



A Review of Explainable Federated Learning Frameworks for Chest X-ray Diagnosis under Heterogeneous Hospital Data

Muhammad Auwal Yusuf
aooval002@gmail.com
Qassim University, Saudi Arabia

ABSTRACT

The application of deep learning in chest X-ray diagnosis has demonstrated promising results in detecting multiple thoracic diseases. However, traditional centralized approaches face significant challenges, including limited generalization across hospitals with heterogeneous patient populations and imaging protocols, compounded by strict privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) that prevent data sharing between institutions. Although centralized deep learning approaches perform well and achieve high local accuracy on their predictions, they are often “black boxes,” which limits clinical trust and interpretability. This review examines existing explainable federated learning frameworks for chest X-ray diagnosis under heterogeneous data conditions. Current approaches enable decentralized training across non-independent and identically distributed (non-IID) hospital environments, utilizing robust aggregation strategies such as Federated Averaging (FedAvg) and Federated Proximal (FedProx) to address label, quantity, and feature skew. To establish clinical trust, explainable Artificial Intelligence (XAI) techniques, such as Gradient weighted Class Activation Mapping (Grad CAM) and SHapley Additive exPlanations (SHAP), have been incorporated to generate interpretable visual explanations. The reviewed frameworks are evaluated on classification performance, robustness under heterogeneity, and stability of generated explanations. However, this review reveals significant gaps: the types of heterogeneity are addressed in isolation, XAI evaluation remains largely qualitative, and explanation stability under non-IID conditions lacks rigorous validation. These findings collectively highlight the need for federated frameworks that unify heterogeneity handling across all its forms simultaneously rather than addressing each in isolation, quantitative XAI assessment, and validation of explanation consistency across diverse hospital environments to enable trustworthy and interpretable clinical deployment.

Keywords: Federated Learning, Explainable AI, Chest X-Ray, Data Heterogeneity, Medical Imaging, Grad-CAM, SHAP, Non-IID, XAI.

1. INTRODUCTION

Deep learning is a fast-evolving artificial intelligence technology that enables computers to learn and extract patterns from complex datasets. In healthcare, it has applications in medical imaging, including analysing chest X-ray images by learning from thousands of data points, and it can automatically detect abnormalities such as thoracic diseases. In recent years, centralized deep learning methods applied to chest X-ray diagnosis have demonstrated promising results in detecting thoracic diseases, including pneumonia, tuberculosis, and cardiomegaly, and have shown high effectiveness when trained on large, well-annotated datasets [1]. Chest X-ray is one of the most common radiological tests for detecting thoracic diseases. Early detection and accurate diagnosis of these thoracic diseases are important for effective patient management, as they improve patient outcomes and reduce mortality. This early detection prevents disease progression, thereby allowing treatment to start before a condition becomes severe or life-threatening, while accurate diagnosis ensures appropriate treatment. Therefore, the use of this centralized deep learning offers the potential for faster, more consistent diagnostic results, improving overall clinical decision-making.

However, as highlighted across reviewed literature, these centralized deep learning approaches don't generalize well across hospitals, it can perform well on their training institution data but fail when deployed in a different clinical environment. Their performance is dependent on the availability of a large, expertly labeled dataset [2]. Strict privacy regulations (e.g., HIPAA, GDPR) and ethical concerns present a major challenge in the collection and management of medical imaging datasets for Artificial Intelligence (AI) applications [3], limiting their performance as different hospitals are not able to collaboratively share data in training the centralized deep learning model. Federated learning has emerged as a privacy-preserving solution that allows hospitals to collaboratively train a model without sharing raw patient data [4], thereby addressing the critical privacy concern in healthcare, and unlike centralized learning, federated learning spreads the training process across multiple clients, which can improve generalization by using data from different sources [1].

Despite its privacy advantages, federated learning introduces new challenges in real-world clinical environments [5]. Standard federated learning approaches (e.g., FedAvg) suffer a performance drop with heterogeneous, non-independent and identically distributed (non-IID) data [5], arising from varying patient demographics and imaging protocols across different hospitals, which manifest as label skew, quantity skew, and feature skew.

Compounding the data heterogeneity challenge is the black box nature of deep learning, which lacks the transparency required for clinical adoption [6]. While explainable Artificial intelligence (XAI) techniques such as Grad-CAM and SHAP improve interpretability, integrating federated learning and explainable Artificial intelligence is highly challenging because distributed, heterogeneous data naturally produces unstable explanations [6]. Also, inherent biases in medical imaging datasets (e.g., class imbalance) can propagate directly through these explainable artificial intelligence outputs, risking visual interpretation that actively misleads clinicians [7]. Therefore, a comprehensive review of XAI stability under heterogeneous federated conditions is essential for its clinical adoption.

This review examines the current state of federated learning frameworks for chest X-ray diagnosis, with particular focus on how they address data heterogeneity and integrate explainable AI. The primary objective of this review is to assess existing privacy-preserving and interpretable artificial intelligence frameworks for chest X-ray diagnosis in terms of their ability to maintain high predictive performance and stable interpretability across heterogeneous, non-IID hospital environments. This review evaluates existing frameworks on their classification accuracy, robustness to data heterogeneity through aggregation strategies such as FedAvg and FedProx, and the stability of generated XAI explanations using feature attribution methods such as Grad-CAM and SHAP across diverse clinical settings.

Through a comprehensive review of the literature, this study synthesizes current approaches to handling data heterogeneity in federated learning, also evaluates the integration and validation of explainable AI techniques in distributed healthcare settings, and identifies critical gaps that need to be addressed to enable trustworthy clinical deployment of these systems.

2. BACKGROUND

Advances in deep learning have changed medical imaging, yet its progress is affected by privacy data regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) [8], as well as fragmented datasets across institutions [4]. Deep learning enables computers to learn and extract patterns from complex datasets. Application of deep learning in medical imaging has demonstrated high effectiveness when they are trained on well-labeled datasets.

Chest X-ray is one of the common radiological methods for detecting thoracic diseases like pneumonia, tuberculosis, and cardiomegaly. The need to accurately and quickly detect these diseases is important because early intervention can be the difference between a life threatening and a mild illness condition [9]. Given this, in many areas of the world, particularly resource limited communities and rural settings, limitation to highly trained radiologist and access to advanced equipment for diagnostic results in patients being at risk. This is due to delayed or inaccurate diagnosis, which in some cases leads to fatal outcomes [9].

Deep learning has demonstrated great success in the classification of thoracic disease from chest x-ray [10]. However, the limitation of an inadequate or large dataset to build an efficient model is a challenge [11]. This is due to the need for data to be in a central location at the traditional (Centralized) deep learning collect data at a central location for training the artificial intelligence model. This challenge arises because hospitals cannot share patient data due to constraints of privacy regulation [8]. This traditional deep learning framework challenges patient privacy and is often not feasible due to legal regulation and infrastructural challenges [8].

2.1 Federated Learning

Federated learning emerges and addresses the issues of privacy concern and legal regulations. Federated learning allows multiple hospitals to collaboratively work together without exchanging patient data. The training data is stored locally without it being in a central location like the traditional deep learning setting. Each user(hospital) store training data locally on their end [12]. The output from each user's end is used to collectively contribute to the model update [12].

Each of the participating user(hospitals) trains its model locally, updates are aggregated to a central server by using aggregation procedures such as the FedAvg (Federated averaging) and FedProx (federated proximal) [3]. The FedAvg simply takes update from all users and averages them, while the FedProx is an improved version of the FedAvg. The aggregation is done without accessing the local data from the participating users [13]. The central server has the global model, which is sent to each user to be trained on their private data and send only updated model weight to the central server, making sure that no raw data is shared [2]. The aggregated global model is redistributed to all users again; the process continues until the global model on the central server converges. Because of this ability to keep local data private, it is particularly useful in healthcare for applications such as medical imaging analysis and predictive diagnostics [3]. Figure 1 below shows the federated learning architecture and demonstrates how the training process occurs.

2.2 Explainable Artificial Intelligence

Explainable Artificial Intelligence (XAI) refers to techniques that make the decision-making process of AI models transparent and interpretable to humans by revealing which input features or image regions most influenced a given prediction [14]. Explainable artificial intelligence techniques, like Grad-CAM, SHAP, and LIME, have the ability to interpret and show a visual representation of disease diagnosis [14]. Making it important in medical imaging because it offers interpretability and transparency in model prediction, thereby addressing the black box nature of AI models.

3. REVIEW METHODOLOGY

3.1 Search Strategy

Search to identify relevant studies on federated learning for chest X-ray diagnosis with focus on data heterogeneity and explainable AI integration was performed. The following databases were searched: IEEE Xplore, PubMed, ScienceDirect, ACM Digital Library, and Google Scholar. The date was sorted to keep result of searches that are only between January 2023 to March 2026.

The following terms were used in combination during the search: ("federated learning" OR "collaborative learning") AND ("chest X-ray" OR "chest radiograph" OR "thoracic imaging" OR "medical imaging") AND ("heterogeneous" OR "non-IID" OR "explainable" OR "Grad-CAM" OR "SHAP" OR "interpretable").

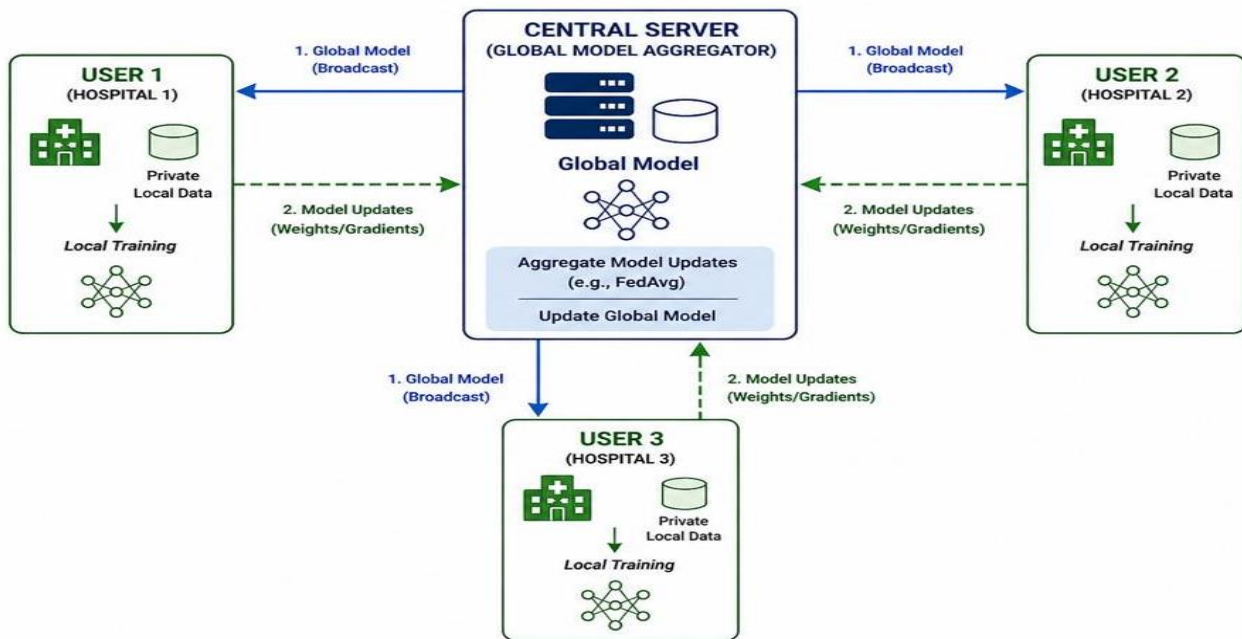


Figure-1: Federated Learning Architecture; Multiple Users Collaboratively Train a Global Model without Sharing Raw Data

3.2 Inclusion and Exclusion Criteria

Studies were selected if they satisfied at least two of the following criteria: (1) addressed federated learning in medical imaging contexts, (2) focused on chest X-ray diagnosis or provided methodology directly applicable to chest radiography, (3) discussed data heterogeneity handling approaches, explainable AI integration strategies, and (4) included experimental evaluation.

Studies were excluded if they: (1) focused exclusively on non-medical federated learning applications, (2) were purely theoretical.

3.3 Study Selection and Data Extraction

A total of 19 studies were found to satisfy criteria for selection above and were included in this review. From each study, information on the federated learning method, the type(s) of heterogeneity handled, the XAI techniques employed, the evaluation metrics reported, and the explicitly stated limitations were extracted. The findings were then synthesized thematically along three dimensions: (1) heterogeneity handling, (2) XAI integration and evaluation, and (3) evaluation frameworks.

4. FEDERATED LEARNING AND HETEROGENEITY

A key challenge in Federated learning is handling heterogeneous data from different sources, which is especially common in multi-institutional healthcare settings [4]. Heterogeneous data, which is often referred to as non-IID data, occurs when datasets held by different users (hospitals) in our context are not independent and identically distributed [15]. That is, instead of having uniform data across all locations, each of the hospitals possesses unique data characteristic that reflects its specific clinical environment heterogeneity, usually appearing in the form of Quantity skew, label distribution skew, and feature distribution skew.

4.1 Types of heterogeneity

Quantity skew occurs when one hospital, such as a large urban center, generates significantly more scans than a smaller rural clinic, resulting in substantial imbalances in the volume of data contributed by each client. Label distribution skew arises when a given hospital has a high concentration of certain disease classes, while partner hospitals have higher concentrations of different conditions. For instance, hospital A may predominantly hold pneumonia and tuberculosis cases, while hospital B holds mostly healthy patients and cardiomegaly scans. Feature distribution skew occurs when, even if two hospitals maintain the same class proportions, the images themselves differ markedly due to differences in equipment; for example, hospital A may use high-resolution modern scanners while hospital B relies on older imaging technology. These hardware differences produce images with substantially different pixel intensity, contrast, and resolution, making it difficult for a global model to learn consistent feature representations.

4.2 Label Skew Solutions

Padmavathi et al. [8] address label skew caused by class imbalance using synthetic enrichment by augmenting training sets with GAN-generated samples together with classical transformation, this improves robustness more particularly for rare disease classes. To account for heterogeneous data across the network, each client gets to personalise the global aggregated model on its local dataset using regularized objectives, thereby improving/balancing local adaptation with global consistency. However, the system is limited because it assumes all clients are the same, it does not cope with dynamic dropouts, and it does not yet provide guarantees for differential privacy, thereby limiting its application to highly controlled clinical settings [8].

Sharma and Guleria [16] also propose a federated learning framework using FedProx with a proximal term of 0.5, which demonstrated that the model, a 5-Client model, is capable of detecting pneumonia in diverse datasets with varying distributions. The framework was trained with pretrained EfficientNet B3 [16]. However, it faces a challenge where the choice of proximal term has a significant impact on the outcome of model performance [16].

Similarly, Mahmood et al. [3] propose a selective aggregation strategy called Federated Performance-based Averaging (FedPA) that allows only sufficiently trained and reliable local models can contribute to global updates. Each client is assigned to exclusively one disease class: normal, covid 19, lung opacity, or pneumonia.

It was implemented where only clients that have at least 85% validation are eligible to take part in global aggregation. The selective strategy filtered out noisy and inferior updates, and only the local models that are well-trained participate in the global parameters. FedPA remove low quality update from the model, reducing the risk of drift in the central/global model a challenge posed by data imbalance. A limitation is that the FedPA was evaluated with focus on accuracy as performance measure which alone will not sufficiently capture the impact of class imbalance [3].

Additionally, Biswas et al. [14] also utilizes Cycle GAN to generate synthetic sample allowing their proposed EDL model to function well with unpaired samples, thereby proving its suitability because the dataset size in a particular class is limited. Even though addressing the label skew issues, it has limitations where the framework was only applicable to binary classification, with the need for validation on a multi-disciplinary dataset.

4.3 Quantity Skew Solutions

Recognizing that the varying number of scans per client (hospital) drastically impacts federated learning convergence (quantity skew) alongside label skew, recent frameworks such as DMFL_net [17] address this by dynamic scheduling. Rather than mathematically altering aggregation weights, DMFL_net accommodates clients with massive data volumes by calculating waiting times based on localized training duration. By using these adaptive timers, the central server skips clients that take too long to train, preventing a heavy model with massive data from stalling the entire global network [17]. Furthermore, clients (hospitals) self-evaluate their local model and may only upload weights if they improve overall performance [17]. A significant challenge is its lack of investigation into temporal changes, which can lead to model bias and performance drift as patient demographics or radiological characteristics of disease evolve over time [17].

Liu et al. [18] address local data insufficiencies by introducing a local contrastive loss mechanism, PCRFed. This regularizes local training and allows clients to optimize for their specific data distribution while simultaneously leveraging the shared knowledge of the global network [18]. While the PCRFed shows great potential, it still highlights remaining challenges in heterogeneity issues, which include protocols, machines, and demographics, which can cause domain shift and reduce the ability of the model to generalize.

4.4 Feature Skew Solutions

Furthermore, feature skew remains one of the least addressed challenges in the review. Most studies focus on image format differences rather than deeper variations in device types and acquisition characteristics. As shown in [10], dataset variability is mainly limited to standard image formats such as PNG, JPG, and JPEG, which are usually standardized during preprocessing and do represent true feature skew across devices.

Similarly, Liu et al. [18], in their federated learning framework named PCRFed, partially addresses the issue of feature skew by introducing a site embedding layer that enhances the model's feature representation. This layer utilizes both avg-pooling and max-pooling operations, which collectively extract a diverse set of features [18]. This helps the model learn better by combining general and more highly important image information. It particularly mentions limitations that must be addressed, including data heterogeneity across medical institutions, due to the variation in image protocol, machine types, and patient demographics. Stating further, this difference may cause a domain shift, thereby reducing the ability to generalize.

4.5 Synthesis: Heterogeneity Gaps

Overall, Babar et al. [19] highlighted a key observation on an experimental study on how heterogeneity affects the performance and convergence of a federated learning algorithm. The results show that there is a struggle with convergence in a non-IID environment across a series of experiments with parameters varied. Further showing that the non-IID distribution significantly degrades global accuracy. Even though some studies try to address this heterogeneity, they tend to focus on a given aspect of it while ignoring the other sides of it, in real life application, any of the classes of heterogeneity might arise, while a system is good at addressing a certain class, if this conditions start to arise, global model convergence will significantly be degraded affecting overall performance of the federated learning framework which will be dangerous in healthcare settings. A certain class of heterogeneity is observed from this review that gets less attention in the federated learning framework: this is the feature skew. Also, quantity skew, even though handled indirectly like the DMFL_net, which mitigates its effect via adaptive scheduling and self-evaluation mechanisms, frames it as a communication efficiency problem rather than a data volume imbalance problem; only one of the reviewed papers directly addresses it. Collectively, these gaps reveal that the existing framework tends to address heterogeneity challenges in isolation, tackling one form at a time under controlled and often idealized experimental conditions. Beyond the challenge of data heterogeneity, a second fundamental barrier to the adoption of federated learning in clinical settings is the lack of transparency in model decision-making. This problem can be mitigated by the use of explainable AI, even though with their own limitations.

5. EXPLAINABLE AI IN FEDERATED LEARNING

The black box nature of AI is another key challenge affecting the adoption of federated learning frameworks in real-world clinical settings. This challenge is being addressed through the use of Explainable AI, which aims to make model decisions transparent and interpretable to clinical users.

5.1 XAI Techniques Overview

Two widely used techniques are the SHAP and Grad-CAM. SHAP measures how each feature of each input contributes to the prediction by measuring the change in output when features are included or excluded [4]. Grad-CAM, secondly, computes gradients by visualizing the most important/contributing region of the predicted class relative to feature maps and produces heatmaps [4].

5.2 XAI Implementation Across Frameworks

In trying to mitigate the black box nature, Tahosin et al. [4] implement SHAP and Grad-CAM in their framework FedVGM, where the SHAP accurately highlights discriminative features, including tumor texture and lesion shapes. The regions align well with clinical knowledge and reinforced that its transparency enhances accuracy, trust, and provides interpretability of the FedVGM [4]. Additionally, the Grad-CAM technique was used to visualize the region model looked at while making a prediction [4]. Colors were used to show how important each region is in model prediction, with prominent red and yellow corresponding to diagnostic features, and green and yellow showing less relevant features [4].

Thereby, enhance model transparency and serve as an added layer that helps clinicians cross-check AI predictions [4]. This promotes the integration of AI in clinical workflow. However, the institutional diversity and demographics across dataset were limited [4]. Given this, the stability of explainable AI has not been critically evaluated on diverse datasets.

While [4] addresses the black box nature of the FL framework by implementing Grad-CAM and SHAP in their framework, Mahamud et al [10] added Grad-CAM++ and LIME in their enhanced DenseNet201 model. The Grad-CAM++ enhances Grad-CAM by adding positive and negative to better show the critical region, while the LIME creates models locally that explain the prediction for certain instances [10]. Overall, contributing to the XAI by giving feature importance, local explanation, and visualization, while providing promising results, it presented a limitation in the implementation where the dataset used may not fully represent the diversity of the global population, and this model needs more validation across diverse populations to ensure generalization [10]. Hence, showing the need for validation of XAI across diverse populations to ensure stability.

Additionally, Biswas et al. [14] integrated XAI in their federated learning-based ensemble network using Grad-CAM and LIME and demonstrated how the XAI helps experts in making decisions. While effective, the framework was only applicable for binary classification, and validation on how it works on multi disciplinary data set needed. Even though not evaluated deeply, Malik et al. [17] added Grad-CAM in their DMFL_Net model for visualization, where highlighted affected areas on the chest are seen with dotted lines. Similarly, Durga et al. [13] incorporated this explainability in interpreting the prediction made by the FLEM model. SHAP and Grad-CAM were incorporated and further added to how FEDXAI offers a privacy-preserving solution explanation [13]. Furthermore, it does not explicitly discuss the limitation of the employed XAI methods; it implicitly highlights the need for improved interpretability of ensemble model decisions as a direction for future work [13]. Also, the variation in Grad-CAM visualization from [13] across hospitals indicates that model explanation may be inconsistent with some clients producing weak or no interpretable activation in the same sample of image, highlighting that XAI might fail in some cases.

Gupta et al. [2] approach the integration of XAI techniques differently, saying most of these techniques are applied after training, that is, post-hoc. The study integrates Grad-CAM and SHAP to generate this visualization interpretation of model prediction [2]. Making it apply to both local (client) and global (server) levels. After each round in the FL setup, the two techniques are applied, generating feature importance scores guaranteeing interpretability at both local and global model levels [2]. Under non-IID data distribution, the study introduces a lightweight federated explanation alignment (L-FEA) module, which addresses the inconsistency of interpretability across clients [2]. Integrating this L-FEA module builds consistency in explanation directly into the FL optimization loop. Evaluated on label skew non-IID condition, performance validation across other heterogeneity remains unattempted, which may degrade explanation stability.

In [9], GRAD-CAM was also integrated as a post-hoc explainability technique. Although qualitative feedback indicated that the highlighted region by this technique is accurate. However, limitations stated that not all highlighted regions may represent pathology directly, and human oversight is needed and remains crucial.

Similarly, Del Cerro et al. [1] evaluated the reliability of interpretability under IID conditions by keeping the same amount of data and increasing the number of clients. Their findings showed that Grad-CAM heatmaps became increasingly fragmented as the number of clients increased, resulting in reduced consistency and generalizability of explanations. This highlights the need for rigorous validation of explainability methods under non-IID clinical settings, which remains a major challenge in federated learning-based medical imaging.

5.3 Synthesis: XAI

Despite incorporating XAI techniques such as SHAP and Grad-CAM in federated learning frameworks, they are often post hoc. They run after model training and are not embedded in the learning process. This limits the credibility because explanations might fail to report on the network's real conduct [2]. Not many studies have tried optimizes interpretability together privacy preserving distributed training, even though Gupta et al. [2] evaluated XAI under non-IID condition, specifically the label skew in isolation, the need for rigorous evaluation on stability of this explanation under non-IID conditions of other class is needed, adding to this, Del Cerro et al. [1] evaluation of this XAI explanations even though under IID conditions shows that this stability degrades as number of clients increases posing an even greater needs for evaluations under non IID settings. Some of the reviewed studies focus particularly on heterogeneity and leaving XAI without proper evaluation, even though incorporated in their framework, given that some of these studies noted that across hospitals, model explanation becomes inconsistent and might fail in some cases, further adding to the need for the proper evaluation of XAI under heterogeneous data conditions.

6. EVALUATION METRICS

6.1 Heterogeneity Evaluation Metrics

Federated learning frameworks in clinical applications are commonly evaluated by their ability to maintain stable performance across participating clients even with the presence of heterogeneous and non-IID distributions. The impact of heterogeneity is mostly assessed through standard classification metrics from the reviewed works. These are accuracy, recall, F1-score, and AUC. Among these metrics, recall and F1-score are important in medical imaging because they help in a more balanced evaluation under class imbalance. From FedPa [3], which evaluated performance with accuracy as the primary metric, shows that relying on accuracy alone may not efficiently capture the impact of class imbalance, which the F1-score and recall can help with. Similarly, Biswas et al. [14] demonstrate this by reporting F1-score alongside accuracy, achieving an F1-score of 98.36% with a recall of 98.13%, which together confirm the framework's ability to minimize both false positives and false negatives simultaneously rather than optimizing for one at the expense of the other.

Several reviewed frameworks extend their evaluation beyond single-metric reporting by assessing performance consistency across clients with varying data distribution and data sizes. Liu et al. [18], through PCRFed, evaluated performance under quantity imbalance, with participating sites holding between 41 and 255 samples, giving a measure of how well the system generalizes when data volumes are different across institutions. However, standardized evaluation protocols specifically designed to measure the effect of feature skew or domain shift remain absent across the reviewed literature. This absence makes direct comparison between frameworks difficult and limits the ability to determine which approaches are most robust under the most clinically realistic heterogeneity conditions.

6.2 Privacy Evaluation Metrics

Although federated learning is regarded as a privacy-preserving learning by keeping raw patient data localized at each participating institution, the evaluation of privacy guarantees remains insufficient across the reviewed frameworks. The decentralized training architecture prevents direct data sharing, which alone does not guarantee privacy, as gradient updates that are shared during aggregation can still leak sensitive information through model inversion and membership inference attacks.

Most reviewed works, including Padmavathi et al. [8], Sharma and Guleria [16], and Liu et al. [18] through PCRFed, assume privacy preservation through the federated architecture itself without quantitatively evaluating vulnerability to such attacks. Padmavathi et al. [8] explicitly acknowledge this gap in their limitations, noting that the framework does not yet provide differential privacy guarantees, which restricts its suitability for deployment in highly regulated clinical environments governed by frameworks such as HIPAA and GDPR.

Across the reviewed literature, it is observed that privacy in most federated medical imaging systems is assumed through architectural design instead of rigorous validation through measurable privacy guarantees. Until privacy evaluation adopts standards of measure compared to those used for prediction performance, the degree of privacy protection offered by the frameworks' realistic adversarial conditions cannot be formally established.

6.3 Explainability Evaluation Metrics

The evaluation of explainable artificial intelligence techniques in the reviewed studies is mostly qualitative. Grad-Cam and SHAP, which appear to be the most used methods, are used across multiple studies reviewed, Gupta et al. [2], Tahosin et al. [4], Biswas et al. [14], and Durga et al. [13], to generate a visual explanation that highlights affected regions in chest x-ray. These visualizations are presented as evidence of the model's transparency and clinical interpretability and assessed by visual inspection against known patterns of a disease pathology.

However, the reviewed literature does not widely report quantitative metrics for evaluating these explanations. Although Gupta et al. [2] partially evaluated explanation consistency across heterogeneous federated clients, broader quantitative XAI metrics such as fidelity, stability, robustness, and localization accuracy are rarely incorporated across the reviewed studies. The study by Del Cerro et al. [1], which evaluated Grad-CAM reliability under increasing client counts even though under IID conditions, is the closest reviewed paper that comes near to a quantitative XAI evaluation, observing that heatmaps become progressively more fragmented and less interpretable as the number of clients increases. This raises concern about explanation stability under non-IID conditions of a real clinical setup; this degradation has not been measured by any of the reviewed frameworks.

This shows a significant limitation because heterogeneous federated environments introduce client-specific data distributions that can cause explanation patterns to diverge across institutions. As observed in the analysis of Durga et al. [13], Grad-CAM visualizations can become inconsistent across clients. The absence of rigorous quantitative XAI evaluation metrics across the reviewed literature, therefore, presents a barrier to the trustworthy deployment of these systems in clinical settings. Table-1 below gives a summary of the evaluation metrics used in this review.

Table-1: Summary of Heterogeneity, Privacy, and Explainability Evaluation Metrics in Reviewed Federated Learning Studies

Evaluation Category	Common Metrics	Purpose	Limitation from the reviewed Studies
Heterogeneity Evaluation	Accuracy, Recall, F1-score, AUC	Measure classification performance under non-IID and imbalanced data distribution	Standardized evaluation for feature skew and domain shift remains limited
Privacy Evaluation	Differential Information Analysis, Privacy, Leakage	Measure privacy preservation and resistance to information leakage	Most frameworks assume privacy through architecture without quantitative validation
Explainability Evaluation	Fidelity, Stability, Localization Accuracy	Measure the interpretability of generated explanations	Evaluation is mostly qualitative, with limited quantitative XAI assessment

7. STATE OF THE ART COMPARISON

Table-2 below presents a structured comparison of the reviewed federated learning frameworks across key dimensions: the federated learning method employed, the type of data heterogeneity addressed, the explainable AI techniques integrated, and key limitations.

Table-2: Comparative Summary of Reviewed Federated Learning Frameworks

Ref	Authors (Year)	Federated Learning Method	Data Heterogeneity Addressed	Explainable AI Technique	Key Limitations
[1]	Del Cerro & Desco (2025)	FedAvg (IID only)	Not addressed	Grad-CAM	IID setting only
[2]	Gupta et al. (2026)	FedAvg + L-FEA module	Label skew	Grad-CAM, SHAP	Explanation evaluated on label skew without validation across feature and quantity skew

Ref	Authors (Year)	Federated Learning Method	Data Heterogeneity Addressed	Explainable AI Technique	Key Limitations
[3]	Mahmood et al. (2025)	Performance-based aggregation	Label skew	None	Accuracy-only evaluation; insufficient for class imbalance measurement
[4]	Tahosin et al. (2026)	FedAvg + Median Agg.	None explicitly addressed	SHAP, Grad-CAM	Limited demographic & institutional diversity; XAI stability not evaluated on diverse datasets
[8]	Padmavathi et al. (2025)	FedAvg + GAN augmentation	Label skew	None	No differential privacy; assumes homogeneous client hardware
[9]	Mathew & Jeba (2025)	FedAvg	None explicitly addressed	Grad-CAM (post-hoc)	Not all highlighted regions map to pathology; human oversight remains essential
[10]	Mahamud et al. (2024)	Standard FedAvg	None explicitly addressed	Grad-CAM++, LIME	Dataset may not represent the diversity of the global population, XAI need diverse validation to ensure generalization.
[13]	Durga et al. (2025)	FedAvg ensemble +	None explicitly addressed	SHAP, Grad-CAM	Inconsistent Grad-CAM across clients; XAI might fail in some cases
[14]	Biswas et al.	FedAvg CycleGAN +	Label Skew	Grad-CAM, LIME	Binary classification only; no multi-disease validation
[16]	Sharma & Guleria (2025)	FedProx ($\mu = 0.5$)	Label skew	None	Manual proximal term tuning required; sensitive to μ choice
[17]	Malik et al. (2023)	Dynamic scheduling + self-eval	Quantity skew (indirect)	Grad-CAM	No temporal drift handling; frame quantity skew as a communication problem
[18]	Liu et al. (2025)	FedAvg contrastive loss +	Quantity skew and Feature skew (partial)	None	Deeper domain shifts from hardware unresolved
[19]	Babar et al. (2024)	Experimental analysis	Label and Quantity skew (analysis only)	None	Analysis only; no proposed solution; confirms degradation under non-IID

7.1 Synthesized Gaps Across Reviewed Frameworks

From the reviewed studies, it has been observed that heterogeneity in FL significantly degrades performance. While these heterogeneity types are being mitigated, they are addressed in isolation, where one form is tackled at a time. Among the heterogeneity types observed, the label skew type gets more attention, and frameworks are evaluated on that given aspect alone, leading to a lack of generalization under real clinical settings. Each of the classes of heterogeneity can arise at any given time, thereby rendering the framework not generalizable and affecting the overall effectiveness of the framework.

Another key gap beyond this heterogeneity challenge, even though XAI techniques are incorporated in FL frameworks, in some of the reviewed studies, is the lack of rigorous evaluation of XAI in the frameworks. XAI on their own are not evaluated deeply in the reviewed frameworks, and this becomes even more problematic when they are introduced to FL settings where heterogeneity exists. As observed from the studies, heterogeneity affects both performance and explanation stability of XAI, so the need to evaluate XAI explanations rigorously under heterogeneous conditions is necessary.

Some studies even show how XAI explanations fail in some cases or become inconsistent across hospitals. These techniques in the reviewed studies are mostly post hoc and are often run after model training and not embedded in the learning process. Overall, these gaps collectively reveal the need for a system that is a generalist, handling heterogeneity effectively and with XAI explanations evaluated rigorously under these heterogeneous conditions.

8. FUTURE WORKS

Across the reviewed studies, some directions for future work can be derived. A unified heterogeneity framework that addresses all three heterogeneity types simultaneously, a framework that jointly addresses and handles quantity, label, and feature skew. Second to that is an explicit feature skew solution, which remains the least addressed across the literature. Future work can explore domain adaptation techniques such as adversarial training, scanner normalization networks, or style transfer embedded within federated learning.

Also, current XAI methods are mostly post-hoc. Future work should embed these methods directly into the FL optimization loop, building on the L-FEA module.

Most importantly, a framework that is rigorously evaluated on XAI stability under non-IID conditions, which will lead to more adaptation of the FL framework incorporated with the XAI method in clinical usage. This will certainly lead to the framework's interpretability being trusted and adopted for the clinical environment.

9. CONCLUSION

This review has examined federated learning frameworks for chest X-ray diagnosis with a focus on data heterogeneity and explainable AI. The reviewed studies showed that federated learning offers a privacy-preserving alternative to centralized training, by allowing models to be trained locally across hospitals without sharing patient data. However, significant gaps remain that currently limit the clinical applicability or deployment of these systems.

Heterogeneity is mostly addressed in isolation, with label skew receiving more attention while feature skew and quantity skew remain relatively underexplored. XAI techniques, while they are being increasingly incorporated in federated learning frameworks, are evaluated qualitatively and have not been rigorously assessed under non-IID conditions that characterize real hospital networks. Privacy, though claimed by all reviewed frameworks, is not formally validated by a lot of frameworks through measurable guarantees.

Addressing these gaps, through unified heterogeneity handling, embedded and rigorously evaluated XAI, and formal privacy validation, represents the critical next step toward federated learning systems that are not only technically robust but also transparent, trustworthy, and ready for real-world clinical deployment.

REFERENCES

- [1] C. F. Del Cerro, M. Desco, and M. Abella, "Evaluation of Client Participation on Federated Learning Scenario for Chest X-ray Imaging," in Proc. 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), Madrid, Spain, 2025, DOI: 10.1109/CBMS65348.2025.00139.
- [2] S. Gupta, M. Gupta, R. Kumar, and A. Abraham, "A Federated Learning and Explainable AI Framework for Privacy-Preserving Brain Tumor Diagnosis Using Multi-Institutional MRI Data," IEEE Access, vol. 14, pp. 22630–22645, 2026, DOI: 10.1109/ACCESS.2026.3660305.
- [3] Mahmood, T. K. Sadique, S. R. Azzuhri, R. Ramli, and L. Ismail, "Federated Performance-based Averaging (FedPA): A Robust and Selective Learning Framework for Chest X-Ray Classification in Heterogeneous Data Environments," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 16, no. 10, pp. 940–951, 2025.
- [4] M. S. Tahosin, Md. A. Sheakh, M. J. Alam, Md. M. Hassan, A. K. Bairagi, S. Abdulla, S. Alshathri, and W. El-Shafai, "FedVGM: Enhancing Federated Learning Performance on Multi-Dataset Medical Images With XAI," IEEE Journal of Biomedical and Health Informatics, vol. 30, no. 2, pp. 1272–1281, Feb. 2026, DOI: 10.1109/JBHI.2025.3600361.
- [5] P. Kulkarni, A. Kanhere, P. H. Yi, and V. S. Parekh, "From Isolation to Collaboration: Federated Class-Heterogeneous Learning for Chest X-Ray Classification," in Proceedings of Machine Learning Research (ML4H), vol. 259, pp. 1–13, 2024.
- [6] A. Amato and D. Branco, "SemFedXAI: A Semantic Framework for Explainable Federated Learning in Healthcare," Information, vol. 16, no. 6, Art. no. 435, 2025, DOI: 10.3390/info16060435.
- [7] F. Ahmed, N. S. Naz, S. Khan, A. U. Rehman, W. M. Ismael, and M. A. Khan, "Explainable artificial intelligence (XAI) in medical imaging: a systematic review of techniques, applications, and challenges," BMC Medical Imaging, vol. 26, Art. no. 37, 2026, DOI: 10.1186/s12880-025-02118-w.
- [8] A. Padmavathi, N. Kushwaha, and S. Varma, "Collaborative Privacy-Preserving Federated Learning for Medical Imaging Across Hospitals," in Proc. 2025 12th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, Sep. 18–19, 2025, DOI: 10.1109/ICRITO66076.2025.11241737.
- [9] M. Mathew and A. Jeba, "Towards Generalizable Pneumonia Detection: A Federated Multi-Modal Explainable Deep Learning Approach," in Proc. 2025 4th International Conference on Automation, Computing and Renewable Systems (ICACRS), Coimbatore, India, 2025, DOI: 10.1109/ICACRS67045.2025.11324386.
- [10] E. Mahamud, N. Fahad, Md. Assaduzzaman, S. M. Zain, K. O. M. Goh, and Md. K. Morol, "An explainable artificial intelligence model for multiple lung diseases classification from chest X-ray images using fine-tuned transfer learning," Decision Analytics Journal, vol. 12, Art. no. 100499, 2024, DOI: 10.1016/j.dajour.2024.100499.
- [11] M. Wakode and G. Kale, "Efficient Medical Image Classification in Model-Heterogeneous Federated Learning," in Proc. 2025 Third International Conference on Industry 4.0 Technology (I4Tech), Pune, India, Sep. 18–20, 2025, DOI: 10.1109/I4Tech64670.2025.11277955.
- [12] N. K. Trivedi, A. K. Agarwal, H. Maheshwari, V. Gautam, and R. G. Tiwari, "Lightweight Federated Learning for COVID-19, Pneumonia, and TB From Chest X-Ray Images," in Proc. 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), 2023, DOI: 10.1109/ICAIIHI57871.2023.10489813.
- [13] S. Durga, E. Daniel, S. Seetha, V. K. Reshma, and V. Sachnev, "FLEM-XAI: Federated learning based real time ensemble model with explainable AI framework for an efficient diagnosis of lung diseases," Frontiers in Computer Science, vol. 7, Art. no. 1633916, Aug. 11, 2025, DOI: 10.3389/fcomp.2025.1633916.
- [14] S. Biswas, R. Mostafiz, M. S. Uddin, and M. S. Uddin, "FLPneXAINet: Federated deep learning and explainable AI for improved pneumonia prediction utilizing GAN-augmented chest X-ray data," PLOS One, vol. 20, no. 7, Art. no. e0324957, Jul. 17, 2025, DOI: 10.1371/journal.pone.0324957.

- [15] J. Sáinz-Pardo Díaz and Á. López García, “Study of the performance and scalability of federated learning for medical imaging with intermittent clients,” arXiv preprint arXiv:2207.08581v3, Nov. 2022.
- [16] S. Sharma and K. Guleria, “A Collaborative Privacy Preserved Federated Learning Framework for Pneumonia Detection using Diverse Chest X-ray Data Silos,” *International Journal of Mathematical, Engineering and Management Sciences*, vol. 10, no. 2, pp. 464–485, 2025, DOI: 10.3389/IJMEMS.2025.10.2.023.
- [17] H. Malik, A. Naeem, R. A. Naqvi, and W.-K. Loh, “DMFL_Net: A Federated Learning-Based Framework for the Classification of COVID-19 from Multiple Chest Diseases Using X-rays,” *Sensors*, vol. 23, no. 2, Art. no. 743, Jan. 2023, DOI: 10.3390/s23020743.
- [18] S. Liu, R. Zhang, M. Fang, H. Li, T. Xun, Z. Wang, W. Shang, J. Tian, and D. Dong, “PCRFed: personalized federated learning with contrastive representation for non-independently and identically distributed medical image segmentation,” *Visual Computing for Industry, Biomedicine, and Art*, vol. 8, no. 6, 2025, DOI: 10.1186/s42492-025-00191-0.
- [19] M. Babar, B. Qureshi, and A. Koubaa, “Investigating the impact of data heterogeneity on the performance of federated learning algorithm using medical imaging,” *PLOS One*, vol. 19, no. 5, Art. no. e0302539, May 15, 2024, DOI: 10.1371/journal.pone.0302539.