Secure Federated Learning for Multi-Institutional Biomedical Image Analysis Through Communication-Efficient Neural Network Optimization and Confidential Model Inference

Blair Stace
Department of Computer Science, George Mason University, Fairfax, VA, USA.
contactblair@gmu.edu

Huseyin Budig
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
huseyinbudig@missouri.edu

Caomhan Kirke
School of Computing, Clemson University, Clemson, SC, USA.
caomhan.kirke458@clemson.edu

Abstract

Multi-institutional biomedical image analysis requires learning from distributed patient data while respecting legal, ethical, and operational constraints that prohibit centralized pooling. Federated learning addresses the data-governance problem, but practical deployment in imaging is constrained by high-dimensional model updates, heterogeneous institutional data, and the need to protect both training updates and post-training inference exchanges. This manuscript develops an integrated methodological framework for secure federated biomedical image analysis that couples communication-efficient neural network optimization with confidential model inference. The training design combines error-feedback compression, adaptive local optimization, secure aggregation, and client-level differential privacy under a threat model that separates honest-but-curious servers, participating institutions, and inference clients. The inference design treats nonlinear activation evaluation as the dominant cryptographic bottleneck and incorporates polynomial approximation and automated activation replacement to reduce private-inference cost while preserving the diagnostic mapping learned by the federated model. The analysis derives communication, convergence, privacy, and inference-cost relationships from explicit assumptions, including non-identity compression, bounded gradient noise, and differential privacy composition. The central insight is that communication efficiency, privacy preservation, and confidential inference are not separable add-ons; they jointly determine the feasible operating region of a multi-institutional imaging system. The framework provides design conditions for selecting compression, privacy, and cryptographic parameters, and it clarifies where approximation error, client drift, secure-aggregation overhead, and private nonlinear evaluation trade against diagnostic reliability.

Keywords

federated learning; biomedical imaging; communication-efficient optimization; secure aggregation; confidential inference; differential privacy

1. Introduction

Multi-institutional biomedical image analysis is constrained by a structural conflict between data scale and data governance. Diagnostic imaging models benefit from diverse scanners, acquisition protocols, patient populations, and annotation practices, yet patient images are rarely transferable across institutions without extensive legal and ethical review. Federated learning provides a coordination mechanism in which institutions train a shared model through local computation and periodic update exchange rather than through raw data transfer [1]. The approach has moved from proof-of-concept demonstrations to a broader research programme covering systems, optimization, privacy, and incentive design, with several open problems remaining under realistic heterogeneity and security assumptions [2]. In medicine, the appeal is direct: multi-institutional collaboration can increase training diversity while keeping patient records within institutional boundaries [3], and imaging consortia have used federated training to build models without centralizing scans [4].

The principal obstacle is that the federated setting changes the cost structure of learning. Model updates are high-dimensional, communication rounds are expensive, and institutions differ in label prevalence, scanner calibration, acquisition protocols, and case mix. Federated optimization methods such as proximal local objectives and adaptive server-side updates were developed to reduce client drift and improve robustness under heterogeneous data [5], [6]. Security introduces a second constraint. If the server observes individual updates, update inversion and membership inference can expose properties of local training data, so secure aggregation is used to reveal only the sum of client updates [7]. Differential privacy provides a complementary guarantee by bounding the influence of any single client or record on released quantities [8], [9]. These mechanisms are necessary but not sufficient for biomedical imaging, where the model itself may be a valuable institutional asset and where inference requests contain sensitive patient images.

Medical-imaging federated learning therefore requires a broader treatment than training-time privacy alone. Analyses of secure and privacy-preserving machine learning in medical imaging identify a tension between formal privacy guarantees, model utility, and practical deployment constraints [10]. Reviews of federated healthcare systems likewise emphasize that communication, heterogeneity, and trust assumptions must be addressed together rather than as isolated components [11]. The communication literature offers sparsification, quantization, and error-feedback methods that reduce the volume of exchanged updates while preserving convergence under bounded compression error [12], [13]. Adaptive optimization methods provide additional tools for handling scale differences across parameters and clients [14]. Yet these methods are usually evaluated with respect to training efficiency, not with respect to the cryptographic inference path that determines whether a trained model can be safely queried by external hospitals or clinical users.

Confidential inference adds a distinct threat surface. Neural networks can memorize training examples and leak information through outputs or intermediate representations [15], and encrypted inference must protect both the queried image and the model weights. Fully homomorphic encryption and secure multi-party computation can evaluate neural networks without revealing inputs or weights, but nonlinear activation functions dominate the computational and communication cost. Recent work on automatic activation replacement shows that carefully chosen replacements for rectified linear units can substantially reduce private-inference cost [16]. This observation is important for biomedical imaging because convolutional networks used in segmentation, detection, and classification contain many

activation layers, and the cost of private nonlinear evaluation grows with feature-map size and network depth.

The central contribution of this manuscript is an integrated framework for secure federated biomedical image analysis that treats communication-efficient training and confidential inference as coupled design problems. The framework formulates the federated training objective, specifies an error-feedback compression and secure-aggregation pipeline, incorporates client-level differential privacy under explicit trust assumptions, and develops a confidential inference path that uses polynomial approximation and activation replacement to control cryptographic cost. The analysis derives convergence and cost relationships from stated assumptions, identifies the conditions under which compression and privacy noise remain compatible with reliable imaging models, and clarifies how approximation error in private inference interacts with diagnostic decision thresholds. The intended contribution is methodological and analytical rather than a report of a multi-site deployment.

2. Related Work and Problem Setting

2.1 Federated learning for biomedical imaging

Federated learning was introduced as a communication-efficient alternative to centralized training over distributed clients [1]. In the biomedical setting, the clients are institutions rather than mobile devices, and the data are collected under clinical protocols. The statistical structure is therefore different from benchmark federated learning: institutions may specialize in particular diseases, use different scanner vendors, and apply different annotation conventions. Federated optimization under non-identical distributions has been studied through proximal terms that penalize large local deviations from the global model [5] and through adaptive server optimizers that adjust to heterogeneous update scales [6]. These methods are relevant to imaging because a small hospital and a large academic centre may contribute very different numbers of studies, label qualities, and acquisition protocols.

The medical-imaging literature has demonstrated the feasibility of federated training for tasks such as brain-tumour segmentation, mammographic density assessment, and chest radiography classification [3], [4]. Feasibility, however, does not settle the design problem. A federated imaging system must decide how often to communicate, how much of each update to transmit, how to protect updates from the server, and how to allow external queries after training. Reviews of federated healthcare architectures emphasize that these decisions interact with regulatory expectations and institutional trust boundaries [11]. The framework developed here treats those decisions as parameters of a single operating policy rather than as independent modules.

2.2 Communication reduction and optimization

Communication reduction in federated learning has relied on sparsification, quantization, low-rank approximation, and error feedback. Deep gradient compression combines sparsification and momentum correction to reduce bandwidth in distributed training [12]. Federated learning strategies based on structured update compression and sketching reduce uplink cost while maintaining approximate convergence [13]. Adaptive optimization methods such as AdaGrad accumulate historical gradient information to adjust per-coordinate step sizes [14], and federated adaptive methods extend this idea to server-side update aggregation [6]. In high-dimensional imaging models, however, compression must be applied with attention to layer structure. Compression of convolutional filters, normalization parameters,

and classifier heads has different effects on representation quality, and aggressive sparsification can remove small but diagnostically relevant updates in rare classes.

The framework adopts error-feedback compression because it provides a tractable analytical device: the compression error is retained locally and reintroduced in later updates, so the effective update sequence remains close to the uncompressed sequence under a bounded compression operator. This is preferable to independent one-shot sparsification, which can discard small class-specific signals repeatedly. The framework also uses a proximal local objective to limit client drift, following the logic of federated proximal optimization [5], and an adaptive server step to handle scale heterogeneity [6]. The contribution here is not a new compression primitive but the integration of these primitives with secure aggregation and differential privacy in a manner that preserves the analytical roles of compression error and privacy noise.

2.3 Confidential inference and activation cost

Confidential inference protects the queried image, the model weights, and sometimes the prediction itself. Cryptographic approaches include homomorphic encryption, secret sharing, garbled circuits, and hybrid protocols. CryptoNets demonstrated that neural networks can be evaluated on encrypted inputs using leveled homomorphic encryption [17]. Gazelle introduced a hybrid design that combines homomorphic encryption for linear operations with symmetric cryptography for nonlinear operations to reduce latency [18]. SecureML showed that secure two-party and three-party computation can support training and inference for linear models and neural networks [19]. These systems establish that private inference is feasible, but they also show that nonlinear activation functions are the principal bottleneck because comparison-based functions such as ReLU are expensive under encryption.

Automated activation replacement addresses this bottleneck by substituting activation functions that are more compatible with cryptographic evaluation [16]. The design trade-off is explicit: a replacement that is cheaper to evaluate may deviate from the original activation, and the resulting approximation error propagates through subsequent layers. In biomedical imaging, this error can shift segmentation boundaries or change predicted probabilities near diagnostic thresholds. The framework therefore treats activation replacement as part of a calibrated inference policy, not as an unconditional optimization. Polynomial approximation is used where it can be bounded on the relevant activation range, and hybrid cryptographic evaluation is retained where approximation would be unsafe.

The threat model for the present framework separates training-time and inference-time adversaries. During training, the server is honest but curious, participating institutions are honest but curious, and an external adversary may observe network traffic. Secure aggregation prevents the server from observing individual updates, and differential privacy bounds the influence of any single client on the aggregate. During inference, the model holder may be an honest-but-curious cloud service, and the querying institution may not wish to reveal the patient image. The inference protocol must therefore protect the input from the model holder and the model from the querying institution, while returning the prediction to the authorized clinician. This separation matters because a training-time privacy guarantee does not automatically protect inference exchanges.

3. Methodological Framework

3.1 System model and training objective

Consider K institutions, each holding a local dataset D_k of n_k biomedical images with labels y_i for a task such as lesion classification, organ segmentation, or disease grading. The global model f_w is parameterized by w. The federated objective is the weighted empirical risk over institutions. This formulation makes the institutional weighting explicit and allows small or specialized sites to be represented according to a chosen policy rather than implicitly by sample count alone.

$$\min_w F(w) = \sum_{k=1}^{K} p_k F_k(w), p_k = \frac{n_k}{\sum_{j=1}^{K} n_j}$$

Here F_k(w) is the local empirical risk on D_k, and p_k is the institution weight. The weight may be set to n_k / ∑ n_j for sample-proportional aggregation or adjusted to prevent a single large institution from dominating. The local risk is the average loss over local samples, typically cross-entropy for classification or a Dice-plus-cross-entropy objective for segmentation. The framework assumes that labels are available locally and that images are not shared. It also assumes that the server cannot inspect individual updates when secure aggregation is active, but the server may collude with up to a bounded fraction of clients only if the secure-aggregation protocol tolerates such collusion.

3.2 Communication-efficient local optimization

At round t, each institution receives the global model w^t and performs local optimization for E epochs. The local update g_k^t is the difference between the locally optimized parameters and the received global parameters, or the accumulated stochastic gradient over the local pass. Direct transmission of g_k^t is expensive for imaging models, which may contain millions of parameters. The framework applies a compression operator C_k that sparsifies and quantizes the update while retaining local error feedback. The error-feedback recursion is:

$$u_k^t = C_k(g_k^t + e_k^t), e_k^{t+1} = g_k^t + e_k^t - u_k^t$$

The variable e_k^t is the residual error carried from previous rounds, and u_k^t is the compressed update transmitted by client k. The operator C_k may combine top-k sparsification with uniform quantization, and it is assumed to satisfy a bounded relative error condition of the form ||u_k^t − (g_k^t + e_k^t)||_2 ≤ δ_k ||g_k^t + e_k^t||_2 for a compression parameter δ_k. This condition is the analytical handle that connects compression to convergence. Error feedback is important because it prevents small but consistent updates, such as those associated with rare pathological findings, from being discarded in every round simply because they fall below a sparsification threshold.

The aggregate update is formed by the server after secure aggregation. The server may apply an adaptive step size or momentum term, following federated adaptive optimization [6]. Local optimization uses a proximal term to limit drift:

$$w_k\^\{t + 1\} = argmin_w\, F_k(w) + (\mu/2)||w - w\^t||\^2$$

The coefficient μ controls the strength of the proximal penalty. A larger μ reduces client drift under heterogeneous data but slows local adaptation. The choice of μ interacts with compression: when updates are heavily compressed, a strong proximal term can stabilize aggregation, but it can also suppress institution-specific adaptations that are clinically meaningful. The framework therefore treats μ as a heterogeneity-control parameter that should be selected with the compression ratio and the expected label imbalance.

3.3 Secure aggregation and differential privacy

Secure aggregation ensures that the server learns only the sum of client updates, not individual contributions. Each client masks its compressed update with a random value r_k^t, and the masks are constructed so that they cancel in the sum. The server observes only the masked sum modulo a large integer R.

$$y^t = \sum_{k=1}^{K} (u_k^t + r_k^t) mod R, \sum_{k=1}^{K} r_k^t = 0 mod R$$

The variable y^t is the aggregate update after mask cancellation, and R is chosen to avoid overflow. The protocol assumes that clients follow the masking procedure and that the server does not collude with enough clients to reconstruct an individual mask. Secure aggregation protects update confidentiality but does not by itself provide a formal privacy guarantee against inference from the aggregate. Differential privacy is therefore layered onto the training pipeline.

Client-level differential privacy is implemented by clipping the local update to a bounded norm C and adding Gaussian noise calibrated to C and a privacy parameter σ. A simplified release mechanism is:

$$g_DP\hat{}t \; = \; clip(g\hat{}t, C) \; + \; N(0, \sigma\hat{}2 \; C\hat{}2 \; I_d)$$

Here g^t is the aggregated update, clip(g^t, C) scales the update to have norm at most C, I_d is the d-dimensional identity matrix, and σ controls the noise scale. In a client-level guarantee, neighboring datasets differ in one institution's contribution, so the sensitivity is bounded by C. The noise scale σ determines the privacy budget through standard composition across rounds [9]. The framework assumes that clipping is applied before secure aggregation and that the noise is added either by clients in a distributed manner or by the server after aggregation if the server is trusted for noise generation. If the server is not trusted for noise addition, a secure noise-generation protocol or a distributed discrete Gaussian mechanism is required.

The interaction between compression and differential privacy requires care. If an update is compressed after noise is added, the compression operator can distort the noise distribution and invalidate the calibration. If noise is added after compression, the sensitivity must be computed with respect to the compressed update. The framework therefore specifies a noise-aware compression order: clip the update, compress with error feedback, and add noise calibrated to the post-compression sensitivity bound. This order preserves the analytical link between compression error and privacy noise while avoiding the common error of treating compression as a privacy mechanism. Compression reduces communication but does not provide a formal privacy guarantee unless it is explicitly incorporated into the sensitivity analysis.

3.4 Confidential inference architecture

After training, the global model is deployed for inference. The inference architecture protects the patient image from the model holder and the model weights from the querying institution. Linear operations such as convolutions and fully connected layers are evaluated using homomorphic encryption or secret-sharing protocols, following the hybrid designs of CryptoNets and Gazelle [17], [18]. Nonlinear operations are the dominant cost. ReLU requires a comparison, which is expensive under encryption, so the framework uses two

strategies: polynomial approximation on bounded activation ranges and automated activation replacement to substitute cryptographically cheaper activations [16]. Secure multi-party computation is retained for layers where approximation error would be clinically unacceptable [19].

For a scalar activation φ on a bounded interval [−B, B], a degree-M polynomial approximation is:

$$sup_{x\in[-B,B]}|\varphi(x) - \sum_{m=0}^{M} a_m x^m| \leq \varepsilon_a ct$$

The coefficients a_m are chosen by a polynomial approximation method such as minimax approximation or Chebyshev interpolation, and ε_act is the maximum approximation error on the interval. The interval bound B must be estimated from the activation statistics of the trained model or derived from layer-wise bounds. If B is too small, activations outside the interval are clipped and the approximation error becomes uncontrolled. If B is too large, the polynomial degree required to achieve a given ε_act increases, which increases multiplicative depth and private-inference latency. The activation replacement strategy addresses the same trade-off at the architectural level: a replacement activation may have a lower-degree polynomial representation or a more encryption-friendly form, but it changes the function computed by the network.

The inference cost model makes the design trade-off explicit. Let L_lin denote the number and size of linear operations, M_l the polynomial degree used at activation layer l, and R_mpc the number of secure multi-party computation rounds for nonlinear operations that are not replaced. The cost is modeled as:

$$C_infer = \alpha L_lin + \beta \sum_\{l = 1\}\^\{L\} M_l + \gamma R_mpc$$

The coefficients α, β, and γ are platform-dependent constants that reflect encryption scheme parameters, network bandwidth, and secure-computation overhead. The model is not a performance prediction for a particular deployment; it is an analytical device for comparing design choices. Increasing M_l reduces approximation error but raises the polynomial-evaluation term. Replacing an activation can reduce M_l or eliminate a comparison, but it may increase the approximation error that must be absorbed by later layers. The framework therefore selects activation replacements layer by layer, subject to a bound on end-to-end output perturbation.

4. Communication-Efficient Neural Network Optimization

4.1 Compression schedule and update structure

The communication cost of federated imaging is dominated by the uplink transmission of model updates. If a model has d parameters, each client transmits at most b·d bits per round without compression. With sparsification ratio s and quantization level q, the nominal uplink cost scales as K·s·q·d, where K is the number of participating clients. This expression is not a performance result; it is a cost accounting identity that clarifies why sparsification and quantization are attractive. The framework uses layer-aware compression because not all parameters are equally sensitive. Normalization parameters and classifier heads are small but influential, while convolutional filters are numerous and often compressible. A practical schedule therefore allocates a higher sparsification ratio to early convolutional layers and a lower ratio to normalization and output layers.

Error feedback changes the effective update sequence. The server receives u_k^t rather than g_k^t, and the local residual e_k^{t+1} carries forward the discarded component. The compression operator is assumed to be unbiased or nearly unbiased and to have bounded relative error. Under these assumptions, the aggregate update remains close to the uncompressed aggregate, and the convergence analysis can separate the optimization error, the gradient noise, and the compression error. The framework also uses a warm-up phase in which compression is disabled for the first few rounds. This allows the global model to move away from initialization before aggressive sparsification is applied. The warm-up is analytically justified because the relative compression error is largest when the update norm is small and the sparsification threshold removes a large fraction of coordinates.

4.2 Heterogeneity control and adaptive aggregation

Institutional heterogeneity in biomedical imaging is not a nuisance parameter; it is often the reason for collaboration. A cancer centre may contribute predominantly malignant cases, a community hospital may contribute screening cases, and a paediatric hospital may contribute a different age distribution. The local objective F_k therefore differs across clients. The proximal term in the local objective reduces the distance between the local model and the global model, which limits client drift. The coefficient μ should be selected relative to the expected heterogeneity. When μ is small, clients adapt strongly to local data but the aggregate update may be unstable. When μ is large, the global model is stable but local specialization is suppressed. The framework treats μ as a policy variable that can be tuned using validation data at each institution without sharing raw images.

Adaptive server aggregation adjusts the global step size based on the history of aggregate updates. Federated adaptive optimization can improve robustness when institutions produce updates of different scales [6]. However, adaptive methods interact with secure aggregation: the server can only observe the aggregate, so per-client adaptive scaling is not possible without additional leakage. The framework therefore applies adaptivity at the server level, using only the aggregate update and a global learning-rate schedule. This preserves the secure-aggregation guarantee while retaining some of the benefits of adaptive optimization.

4.3 Convergence trade-off

Under L-smoothness of the global objective, bounded gradient variance σ^2, and bounded compression error δ, a standard analysis of stochastic federated optimization with error feedback yields a convergence bound of the following form. The bound is stated to expose the trade-off among optimization error, privacy noise, and compression error, not to claim a particular empirical rate.

$$(1/T)\sum E||\nabla F(w^t)||^2 \le 2\Delta_0/(\eta T) + 2\eta L\sigma^2 + L^2\eta^2\delta^2$$

Here T is the number of aggregation rounds, η is the server learning rate, L is the smoothness constant, σ^2 is the gradient or privacy-noise variance, δ is the compression-error parameter, and Δ_0 = F(w^0) − F is the initial optimality gap. The first term decreases with T and reflects the optimization path. The second term grows with η and σ^2 and reflects the noise floor introduced by stochastic gradients and differential privacy. The third term grows with η and δ^2 and reflects the compression error. The bound implies that increasing the privacy noise σ or the compression error δ cannot be offset indefinitely by taking more rounds; there is an error floor. It also implies that a smaller learning rate reduces the noise and compression terms but slows the optimization term. The practical design question is therefore

not how to minimize communication alone, but how to choose η, σ, and δ so that the error floor remains below the variability of the clinical endpoint.

The bound also clarifies why error feedback is preferable to independent compression. Without error feedback, the compression error term can accumulate across rounds because discarded components are not reintroduced. With error feedback, the compression error appears as a bounded variance term rather than as a systematic bias. This distinction matters for rare classes: repeated one-shot sparsification can systematically remove updates associated with small lesion classes, whereas error feedback retains those components and transmits them in later rounds when they become relatively large.

5. Confidential Model Inference and Security Analysis

5.1 Threat model for inference

The inference threat model differs from the training threat model. During training, the adversary may be an honest-but-curious server or a participating institution. During inference, the adversary may be a cloud service that holds the model, a querying institution that holds the patient image, or an external observer of network traffic. The querying institution may not trust the model holder with the image, and the model holder may not trust the querying institution with the model weights. The inference protocol must therefore provide input confidentiality, model confidentiality, and output confidentiality. Input confidentiality prevents the model holder from reconstructing the image. Model confidentiality prevents the querying institution from extracting weights or architecture details. Output confidentiality may be required if the prediction itself is sensitive, although in many clinical workflows the prediction is returned to the authorized clinician and output privacy is less restrictive than input privacy.

Cryptographic inference protocols provide these guarantees under computational assumptions. Homomorphic encryption protects the input and intermediate ciphertexts from the model holder, but the model holder must evaluate the network on ciphertexts. Secret-sharing protocols distribute the computation across multiple non-colluding parties. Hybrid protocols combine the two to reduce cost. The framework assumes an honest-but-curious model holder and a querying institution that follows the protocol. It does not protect against a malicious model holder that returns incorrect results, although verifiable computation could be added. It also assumes that cryptographic keys are managed by the querying institution and that the model holder does not collude with the key server.

5.2 Activation replacement and approximation error

Activation replacement is the main architectural lever for reducing private-inference cost. A ReLU layer requires a comparison, which is expensive under homomorphic encryption because it requires bootstrapping or a secure comparison protocol. A polynomial activation can be evaluated using additions and multiplications, which are native to homomorphic encryption. Automated activation replacement methods search for replacements that preserve accuracy while reducing private-inference cost [16]. The framework uses this idea in a constrained form: replacements are accepted only if the induced output perturbation is bounded on a validation set or on a certified input region. The validation set may be a small institutional sample that is not shared across sites; the bound is computed locally and aggregated as a scalar certificate rather than as raw data.

The approximation error ε_{act} in the polynomial activation layer propagates through the network. If the network is Lipschitz continuous with constant L_f, the output perturbation is bounded by L_f times the sum of per-layer activation errors. This bound is conservative but useful: it identifies layers where approximation error is amplified. In convolutional networks, the Lipschitz constant can be controlled by spectral normalization or weight clipping. The framework therefore recommends layer-wise sensitivity analysis before activation replacement. Layers with large Lipschitz constants are evaluated with secure multi-party computation, while layers with small constants can use polynomial approximation or replacement.

5.3 Hybrid protocol design

The hybrid inference protocol assigns linear operations to homomorphic encryption and nonlinear operations to either polynomial evaluation or secure multi-party computation. Convolutions and matrix multiplications are evaluated on ciphertexts using packed encoding. Nonlinear operations that have been replaced by polynomials are also evaluated on ciphertexts. Operations that cannot be approximated safely are evaluated using a secure multi-party computation protocol with secret-shared inputs. The protocol must convert between ciphertext and secret-shared representations at layer boundaries. Each conversion incurs communication and latency, so the number of conversions is a primary design variable. The cost model in Equation 7 makes this explicit: the term γR_{mpc} penalizes conversion rounds, and the term $\beta \sum M_l$ penalizes high-degree polynomial evaluation. The optimal design minimizes the total cost subject to an end-to-end error bound.

The protocol also interacts with the federated training model. If activation replacement is applied during training, the federated model learns with the replacement activation. If it is applied only at inference time, the model must be robust to the approximation error. The framework prefers training with the replacement activation when the replacement is architecture-preserving and does not change the optimization landscape qualitatively. This reduces the train-inference mismatch. When replacement changes the activation shape significantly, the model is trained with the original activation and the replacement is applied only in the private-inference path, with a certified error bound.

5.4 Privacy leakage accounting

Differential privacy provides a formal bound on the influence of any one institution on the released aggregate. Secure aggregation prevents the server from observing individual updates. Together, they reduce the risk of training-time leakage. The privacy budget accumulates across rounds, so the number of rounds and the noise scale must be chosen jointly. A smaller noise scale improves utility but consumes the privacy budget faster. The convergence bound in Equation 6 shows that privacy noise also raises the error floor, so there is a three-way trade-off among privacy, communication, and accuracy.

Inference-time leakage is controlled by cryptographic assumptions rather than by differential privacy. Homomorphic encryption hides the input and intermediate values from the model holder under the hardness of the underlying lattice problem. Secret sharing hides the input from any single party under the non-collusion assumption. The framework does not claim that cryptographic inference provides a differential privacy guarantee for the training data. It provides input and model confidentiality for the inference exchange. If the prediction itself must be privatized, a separate output perturbation mechanism is required. The

distinction is important because biomedical systems often need both training-time privacy and inference-time confidentiality, and the two guarantees address different adversaries.

6. Technical Discussion and Boundary Conditions

The proposed framework has several boundary conditions. First, secure aggregation requires a bounded fraction of colluding clients. If the server colludes with enough clients to reconstruct a mask, update confidentiality is lost. The framework assumes that the secure-aggregation protocol tolerates a threshold of collusion and that institutions enforce authentication. Second, differential privacy at the client level may be too coarse for small institutions. A small hospital with a few hundred studies contributes a large fraction of its dataset to each update, so client-level noise must be calibrated to protect the institution as a whole. If record-level privacy is required, a nested mechanism is needed. Third, compression can interact with rare classes. Error feedback mitigates this risk, but aggressive sparsification can still slow the learning of underrepresented findings. The framework recommends class-aware validation at each institution and a compression schedule that relaxes sparsification for output layers.

Private inference introduces a separate boundary condition: approximation error must be evaluated against clinical decision thresholds. A segmentation model may tolerate small pixel-level perturbations but not boundary shifts that change volume measurements. A classification model may tolerate probability perturbations except near the operating threshold. The framework therefore requires layer-wise and output-level error bounds before activation replacement is accepted. These bounds can be computed analytically for polynomial approximations and estimated locally for replacement activations. The bounds are not a substitute for clinical validation, but they provide a principled screening criterion.

Communication efficiency and privacy also interact with institutional incentives. A hospital that contributes many studies may expect a model that performs well on its population, while a small hospital may need stronger privacy guarantees. The weighted objective p_k can be adjusted to reflect contribution or need, but the adjustment affects the global model's performance across sites. The framework does not prescribe a single weighting policy. It makes the policy explicit so that institutions can negotiate the trade-off. Real-world federated implementations for breast-density classification illustrate that governance, data harmonization, and model evaluation are as important as the optimization algorithm [20]. The present framework is compatible with such implementations because it separates the training, privacy, and inference layers while keeping their interfaces explicit.

A final boundary condition concerns the cryptographic cost model. The coefficients α, β, and γ in Equation 7 are platform-dependent and are not estimated here. The model is intended to guide design choices, not to predict latency. In practice, the coefficients would be measured on the target deployment and used to select polynomial degrees, replacement layers, and secure-computation rounds. The analytical value of the model is that it identifies the terms that dominate cost and the terms that can be reduced by architectural choices. It also shows why communication-efficient training and confidential inference cannot be optimized independently: a model that is small enough for efficient federated training may still have activation layers that are expensive to evaluate privately, and a model that is friendly to private inference may have a different parameter structure that affects compression.

7. Conclusion

This manuscript developed an integrated methodological framework for secure federated learning in multi-institutional biomedical image analysis. The framework formulates the federated training objective, combines error-feedback compression with proximal local optimization and adaptive server aggregation, protects individual updates through secure aggregation, and bounds training-time leakage through client-level differential privacy. It then specifies a confidential inference path that uses homomorphic encryption and secure multi-party computation for linear and nonlinear operations, with polynomial approximation and automated activation replacement used to control the cost of private nonlinear evaluation. The analysis derives convergence, communication, privacy, and inference-cost relationships from explicit assumptions and identifies the error floors introduced by compression and privacy noise.

The main conclusion is that communication efficiency, privacy preservation, and confidential inference form a coupled design space. Compression reduces the cost of federated training but introduces an error term that interacts with privacy noise and client heterogeneity. Secure aggregation hides individual updates but does not replace differential privacy. Confidential inference protects patient images and model weights but introduces approximation error and cryptographic cost that must be evaluated against clinical decision thresholds. The framework provides a vocabulary and a set of analytical relationships for making these trade-offs explicit. It also clarifies where further theoretical and empirical work is needed: tighter bounds on activation replacement error, privacy accounting for compressed updates, and deployment studies that measure cryptographic cost under realistic imaging workloads. The contribution is a coherent design basis for multi-institutional imaging systems that require both collaborative learning and confidential inference.

References

1. McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (pp. 1273–1282). PMLR.

2. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends in Machine Learning, 14(1–2), 1–210.

3. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3, 119.

4. Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., ... & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. Scientific Reports, 10, 12598.

5. Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Talwalkar, A., & Smith, V. (2020). Federated optimization in heterogeneous networks. Proceedings of Machine Learning and Systems, 2, 429–450.

6. Reddi, S. J., Charles, Z., Zaheer, M., Garrett, Z., Rush, K., Konečný, J., ... & McMahan, H. B. (2021). Adaptive federated optimization. International Conference on Learning Representations (ICLR).

7. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS) (pp. 1175–1191).

8. Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv:1712.07557.

9. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS) (pp. 308–318).

10. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311.

11. Antunes, R. S., André da Costa, C., Küderle, A., Yari, I. A., & Eskofier, B. (2022). Federated learning for healthcare: Systematic review and architecture proposal. ACM Transactions on Computing for Healthcare, 3(4), 1–23.

12. Lin, Y., Han, S., Mao, H., Wang, Y., & Dally, W. J. (2018). Deep gradient compression: Reducing the communication bandwidth for distributed training. International Conference on Learning Representations (ICLR).

13. Konečný, J., McMahan, H. B., Yu, F. X., Richtárik, P., Suresh, A. T., & Bacon, D. (2016). Federated learning: Strategies for improving communication efficiency. arXiv:1610.05492.

14. Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12, 2121–2159.

15. Carlini, N., Liu, C., Erlingsson, Ú., Song, D., & Terzis, A. (2019). The secret sharer: Evaluating and testing unintended memorization in neural networks. In Proceedings of the 28th USENIX Security Symposium (pp. 267–284).

16. Peng, H., Huang, S., Zhou, T., Luo, Y., Wang, C., Wang, Z., ... & Ding, C. (2023, October). Autorep: Automatic relu replacement for fast private network inference. In 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 5155-5165). IEEE.

17. Gilad-Bachrach, R., Dowlin, N., Laine, K., Lauter, K., Naehrig, M., & Wernsing, J. (2016). CryptoNets: Applying neural networks to encrypted data with high throughput and accuracy. International Conference on Machine Learning, 201–210.

18. Juvekar, C., Vaikuntanathan, V., & Chandrakasan, A. (2018). Gazelle: A low latency framework for secure neural network inference. In Proceedings of the 27th USENIX Security Symposium (pp. 1651–1669).

19. Mohassel, P., & Zhang, Y. (2017). SecureML: A system for scalable privacy-preserving machine learning. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 19–38). IEEE.

20. Roth, H. R., Chang, K., Singh, P., Neumark, N., Li, W., Gupta, V., ... & Xu, D. (2020). Federated learning for breast density classification: A real-world implementation. In

Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning (pp. 181–191). Springer.