xHetMed-FL: Explainable Federated Shared Representation Learning Across Heterogeneous Medical Imaging Modalities

Anonymous authors
Paper under double-blind review

Abstract

Medical images are usually stored in different hospitals and with different imaging modalities and are often subject to different storage formats, privacy concerns, modality-specific characteristics and different security restrictions, which hinders centralized learning. Current federated learning methods usually assume a uniform class or learning goal across clients and the raw patient data cannot be shared. To address these challenges, we propose xHetMed-FL, a multimodal federated learning scheme using 2D/3D CNNs to create a shared 128-D latent representation while federating only a few projection parameters. The proposed framework is organized around four heterogeneous clients representing SPECT, Chest X-ray, MRI, and CT+PET imaging. Rather than forcing these modalities into a common input architecture, each client retains a modality-specific preprocessing pipeline and encoder. The SPECT client uses verified pathological and non-pathological labels for supervised disease classification, whereas the Chest X-ray, MRI, and CT+PET clients learn representations without disease labels through a cosine-similarity-based self-supervised objective. Only the shared projection is communicated and aggregated by the federated server, while modality-specific encoders, client-specific components, prediction heads, and Batch Normalization parameters remain local. To address optimization under heterogeneous client conditions, FedProx is incorporated into local training, providing proximal regularization with respect to the global shared parameters, while FedBN preserves client-specific Batch Normalization statistics. Following federated training, the SPECT client's validation accuracy, F1 score, and ROC-AUC were 82.35%, 82.86%, and 91.77%, respectively. The cosine similarity in the Chest X-ray, MRI, and CT+PET clients was 1.0000, 1.0000, and 0.9992, respectively, while the variance in their latent representations was 0.3185, 0.3168, and 0.2965, respectively. Grad-CAM is used to interpret the trained network at the post-training stage, providing modality-specific explanations while preserving the heterogeneous local components. Overall, the proposed framework demonstrates the viability of learning shared representations from heterogeneous medical imaging modalities with different learning goals while keeping modality-specific components local and the models interpretable.

1 Introduction

Medical imaging plays a fundamental role in modern healthcare by supporting disease diagnosis, screening, treatment planning, and clinical decision-making. Recent advances in deep learning, particularly convolutional neural networks (CNNs), have enabled effective extraction of discriminative features from medical images. Three-dimensional (3D) CNNs are well suited to volumetric modalities such as CT, MRI, and SPECT because they exploit spatial information across multiple dimensions, whereas two-dimensional (2D) CNNs remain appropriate for planar modalities such as chest X-ray Singh et al. (2020); Shan et al. (2018); Yu et al. (2020); Setio et al. (2017); Zhou et al. (2021); Wang et al. (2024). However, conventional centralized deep learning requires medical images to be collected at a common location, which is challenging because clinical data are sensitive and distributed across hospitals and institutions. Federated learning (FL) addresses this limitation

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by retaining data locally and collaboratively optimizing model parameters through a central server. FedAvg established this paradigm by enabling decentralized clients to train a shared model without directly exchanging local data McMahan et al. (2017). Nevertheless, conventional FL generally assumes compatible model architectures and a common learning objective, assumptions that are difficult to satisfy in heterogeneous medical imaging environments. Medical modalities differ substantially in dimensionality, intensity characteristics, acquisition mechanisms, anatomical information, and image distributions. Consequently, a common local architecture may be inappropriate: 3D CNNs are naturally suited to volumetric images, whereas 2D CNNs are more appropriate for planar images Singh et al. (2020); Shan et al. (2018); Yu et al. (2020). In addition, modality-specific preprocessing, including normalization, intensity processing, and spatial transformation, can substantially influence learned representations Masoudi et al. (2021); Botnari et al. (2024); Trojani et al. (2024). These differences create both architectural and statistical heterogeneity among federated clients. A further challenge is the availability of reliable clinical labels. Some medical imaging modalities contain verified disease annotations, whereas others may be largely unlabeled. Self-supervised learning can exploit such data by constructing representation-learning objectives directly from the available images Taleb et al. (2020); Jaiswal et al. (2021); Anton et al. (2022). Federated self-supervised learning further enables distributed unlabeled data to contribute without centralized access Zhuang et al. (2022); het (2023). Heterogeneous federated learning has consequently explored shared representations, partial parameter sharing, and model-heterogeneous architectures, allowing clients with different local models to participate through compatible components Collins et al. (2021); Xu et al. (2023); Alam et al. (2022); Sun et al. (2021); Xie et al. (2024); Wang et al. (2025). Optimization heterogeneity also requires consideration because different client objectives and distributions can cause divergence of shared parameters. FedProx constrains local optimization through a proximal term Li et al. (2020b); Hanzely et al. (2020), while FedBN retains Batch Normalization parameters and statistics locally to accommodate client-specific feature distributions Li et al. (2021).

Based on these challenges, we propose **xHetMed-FL**, an explainable federated shared-representation framework for heterogeneous medical imaging. The framework comprises four clients: SPECT, Chest X-ray, MRI, and CT+PET. Each client uses a modality-specific preprocessing pipeline and encoder. SPECT performs supervised pathological/non-pathological classification, whereas Chest X-ray, MRI, and CT+PET perform self-supervised representation learning using cosine similarity. Despite heterogeneous local architectures, all encoders produce 128-dimensional features, which are mapped into a common 128-dimensional latent space through a shared projection Collins et al. (2021). Only this projection is federated; modality-specific encoders, prediction heads, and Batch Normalization components remain local. FedProx is incorporated for optimization under heterogeneity Li et al. (2020b), while FedBN preserves modality-specific normalization statistics Li et al. (2021). The server performs sample-weighted aggregation of the shared projection parameters following each local training round McMahan et al. (2017), with training conducted for ten communication rounds. Following training, each client is evaluated according to its learning objective. SPECT is evaluated using accuracy, precision, recall, F1-score, MCC, and AUC, while Chest X-ray, MRI, and CT+PET are evaluated using representation loss and cosine similarity. Finally, Grad-CAM provides modality-specific explanations Selvaraju et al. (2017). Class-specific 3D Grad-CAM is used for SPECT, while 2D/3D Grad-CAM with the latent representation norm as the target is used for the self-supervised clients.

The major contributions of the proposed framework are:

- A heterogeneous multimodal federated framework supporting 2D and 3D medical imaging clients without requiring a common local encoder architecture.
- A shared 128-dimensional representation providing a common interface between heterogeneous modality-specific encoders.
- A mixed supervised/self-supervised federated setting in which SPECT performs supervised disease classification while Chest X-ray, MRI, and CT+PET perform self-supervised representation learning.
- Selective federation of only the shared projection parameters while modality-specific encoders, prediction heads, and Batch Normalization parameters remain local.
- Integration of FedProx and FedBN to address optimization and statistical heterogeneity.
- A Grad-CAM-based explainability framework for supervised disease prediction and self-supervised representation learning.

- Quantitative analysis of attribution maps and latent representations in addition to qualitative visualization.

2 Related Work

2.1 Deep Learning for Medical Image Analysis

Deep learning has become an important approach for medical image analysis because CNNs can learn hierarchical representations directly from imaging data. 3D CNNs exploit volumetric information and have been widely investigated for modalities such as CT and MRI Singh et al. (2020); Shan et al. (2018); Setio et al. (2017). Comparisons between 2D and 3D CNNs further indicate that architecture selection depends on the dimensional characteristics of the imaging task Yu et al. (2020). Consequently, heterogeneous medical imaging systems require modality-appropriate architectures: MRI, CT, and SPECT benefit from volumetric processing, whereas chest X-ray is naturally represented as a 2D image Zhou et al. (2021); Collins et al. (2021); Xie et al. (2024).

2.2 Medical Image Preprocessing

Medical modalities differ in intensity distributions, spatial dimensions, acquisition protocols, and noise characteristics, making preprocessing an important component of deep learning pipelines. Intensity normalization, spatial transformation, and modality-specific preprocessing can influence downstream performance Masoudi et al. (2021); Botnari et al. (2024); Trojani et al. (2024). PET/CT introduces additional complexity because CT and PET provide complementary anatomical and functional information with different intensity characteristics, motivating appropriate modality-specific processing Fallahpoor et al. (2024). These observations support separate preprocessing pipelines for heterogeneous federated clients.

2.3 Federated Learning for Heterogeneous Data

FedAvg enables collaborative learning without transferring local datasets by aggregating locally optimized model parameters McMahan et al. (2017). However, standard FL generally assumes compatible client models and objectives. Shared representation learning can instead separate local feature extraction from a common component that is exchanged across clients Collins et al. (2021), while personalized FL investigates collaboration among heterogeneous local models Xu et al. (2023). FedRolex and PartialFed further demonstrate that complete model equivalence is not necessary for federated collaboration Alam et al. (2022); Sun et al. (2021). In medical applications, MH-pFLID investigates model-heterogeneous personalized FL through parameter- and representation-level collaboration Xie et al. (2024), motivating designs where modality-specific encoders remain local while a compatible representation component is federated.

2.4 Federated Self-Supervised Learning

High-quality medical annotations are expensive and require clinical expertise. Self-supervised learning addresses this limitation by learning representations from unlabeled data through pretext or representation-learning objectives **?**Anton et al. (2022). 3D self-supervised learning has also been explored to exploit volumetric medical information without manual labels Taleb et al. (2020). Federated self-supervised learning extends this approach to distributed datasets, enabling collaborative representation learning without centralized data Zhuang et al. (2022). Heterogeneous federated self-supervised methods additionally consider differences in client data characteristics and learning environments het (2023).

2.5 Federated Optimization under Heterogeneity

FedProx introduces a proximal regularization term that constrains local updates relative to the global model, helping address heterogeneous client objectives and distributions Li et al. (2020b); Hanzely et al. (2020). FedBN retains Batch Normalization parameters and running statistics locally to address client-specific feature distributions Li et al. (2021). Other approaches investigate feature decorrela-

tion and distribution-aware federated learning Shi et al. (2023); Li et al. (2023; 2020a). xHetMed-FL combines the principles of FedProx and FedBN with selective sharing of the common projection.

2.6 Explainable Medical Artificial Intelligence

Explainability is important in medical imaging because model predictions can influence clinical decision-making. Grad-CAM generates spatial attribution maps by weighting convolutional feature maps according to gradients of a selected target Selvaraju et al. (2017), while Captum's LayerGradCam provides an implementation for convolutional networks Captum Contributors (2025). Quantitative evaluation of explanations has been investigated beyond visual inspection Samek et al. (2017); Khan et al. (2024); Dudeja & Sharma (2026), and sanity checks emphasize verifying that attribution methods respond to learned model parameters rather than merely image structure Adebayo et al. (2018). In xHetMed-FL, Grad-CAM is applied after training to both supervised and self-supervised clients.

3 Proposed Methodology

The proposed methodology is divided into various subsections as discussed below:

3.1 Framework Overview

The proposed xHetMed-FL framework consists of four heterogeneous medical imaging clients:

1. **SPECT client:** a 3D CNN using supervised pathological/non-pathological classification;
2. **Chest X-ray client:** a 2D CNN using self-supervised representation learning;
3. **MRI client:** a 3D CNN processing paired T1/T2 volumes using self-supervised representation learning;
4. **CT+PET client:** a 3D CNN processing paired CT/PET volumes using self-supervised representation learning.

The framework is designed so that the local encoders can be architecturally different. For client m, the local encoder is represented by E_m:

$$h_m = E_m(X_m), \tag{1}$$

where X_m denotes the local modality-specific input and h_m is a 128-dimensional feature vector. A shared projection maps the local representation into the common latent space:

$$z_m = P(h_m). \tag{2}$$

The resulting architecture can therefore be expressed as

$$X_m \to E_m \to h_m \to P \to z_m. \tag{3}$$

The local encoders remain at their corresponding clients, while P is the only model component exchanged with the server. The overall architecture of the proposed methodology is illustrated in Figure 1.

3.2 Modality-Specific Encoders

The heterogeneous clients require different feature extraction architectures. SPECT, MRI, and CT+PET use 3D CNNs because their inputs are volumetric, whereas Chest X-ray uses a 2D CNN. The 3D CNN consists of four convolutional blocks. The first three blocks contain 3D convolution, Batch Normalization, ReLU activation, and 3D max-pooling. The fourth block contains 3D convolution, Batch Normalization, and ReLU. The feature channels increase progressively from 16 to 32, 64, and 128. Adaptive global average pooling reduces the final feature map to a compact representation, which is then mapped to 128 dimensions using a fully connected layer. The Chest X-ray client uses a 2D CNN with modality-appropriate convolutional processing, Batch Normalization, ReLU activation, and spatial pooling. Global average pooling and a fully connected layer produce the same 128-dimensional feature interface. The detailed layer-level architecture is moved to Appendix C so that the main paper does not spend several pages repeating implementation-level architecture information.

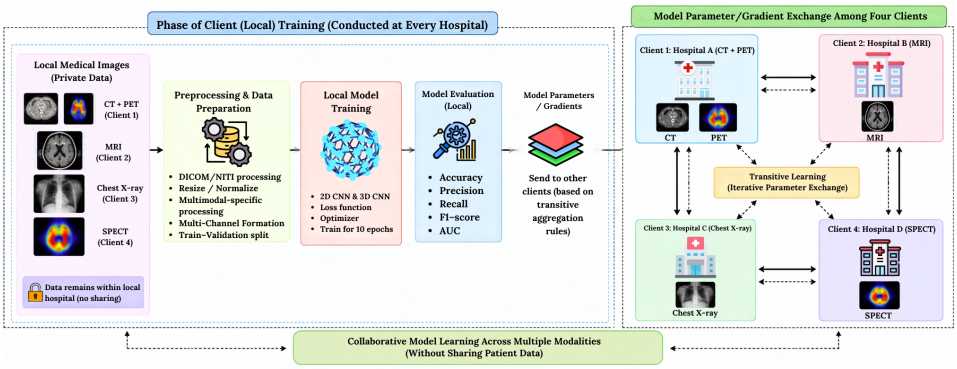


Figure 1: Overall Architecture of the Proposed Methodology

3.3 Shared Latent Projection

Although the local encoders differ, they all produce a 128-dimensional feature vector. The shared projection is implemented using fully connected layers, Layer Normalization, and ReLU:

$$z_m = W_2 \,\mathrm{ReLU}\left[\mathrm{LN}(W_1 h_m + b_1)\right] + b_2. \tag{4}$$

This projection acts as the interface between the heterogeneous local encoders and the federated server. The complete parameter decomposition is

$$\Theta_m = \{\Theta_m^E, \Theta_m^P, \Theta_m^H, \Theta_m^{BN}\}, \tag{5}$$

where Θ_m^E represents the local encoder, Θ_m^P the shared projection, Θ_m^H the local prediction head, and Θ_m^{BN} the Batch Normalization parameters. Only Θ_m^P is federated.

3.4 Supervised SPECT Learning

The SPECT client has verified pathological and non-pathological labels. Its shared representation is passed to a local classification head:

$$\hat{y} = H_{\mathrm{SPECT}}(z). \tag{6}$$

The supervised objective is the cross-entropy loss:

$$\mathcal{L}_{SPECT} = -\sum_c y_c \log(\hat{y}_c). \tag{7}$$

The classification head is local and is not communicated to the server.

3.5 Self-Supervised Representation Learning

The Chest X-ray, MRI, and CT+PET clients do not use verified disease labels in the implemented experimental setting. Instead, two representations generated from corresponding views are encouraged to be similar. The cosine similarity is

$$\mathrm{CosSim} = \frac{z^{(1)T} z^{(2)}}{\|z^{(1)}\|_2 \|z^{(2)}\|_2}. \tag{8}$$

The corresponding representation-learning loss is

$$\mathcal{L}_{SSL} = 1 - \mathrm{CosSim}\,. \tag{9}$$

This allows unlabeled clients to contribute to the shared representation without introducing artificial disease labels.

3.6 FEDPROX OPTIMIZATION

The local objective is augmented with a FedProx proximal term:

$$\mathcal{L}_m^{FedProx} = \mathcal{L}_m + \frac{\mu}{2} \left\| \Theta_m^P - \Theta_G^{(t)} \right\|_2^2, \tag{10}$$

where $\mu = 0.01$ and $\Theta_G^{(t)}$ denotes the shared global projection received at round t. FedProx is applied to the shared component while the modality-specific parameters continue to be optimized locally.

3.7 FEDBN AND SELECTIVE PARAMETER SHARING

Batch Normalization is retained locally. Therefore,

$$\Theta_m^{local} = \{\Theta_m^E, \Theta_m^H, \Theta_m^{BN}\}. \tag{11}$$

The communicated parameter set is

$$\Theta_m^{Fed} = \Theta_m^P. \tag{12}$$

This design prevents Batch Normalization statistics from one modality from being averaged with statistics belonging to a different modality.

3.8 FEDERATED AGGREGATION

After local training, only the updated shared projection parameters are transmitted to the server. The aggregation weight is

$$w_m = \frac{n_m}{\sum_{j=1}^4 n_j}, \tag{13}$$

where n_m is the number of local training samples. The global shared projection is updated as

$$\Theta_G^{(t+1)} = \sum_{m=1}^4 w_m \Theta_{m,P}^{(t+1)}. \tag{14}$$

The experimentally obtained aggregation weights are Table 1. The aggregated projection is redistributed to all clients before the next local training round.

3.9 FEDERATED TRAINING

The federated experiment consists of ten communication rounds with one local training epoch per round. As presented in Table 3, the learning rate is 10^{-4} and the FedProx coefficient is 0.01. The random seed is 42. Only the shared projection parameters are communicated. At each round, the server sends the current shared projection to all clients. Each client combines this projection with its private encoder and local components, performs one local optimization epoch according to its learning objective, and sends only the updated projection back to the server. The server computes sample-based aggregation weights and produces the next global projection. The complete algorithm with every local update, parameter-preservation step, aggregation operation, and evaluation step is provided in Appendix G.

3.10 EXPLAINABILITY

Following federated training, Grad-CAM is used to interpret the trained local networks Selvaraju et al. (2017). For SPECT, the target is the predicted disease class. For the self-supervised clients, the target is the magnitude of the learned representation,

$$s = \|z\|_2. \tag{15}$$

The corresponding 2D or 3D attribution maps are generated from the final convolutional feature maps. Detailed Grad-CAM equations, thresholding procedures, modality-specific visualization procedures, and quantitative attribution statistics are provided in Appendix I.

Table 1: Aggregation weights.

Client	Weight
SPECT	0.2386
X-ray	0.5258
MRI	0.2191
CT+PET	0.0165

Table 2: Dataset statistics and partitioning.

Client	Train	Val.	Test	Input	Obj.
SPECT	953	204	205	2×96^3	Sup.
X-ray	2100	450	450	224^2	SSL
MRI	875	188	188	2×128^3	SSL
CT+PET	66	14	16	2×128^3	SSL

Table 3: Federated training configuration.

Parameter	Value
Clients	4
Rounds	10
Local epochs/round	1
Learning rate	10^{-4}
FedProx μ	0.01
Latent dimension	128
Aggregation	FedProx + FedBN
Federated	Shared projection
Local	Encoder, head, BN
Seed	42
Training time	1570.01 min

Table 4: Final SPECT validation performance.

Metric	Value
Loss	0.3921
Accuracy	82.35%
Precision	88.78%
Recall	77.68%
F1-score	82.86%
MCC	0.6546
ROC-AUC	0.9177

4 Experimental Setup

4.1 Datasets

The proposed framework was evaluated using four heterogeneous medical imaging datasets distributed across four clients. Each client represents a distinct imaging modality and retains its raw data locally. Table 2 shows dataset statistics and partitioning.

4.1.1 SPECT

The SPECT dataset contains 1,362 preprocessed volumetric samples stored in NumPy format BENKHADDA (2026). Each sample contains two image channels. The original spatial size is $96 \times 96 \times 64$. The dataset contains 747 pathological and 615 non-pathological samples. After stratified partitioning using random seed 42, 953 samples were assigned to training, 204 to validation, and 205 to testing.

4.1.2 Chest X-ray

The Chest X-ray client uses a locally prepared subset of 3,000 images from the VinDr Chest X-ray dataset distributed through Zenodo Indian Institute of Science Bangalore (2024). The experimental split contains 2,100 training, 450 validation, and 450 test images. Images are converted to grayscale, resized to 224×224, and normalized. Because no verified disease target is used, the client performs self-supervised representation learning.

4.1.3 MRI

The MRI client uses BraTS2021 data distributed through Zenodo Baltruschat (2023). The data contain 2,502 NIfTI files corresponding to 1,251 subjects with paired T1 and T2 modalities. The experimental split contains 875 training, 188 validation, and 188 test subjects. Patient-level partitioning ensures that paired T1/T2 images belonging to the same subject remain within the same subset.

4.1.4 CT+PET

The CT+PET client uses the QUADRA healthy-control dataset distributed through Zenodo Medical University of Vienna (2025). The dataset contains paired CT and PET scans. PET is aligned to CT. CT intensity is clipped to $[-1000, 1000]$ HU, PET values below zero are clamped to zero, and paired volumes are resized to 128^3. The experimental split consists of 66 training, 14 validation, and 16 test studies.

Table 5: Final representation-learning performance.

Client	Loss	Cosine Sim.
Chest X-ray	0.0000	1.0000
MRI	0.0000	1.0000
CT+PET	0.0008	0.9992
Average	0.0003	0.9997

Table 6: Selected SPECT results during federated training.

Round	Loss	Acc.	Prec.	Recall	F1	AUC
1	0.5459	0.8235	0.8958	0.7679	0.8269	0.8971
2	0.4758	0.8333	0.9333	0.7500	0.8317	0.8999
4	0.5533	0.6961	0.6488	0.9732	0.7786	0.9043
7	0.3790	0.8480	0.9263	0.7857	0.8502	0.9058
10	0.3921	0.8235	0.8878	0.7768	0.8286	0.9177

The training process required a total of 1,570.01 minutes for the ten federated rounds, and the final trained models were saved separately for the four clients.

5 Experimental Results

5.1 SPECT Classification Performance

The final model was evaluated after the tenth federated communication round. Since SPECT is the client with a verified supervised disease target, classification metrics are reported for this client. As shown in Table 4, the model obtained a validation loss of 0.3921 and accuracy of 82.35%. Precision, recall, and F1-score were 88.78%, 77.68%, and 82.86%, respectively. The MCC was 0.6546, while ROC-AUC reached 0.9177.

5.2 Self-Supervised Representation Performance

For the unlabeled clients, representation loss and cosine similarity were used to evaluate the learned common representation as presented in Table 5. The Chest X-ray and MRI clients achieve cosine similarity of 1.0000, while CT+PET obtains 0.9992. The corresponding average cosine similarity is 0.9997.

5.3 Round-Wise Federated Performance

The SPECT validation results exhibit variation across communication rounds rather than strictly monotonic convergence. Selected results are shown in Table 6; the complete round-wise results, including all reported metrics, are provided in Appendix K.

The final model therefore maintains useful supervised discriminative information while simultaneously optimizing a common representation for heterogeneous unlabeled clients.

5.4 Latent Representation Statistics

The final latent representations of the self-supervised clients have variances of 0.318496, 0.316800, and 0.296452 for Chest X-ray, MRI, and CT+PET, respectively as shown in Table 7.

5.5 Explainability Results

Grad-CAM was generated after federated training to provide modality-specific spatial explanations. As shown in Table 8 for SPECT, the target was the predicted disease class. For the self-supervised clients, the target was the magnitude of the learned latent representation. The representative SPECT sample had ground-truth class 1 and predicted class 1. Its mean attribution was 0.069746 with standard deviation 0.154910. The number of important voxels was 81,367, representing 3.88% of the volume. For Chest X-ray, the mean attribution was 0.057282 with standard deviation 0.129690. A total of 1,341 pixels were classified as important, corresponding to 2.67% of the image. For MRI, the mean attribution was 0.013345 and standard deviation was 0.055196, with 2,626 important voxels corresponding to 0.13%. For CT+PET, the mean attribution was 0.101050 and standard deviation was 0.105035, with 1,626 important voxels corresponding to 0.08%. The qualitative explanations demonstrate that the shared representation can be interpreted at the local client level despite differences in imaging modality and architecture. Detailed attribution equations, thresholding procedures, modality-specific maps, and all visualization results are retained in Appendix I.

Table 7: Latent representation statistics.

Client	Mean	Std.	Variance
Chest X-ray	0.021806	0.564355	0.318496
MRI	-0.034253	0.562850	0.316800
CT+PET	-0.062608	0.544474	0.296452

Table 8: Representative quantitative Grad-CAM statistics.

Client	Mean	Std.	Important	Ratio	Latent Var.
SPECT	.069746	.154910	81367	3.88%	.240568
X-ray	.057282	.129690	1341	2.67%	.318496
MRI	.013345	.055196	2626	.13%	.316800
CT+PET	.101050	.105035	1626	.08%	.296452

6 Discussion

The experimental results demonstrate three important properties of xHetMed-FL. First, heterogeneous local architectures can participate in federated learning without requiring direct parameter-level compatibility. The SPECT, MRI, and CT+PET clients process volumetric data through 3D CNNs, while the Chest X-ray client uses a 2D CNN. The common 128-dimensional representation provides a compact interface between these otherwise incompatible local architectures. Second, the framework supports heterogeneous learning objectives. The SPECT client performs supervised disease classification using verified labels, while the remaining clients optimize representation learning without disease labels. The resulting cosine similarities close to one indicate that the self-supervised clients successfully learn aligned representations under the shared projection. Third, keeping the encoders and Batch Normalization components local avoids directly averaging parameters that are strongly tied to modality-specific distributions. FedProx further constrains the shared local updates relative to the global projection. The explainability results provide an additional perspective on the learned representations. The attribution distributions differ across modalities, which is expected because the input dimensionality and convolutional representations differ. The quantitative attribution statistics complement the visual maps by reporting the intensity and spatial concentration of the explanations. However, the present experiments should be interpreted within their current scope. The SPECT client has a verified disease target, whereas the other three clients are evaluated using representation-learning metrics. Therefore, a direct comparison of clinical classification performance across all four modalities is not currently supported by the experimental setting.

7 Conclusion and Limitations

This work proposed xHetMed-FL, an explainable federated shared-representation framework for heterogeneous medical imaging modalities. The framework enables SPECT, Chest X-ray, MRI, and CT+PET clients to participate in a common federated learning process while retaining modality-specific preprocessing, local encoders, local prediction heads, and Batch Normalization statistics. The shared 128-dimensional projection provides a common representation interface, while selective parameter federation avoids requiring complete architectural compatibility. The SPECT client performs supervised disease classification, whereas the Chest X-ray, MRI, and CT+PET clients perform self-supervised representation learning. FedProx and FedBN are incorporated to address optimization and feature-distribution heterogeneity. After ten communication rounds, the SPECT client achieved 82.35% validation accuracy, 82.86% F1-score, and 91.77% ROC-AUC. The Chest X-ray, MRI, and CT+PET clients achieved cosine similarities of 1.0000, 1.0000, and 0.9992, respectively. Grad-CAM provided modality-specific explanations while retaining the local heterogeneous components. The current study has several limitations. The framework is evaluated using four clients and should be investigated with a larger number of institutions and modalities. Only SPECT has a verified supervised disease target in the implemented experiment, while the remaining clients are evaluated using representation-learning objectives. Future work can investigate downstream clinical tasks on the self-supervised clients, adaptive latent dimensions, modality-aware aggregation, larger multi-institutional datasets, additional privacy mechanisms, and clinical validation. More comprehensive evaluation of explanation faithfulness, robustness, and stability can also strengthen the interpretability analysis.

8 Acknowledgment

The authors thank their respective institutions for providing support and computational resources for the experimental work. The authors also thank the Editor-in-Chief and Associate Editors for their time and consideration of the submitted manuscript.

REFERENCES

Federated self-supervised learning for heterogeneous clients. In *International Conference on Learning Representations*, 2023.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, pp. 9505–9515, 2018.

Samiul Alam, Luyang Liu, Ming Yan, and Mi Zhang. Fedrolex: Model-heterogeneous federated learning with rolling sub-model extraction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Jonah Anton, Quinn O'Neil, Akshay Chaudhari, et al. How well do self-supervised models transfer to medical imaging? *Journal of Imaging*, 8(12):320, 2022. doi: 10.3390/jimaging8120320.

Ivo Baltruschat. Presorted brats2021 dataset, 2023. URL https://doi.org/10.5281/zenodo.8141526.

BENKHADDA. Dat parkinson, 2026. URL https://doi.org/10.5281/zenodo.21875852.

A. Botnari, M. Kadar, and J. M. Patrascu. Considerations on image preprocessing techniques required by deep learning models. the case of the knee mris. *Maedica*, 19(3):526–535, 2024. doi: 10.26574/maedica.2024.19.3.526.

Captum Contributors. Layergradcam. https://captum.ai/api/layer.html, 2025. Captum Model Interpretability for PyTorch.

Liam Collins, Hamed Hassani, Aryan Mokhtari, and Sanjay Shakkottai. Exploiting shared representations for personalized federated learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2089–2099. PMLR, 2021.

Tushar Dudeja and Prachi Sharma. Geometry-aware localization evaluation of grad-cam for wafer map defect classification. *Scientific Reports*, 16(1):23691, 2026. doi: 10.1038/s41598-026-52885-x. URL https://doi.org/10.1038/s41598-026-52885-x.

Mahsa Fallahpoor et al. Deep learning techniques in pet/ct imaging: A comprehensive review from sinogram to image space. *Computer Methods and Programs in Biomedicine*, 243:107880, 2024. doi: 10.1016/j.cmpb.2023.107880.

Filip Hanzely, Slavomir Hanzely, Samuel Horvath, and Peter Richtarik. Lower bounds and optimal algorithms for personalized federated learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Indian Institute of Science Bangalore. Vindr chest x-ray dataset, 2024. URL https://doi.org/10.5281/zenodo.12721389.

Ashima Jaiswal, Ashwin Kumar Babu, Mohammad Zaki Zadeh, Debarun Banerjee, and Fillia Makedon. A survey on self-supervised learning: Auxiliary pretext tasks and contrastive learning methods in imaging. *IEEE Access*, 9:144248–144278, 2021.

A. Khan et al. Saliency-driven explainable deep learning in medical imaging: bridging visual explainability and statistical quantitative analysis. *BMC Bioinformatics*, 2024.

Haoliang Li, Yufei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex C. Kot. Domain generalization for medical imaging classification with linear-dependency regularization. In *Advances in Neural Information Processing Systems*, volume 33, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020b.

Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2021.

Zexi Li, Tao Lin, Xinyi Shang, and Chao Wu. Revisiting weighted aggregation in federated learning with neural networks. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19767–19788. PMLR, 2023.

Samira Masoudi, Stephanie A. Harmon, Sherif Mehralivand, Stephanie M. Walker, and Baris Turkbey. Quick guide on radiology image pre-processing for deep learning applications in prostate cancer research. *Journal of Medical Imaging*, 8(1):010901, 2021. doi: 10.1117/1.JMI.8.1.010901.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 2017.

Medical University of Vienna. Total-body [18f]fdg-pet/ct imaging of healthy controls: Test/retest data for systemic, multi-organ analysis, 2025. URL `https://doi.org/10.5281/zenodo.16588733`.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.2599820.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017.

Arnaud Arindra Setio, Alberto Traverso, Thomas de Bel, Alexander S. Berens, Vincent E. van den Bogaard, Piergiorgio Cerello, Hao Chen, Qi Dou, Maria Elena Fantacci, Jacobus Geleijns, et al. Lung nodule detection in ct using 3d convolutional neural networks. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, 2017. doi: 10.1109/ISBI.2017.7950542.

Hongming Shan, Yi Zhang, Qingsong Yang, Uwe Kruger, Mannudeep K. Kalra, Ling Sun, Wenxiang Cong, and Ge Wang. 3-d convolutional encoder-decoder network for low-dose ct via transfer learning from a 2-d trained network. *IEEE Transactions on Medical Imaging*, 37(6):1522–1534, 2018. doi: 10.1109/TMI.2018.2832217.

Yujun Shi, Jian Liang, Wenqing Zhang, Vincent Y. F. Tan, and Song Bai. Towards understanding and mitigating dimensional collapse in heterogeneous federated learning. In *International Conference on Learning Representations*, 2023.

Satya P. Singh, Lipo Wang, Sukrit Gupta, Haveesh Goli, Parasuraman Padmanabhan, and Balazs Gulyas. 3d deep learning on medical images: A review. *Sensors*, 20(18):5097, 2020. doi: 10.3390/s20185097.

Benyuan Sun, Hongxing Huo, Yi Yang, and Bo Bai. Partialfed: Cross-domain personalized federated learning via partial initialization. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Aiham Taleb, Winfried Loetzsch, Noel Danz, Julius Severin, Thomas Gaertner, Benjamin Bergner, and Christoph Lippert. 3d self-supervised methods for medical imaging. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Valeria Trojani, Maria Chiara Bassi, Laura Verzellesi, and Marco Bertolini. Impact of preprocessing parameters in medical imaging-based radiomic studies: A systematic review. *Cancers*, 16(15): 2668, 2024. doi: 10.3390/cancers16152668.

Hao Wang et al. Improving generalization and personalization in model-heterogeneous federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):88–101, 2025. doi: 10.1109/TNNLS.2024.3405190.

Yancheng Wang, Rajeev Goel, Utkarsh Nath, Alvin C. Silva, Teresa Wu, and Yingzhen Yang. Learning low-rank feature for thorax disease classification. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3719.

Luyuan Xie, Manqing Lin, Tianyu Luan, Cong Li, Yuejian Fang, Qingni Shen, and Zhonghai Wu. Mh-pflid: Model heterogeneous personalized federated learning via injection and distillation for medical data analysis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54561–54575. PMLR, 2024.

Jian Xu, Xinyi Tong, and Shao-Lun Huang. Personalized federated learning with feature alignment and classifier collaboration. In *International Conference on Learning Representations*, 2023.

Juezhao Yu, Bohan Yang, Jing Wang, Joseph Leader, David Wilson, and Jiantao Pu. 2d cnn versus 3d cnn for false-positive reduction in lung cancer screening. *Journal of Medical Imaging*, 7(5): 051202, 2020. doi: 10.1117/1.JMI.7.5.051202.

Zongwei Zhou et al. Reinventing 2d convolutions for 3d images. *IEEE Transactions on Medical Imaging*, 2021.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. An empirical study of federated self-supervised learning. In *International Conference on Learning Representations*, 2022.

APPENDIX

A DETAILED DATASET STATISTICS

This appendix provides the complete dataset descriptions, partitioning strategy, preprocessing details, architectural configuration, federated training procedure, evaluation metrics, extended numerical results, and additional explainability analysis associated with the proposed xHetMed-FL framework.

A.1 OVERVIEW OF THE FOUR FEDERATED CLIENTS

The proposed framework consists of four heterogeneous medical-imaging clients. Each client corresponds to a different imaging modality and retains its own modality-specific preprocessing and feature-extraction pipeline. The four clients are:

1. **Client 1 – SPECT:** supervised pathological versus non-pathological classification.
2. **Client 2 – Chest X-ray:** self-supervised representation learning.
3. **Client 3 – MRI:** self-supervised representation learning using paired T1- and T2-weighted volumes.
4. **Client 4 – CT+PET:** self-supervised representation learning using paired CT and PET volumes.

The main characteristics of the four clients are summarized in Table 9.

Table 9: Complete overview of the heterogeneous federated clients.

Client	Modality	Dim.	Channels	Input Size	Learning	Latent Dim.
1	SPECT	3D	2	128^3	Supervised	128
2	Chest X-ray	2D	1	224×224	Self-supervised	128
3	MRI	3D	2	128^3	Self-supervised	128
4	CT+PET	3D	2	128^3	Self-supervised	128

A.2 CLIENT 1: SPECT DATASET

The SPECT client uses the SPECT dataset distributed through Zenodo BENKHADDA (2026). The dataset contains 1,362 preprocessed volumetric samples stored in NumPy format. Each sample contains two image channels with an original spatial size of $96 \times 96 \times 64$. The corresponding label file contains 1,362 records with the binary target variable `is_pathologic`. Among these samples, 747 are pathological and 615 are non-pathological. A stratified partition was performed using a fixed random seed of 42. The resulting dataset contains 953 training samples, 204 validation samples, and 205 testing samples. The class distribution was preserved across the partitions.

Table 10: Detailed SPECT dataset statistics.

Property	**Value**
Total samples	1,362
Pathological samples	747
Non-pathological samples	615
Original spatial size	$96 \times 96 \times 64$
Number of channels	2
Training samples	953
Validation samples	204
Testing samples	205
Learning paradigm	Supervised
Target variable	`is_pathologic`
Random seed	42

The SPECT samples were converted to the required volumetric representation and resized to $128 \times 128 \times 128$. Non-zero voxels were standardized using a z-score transformation. Since verified pathological and non-pathological labels are available, this client performs supervised binary classification.

A.3 CLIENT 2: CHEST X-RAY DATASET

Client 2 uses the VinDr Chest X-ray dataset distributed through Zenodo Indian Institute of Science Bangalore (2024). The source repository reports approximately 48.0 GB across the principal dataset archives. A locally prepared experimental subset of 3,000 images was used in the implemented experiment. The images were converted to grayscale and resized to 224×224 pixels. The experimental split contains 2,100 training images, 450 validation images, and 450 testing images. Because a verified disease classification target was not used for this client in the implemented federated objective, the Chest X-ray client performs self-supervised representation learning.

Table 11: Detailed Chest X-ray dataset statistics.

Property	**Value**
Experimental samples	3,000
Input modality	Chest X-ray
Input representation	Grayscale
Input size	224×224
Training samples	2,100
Validation samples	450
Testing samples	450
Learning paradigm	Self-supervised

A.4 CLIENT 3: MRI DATASET

Client 3 uses the presorted BraTS2021 MRI dataset distributed through Zenodo Baltruschat (2023). The experimental dataset contains 2,502 NIfTI files corresponding to 1,251 subjects. T1-weighted and T2-weighted MRI volumes are used as two complementary input channels. Thus, each subject contributes a paired two-channel volumetric input. The dataset is partitioned at the patient level so

that T1 and T2 volumes from the same subject always remain in the same subset. The resulting partition consists of 875 training subjects, 188 validation subjects, and 188 testing subjects.

Table 12: Detailed MRI dataset statistics.

Property	Value
Total NIfTI files	2,502
Unique subjects	1,251
Modalities	T1 + T2
Number of channels	2
Input size	$128 \times 128 \times 128$
Training subjects	875
Validation subjects	188
Testing subjects	188
Learning paradigm	Self-supervised
Partition level	Patient

The patient-level partitioning prevents the same subject from contributing data to multiple subsets and thereby reduces the possibility of subject-level information leakage.

A.5 Client 4: CT+PET Dataset

Client 4 uses the QUADRA healthy-control CT/PET dataset distributed through Zenodo Medical University of Vienna (2025). The source dataset contains CT and PET scans together with segmentation information from healthy subjects acquired using a Siemens Biograph Vision QUADRA system. The Zenodo release contains the `QUADRA_HC.zip` archive with a reported size of approximately 20.2 GB. CT and PET are processed as complementary imaging channels. PET volumes are spatially aligned with the CT representation. CT intensity values are clipped to the range $[-1000, 1000]$ HU, while PET values below zero are clamped to zero. The resulting volumes are resized to $128 \times 128 \times 128$. The experimental partition consists of 66 training studies, 14 validation studies, and 16 testing studies. Since verified disease labels are not used for this client, CT+PET performs self-supervised representation learning.

Table 13: Detailed CT+PET dataset statistics.

Property	Value
Imaging modalities	CT + PET
Number of channels	2
Input size	$128 \times 128 \times 128$
Training studies	66
Validation studies	14
Testing studies	16
Learning paradigm	Self-supervised
CT clipping range	$[-1000, 1000]$ HU
PET preprocessing	Negative values $\rightarrow 0$

A.6 Dataset Partitioning Strategy

For each client m, the local dataset is represented as

$$D_m = D_m^{\text{train}} \cup D_m^{\text{val}} \cup D_m^{\text{test}}, \tag{16}$$

where

$$D_m^{\text{train}} \cap D_m^{\text{val}} = D_m^{\text{train}} \cap D_m^{\text{test}} = D_m^{\text{val}} \cap D_m^{\text{test}} = \emptyset. \tag{17}$$

Only the training partition is used for local optimization. The validation partition is used for monitoring the federated learning process, while the test partition is reserved for final evaluation and explainability analysis.

For SPECT, stratification is used to preserve the pathological and non-pathological class proportions. For MRI, patient-level separation is used to ensure that paired T1/T2 volumes from a subject remain within the same partition. For CT+PET, subject/study-level separation is used so that paired CT/PET studies are not distributed across different subsets.

Table 14: Complete dataset partition used by the four federated clients.

Client	Training	Validation	Testing	Total
SPECT	953	204	205	1,362
Chest X-ray	2,100	450	450	3,000
MRI	875	188	188	1,251 subjects
CT+PET	66	14	16	96 studies

B Complete Modality-Specific Preprocessing

A central characteristic of xHetMed-FL is that the input processing pipeline is retained locally at each client. This is necessary because the four modalities have substantially different dimensionality and intensity characteristics.

B.1 SPECT Preprocessing

The SPECT data contain two volumetric channels with original dimensions $96 \times 96 \times 64$. The preprocessing pipeline consists of:

1. loading the two-channel volumetric representation;
2. identifying the non-zero voxel region;
3. applying z-score normalization to the non-zero voxels;
4. resizing the volume to $128 \times 128 \times 128$;
5. retaining the two channels as the input to the 3D CNN.

The final SPECT input therefore has the form $X_{\mathrm{SPECT}} \in \mathbb{R}^{2\times128\times128\times128}$.

B.2 Chest X-ray Preprocessing

The Chest X-ray preprocessing pipeline consists of:

1. loading the X-ray image;
2. converting the image to grayscale;
3. resizing the image to 224×224;
4. applying min–max intensity normalization;
5. supplying the resulting single-channel image to the 2D CNN.

The final input is therefore $X_{\mathrm{Xray}} \in \mathbb{R}^{1\times224\times224}$.

B.3 MRI Preprocessing

The MRI pipeline uses paired T1- and T2-weighted volumes. The two modalities are retained as separate input channels: $X_{\mathrm{MRI}} = [X_{\mathrm{T1}}, X_{\mathrm{T2}}]$. The volumes are resized to $128 \times 128 \times 128$ and supplied jointly to the local 3D CNN. The resulting input has dimensions $X_{\mathrm{MRI}} \in \mathbb{R}^{2\times128\times128\times128}$. The T1 and T2 volumes belonging to the same patient are always processed together.

B.4 CT+PET PREPROCESSING

CT and PET are processed as paired complementary channels. The CT intensity range is restricted to $I_{\mathrm{CT}} \in [-1000, 1000]$ HU. PET values below zero are clamped to zero. The processed CT and PET volumes are spatially aligned and resized to 128^3. The final representation is $X_{\mathrm{CTPET}} \in \mathbb{R}^{2\times128\times128\times128}$. This modality-specific preprocessing is performed locally and is never communicated to the federated server.

C COMPLETE MODALITY-SPECIFIC ENCODER ARCHITECTURE

C.1 GENERAL ARCHITECTURE

The four clients use heterogeneous encoders because the dimensionality of their imaging data differs. SPECT, MRI, and CT+PET use 3D CNN encoders, while Chest X-ray uses a 2D CNN encoder. Despite this architectural heterogeneity, every encoder produces a 128-dimensional feature vector:

$$h_m = E_m(X_m), \qquad h_m \in \mathbb{R}^{128}. \tag{18}$$

The encoders remain local throughout federated training.

C.2 3D CNN ARCHITECTURE

The SPECT, MRI, and CT+PET clients use a common structural 3D CNN design. The encoder contains four convolutional blocks. The first three blocks consist of 3D Conv $\rightarrow$ BatchNorm $\rightarrow$ ReLU $\rightarrow$ 3D MaxPool.

The fourth block consists of 3D Conv $\rightarrow$ BatchNorm $\rightarrow$ ReLU.

The feature channels increase progressively: $16 \rightarrow 32 \rightarrow 64 \rightarrow 128$.

The final convolutional layer therefore generates a 128-channel three-dimensional feature map. This feature map is retained because it is also required by the Grad-CAM explainability module. Adaptive global average pooling is subsequently applied:

$$F_{\mathrm{GAP}} = \mathrm{AdaptiveAvgPool3D}(F), \tag{19}$$

reducing the spatial dimensions to $1 \times 1 \times 1$. The resulting feature vector is flattened and passed through a fully connected layer to produce the 128-dimensional representation.

C.3 2D CNN ARCHITECTURE FOR CHEST X-RAY

The Chest X-ray client uses a two-dimensional CNN because the input is a planar image. The local architecture consists of stacked 2D convolutional processing blocks with Batch Normalization, nonlinear activation, and spatial downsampling.

The final convolutional feature map is retained for the 2D Grad-CAM calculation. Adaptive global average pooling converts the feature map into a compact representation, followed by a fully connected layer that produces the shared 128-dimensional feature vector.

Thus,

$$h_{\mathrm{Xray}} = E_{\mathrm{Xray}}(X_{\mathrm{Xray}}) \in \mathbb{R}^{128}. \tag{20}$$

The complete encoder remains local and is never averaged with the three-dimensional encoders.

C.4 SHARED REPRESENTATION INTERFACE

Although the local encoders are heterogeneous, they all expose the same interface:

$$E_m : X_m \to h_m, \qquad h_m \in \mathbb{R}^{128}. \tag{21}$$

This common feature dimensionality allows the same shared projection architecture to be used by all four clients.

D Complete Shared Projection and Learning Objectives

The modality-specific feature vector is transformed into a common latent representation using the shared projection

$$z_m = P(h_m). \tag{22}$$

The implemented projection is

$$P(h) = W_2 \,\mathrm{ReLU}\,(\mathrm{LN}(W_1 h + b_1)) + b_2, \tag{23}$$

where W_1 and W_2 are learnable projection matrices, b_1 and b_2 are bias terms, and LN denotes Layer Normalization.

The complete local representation pipeline is

$$X_m \to E_m \to h_m \to P \to z_m. \tag{24}$$

The projection output has dimensionality $z_m \in \mathbb{R}^{128}$. Only the parameters of P participate in federated aggregation.

D.1 SPECT Supervised Objective

For SPECT,

$$\hat{y}_{\mathrm{SPECT}} = H_{\mathrm{SPECT}}(z_{\mathrm{SPECT}}), \tag{25}$$

where H_{SPECT} is a two-class local prediction head.

The classification objective is

$$\mathcal{L}_{\mathrm{SPECT}} = -\sum_{c=1}^{2} y_c \log(\hat{y}_c). \tag{26}$$

D.2 Self-Supervised Objective

For the three unlabeled clients, two corresponding representations are obtained $z_m^{(1)}, z_m^{(2)} \in \mathbb{R}^{128}$.

Their cosine similarity is

$$\mathrm{CosSim} = \frac{z_m^{(1)\top} z_m^{(2)}}{\|z_m^{(1)}\|_2 \|z_m^{(2)}\|_2}. \tag{27}$$

The corresponding representation loss is

$$\mathcal{L}_{\mathrm{rep}} = 1 - \mathrm{CosSim}\left(z_m^{(1)}, z_m^{(2)}\right). \tag{28}$$

Therefore,

$$\mathcal{L}_m = \begin{cases} \mathcal{L}_{\text{SPECT}}, & m = \text{SPECT}, \\ \mathcal{L}_{\text{rep}}, & m \in \{\text{Xray}, \text{MRI}, \text{CTPET}\}. \end{cases} \tag{29}$$

E Complete FedProx and FedBN Formulation

The heterogeneous client setting requires the local optimization process to account for differences in both data distributions and model architectures. FedProx is therefore applied to the shared projection parameters.

Let θ_m denote the locally updated shared projection parameters and $\theta^{(t)}$ denote the global parameters at round t. The FedProx objective is

$$\mathcal{L}_m^{\text{FedProx}} = \mathcal{L}_m + \frac{\mu}{2} \left\| \theta_m - \theta^{(t)} \right\|_2^2, \tag{30}$$

where $\mu = 0.01$.

The complete client objective is consequently

$$\mathcal{L}_m^{\text{FedProx}} = \begin{cases} \mathcal{L}_{\text{SPECT}} + \frac{\mu}{2} \|\theta_m - \theta^{(t)}\|_2^2, & m = 1, \\ \mathcal{L}_{\text{rep}} + \frac{\mu}{2} \|\theta_m - \theta^{(t)}\|_2^2, & m = 2, 3, 4. \end{cases} \tag{31}$$

E.1 Parameter Decomposition

The complete parameter set of client m is

$$\Theta_m = \left\{ \Theta_m^E, \Theta_m^P, \Theta_m^H, \Theta_m^{BN} \right\}. \tag{32}$$

The federated and local parameter sets are

$$\Theta_m^{\text{fed}} = \Theta_m^P, \tag{33}$$

and

$$\Theta_m^{\text{local}} = \left\{ \Theta_m^E, \Theta_m^H, \Theta_m^{BN} \right\}. \tag{34}$$

Thus, the modality-specific encoder, local head, and Batch Normalization parameters never leave their respective clients.

E.2 FedBN

Batch Normalization parameters and running statistics are maintained locally $\Theta_m^{BN} \notin \Theta^{\text{fed}}$. This allows each modality to preserve its own feature-distribution statistics and avoids averaging normalization statistics across fundamentally different medical imaging modalities.

F Complete Weighted Federated Aggregation

Let n_m denote the number of local training samples for client m. The total number of training samples is

$$N = \sum_{m=1}^{4} n_m. \tag{35}$$

The aggregation weight is

$$w_m = \frac{n_m}{\sum_{j=1}^{4} n_j}. \tag{36}$$

The implemented aggregation weights are $w_{\mathrm{SPECT}} = 0.2386$, $w_{\mathrm{Xray}} = 0.5258$, $w_{\mathrm{MRI}} = 0.2191$, $w_{\mathrm{CTPET}} = 0.0165$. The global shared projection is updated according to

$$\Theta_G^{(t+1)} = \sum_{m=1}^{4} w_m \Theta_{m,P}^{(t+1)}. \tag{37}$$

The aggregated projection is then redistributed:

$$\Theta_{m,P}^{(t+1)} \leftarrow \Theta_G^{(t+1)}. \tag{38}$$

No encoder, prediction-head, or Batch Normalization parameter is included in this aggregation.

G Complete Federated Training Algorithm

The complete federated procedure is given in Algorithm 1.

Algorithm 1 xHetMed-FL Training with FedProx and FedBN

Require: Clients $\mathcal{C} = \{C_1, C_2, C_3, C_4\}$, datasets $\{\mathcal{D}_m\}_{m=1}^4$, rounds $T = 10$, local epochs $E = 1$, learning rate $\eta = 10^{-4}$, FedProx coefficient $\mu = 0.01$
Ensure: Global shared projection Θ_G

Initialize $\Theta_G^{(0)}$
for $t = 0$ to $T - 1$ **do**
  for each client $m \in \{1, 2, 3, 4\}$ **do**
    Receive $\Theta_G^{(t)}$
    Load $\Theta_G^{(t)}$ into local shared projection
    Preserve local encoder E_m
    Preserve local head H_m
    Preserve local BatchNorm parameters Θ_m^{BN}
    for $e = 1$ to E **do**
      for each local batch B_m **do**
        Generate local feature: $h_m = E_m(X_m)$
        Generate shared representation: $z_m = P(h_m)$
        if $m = 1$ **then**
          Generate SPECT prediction: $\hat{y}_m = H_m(z_m)$
          Compute $\mathcal{L}_m = \mathcal{L}_{\text{SPECT}}$
        else
          Generate $z_m^{(1)}$ and $z_m^{(2)}$
          Compute $\mathcal{L}_m = 1 - \text{CosSim}(z_m^{(1)}, z_m^{(2)})$
        end if
        Add FedProx term:

$$\mathcal{L}_m^{\text{FedProx}} = \mathcal{L}_m + \frac{\mu}{2}\|\Theta_{m,P} - \Theta_G^{(t)}\|_2^2$$

        Update local parameters using learning rate η
      end for
    end for
    Retain E_m, H_m, and Θ_m^{BN} locally
    Send only $\Theta_{m,P}^{(t+1)}$ to server
  end for
  Compute sample-based weights w_m
  Aggregate:

$$\Theta_G^{(t+1)} = \sum_{m=1}^{4} w_m \Theta_{m,P}^{(t+1)}$$

  Redistribute $\Theta_G^{(t+1)}$
end for
return $\Theta_G^{(T)}$

H Complete Evaluation Metrics

H.1 Classification Metrics

For SPECT, the following metrics are calculated.

H.1.1 Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{39}$$

H.1.2 PRECISION

$$\text{Precision} = \frac{TP}{TP + FP}. \tag{40}$$

H.1.3 RECALL

$$\text{Recall} = \frac{TP}{TP + FN}. \tag{41}$$

H.1.4 F1-SCORE

$$\text{F1} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{42}$$

H.1.5 MATTHEWS CORRELATION COEFFICIENT

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{43}$$

H.1.6 ROC-AUC

The receiver operating characteristic curve is generated from the predicted probabilities, and its area under the curve is reported as ROC-AUC.

H.2 REPRESENTATION METRICS

For the self-supervised clients, the primary evaluation quantities are representation loss and cosine similarity:

$$\mathcal{L}_{\text{rep}} = 1 - \text{CosSim}(z^{(1)}, z^{(2)}), \tag{44}$$

where

$$\text{CosSim} = \frac{z^{(1)\top} z^{(2)}}{\|z^{(1)}\|_2 \|z^{(2)}\|_2}. \tag{45}$$

A lower representation loss and higher cosine similarity indicate greater agreement between the corresponding learned representations under the adopted self-supervised objective.

The evaluation therefore follows:

$$\boxed{\begin{aligned} \text{SPECT} &\rightarrow \text{Classification Metrics}, \\ \text{Xray},\ \text{MRI},\ \text{CT} + \text{PET} &\rightarrow \text{Representation Metrics}. \end{aligned}} \tag{46}$$

I COMPLETE GRAD-CAM FORMULATION

The xHetMed-FL framework uses Captum's `LayerGradCam` implementation to generate post-training explanations Selvaraju et al. (2017); Captum Contributors (2025).

The Grad-CAM target depends on the client:

$$s_m = \begin{cases} y^c_{\text{SPECT}}, & m = \text{SPECT}, \\ \|z_m\|_2, & m \in \{\text{Xray}, \text{MRI}, \text{CTPET}\}. \end{cases} \tag{47}$$

For activation map A^k, its importance weight is

$$\alpha_k = \frac{1}{Z} \sum_i \sum_j \sum_l \frac{\partial s}{\partial A^k_{ijl}}. \tag{48}$$

The Grad-CAM map is

$$L_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha_k A^k \right). \tag{49}$$

The resulting map is upsampled to the corresponding input resolution and normalized to $[0, 1]$.

I.1 SPECT GRAD-CAM

For SPECT,

$$s_{\text{SPECT}} = y^c_{\text{SPECT}}. \tag{50}$$

The final 3D convolutional feature map is used as the target layer. The input has dimensions $2 \times 128 \times 128 \times 128$. The resulting attribution volume is resized to $128 \times 128 \times 128$.

The two input channels are visualized separately.

I.2 CHEST X-RAY GRAD-CAM

For Chest X-ray,

$$s_{\text{Xray}} = \|z_{\text{Xray}}\|_2. \tag{51}$$

The resulting attribution map is $L_{\text{Xray}} \in \mathbb{R}^{224 \times 224}$.

The map is normalized and overlaid on the original Chest X-ray.

I.3 MRI GRAD-CAM

For MRI,

$$s_{\text{MRI}} = \|z_{\text{MRI}}\|_2. \tag{52}$$

The resulting three-dimensional attribution volume is resized to $128 \times 128 \times 128$ and visualized separately for the T1 and T2 channels, typically using the central axial slice.

I.4 CT+PET GRAD-CAM

For CT+PET,

$$s_{\text{CTPET}} = \|z_{\text{CTPET}}\|_2. \tag{53}$$

The attribution volume is generated from the final 3D convolutional layer and visualized separately for CT and PET.

J QUANTITATIVE EXPLAINABILITY ANALYSIS

Visual Grad-CAM maps are complemented by quantitative attribution statistics.

Let G represent a normalized attribution map. The threshold is defined as

Table 15: Complete Grad-CAM configuration across clients.

Client	Grad-CAM	Target
SPECT	3D	Predicted disease class
Chest X-ray	2D	$\|z\|_2$
MRI	3D	$\|z\|_2$
CT+PET	3D	$\|z\|_2$

$$T = 0.5 \max(G). \tag{54}$$

An attribution value is classified as important when

$$G_i \geq T. \tag{55}$$

The number of important pixels or voxels is

$$N_{\text{important}} = \sum_i \mathbb{I}(G_i \geq T). \tag{56}$$

The important-region ratio is

$$R_{\text{important}} = \frac{N_{\text{important}}}{N_{\text{total}}} \times 100. \tag{57}$$

The mean attribution is

$$\mu_G = \frac{1}{N} \sum_{i=1}^{N} G_i, \tag{58}$$

and the standard deviation is

$$\sigma_G = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (G_i - \mu_G)^2}. \tag{59}$$

For the latent representation $z \in \mathbb{R}^{128}$,

$$\mu_z = \frac{1}{128} \sum_{j=1}^{128} z_j, \tag{60}$$

and

$$\text{Var}(z) = \frac{1}{128} \sum_{j=1}^{128} (z_j - \mu_z)^2. \tag{61}$$

The complete quantitative explainability descriptor is

$$\boxed{\{\mu_G, \sigma_G, N_{\text{important}}, R_{\text{important}}, \mu_z, \sigma_z, \text{Var}(z), z_{\min}, z_{\max}\}.} \tag{62}$$

Table 16: Extended communication-round validation results for the SPECT client.

Round	Loss	Accuracy	Precision	Recall	F1	AUC	MCC
1	0.5459	0.8235	0.8958	0.7679	0.8269	0.8971	0.6571
2	0.4758	0.8333	0.9333	0.7500	0.8317	0.8999	0.6862
3	0.4375	0.8186	0.8788	0.7768	0.8246	0.9058	0.6435
4	0.5533	0.6961	0.6488	0.9732	0.7786	0.9043	–
7	0.3790	0.8480	0.9263	0.7857	0.8502	0.9058	0.7079
8	0.3741	0.8186	0.8440	0.8214	0.8326	0.9171	0.6351
9	0.4549	0.7990	0.7630	0.9196	0.8340	0.9136	–
10	0.3921	0.8235	0.8878	0.7768	0.8286	0.9177	0.6546

K Complete Round-Wise Experimental Results

The supervised SPECT client was evaluated after each reported communication round. The results are presented in Table 16.

The reported results show that federated training does not produce strictly monotonic changes in every metric. For example, the validation accuracy reaches 84.80% at round 7 but decreases to 82.35% at round 10. The ROC-AUC, however, reaches 0.9177 at the final reported round.

The observed variation is consistent with the heterogeneous nature of the four clients and the fact that only the shared projection is aggregated while modality-specific components remain local.

K.1 Final SPECT Results

The final validation results after ten communication rounds are:

Table 17: Final SPECT validation results.

Metric	Value
Loss	0.3921
Accuracy	82.35%
Precision	88.78%
Recall	77.68%
F1-score	82.86%
MCC	0.6546
ROC-AUC	0.9177

K.2 Self-Supervised Client Results

The final representation-learning results are presented in Table 18.

Table 18: Final representation-learning results for self-supervised clients.

Client	Representation Loss	Cosine Similarity
Chest X-ray	0.0000	1.0000
MRI	0.0000	1.0000
CT+PET	0.0008	0.9992
Average	0.0003	0.9997

The results indicate very high similarity between the corresponding representations under the adopted self-supervised objective. These values should be interpreted as measurements of representation agreement under the implemented cosine-similarity objective rather than as disease classification performance.

L Quantitative Grad-CAM Results

The representative test samples from each client were analyzed using the quantitative Grad-CAM procedure described above.

Table 19: Quantitative Grad-CAM and latent-representation statistics for representative test samples.

Client	Mean Attribution	Std. Attribution	Important Pixels/Voxels	Ratio (%)	Latent Mean	Latent Variance
SPECT	0.069746	0.154910	81,367	3.88	–	0.240568
Chest X-ray	0.057282	0.129690	1,341	2.67	0.021806	0.318496
MRI	0.013345	0.055196	2,626	0.13	-0.034253	0.316800
CT+PET	0.101050	0.105035	1,626	0.08	-0.062608	0.296452

L.1 SPECT Attribution Statistics

The representative SPECT sample produced a mean attribution of 0.069746 and standard deviation of 0.154910. A total of 81,367 voxels were identified as important using the threshold of 0.5 times the maximum attribution. This corresponds to 3.88% of the complete 128^3 spatial volume.

The associated latent representation had a variance of 0.240568. Since the Grad-CAM target for SPECT is the predicted disease class, the attribution map is interpreted as a class-oriented explanation.

L.2 Chest X-ray Attribution Statistics

For the representative Chest X-ray sample, the mean attribution was 0.057282 and the standard deviation was 0.129690. A total of 1,341 pixels were classified as important out of 50,176 image pixels, corresponding to 2.67%.

The associated latent representation had $\mu_z = 0.021806$, and $\mathrm{Var}(z) = 0.318496$.

Because this client uses a representation-level Grad-CAM target, the attribution should be interpreted as indicating regions contributing to the learned representation rather than a disease-class decision.

L.3 MRI Attribution Statistics

For the representative MRI sample, the mean attribution was 0.013345 and the standard deviation was 0.055196. A total of 2,626 voxels were identified as important out of 2,097,152 voxels, corresponding to 0.13%. The latent representation statistics were $\mu_z = -0.034253$, $\sigma_z = 0.562850$, and $\mathrm{Var}(z) = 0.316800$.

L.4 CT+PET Attribution Statistics

For CT+PET, the mean attribution was 0.101050 and the standard deviation was 0.105035. A total of 1,626 voxels were identified as important out of 2,097,152 voxels, corresponding to 0.08%. The latent representation statistics were $\mu_z = -0.062608$, $\sigma_z = 0.544474$, and $\mathrm{Var}(z) = 0.296452$.

M Additional Explainability Figures

M.1 SPECT Grad-CAM

The representative SPECT sample was correctly classified, with both ground-truth and predicted classes equal to 1. Because SPECT is the supervised client, the Grad-CAM target is the predicted disease class.

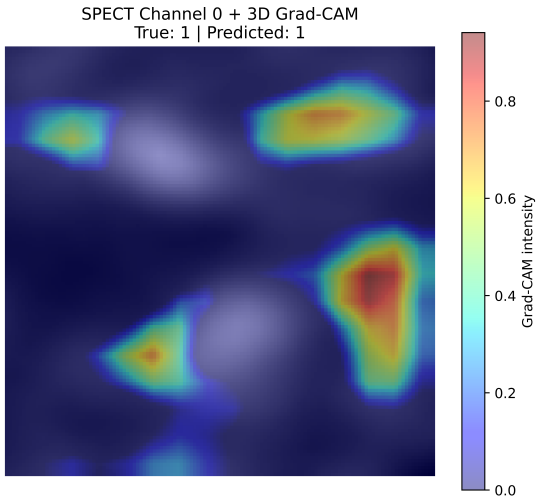


Figure 2: Representative 3D Grad-CAM visualization for the SPECT client. The sample has ground-truth class 1 and predicted class 1. The highlighted regions correspond to spatial regions contributing to the predicted disease-class output.



(a) Intensity heatmap (b) Heatmap overlay (c) Normalized representation

Figure 3: Qualitative visualization of the Chest X-ray representation, including the intensity heatmap, heatmap overlay, and normalized representation. The visualization corresponds to the representation-level Grad-CAM target used by the self-supervised Chest X-ray client.

M.2 Chest X-ray Grad-CAM

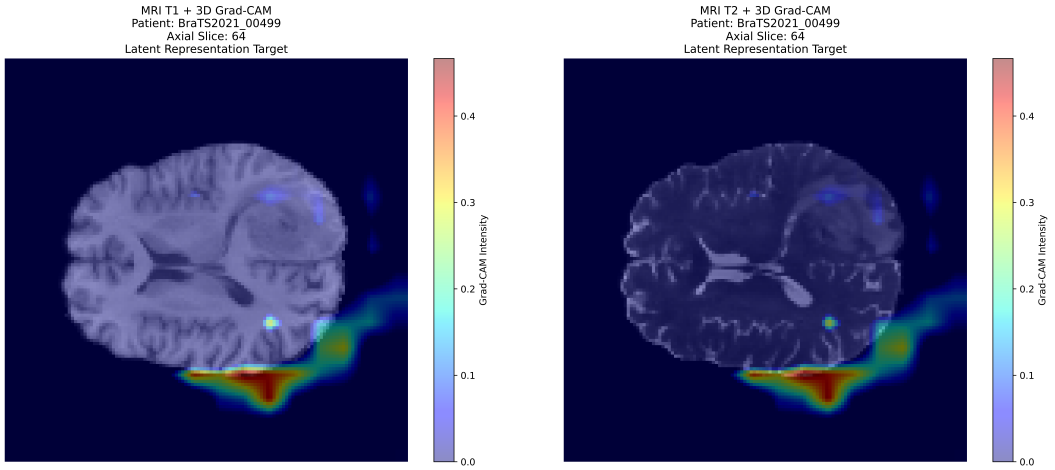
The Chest X-ray client uses a two-dimensional CNN. The magnitude of the 128-dimensional latent representation is used as the scalar Grad-CAM target. Consequently, the visualization identifies spatial regions contributing to the learned representation.

M.3 MRI Grad-CAM

The T1 and T2 volumes are processed jointly as a two-channel input while their attribution maps are visualized independently. Because the MRI client is self-supervised, the highlighted regions correspond to the learned representation rather than a disease-specific classification output.

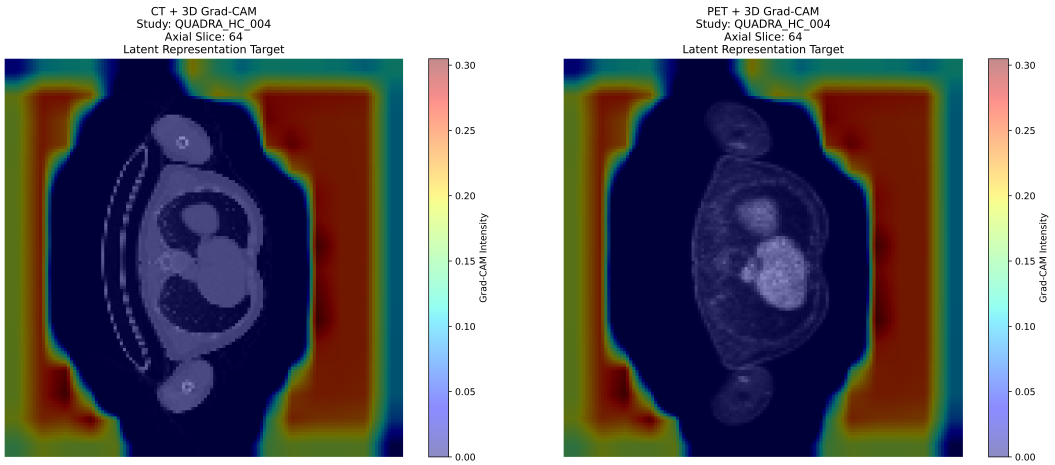
M.4 CT+PET Grad-CAM

Although CT and PET are jointly processed as a two-channel input, their Grad-CAM maps are visualized separately to examine modality-specific spatial contributions.



(a) T1-weighted MRI (b) T2-weighted MRI

Figure 4: Representative 3D Grad-CAM visualization for BraTS2021 subject 00499. The T1- and T2-weighted channels are shown at the same axial slice. The highlighted regions represent spatial contributions to the learned latent representation.



(a) CT (b) PET

Figure 5: Representative 3D Grad-CAM visualization for the CT+PET client using the QUADRA_HC_004 study. Attribution maps are shown separately for the CT and PET channels and correspond to contributions to the learned latent representation.

N Implementation Configuration

The principal experimental configuration is summarized in Table 20.

O Federated Communication Details

At the beginning of every communication round, the central server broadcasts the current shared projection to all four clients. Each client loads the received projection into its local architecture while retaining its own modality-specific encoder and Batch Normalization statistics.

The communication sequence is

Table 20: Complete implementation and federated-training configuration.

Configuration	Value
Number of clients	4
Federated rounds	10
Local epochs per round	1
Learning rate	10^{-4}
FedProx coefficient	0.01
Shared latent dimension	128
Random seed	42
SPECT encoder	3D CNN
Chest X-ray encoder	2D CNN
MRI encoder	3D CNN
CT+PET encoder	3D CNN
Federated parameters	Shared projection
Local parameters	Encoder + head + BatchNorm
Aggregation	Sample-weighted
Normalization strategy	FedBN
Optimization strategy	FedProx
Explainability	Grad-CAM / LayerGradCam
Total training time	1570.01 min

$$\Theta_G^{(t)} \rightarrow \{\Theta_{1,P}^{(t)}, \Theta_{2,P}^{(t)}, \Theta_{3,P}^{(t)}, \Theta_{4,P}^{(t)}\}. \tag{63}$$

After local optimization,

$$\{\Theta_{1,P}^{(t+1)}, \Theta_{2,P}^{(t+1)}, \Theta_{3,P}^{(t+1)}, \Theta_{4,P}^{(t+1)}\} \rightarrow \text{Server}. \tag{64}$$

The server then performs

$$\Theta_G^{(t+1)} = \sum_m w_m \Theta_{m,P}^{(t+1)}. \tag{65}$$

The resulting global projection is redistributed for the subsequent round.

Consequently, the complete communication architecture can be represented as

$$\boxed{\text{Local Encoder} + \text{Local BN} + \text{Local Head} + \text{Shared Projection}}$$

with only the shared projection participating in inter-client communication.