achieving accuracies of 98.65%, 97.30%, and 95.59% for binary, multi-class, and BI-RADS classification, respectively, with an AUC of 0.970. Similarly, Alhussan et al. [31] developed a federated framework for 3D mammographic image classification and reported 97.37% accuracy under federated training using CNN-based models, demonstrating that FL can maintain competitive predictive performance while preserving patient privacy. Jiménez-Sánchez et al. [32] introduced a memory-aware curriculum federated learning strategy combined with domain adaptation, achieving improvements of approximately 5% in ROC-AUC and 6% in PR-AUC on multi-site clinical datasets. However, these studies generally involved a limited number of clients, did not extensively evaluate robustness under highly heterogeneous federated environments, and primarily focused on model architecture or training strategy optimization rather than systematically comparing aggregation behavior under severe non-IID conditions and varying levels of statistical heterogeneity.

Despite its promise, FL in healthcare faces several technical challenges, including data heterogeneity, high communication overhead, and uneven computational resources across institutions. Differences in imaging protocols and demographics lead to non-IID data, which may degrade model convergence and accuracy. Recent surveys further emphasize that statistical heterogeneity, communication constraints, and client drift remain major barriers to reliable federated deployment in medical imaging applications [33]. In clinical deployment scenarios, unreliable convergence under heterogeneous distributions may reduce diagnostic sensitivity and increase the risk of false-negative breast cancer predictions across institutions.

Although existing studies demonstrate the potential of federated learning for breast cancer prediction and medical imaging, many focus on limited aggregation strategies or relatively simplified federated settings. Most evaluations consider moderate heterogeneity or near-IID data distributions, which do not fully reflect real-world clinical environments where institutions differ in dataset size, class balance, imaging protocols, acquisition quality, and patient demographics. Consequently, the robustness and convergence behavior of aggregation methods under severe heterogeneity remain insufficiently explored. In particular, limited attention has been given to progressively challenging scenarios combining class imbalance, non-IID multi-source distributions, and client-specific image distortions. These limitations motivate the present study, which provides a comparative evaluation of multiple federated aggregation methods across five heterogeneous breast cancer imaging configurations designed to better emulate realistic multi-institutional healthcare environments.

Table 1 highlights that most existing federated learning studies in breast cancer prediction and medical imaging primarily focus on predictive performance under relatively simplified federated settings. Many recent works evaluate only a single aggregation strategy or consider limited levels of statistical heterogeneity. In contrast, the proposed study systematically investigates the robustness and convergence behavior of multiple aggregation methods under progressively challenging and realistic healthcare scenarios involving severe class imbalance, non-IID multi-source distributions. This broader evaluation framework enables a more comprehensive analysis of aggregation stability and clinical reliability in decentralized medical imaging environments.

Table 1. Comparison of recent federated learning studies for medical imaging and breast cancer classification.

Study	Year	Application	FL Method	Heterogeneity Setting	Main Limitation
Sheller et al. [27]	2020	Brain tumor segmentation	FedAvg	Multi-institutional	Limited aggregation comparison and imbalance analysis
Li et al. [29]	2022	Histopathological breast cancer classification	FedAvg	Non-IID datasets	Limited heterogeneity scenarios
Almufareh et al. [16]	2023	Breast cancer prediction	Federated framework	Multi-hospital setup	Limited robustness analysis under severe non-IID conditions
Briola et al. [34]	2024	Explainable breast cancer FL	XAI-based FL	Moderate heterogeneity	Focused mainly on interpretability rather than aggregation robustness
Gupta et al. [17]	2025	Breast cancer pathology classification	FedAvg ensemble	Limited non-IID	No analysis of client-specific distortions or severe imbalance
Al-Hejri et al. [30]	2025	Mammogram classification	Hybrid ViT-CNN FL	Multi-class FL setting	Primarily architecture-focused with limited aggregation comparison
Proposed Work	2026	Breast cancer mammogram classification	FedAvg, FedProx, FedNova, FedDyn, SCAFFOLD	Severe imbalance, non-IID multi-source data, and client-specific distortions	Comprehensive comparative robustness analysis under realistic federated healthcare conditions

2.3. Key Federated Learning Algorithms for Medical Imaging

Several FL algorithms have been proposed to address challenges such as data heterogeneity, communication overhead, and convergence instability in medical imaging. McMahan et al. [13] introduced the FedAvg algorithm, which enables decentralized training through iterative local updates and global parameter averaging. Although FedAvg reduces communication costs and performs effectively in relatively balanced environments, it often suffers from client drift and unstable convergence under highly heterogeneous non-IID settings.

FedProx, introduced by Li et al. [19], extends FedAvg by adding a proximal regularization term that limits excessive divergence between local and global models. This improves optimization stability under heterogeneous and non-IID settings, where unrestricted local updates can negatively affect convergence. The method achieved up to 22% higher test accuracy than FedAvg in heterogeneous environments. Recent theoretical and optimization studies have further examined the role of regularization and control-inspired design principles in improving convergence stability in federated learning under heterogeneous and non-smooth settings [35,36].

To address objective inconsistency in heterogeneous federated learning, Li et al. [20] proposed FedNova, a normalized averaging method that mitigates convergence bias caused by variable local updates across clients. Recent federated optimization studies have further highlighted the importance of normalization and adaptive aggregation strategies for improving convergence stability in highly heterogeneous environments [37]. While the method provides theoretical improvements in federated optimization, its performance may still degrade under extreme imbalance and highly skewed distributions.

Acar et al. [21] introduced FedDyn, a dynamic regularization method that adjusts optimization per device and communication round to better align local and global objectives. Recent studies on adaptive federated optimization have shown that dynamic regulariza-

tion mechanisms can significantly improve convergence robustness under severe non-IID conditions and partial client participation [38]. This approach demonstrated robustness in heterogeneous and partially participating environments with both convex and non-convex objectives. In contrast, Karimireddy et al. [22] proposed SCAFFOLD, which explicitly addresses client drift through the use of control variates that reduce update variance across clients. The algorithm demonstrated faster and more stable convergence, particularly in highly heterogeneous settings.

These aggregation methods differ substantially in their mechanisms for handling statistical heterogeneity and client drift. FedAvg relies on simple parameter averaging and generally performs well under balanced IID conditions, whereas FedProx, FedDyn, FedNova, and SCAFFOLD introduce various regularization and correction strategies to improve convergence in heterogeneous federated environments. Despite these advances, relatively few studies provide comprehensive comparative analyses of these methods under realistic medical imaging scenarios involving simultaneous imbalance, non-IID distributions.

Unlike previous studies that primarily focus on single aggregation methods or simplified federated settings, this work systematically compares five widely used aggregation strategies under progressively complex and realistic breast cancer federated learning scenarios. The proposed framework incorporates varying levels of class imbalance, non-IID multi-source distributions, and client-specific image distortions to better emulate real-world healthcare environments. This framework enables a deeper analysis of aggregation robustness, stability, and clinical reliability under challenging decentralized medical imaging conditions.

3. Methodology

The work focuses on the development and assessment of a federated learning approach for the classification of mammogram images into benign and malignant categories. The methodology is organized around three core components: the use of public mammographic datasets, the implementation of a cross-silo federated learning framework, and the deployment of a lightweight deep learning model (MobileNetV2) across multiple clients. To simulate real-world hospital collaboration, experiments were conducted under various data distribution scenarios using different aggregation algorithms. The goal is to assess the robustness and accuracy of the global model in decentralized and privacy-preserving environments.

3.1. Dataset Description

3.1.1. Data Sources

This study uses multiple publicly available mammography datasets to simulate realistic federated learning environments for breast cancer classification. No private clinical data were used in this work. The selected datasets originate from different institutions and acquisition settings, allowing the experiments to reflect real-world heterogeneity commonly encountered in cross-silo federated learning scenarios.

The public datasets used include the following:

The INbreast database, known as the "Full-field Digital Mammography Database", provides high-quality annotated full-field digital mammographic images [39].

The Mammographic Image Analysis Society (MIAS) dataset contains digitized mammograms widely used in breast cancer detection and classification research [40].

The Digital Database for Screening Mammography (DDSM) includes scanned film mammograms with associated pathology information and lesion annotations [41].

Additionally, a CLAHE-enhanced version of the DDSM dataset was generated by applying the Contrast-Limited Adaptive Histogram Equalization (CLAHE) technique to improve image contrast and visibility of breast tissue structures [42].

Finally, the Radiological Society of North America (RSNA) Breast Cancer Dataset, released as part of the RSNA 2022 Breast Cancer Screening Challenge, was incorporated to further increase data diversity and simulate large-scale clinical screening environments [43].

The datasets differ in several aspects, including image acquisition devices, image quality, resolution, preprocessing characteristics, class distributions, and demographic variability. These differences introduce natural heterogeneity across clients and better reflect realistic healthcare federated learning settings where hospitals may rely on different imaging protocols and patient populations.

Across the five experimental configurations, clients were assigned different datasets, class imbalance ratios, and preprocessing pipelines to simulate IID, non-IID, imbalanced, and heterogeneous cross-silo environments. While all clients perform the same binary classification task (benign vs. malignant), the underlying data distributions and image characteristics vary significantly between clients in the heterogeneous configurations.

3.1.2. Preprocessing Steps

To ensure consistent input and facilitate model training, the following preprocessing operations were applied:

- Resizing: All images were resized to 224 × 224 pixels to match the input size required by CNN-based architectures like MobileNetV2.
- Channel conversion: Grayscale images were converted to RGB to match pretrained model expectations.
- Normalization: Pixel intensities were scaled to a [0, 1] range.
- Augmentation: Techniques such as rotation, flipping, zooming, and contrast adjustments were used to increase dataset diversity and model robustness.
- CLAHE (for contrast enhancement): Applied on selected datasets to enhance visibility of tissue structures [42].

3.2. Data and Client Distribution Configurations

The federated learning experiments were designed to assess model robustness across various simulated real-world data distribution scenarios, including IID and non-IID setups, balanced and imbalanced datasets, and varying image distortions. The datasets were configured into multiple experimental scenarios to simulate realistic challenges encountered in federated healthcare environments.

For reproducibility and fair comparison, the same data partitions, client assignments, and class distributions were maintained across all aggregation method experiments within each configuration. This ensured that performance differences observed between aggregation methods were attributable to the aggregation strategies themselves rather than variations in data partitioning.

After the client-level data distributions were defined for each experimental configuration, each local client dataset was further divided into training, validation, and testing subsets using a 70%-10%-20% split, respectively. The training subsets were used for local federated optimization, while the validation subsets were used to monitor convergence and reduce overfitting during local training. The testing subsets were reserved exclusively for evaluation purposes and were combined during global assessment to ensure consistent, reproducible, and fair comparison of aggregation method performance across all experimental configurations.

In the implemented federated learning setup, each client stores its own local mammogram image dataset organized into benign and malignant classes. Depending on the experimental configuration, the client distributions follow IID, imbalanced, or non-IID settings to simulate varying levels of heterogeneity across healthcare institutions. Dur-

ing training, data remain locally stored on each client, and only model parameters are exchanged with the central server during aggregation.

All images were resized to (224×224) pixels and converted into tensor format before training. Local training was performed using shuffled mini-batches with a batch size of 32 to improve convergence stability during federated optimization.

3.2.1. Configuration 1: IID and Class-Balanced Data Across Clients

Table 2 presents the distribution of DDSM mammograms enhanced using Contrast-Limited Adaptive Histogram Equalization (CLAHE) across the five federated learning clients, as shown in Figure 1. Each client—Client1 through Client5—has been assigned an equal number of image cases, totaling 1000 per client. These are evenly split between two classes: 500 benign cases and 500 malignant cases. Due to the small and balanced nature of the dataset, training was limited to 20 epochs to evaluate performance under ideal IID conditions. This balanced distribution ensures uniformity across clients, making it suitable for evaluating models under ideal conditions where data is independent and identically distributed (IID).

Table 2. Data Distribution Across 5 Clients—Configuration 1 (IID and Balanced).

Client	Total Cases	Benign	Malignant
Client1	1000	500	500
Client2	1000	500	500
Client3	1000	500	500
Client4	1000	500	500
Client5	1000	500	500

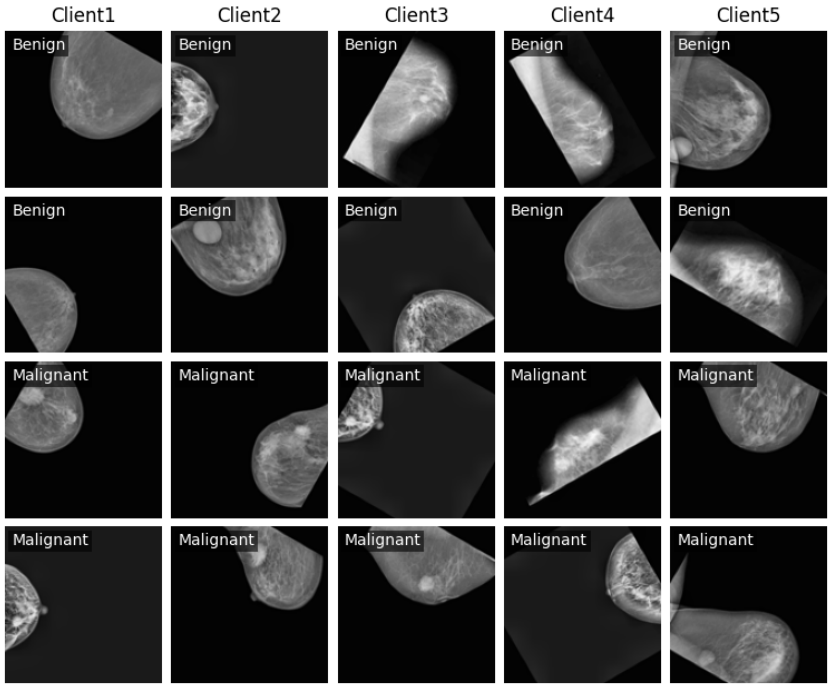

Figure 1. Sample Images from Configuration 1 (Balanced and IID).

3.2.2. Configuration 2: Imbalanced Data Skewed Toward the Benign Class

Table 3 illustrates an imbalanced distribution of CLAHE-enhanced medical image cases from the DDSM dataset among five clients in a federated learning setup (Figure 2). Each client possesses a different total number of cases, ranging from 1350 to 1500. Notably, the class distribution varies significantly across clients, introducing both imbalance and

non-IID (non-identically distributed) conditions. Client1 has only benign cases (1350) with no malignant cases, while Clients 2 through 5 have a progressively increasing number of malignant cases (from 300 to 750) and a corresponding decrease in benign cases (from 1200 to 750). This configuration reflects a more realistic and challenging federated learning scenario, where data heterogeneity among clients can impact model training and convergence. To account for the increased complexity introduced by class imbalance, training was extended to 100 epochs.

Table 3. Data Distribution Across 5 Clients—Configuration 2 (Class Imbalance).

Client	Total Cases	Benign	Malignant
Client1	1350	1350	0
Client2	1500	1200	300
Client3	1500	1050	450
Client4	1500	900	600
Client5	1500	750	750

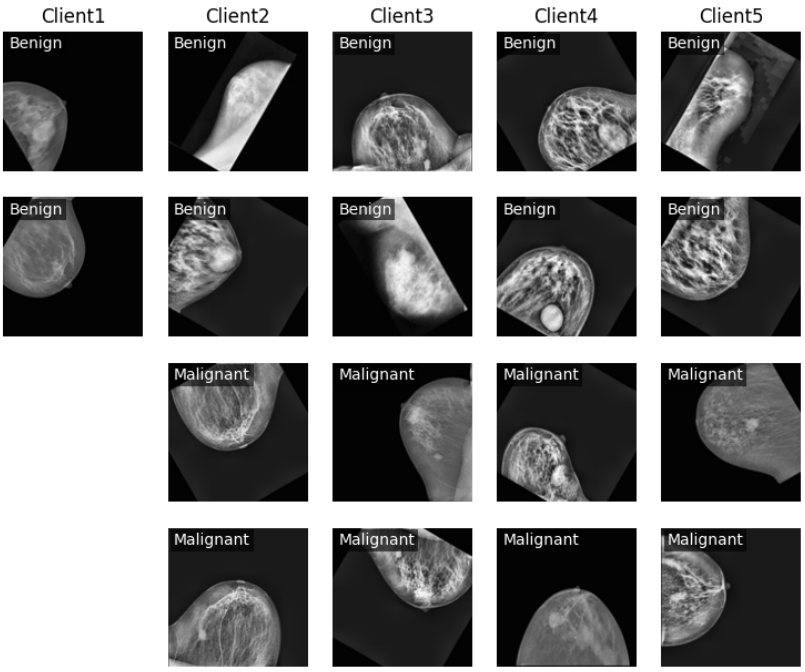

Figure 2. Sample Images from Configuration 2 (Imbalanced Distribution).

3.2.3. Configuration 3: Non-IID Multi-Source Data Across Clients with Varying Class Imbalances

Table 4 presents a heterogeneous setup for federated learning using medical images sourced from different datasets, DDSM, INBreast, MIAS, and the CLAHE-enhanced DDSM images, trained over 100 epochs, as illustrated in Figure 3. Each client is associated with a distinct dataset or a different subset of one. The number of total cases varies significantly among clients, ranging from 1050 to 4000. Class distribution is also uneven, indicating a strong imbalance and non-homogeneity. For example, Client1 (DDSM) has a dominant number of benign cases (3400) and fewer malignant cases (600), while Client2 (INBreast) shows the opposite, with 2700 malignant and only 300 benign cases. This configuration is designed to reflect realistic cross-silo federated learning conditions, where each client (e.g., hospital or center) uses a different data source with varying class distributions and volumes, posing challenges for model generalization and training consistency.

Table 4. Data Distribution Across 5 Clients—Configuration 3 (Non-IID by Source).

Dataset/Client	Total Cases	Benign	Malignant
Client1 (DDSM)	4000	3400	600
Client2 (INbreast)	3000	300	2700
Client3 (MIAS)	2400	1500	900
Client4 (DDSM)	1050	900	150
Client5 (CLAHE)	3150	2700	450

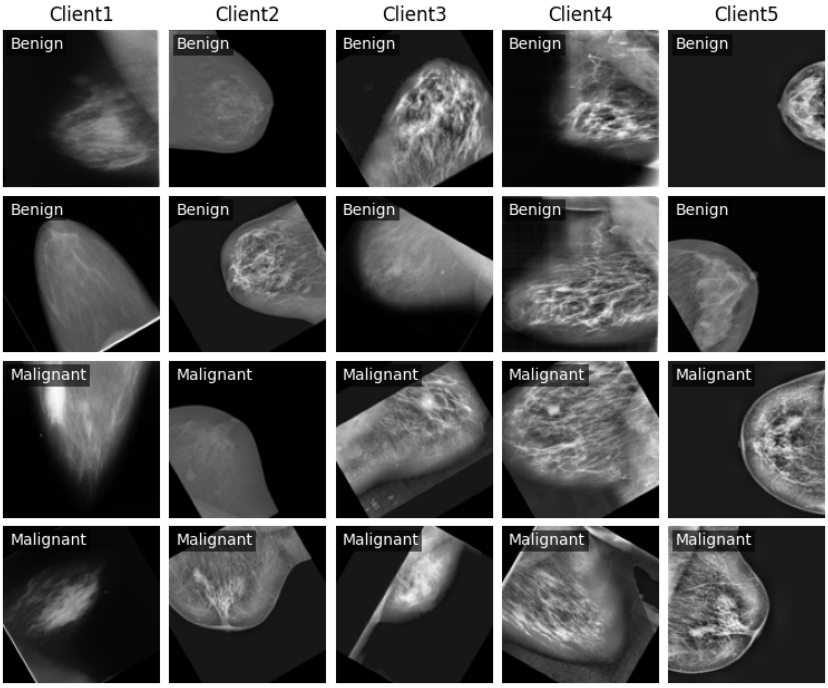

Figure 3. Sample Images from Configuration 3 (Non-IID Sources).

3.2.4. Configuration 4: Highly Imbalanced, Non-Homogeneous, and Non-IID Multi-Source Data Skewed Toward Benign Class

Table 5 displays a highly imbalanced and non-homogeneous data distribution across five clients, each using different datasets, DDSM, INBreast, RSNA, MIAS, and the CLAHE-enhanced DDSM images (Figure 4), trained over 100 epochs. The total number of image cases per client varies widely—from as few as 247 in Client4 (MIAS) to as many as 5470 in Client3 (RSNA). A stark class imbalance is evident across all clients, with benign cases overwhelmingly outnumbering malignant ones. For instance, Client3 (RSNA) has 5355 benign images and only 115 malignant (2.1%), while Client5 (INBreast) has a slightly higher malignancy rate of 11.58%. This configuration simulates a realistic and challenging federated learning environment where data disparity arises not only from varying class ratios but also from differences in dataset size and source, making it a rigorous test for model robustness and generalizability.

Table 5. Data Distribution Across 5 Clients—Configuration 4 (Combined Challenges).

Dataset/Client	Total Cases	Benign	Malignant (%)
Client1 (DDSM)	378	336	42 (11.1%)
Client2 (CLAHE)	1180	1086	94 (8.0%)
Client3 (RSNA)	5470	5355	115 (2.1%)
Client4 (MIAS)	247	238	9 (3.6%)
Client5 (INbreast)	285	252	33 (11.6%)

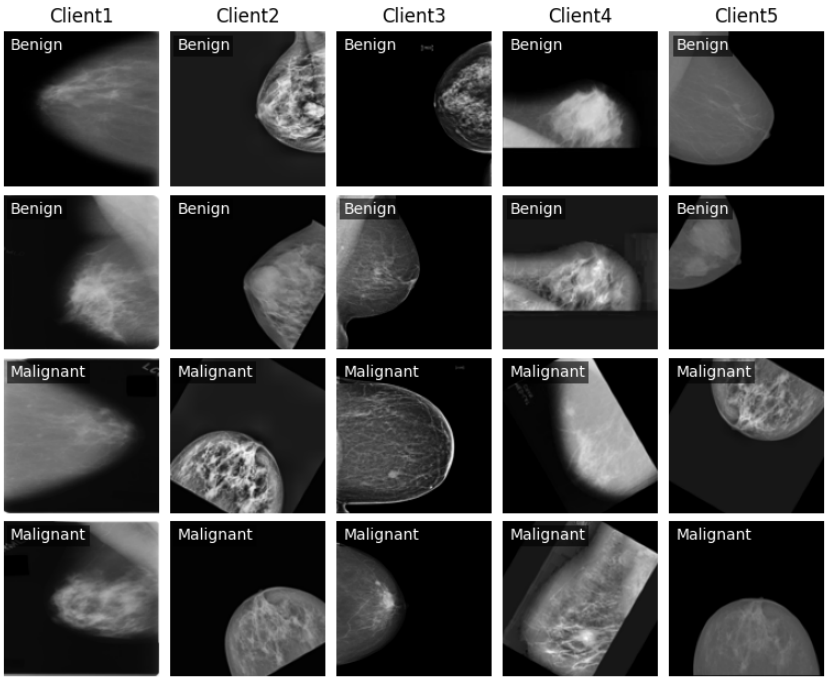

Figure 4. Sample Images from Configuration 4 (Combined Distribution).

3.2.5. Configuration 5: Non-IID Multi-Source Data with Client Specific Image Distortions

This configuration maintains the same data distribution as Configuration 4 but introduces additional image augmentation techniques to intensify the non-IID characteristics of the dataset. The goal is to simulate client-level data heterogeneity not only through varying dataset sources and class imbalances but also by applying distinct visual distortions to each client's images (Figure 5). This is achieved by customizing a unique transformation pipeline for each client, introducing effects such as noise, contrast variation, resizing, and color alterations. The transformation details per client are as follows:

1. Client 1:
    - RandomRotation (15°): Rotates the image randomly within ±15 degrees.
    - ColorJitter (brightness & contrast adjustment): Slightly changes brightness and contrast to mimic different lighting conditions.
2. Client 2:
    - Resize (200, 200): Resizes images to a slightly smaller dimension than default (224, 224).
    - RandomHorizontalFlip: Flips images horizontally with a 50% probability.
    - Random Gaussian Blur: Applies a Gaussian blur with a probability of 50%.
3. Client 3:
    - Resize (256, 256): Increases image size before training.
    - RandomVerticalFlip: Flips images vertically with a 50% probability.
    - RandomAffine (Shear transformation): Applies random shearing to distort the image.
4. Client 4:
    - RandomResizedCrop (224, 224, scale = (0.8, 1.0)): Crops a portion of the image randomly while maintaining aspect ratio.
    - AdjustSharpness (Factor = 2): Enhances sharpness to create variations in edge clarity.

5. Client 5:
    - ColorJitter (Saturation boost): Increases saturation in colors randomly with a 50% probability.
    - RandomGrayscale: Converts some images to grayscale with a probability of 30%.

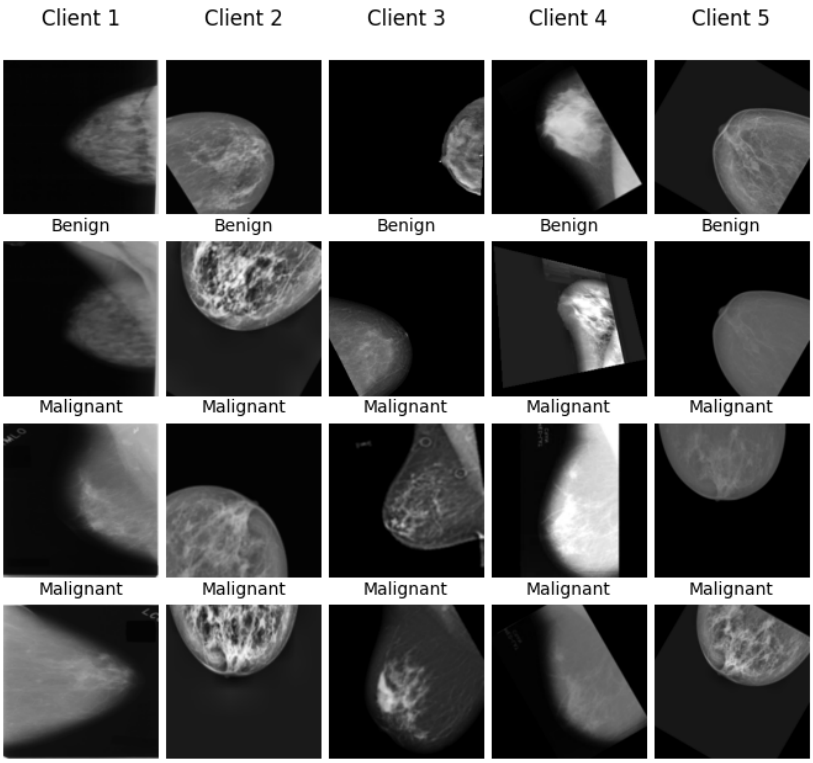

Figure 5. Sample Images from Configuration 5 (Image Distortion).

3.3. Federated Aggregation Methods

Five federated aggregation strategies were implemented to address the challenges of decentralized, heterogeneous, and potentially imbalanced datasets.

1. FedAvg

In the FedAvg algorithm, the global model weights are updated by computing a weighted average of the client models' weights based on the number of samples at each client [13]. Let w_g denote the global model parameters, w_i the parameters of client i, n_i the number of samples at client i, and $N = \sum_{i=1}^{K} n_i$ the total number of samples across all clients. Then, the global model is updated as:

$$w_g = \sum_{i=1}^{K} \frac{n_i}{N} w_i \tag{1}$$

where K is the total number of clients.

2. FedProx

FedProx extends FedAvg by adding a proximal term to the local client loss to limit deviation from the global model [19]. The local objective for client i is:

$$\mathcal{L}_i^{\text{FedProx}}(w_i) = \mathcal{L}_i(w_i) + \frac{\mu}{2} \|w_i - w_g\|^2 \tag{2}$$

where w_i are the local model parameters, w_g are the current global model parameters, and μ is a hyperparameter controlling the proximal term strength.

After local training, the global model is updated using a weighted average of client parameters:

$$w_g = \sum_{i=1}^{K} \frac{n_i}{N} w_i \tag{3}$$

where K is the total number of clients, n_i is the number of samples in client i, and $N = \sum_{i=1}^{K} n_i$.

3. FedNova

In the FedNova algorithm, the global model weights are updated using a weighted average of the client models' weights, where the weights are proportional to the number of local updates (gradient steps) performed on each client [20]. Let w_g denote the global model parameters, w_i the parameters of client i, u_i the number of local updates performed by client i, and $U = \sum_{i=1}^{K} u_i$ the total number of updates across all clients. Then, the global model is updated as:

$$w_g = \sum_{i=1}^{K} \frac{u_i}{U} w_i \tag{4}$$

where K is the total number of clients.

4. FedDyn

FedDyn extends FedAvg by adding a dynamic regularization term during client training to reduce the drift from the global model [21]. For client i, the local objective is augmented as:

$$\mathcal{L}_i^{\text{FedDyn}}(w_i) = \mathcal{L}_i(w_i) + \lambda \sum_j \|w_i^j - w_g^j\|^2 \tag{5}$$

where $\mathcal{L}_i(w_i)$ is the local empirical loss, w_i^j represents the parameters of client i at layer j, w_g^j are the corresponding global parameters, and λ is a hyperparameter controlling the strength of the regularization.

After local updates, the global model is aggregated using the same weighted averaging as FedAvg:

$$w_g = \sum_{i=1}^{K} \frac{n_i}{N} w_i \tag{6}$$

where n_i is the number of samples at client i, $N = \sum_{i=1}^{K} n_i$, and K is the number of clients.

5. SCAFFOLD

The SCAFFOLD algorithm addresses client drift in federated learning by introducing control variates that correct local updates [22]. Let w_g denote the global model parameters, w_i the local parameters of client i, c_i the control variate for client i, c_g the global control variate, and K the total number of clients. The global model is updated as follows:

$$w_g \leftarrow w_g + \frac{1}{K} \sum_{i=1}^{K} (w_i - w_g + c_i - c_g) \tag{7}$$

Here, the term $c_i - c_g$ acts as a correction factor to mitigate the effect of client-specific updates diverging from the global objective, which is particularly useful in heterogeneous (non-IID) data scenarios.

The selected aggregation methods differ significantly in their ability to handle heterogeneous and non-IID client distributions. FedAvg relies on simple weighted averaging of local model parameters, which performs well when client data distributions are relatively homogeneous. However, under highly non-IID conditions, local client updates may diverge substantially due to differences in class distributions, imaging sources, and local feature representations. This divergence can destabilize global convergence and bias the aggregated model toward dominant client distributions or majority classes. FedNova

partially addresses optimization inconsistency by normalizing client updates according to the number of local optimization steps. While this improves convergence stability under moderate heterogeneity, the method remains sensitive to severe distribution shifts because it does not explicitly constrain local model divergence from the global objective. In contrast, FedProx introduces a proximal regularization term that penalizes excessive deviation between local and global model parameters during training. This additional constraint helps reduce client drift, stabilizes optimization, and improves robustness under heterogeneous and imbalanced federated settings. Similarly, FedDyn and SCAFFOLD incorporate adaptive correction mechanisms designed to compensate for inconsistencies between local and global optimization objectives. FedDyn dynamically regularizes client updates to better align local and global gradients, whereas SCAFFOLD uses control variates to reduce update variance caused by client drift. These mechanisms are particularly important in federated medical imaging environments where data distributions may vary substantially across institutions.

4. Experimental Protocol

4.1. Federated Learning Setup

As part of a cross-silo federated learning approach, each of the five clients (hospitals) locally trains a MobileNetV2 model on its own private dataset and shares only the updated model weights with a central server. The server then aggregates these updates to produce a new global model, which is redistributed to all clients for the next training round. This setup ensures that no raw patient data is exchanged, thus preserving data privacy and compliance with healthcare regulations.

Model Architecture

The model architecture used for all experiments is MobileNetV2, a lightweight convolutional neural network designed for efficient computation while maintaining strong performance on image classification tasks. The model is based on depthwise separable convolutions, inverted residuals, and linear bottlenecks, which help reduce the number of parameters and improve inference speed [6]. Its efficiency makes it well-suited for deployment in federated learning environments where computational resources and communication capacity at client devices may be limited [13,18].

In addition to its computational efficiency, MobileNetV2 was selected because the primary objective of this study is to evaluate the behavior of different federated aggregation methods under varying data heterogeneity conditions while maintaining a consistent model architecture across all experiments. Using a single backbone architecture allows for a fair comparison between aggregation strategies by minimizing variability introduced by differences in model design. Furthermore, lightweight architectures such as MobileNetV2 are particularly relevant in healthcare federated learning scenarios, where participating institutions may have heterogeneous hardware capabilities and limited computational resources.

4.2. Training Strategy

Training was conducted locally using Anaconda Lab on a Windows 11 Pro machine equipped with an Intel Core i5-8350U CPU at 1.70 GHz and 16 GB RAM. The implementation relied on Python 3.11.7 and utilized libraries such as Torch, Torchvision, and EasyFL for managing federated learning experiments, as well as PIL (Pillow), Matplotlib 3.8.0, and Scikit-learn for data handling and evaluation.

4.2.1. Training Parameters

All clients trained the MobileNetV2 model using the Adam optimizer with a learning rate of 0.001 and a batch size of 32. The selected hyperparameter values were chosen based on commonly adopted settings in federated deep learning and medical image classification literature, while also considering convergence stability and computational efficiency in heterogeneous federated environments. The model was initialized using pretrained ImageNet weights to improve feature extraction capability and accelerate convergence during federated training. All input mammogram images were resized to (224×224) pixels to match the standard input dimensions required by the proposed architecture. Each client performed 100 local training epochs per communication round, except in Configuration 1, where training was limited to 20 epochs due to the smaller and more homogeneous IID dataset. The increased number of epochs in the remaining configurations was necessary to allow more stable convergence under imbalanced and highly heterogeneous non-IID conditions. The loss function used was cross-entropy loss, which is appropriate for binary classification tasks such as distinguishing benign from malignant mammogram images.

4.2.2. Evaluation Metrics

Model performance was evaluated using accuracy, loss, recall, specificity, precision, and F1-score. Accuracy reflects the overall correctness of predictions, while loss measures the discrepancy between predicted and true labels. Recall is especially critical in breast cancer prediction, as it represents the ability to correctly identify malignant cases, thereby minimizing false negatives, which is vital in a medical context where missing a positive diagnosis can delay treatment and have serious health consequences. Specificity measures the ability to correctly identify benign cases, reducing false positives. Precision evaluates the proportion of correctly predicted malignant cases among all positive predictions, reflecting the reliability of the model's positive outputs. The F1-score provides a balanced measure between precision and recall, making it particularly useful under imbalanced data distributions. Together, these metrics provide a comprehensive evaluation of model performance across different federated learning configurations.

5. Results and Discussion

This section presents and analyzes the results obtained from our federated learning experiments using the MobileNetV2 architecture for breast cancer classification. We report the performance of the global model under each data configuration using five aggregation methods: FedAvg, FedProx, FedNova, FedDyn and SCAFFOLD. A comparative analysis is provided to highlight how each method performs in terms of loss, accuracy, recall, specificity, precision, and F1-score across the different data distributions. The results demonstrate the impact of data heterogeneity on model performance and emphasize the effectiveness of specific aggregation strategies in addressing the challenges of federated medical image classification.

5.1. Aggregation Methods Comparison

5.1.1. Configuration 1: Results of IID and Class-Balanced Setting

The results in Table 6 show clear differences in performance among the aggregation methods under the balanced and homogeneous IID configuration. FedAvg achieved the best overall performance, obtaining the lowest loss, highest accuracy, and highest recall (0.994), demonstrating excellent capability in correctly detecting malignant cases. It also maintained high precision, specificity, and F1-score values, confirming balanced classification performance across both classes. This strong performance is expected since FedAvg is particularly effective in IID environments due to its simple averaging strategy.

Table 6. Performance of the model across different clients (Configuration 1).

Aggregation Method	Loss	Accuracy (%)	Recall	Specificity	Precision	F1-Score
FedAvg	0.019	99.32	0.994	0.992	0.992	0.993
FedProx	0.052	98.00	0.968	0.992	0.991	0.980
FedNova	0.043	98.36	0.992	0.975	0.975	0.984
FedDyn	0.042	98.44	0.992	0.977	0.977	0.985
SCAFFOLD	0.042	99.28	0.990	0.996	0.996	0.993

SCAFFOLD and FedDyn also produced strong results, with high recall and competitive accuracy values. SCAFFOLD achieved the highest specificity and precision, indicating fewer false positive predictions. In contrast, FedProx and FedNova showed the lowest recall and F1-score, suggesting weaker sensitivity in detecting malignant samples, as their regularization and normalization strategies—designed for non-IID data—can unnecessarily restrict model updates when data is already well-distributed. Overall, the results indicate that simpler aggregation methods such as FedAvg are more suitable for balanced federated settings.

The confusion matrix in Table 7 supports these findings by illustrating how each method classified benign and malignant samples. FedAvg stands out with the fewest misclassifications in both classes, aligning with its high accuracy. SCAFFOLD and FedDyn also maintain strong classification performance, with slightly higher but still low error rates. In contrast, FedProx and FedNova show a notable increase in false positives for benign cases, reflecting the unnecessary constraints their mechanisms introduce in an already balanced setup. These results further confirm that methods tailored for non-IID data may underperform in homogeneous settings.

Table 7. Confusion matrix for different aggregation methods (Configuration 1).

Label		Predicted Label		Aggregation Method
		Benign	Malignant	
True Label	Benign	2481	19	FedAvg
		2479	21	FedProx
		2437	63	FedNova
		2442	58	FedDyn
		2490	10	SCAFFOLD
	Malignant	15	2485	FedAvg
		79	2421	FedProx
		19	2481	FedNova
		20	2480	FedDyn
		26	2474	SCAFFOLD

Note: Each row shows the performance of a specific aggregation method in terms of true/false positives and negatives.

5.1.2. Configuration 2: Results of Imbalanced Data Setting

In the second configuration (Table 8), where the data is imbalanced toward the benign class, the aggregation methods show more noticeable differences in performance. FedAvg achieved the highest accuracy (99.22%), highest specificity (0.9897), precision (0.9750), and F1-score (0.9866), while also maintaining a very high recall (0.9986). This indicates that FedAvg handled both classes effectively despite the imbalance. In contrast, FedProx, FedNova, FedDyn, and SCAFFOLD achieved perfect or near-perfect recall ($\geq$0.999), demonstrating excellent sensitivity to malignant cases, which is particularly important in medical

diagnosis. However, these methods exhibited lower specificity and precision, indicating a higher number of false positives caused by the skewed class distribution. Among them, SCAFFOLD provided a better compromise, achieving perfect recall (1.000) with improved specificity (0.8783) and F1-score (0.8681) compared to FedProx and FedDyn. Overall, the results highlight the trade-off between maximizing minority class detection and maintaining balanced overall classification performance under imbalanced data conditions.

Table 8. Performance of the model across different clients (Configuration 2).

Aggregation Method	Loss	Accuracy (%)	Recall	Specificity	Precision	F1-Score
FedAvg	0.024	99.224	0.9986	0.9897	0.9750	0.9866
FedProx	0.454	91.156	1.0000	0.8762	0.7636	0.8660
FedNova	0.331	92.871	0.9995	0.9004	0.8005	0.8890
FedDyn	0.407	89.755	0.9995	0.8568	0.7365	0.8481
SCAFFOLD	0.181	91.306	1.0000	0.8783	0.7666	0.8681

As shown in Table 9, class-wise predictions reveal how different methods behave under imbalance. FedAvg maintains strong overall performance with minimal errors across both classes, reflecting its accuracy advantage. However, FedProx, FedNova, and FedDyn show a clear bias toward detecting malignant cases—correctly classifying almost all of them—while mislabeling a large number of benign samples. This shift suggests a prioritization of sensitivity over specificity. SCAFFOLD stands out by achieving perfect detection of malignant cases while keeping false positives relatively lower than the other specialized methods, striking a better balance between the two classes in imbalanced conditions.

Table 9. Confusion matrix for different aggregation methods (Configuration 2).

Label		Predicted Label		Aggregation Method
		Benign	Malignant	
True Label	Benign	5196	54	FedAvg
		4600	650	FedProx
		4727	523	FedNova
		4498	752	FedDyn
		4611	639	SCAFFOLD
	Malignant	3	2097	FedAvg
		0	2100	FedProx
		1	2099	FedNova
		1	2099	FedDyn
		0	2100	SCAFFOLD

5.1.3. Configuration 3: Results of Non-IID Multi-Source, Imbalanced Data Setting

In the third configuration (Table 10), where the data is both non-IID and imbalanced across multiple sources, the aggregation methods exhibit substantially different behaviors due to the increased divergence between local client updates. FedAvg and FedNova perform poorly, with accuracies around 65% and extremely low recall values (0.009 and 0.030), indicating a near-complete failure to detect malignant cases despite achieving perfect precision and specificity. Under severe heterogeneity, local updates become biased toward dominant benign distributions, causing unstable global convergence and poor minority-class generalization. Although FedNova introduces normalization to stabilize optimization, it remains highly sensitive to extreme distribution shifts.

Table 10. Performance of the model across different clients (Configuration 3).

Aggregation Method	Loss	Accuracy (%)	Recall	Specificity	Precision	F1-Score
FedAvg	2.384	65.022	0.009	1.000	1.000	0.017
FedProx	0.298	86.618	0.773	0.917	0.836	0.803
FedNova	2.189	65.764	0.030	1.000	1.000	0.058
FedDyn	0.309	84.846	0.591	0.989	0.966	0.733
SCAFFOLD	0.480	75.434	0.623	0.827	0.661	0.642

In contrast, FedProx significantly outperforms the other methods, achieving 86.62% accuracy, 0.773 recall, 0.836 precision, and the highest F1-score (0.803). Its proximal regularization term limits excessive deviation between local and global models, reducing client drift and improving optimization stability under heterogeneous conditions. FedDyn also demonstrates strong robustness, achieving very high precision (0.966) and specificity (0.989), although its lower recall (0.591) indicates that some malignant cases are still missed. SCAFFOLD achieves a more balanced behavior with recall of 0.623, but its lower specificity (0.827) and precision (0.661) indicate a higher false-positive rate. Overall, these results highlight the importance of aggregation strategies specifically designed to handle non-IID federated environments while maintaining a balance between sensitivity and classification reliability.

The confusion matrix in Table 11 further highlights the limitations of certain aggregation methods under highly heterogeneous conditions. FedAvg and FedNova exhibit a strong bias toward the benign class, correctly classifying nearly all benign samples while misclassifying most malignant cases. This behavior suggests that the global model becomes dominated by majority-class gradients, leading to poor minority-class sensitivity. In contrast, FedProx achieves a more balanced distribution of predictions across both classes, reflecting its improved ability to reduce client drift and stabilize optimization. FedDyn and SCAFFOLD provide partial improvements but continue to misclassify a substantial number of malignant samples, indicating that their adaptive correction mechanisms remain insufficient under extreme imbalance and multi-source heterogeneity.

Table 11. Confusion matrix for different aggregation methods (Configuration 3).

Label		Predicted Label		Aggregation Method
		Benign	Malignant	
True Label	Benign	8800	0	FedAvg
		8071	729	FedProx
		8800	0	FedNova
		8701	99	FedDyn
		7267	1533	SCAFFOLD
	Malignant	4757	43	FedAvg
		1091	3709	FedProx
		4656	144	FedNova
		1962	2838	FedDyn
		1808	2992	SCAFFOLD

5.1.4. Configuration 4: Results of Highly Imbalanced, Non-Homogeneous, and Non-IID Multi-Source Data Setting

In the 4th configuration (Table 12), where client data is non-IID and heavily skewed toward benign cases, aggregation methods show different abilities to detect the underrepresented malignant class. Although FedAvg and FedNova achieve relatively high

accuracy (around 97.6%), their low recall (0.392 for both methods) indicates poor sensitivity to malignant cases and a strong bias toward the dominant benign class. Despite their high specificity (0.999) and precision (0.966), many malignant samples remain undetected, resulting in lower F1-scores (0.558). In breast cancer diagnosis, such low recall values are problematic because they correspond to a high number of false negatives, meaning malignant cases may remain undetected. Missed cancer diagnoses can delay treatment and negatively affect patient outcomes, making accuracy alone insufficient for evaluating clinical reliability in highly imbalanced datasets. FedProx improves recall to 0.567 with a higher F1-score (0.712), suggesting better handling of imbalance and heterogeneity. However, FedDyn and SCAFFOLD achieve the best overall performance, combining high accuracy with significantly better recall values (0.823 and 0.700, respectively). FedDyn obtains the highest F1-score (0.900), while SCAFFOLD achieves perfect precision and specificity (1.000). These results indicate that advanced aggregation methods such as FedDyn and SCAFFOLD are more reliable in highly imbalanced federated settings, providing a better balance between detecting malignant cases and maintaining overall classification performance.

Table 12. Performance of the model across different clients (Configuration 4).

Aggregation Method	Loss	Accuracy (%)	Recall	Specificity	Precision	F1-Score
FedAvg	0.324	97.593	0.392	0.9990	0.966	0.558
FedProx	0.078	98.228	0.567	0.9990	0.960	0.712
FedNova	0.298	97.646	0.392	0.9990	0.966	0.558
FedDyn	0.049	99.286	0.823	0.9997	0.992	0.900
SCAFFOLD	0.053	98.836	0.700	1.0000	1.000	0.823

The confusion matrix in Table 13 further confirms these observations. While FedAvg and FedNova classify benign samples almost perfectly, they misclassify a large proportion of malignant cases as benign (178 out of 293), indicating poor minority-class sensitivity. This behavior reflects the strong influence of majority-class gradients during aggregation under highly imbalanced non-IID conditions. FedProx provides a moderate reduction in false negatives, demonstrating improved robustness to client heterogeneity. In contrast, FedDyn and SCAFFOLD maintain strong benign classification performance while substantially reducing malignant misclassifications, highlighting their improved ability to manage client drift and preserve minority-class detection under skewed federated distributions.

Table 13. Confusion matrix for different aggregation methods (Configuration 4).

Label		Predicted Label		Aggregation Method
		Benign	Malignant	
True Label	Benign	7263	4	FedAvg
		7260	7	FedProx
		7263	4	FedNova
		7265	2	FedDyn
		7267	0	SCAFFOLD
	Malignant	178	115	FedAvg
		127	166	FedProx
		178	115	FedNova
		52	241	FedDyn
		88	205	SCAFFOLD

5.1.5. Configuration 5: Results of Non-IID Multi-Source Data with Client Specific Image Distortions

In this most challenging configuration (Table 14), where the data is highly imbalanced, non-IID, and affected by client-specific image distortions, most aggregation methods exhibit severe degradation in malignant case detection. FedAvg, FedNova, and SCAFFOLD achieve zero recall, precision, and F1-score despite maintaining overall accuracies around 96% and perfect specificity. This behavior indicates that the models converge toward predicting nearly all samples as benign, completely failing to identify malignant cases. The high specificity values therefore reflect a strong bias toward the dominant benign class rather than robust classification performance. From a clinical perspective, these results are highly concerning because all malignant samples are misclassified as benign, leading to an unacceptable number of false negatives. Such behavior reinforces that accuracy alone can be misleading in highly imbalanced medical datasets and emphasizes the importance of recall and F1-score when evaluating clinical reliability in federated healthcare systems. The collapse in recall can be attributed to the combined effects of severe class imbalance, heterogeneous client distributions, and client-specific distortions, which collectively destabilize the aggregation process. Under these conditions, local updates become dominated by benign-class gradients, while conflicting client objectives increase gradient inconsistency and client drift across the federation. Aggregation methods such as FedAvg and FedNova, which rely heavily on direct parameter averaging, struggle to reconcile these divergent updates, causing the global model to converge toward majority-class predictions and fail to learn discriminative malignant representations. FedProx clearly outperforms the other methods, achieving the highest accuracy (98.47%), recall (0.604), and F1-score (0.753), while maintaining perfect specificity and precision. Its proximal regularization mechanism constrains excessive local model divergence, thereby improving optimization stability under highly heterogeneous and distorted federated environments. FedDyn provides only marginal improvement, suggesting that dynamic regularization alone is insufficient to fully compensate for extreme imbalance and feature inconsistency across clients.

Table 14. Performance of the model across different clients (Configuration 5).

Aggregation Method	Loss	Accuracy (%)	Recall	Specificity	Precision	F1-Score
FedAvg	0.166	96.124	0.000	1.000	0.000	0.000
FedProx	0.098	98.466	0.604	1.000	1.000	0.753
FedNova	0.172	96.124	0.000	1.000	0.000	0.000
FedDyn	0.118	96.217	0.024	1.000	1.000	0.047
SCAFFOLD	0.151	96.124	0.000	1.000	0.000	0.000

As shown in Table 15, the combined effects of class imbalance, non-IID client distributions, and image distortions are clearly reflected in the classification outcomes. While all methods classify benign samples almost perfectly, FedProx is the only aggregation method capable of correctly identifying a substantial proportion of malignant cases. FedDyn detects only a limited number of malignant samples, whereas FedAvg, FedNova, and SCAFFOLD completely collapse toward majority-class predictions, failing to recognize any malignant instances. This behavior suggests that under extreme heterogeneity, the optimization process becomes dominated by benign-class updates, reducing the model's ability to learn robust minority-class representations. Furthermore, client-specific distortions introduce additional feature inconsistency across local datasets, increasing gradient divergence and destabilizing global convergence. These findings demonstrate the limitations of conventional aggre-

gation strategies in highly heterogeneous federated medical imaging environments and further highlight the importance of robust aggregation mechanisms for clinically reliable minority-class detection.

Table 15. Confusion matrix for different aggregation methods (Configuration 5).

Label		Predicted Label		Aggregation Method
		Benign	Malignant	
True Label	Benign	7267	0	FedAvg
		7267	0	FedProx
		7267	0	FedNova
		7267	0	FedDyn
		7267	0	SCAFFOLD
	Malignant	293	0	FedAvg
		116	177	FedProx
		293	0	FedNova
		286	7	FedDyn
		293	0	SCAFFOLD

5.2. Performance Summary Across All Configurations

Across all configurations, the performance of aggregation methods varied significantly depending on the level of data imbalance, heterogeneity, and distortion. In the first balanced setup, all methods performed well, with FedAvg and SCAFFOLD slightly outperforming the others in terms of accuracy, specificity, precision and F1-score. As class imbalance was introduced, methods such as FedProx, FedDyn, and SCAFFOLD demonstrated improved sensitivity toward the minority malignant class by achieving higher recall values, although this sometimes occurred at the expense of overall accuracy. In highly non-IID and imbalanced scenarios, FedAvg and FedNova experienced substantial performance degradation, characterized by poor recall and limited generalization capability. This behavior is primarily attributed to the divergence of local client updates under heterogeneous data distributions, which destabilizes global model aggregation. In contrast, FedProx remained more robust due to its proximal regularization mechanism, which constrains excessive local model deviations and mitigates client drift during training. FedDyn and SCAFFOLD also improved robustness under moderately heterogeneous settings through adaptive correction mechanisms, although their performance became less stable under extreme imbalance combined with image distortions. Overall, FedProx emerged as the most stable and resilient aggregation method across the evaluated real-world-like federated settings involving client diversity and data irregularities.

An important observation across all experiments is that high classification accuracy does not necessarily imply clinically reliable performance in federated medical imaging systems. In several highly imbalanced configurations, some aggregation methods achieved strong overall accuracy while exhibiting extremely low or even zero recall for malignant cases. This behavior indicates that the optimization process became dominated by majority-class gradients, causing the global models to converge toward benign-biased decision boundaries that inadequately captured malignant patterns. In clinical practice, such behavior is particularly dangerous because false negatives correspond to missed cancer diagnoses, potentially delaying treatment and negatively affecting patient outcomes. These findings are particularly relevant in decentralized healthcare systems where hospitals may possess highly heterogeneous imaging distributions acquired using different scanners, acquisition protocols, and patient populations. Therefore, recall and sensitivity should be considered essential evaluation metrics in medical diagnosis tasks, especially in breast cancer screen-

ing applications where minimizing false negatives is critical for patient safety and timely clinical intervention.

Although this study primarily focused on classification robustness under heterogeneous federated settings, computational efficiency remains an important practical consideration in federated learning systems. FedAvg maintains low computational complexity and communication overhead due to its simple parameter averaging strategy, whereas FedProx introduces additional proximal regularization computations to stabilize local updates under heterogeneous conditions. Similarly, FedDyn and SCAFFOLD rely on adaptive correction and synchronization mechanisms that may increase computational and communication costs despite improving robustness against client drift and distribution shifts. MobileNetV2 was selected partly because of its lightweight architecture and reduced parameter count, which help limit communication overhead and computational burden in distributed federated environments.

5.3. Best Performing Aggregation Methods

Based on the experimental results across all configurations, FedProx stands out as the most consistently effective aggregation method, achieving recall values between 0.567 and 1.000 across configurations, and maintaining accuracy as high as 98.47% even under extreme distortions. It maintained strong performance even in the most challenging scenarios involving non-IID, imbalanced data, and image distortions, showing higher recall and accuracy compared to other methods. SCAFFOLD and FedDyn also delivered competitive results in moderately heterogeneous setups, particularly in maintaining a balance between accuracy and recall. In contrast, FedAvg and FedNova performed well only in balanced scenarios but failed under more complex distributions. Overall, FedProx proves to be the most robust and adaptable, making it the best choice for real-world federated learning environments with diverse and imbalanced client data.

6. Conclusions

In this work, we explored the impact of various data distribution configurations on federated learning performance in a binary image classification task. Five distinct configurations were designed to reflect increasingly realistic and complex federated settings: from a balanced and homogeneous dataset (Configuration 1), to imbalanced (Configuration 2), non-homogeneous with diverse data sources (Configuration 3), and finally, a highly non-IID and heterogeneous scenario with client-specific visual transformations (Configuration 4 and Configuration 5). As a result, FedProx is identified as the most effective aggregation method, achieving up to 98.47% accuracy and high recall under extreme conditions. Its proximal term penalizes deviations from the global model. By doing that, it reduces drift, stabilizes updates, and ensures more consistent aggregations, outperforming other methods in non-IID and imbalanced scenarios. SCAFFOLD and FedDyn showed competitive results in moderately heterogeneous environments, while FedAvg excelled in balanced conditions. By simulating diverse real-world challenges—such as varying dataset sizes, class imbalance, and visual inconsistencies—this study establishes a strong foundation for assessing the adaptability of federated models in decentralized medical image analysis.

Despite the promising results obtained in this study, several limitations of the current experimental setup should be acknowledged. The experiments were conducted in a simulated cross-silo federated learning environment involving five clients and limited computational resources. Although this setup enables controlled evaluation of aggregation methods under heterogeneous medical imaging conditions, it does not fully capture the complexity of large-scale real-world federated healthcare systems involving numerous distributed institutions. In particular, scalability-related challenges such as communication

latency, synchronization overhead, bandwidth constraints, asynchronous client participation, and large-scale distributed training were not evaluated in the current framework. Future research may therefore examine hyperparameter sensitivity, varying client participation rates, and dataset-specific aggregation behavior to better understand the stability and scalability of federated learning systems under realistic healthcare environments. Furthermore, the use of the lightweight MobileNetV2 architecture was partly motivated by computational efficiency considerations within the available hardware environment. Future work will therefore focus on extending the proposed framework toward larger-scale federated environments, exploring more communication-efficient distributed learning strategies, and investigating additional CNN architectures and personalized federated learning approaches to further evaluate the generalizability, robustness, and consistency of the obtained results across diverse classification models. Moreover, integrating multimodal data by combining mammographic images with clinical information could improve diagnostic relevance and decision-making capabilities in breast cancer prediction systems. Although federated learning improves privacy by keeping sensitive medical data localized at each client, the current framework does not integrate advanced privacy-preserving mechanisms such as secure aggregation, differential privacy, or encryption-based protections. Consequently, exchanged model updates may still be vulnerable to potential information leakage or gradient-based inference attacks under adversarial settings. Addressing these security limitations remains an important challenge for real-world clinical deployment. In addition, the integration of Explainable Artificial Intelligence (XAI) techniques within federated learning frameworks represents a promising direction for improving model interpretability, transparency, and clinician trust in decentralized breast cancer diagnosis systems.

Author Contributions: Conceptualization, N.S.L., M.M. and S.M.; methodology, N.S.L., M.M. and S.M.; software, N.S.L.; validation, N.S.L., M.M. and S.M.; formal analysis, N.S.L.; investigation, N.S.L., M.M. and S.M.; resources, N.S.L.; data curation, N.S.L.; writing—original draft preparation, N.S.L.; writing—review and editing, N.S.L., M.M. and S.M.; visualization, N.S.L.; supervision, M.M. and S.M.; project administration, M.M. and S.M.; funding acquisition, M.M. and S.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted using publicly available and fully anonymized datasets for breast cancer image classification. All these datasets have relevant ethical statements. No direct involvement of human participants was performed, and no identifiable personal data were used. Institutional Review Board approval is not required according to the Declaration of Helsinki (World Medical Association, 2013 revision).

Informed Consent Statement: Informed consent was not required for this study, as all data used are anonymized and publicly accessible.

Data Availability Statement: All data sources used in this study are publicly available and have been cited within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FL	Federated Learning
CNN	Convolutional Neural Network
AI	Artificial Intelligence
XAI	Explainable Artificial Intelligence
AUC	Area Under the Curve

HIPAA	Health Insurance Portability and Accountability Act
GDPR	General Data Protection Regulation
CLAHE	Contrast-Limited Adaptive Histogram Equalization
RSNA	Radiological Society of North America
DDSM	Digital Database for Screening Mammography
MIAS	Mammographic Image Analysis Society
INbreast	Full-field Digital Mammography Database
FedAvg	Federated Averaging
FedProx	Federated Proximal
FedNova	Federated Normalized Averaging
FedDyn	Federated Dynamics
SCAFFOLD	Stochastic Controlled Averaging for Federated Learning
IID	Independent and Identically Distributed
Non-IID	Non-Independent and Identically Distributed

References

1. World Health Organization. Breast Cancer: Prevention and Control. 2023. Available online: https://www.who.int/ (accessed on 30 June 2025).
2. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
4. Litjens, G.; Kooi, T.; Bejnordi, B.E.; Setio, A.A.A.; Ciompi, F.; Ghafoorian, M.; van der Laak, J.A.W.M.; van Ginneken, B.; Sánchez, C.I. A survey on deep learning in medical image analysis. *Med. Image Anal.* **2017**, *42*, 60–88. [CrossRef] [PubMed]
5. Esteva, A.; Kuprel, B.; Novoa, R.A.; Ko, J.; Swetter, S.M.; Blau, H.M.; Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **2017**, *542*, 115–118. [CrossRef] [PubMed]
6. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520. [CrossRef]
7. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *npj Digit. Med.* **2020**, *3*, 119. [CrossRef]
8. Health Insurance Portability and Accountability Act (HIPAA). U.S. Federal Law 1996. 1996. Available online: https://www.hhs.gov/hipaa/index.html (accessed on 30 June 2025).
9. General Data Protection Regulation (GDPR). EU Regulation 2016/679. 2016. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 30 June 2025).
10. Lundervold, A.S.; Lundervold, A. An overview of deep learning in medical imaging focusing on MRI. *Z. Med. Phys.* **2019**, *29*, 102–127. [CrossRef]
11. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241. [CrossRef]
12. Sheller, M.J.; Edwards, B.; Reina, G.A.; Martin, J.; Bakas, S. Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Sci. Rep.* **2020**, *10*, 12598. [CrossRef]
13. McMahan, H.B.; Moore, E.; Ramage, D.; Hampson, S.; y Arcas, B.A. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), Lauderdale, FL, USA, 20–22 April 2017.
14. Yang, Q.; Liu, Y.; Chen, T.; Tong, Y. Federated Machine Learning: Concept and Applications. *ACM Trans. Intell. Syst. Technol.* **2019**, *10*, 12. [CrossRef]
15. Kaissis, G.A.; Makowski, M.R.; Rückert, D.; Braren, R.F. Secure, privacy-preserving and federated machine learning in medical imaging. *Nat. Mach. Intell.* **2020**, *2*, 305–311. [CrossRef]
16. Almufareh, M.F.; Tariq, N.; Humayun, M.; Almas, B. A Federated Learning Approach to Breast Cancer Prediction in a Collaborative Learning Framework. *Healthcare* **2023**, *11*, 3185. [CrossRef]
17. Gupta, C.; Gill, N.S.; Gulia, P.; Alduaiji, N.; Shreyas, J.; Shukla, P.K. Applying YOLOv6 as an ensemble federated learning framework to classify breast cancer pathology images. *Sci. Rep.* **2025**, *15*, 12345. [CrossRef]

18. Kairouz, P.; McMahan, H.B.; Avent, B.; Bellet, A.; Bennis, M.; Nitin Bhagoji, A.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. Advances and Open Problems in Federated Learning. *Found. Trends. Mach. Learn.* **2021**, *14*, 1–210. [CrossRef]
19. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *Proc. Mach. Learn. Syst.* **2020**, *2*, 429–450.
20. Wang, J.; Liu, Q.; Liang, H.; Joshi, G.; Poor, H.V. Tackling the objective inconsistency problem in heterogeneous federated optimization. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20), Red Hook, NY, USA, 6–12 December 2020.
21. Acar, D.A.E.; Zhao, Y.; Matas, R.; Mattina, M.; Whatmough, P.N.; Saligrama, V. Federated Learning Based on Dynamic Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
22. Karimireddy, S.P.; Kale, S.; Mohri, M.; Reddi, S.J.; Stich, S.U.; Suresh, A.T. SCAFFOLD: Stochastic controlled averaging for federated learning. In Proceedings of the 37th International Conference on Machine Learning (ICML '20), Virtual Event, 13–18 July 2020.
23. Spanhol, F.A.; Oliveira, L.S.; Petitjean, C.; Heutte, L. A Dataset for Breast Cancer Histopathological Image Classification. *IEEE Trans. Biomed. Eng.* **2016**, *63*, 1455–1462. [CrossRef]
24. Ragab, D.A.; Attallah, O.; Sharkas, M.; Ren, J.; Marshall, S. A framework for breast cancer classification using Multi-DCNNs. *Comput. Biol. Med.* **2021**, *131*, 104245. [CrossRef]
25. Arevalo, J.; González, F.A.; Ramos-Pollán, R.; Oliveira, J.L.; Guevara Lopez, M.A. Representation learning for mammography mass lesion classification with convolutional neural networks. *Comput. Methods Programs Biomed.* **2016**, *127*, 248–257. [CrossRef]
26. Castro, D.; Walker, I.; Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **2020**, *11*, 3673. [CrossRef] [PubMed]
27. Sheller, M.J.; Reina, G.A.; Edwards, B.; Martin, J.; Bakas, S. Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. In *International MICCAI Brainlesion Workshop*; Springer: Cham, Switzerland, 2019; Volume 11383, pp. 92–104. [CrossRef]
28. Li, X.; Gu, Y.; Dvornek, N.; Staib, L.H.; Ventola, P.; Duncan, J.S. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. *Med. Image Anal.* **2020**, *65*, 101765. [CrossRef]
29. Li, L.; Xie, N.; Yuan, S. A Federated Learning Framework for Breast Cancer Histopathological Image Classification. *Electronics* **2022**, *11*, 3767. [CrossRef]
30. Al-Hejri, A.M.; Sable, A.H.; Al-Tam, R.M.; Al-antari, M.A.; Alshamrani, S.S.; Alshmrany, K.M.; Alatebi, W. A hybrid explainable federated-based vision transformer framework for breast cancer prediction via risk factors. *Sci. Rep.* **2025**, *15*, 18453. [CrossRef] [PubMed]
31. Alhussan, A.A.; Nhidi, W.; Filali, I.; Benhmida, F.; Ejbali, R. Federated Learning Architecture for 3D Breast Cancer Image Classification. *Cancers* **2025**, *17*, 3450. [CrossRef]
32. Jiménez-Sánchez, A.; Tardy, M.; González Ballester, M.A.; Mateus, D.; Piella, G. Memory-aware curriculum federated learning for breast cancer classification. *Comput. Methods Programs Biomed.* **2022**, *226*, 107318. [CrossRef]
33. Guan, H.; Yap, P.-T.; Bozoki, A.; Liu, M. Federated Learning for Medical Image Analysis: A Survey. *Pattern Recognit.* **2024**, *151*, 110424. [CrossRef] [PubMed]
34. Briola, E.; Nikolaidis, C.C.; Perifanis, V.; Pavlidis, N.; Efraimidis, P. A Federated Explainable AI Model for Breast Cancer Classification. In Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference (EICC '24), New York, NY, USA, 5–6 June 2024; pp. 194–201. [CrossRef]
35. Rai, A.; Chen, X.; Mou, S. Control-Inspired Federated Learning: A Projection-Based Approach. *IFAC-PapersOnLine* **2025**, *59*, 109–114. [CrossRef]
36. Yuan, X.; Li, P. On Convergence of FedProx: Local Dissimilarity Invariant Bounds, Non-smoothness and Beyond. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2022; Volume 35, pp. 10752–10765.
37. Kang, M.; Kim, S.; Jin, K.H.; Adeli, E.; Pohl, K.M.; Park, S.H. FedNN: Federated Learning on Concept Drift Data Using Weight and Adaptive Group Normalizations. *Pattern Recognit.* **2024**, *149*, 110230. [CrossRef]
38. Yao, T.; Li, J.; Liu, J. FedAWR: Aggregation Optimization in Federated Learning with Adaptive Weights and Learning Rates. *Future Internet* **2026**, *18*, 106. [CrossRef]
39. Moreira, I.C.; Amaral, I.; Domingues, I.; Cardoso, A.; Cardoso, M.J.; Cardoso, J.S. INbreast: Toward a Full-Field Digital Mammographic Database. *Acad. Radiol.* **2012**, *19*, 236–248. [CrossRef] [PubMed]
40. Suckling, J.; Parker, J.; Dance, D.; Astley, S.; Hutt, I.; Boggis, C.; Ricketts, I.; Stamatakis, E.; Cerneaz, N.; Kok, S.; et al. *Mammographic Image Analysis Society (MIAS) Database v1.21*; Apollo—University of Cambridge Repository: Cambridge, UK, 2015. [CrossRef]
41. Heath, M.; Bowyer, K.; Kopans, D.; Moore, R.; Kegelmeyer, W.P. The digital database for screening mammography. In Proceedings of the 5th International Workshop on Digital Mammography, Toronto, ON, Canada, 11–14 June 2001.

42. Pisano, E.D.; Zong, S.; Hemminger, B.M.; DeLuca, M.; Johnston, R.E.; Muller, K.; Braeuning, M.P.; Pizer, S.M. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J. Digit. Imaging* **1998**, *11*, 193–200. [CrossRef]
43. The Radiological Society of North America. RSNA Screening Mammography Breast Cancer Detection Dataset. 2022. Available online: https://www.kaggle.com/competitions/rsna-breast-cancer-detection (accessed on 2 July 2025).