

Privacy-Preserving and Explainable Federated Clinical Intelligence Using Secure Aggregation, Differential Privacy and Multimodal XAI

Maria Archana Devi.P

Research Scholar

Department of Computer science and Engineering
Arni University, Himachal Pradesh, India

archana.sap7@gmail.com

Dr. Gurvinder Singh

Professor CSE Department
Arni University, Himachal Pradesh, India

Abstract

Federated clinical artificial intelligence reduces the need to centralize sensitive patient records, but data locality alone does not guarantee privacy, secure collaboration, calibrated confidence, or interpretable predictions. This paper presents a trustworthiness-focused extension of a multimodal clinical intelligence framework that integrates privacy-preserving aggregation, differential privacy, probability calibration, and modality-specific explainable artificial intelligence. Hospitals retain chest radiographs, structured electronic health record variables and vital signs, clinical text, patient identifiers, and local graph structures within their own environments. Only model updates participate in federated optimization. Secure aggregation protects the visibility of individual hospital updates under an honest-but-curious server assumption, while optional differential privacy clips and perturbs updates to provide a formal privacy-utility trade-off. The prediction pipeline is complemented by modality-specific explanations: image localization for chest radiographs, feature attribution for structured clinical variables, token-level attribution for clinical text, and graph-attention evidence for local relational reasoning. Probability calibration is evaluated separately from discrimination so that confidence reliability is not conflated with classification performance. The paper proposes an evaluation framework spanning predictive utility, privacy budget, communication overhead, calibration, explanation plausibility and stability, and site-level robustness. Illustrative values from the source thesis show the expected reporting pattern, including a decline from 95.2% accuracy without differential privacy to 94.6% at a light-DP setting, secure-aggregation overhead without predictive degradation, and an expected calibration error reduction from 0.047 to 0.018 after temperature scaling. These values are examples from the thesis and must be replaced by implementation-derived results before publication.

Keywords— federated learning, secure aggregation, differential privacy, explainable AI, multimodal clinical AI, calibration, privacy–utility trade-off, healthcare decision support.

1. Introduction

Healthcare artificial intelligence increasingly combines imaging, structured measurements, and free-text documentation. Such multimodal models can represent a patient more completely than a single-source classifier, yet they also increase the sensitivity and dimensionality of the information being processed. Central pooling of radiographs, EHR variables, clinical notes, and derived representations can conflict with institutional governance and privacy constraints. Federated learning addresses part of this challenge by

keeping raw records at participating hospitals, but federated learning should not be equated with automatic anonymization. Model parameters and gradients may still reveal information correlated with local training examples.

The source framework therefore separates four concepts that are sometimes incorrectly treated as interchangeable: local data retention, protected communication, secure aggregation, and differential privacy. Raw clinical data remain local. Transport protection secures communication channels. Secure aggregation prevents the coordinating server from directly inspecting an individual client update, while differential privacy limits the influence of protected training units through clipping and calibrated random noise. These layers address different threat surfaces and should be evaluated independently.

Trustworthiness also depends on what the model communicates after training. A disease probability is more useful when its confidence is calibrated and when the clinical evidence associated with the prediction can be inspected. The proposed framework therefore combines privacy protection with modality-specific explainability and post-hoc calibration. The objective of this paper is to isolate and systematically present this privacy-and-trustworthiness contribution rather than repeat the architectural and optimization questions addressed by the preceding multimodal, graph, and federated-learning papers.

2. Related Work and Research Gap

Practical secure aggregation was established by Bonawitz et al. as a method for recovering an aggregate of client vectors without revealing individual contributions to the server. In a multi-hospital environment, this mechanism is relevant when the server follows the protocol but should not directly inspect a hospital's update. Differentially private deep learning, as demonstrated by Abadi et al., uses norm clipping, calibrated noise, and privacy accounting to bound the influence of protected training units. These methods are complementary: secure aggregation protects update visibility during aggregation, whereas differential privacy provides a statistical guarantee defined by an explicit privacy budget.

Clinical explainability is likewise modality dependent. Grad-CAM and related localization techniques can identify image regions associated with a class; integrated gradients and SHAP-style methods can attribute predictions to structured variables; token-level attribution can highlight important phrases in clinical text; and graph-attention coefficients can identify influential neighboring representations. The source thesis explicitly warns that such explanations are model evidence rather than causal proof. Calibration is a distinct problem: a model may discriminate well while assigning systematically overconfident or underconfident probabilities. Temperature scaling offers a simple validation-set-based correction without changing most class rankings.

The research gap addressed here is therefore not the absence of any one privacy or explanation technique. The gap is the lack of a single evaluation structure that reports privacy assumptions, privacy budget, predictive utility, cryptographic overhead, calibration reliability, and multimodal explanation quality together. The proposed paper treats these dimensions as linked aspects of trustworthy federated clinical decision support.

3. Proposed Privacy-Preserving and Explainable Framework

Each hospital runs the same local multimodal disease-prediction model. The local model may contain the modality-specific encoders, multimodal Transformer, local clinical graph, GAT, and prediction head

defined in the overall thesis. This paper focuses on what happens at the privacy boundary and at the output interpretation layer. Patient-level CXR images, EHR rows, notes, identifiers, and graph structures remain within the hospital environment. After local optimization, the hospital produces a parameter update. That update can be clipped and perturbed for differential privacy, masked for secure aggregation, and transmitted to the coordinating server. The server reconstructs only the aggregate required for the federated update under the secure-aggregation protocol.

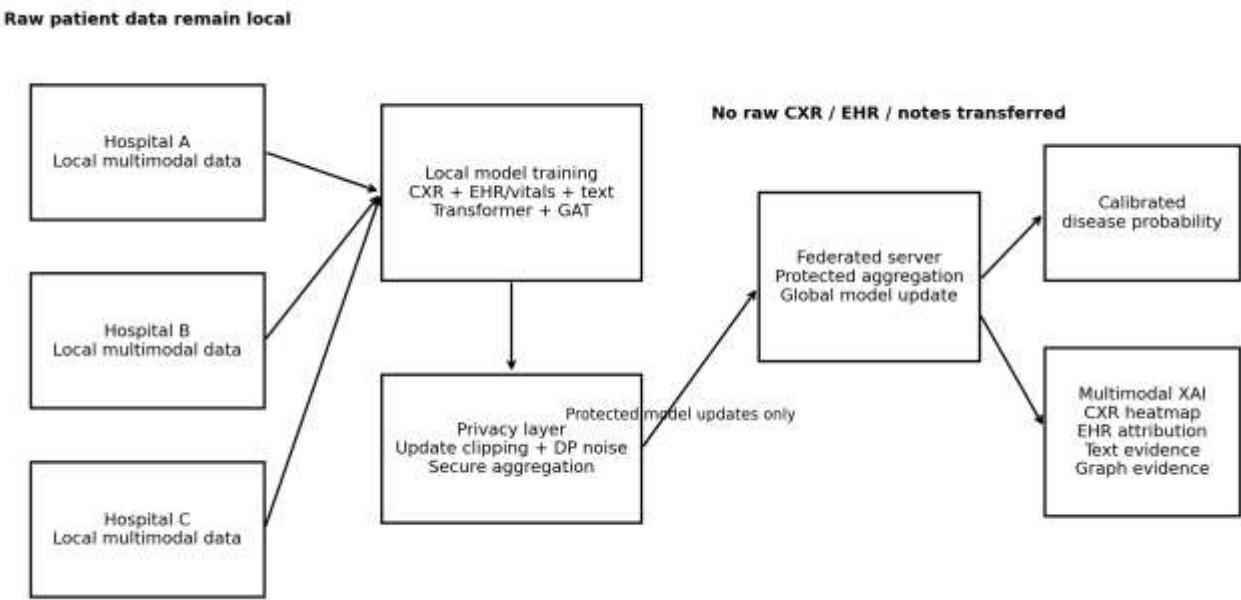


Figure 1. Privacy-preserving and explainable federated clinical intelligence workflow.

3.1 Threat Model and Layered Privacy Design

The baseline threat model in the source is an honest-but-curious coordinating server. The server is assumed to execute the aggregation protocol correctly while potentially attempting to learn from information it receives. Secure aggregation is intended to reduce the visibility of individual hospital updates under this model. The framework does not claim that secure aggregation prevents malicious clients, poisoning, or Byzantine behavior. Authentication, secure key exchange, model-version checking, update validation, access control, audit logging, and stronger robust aggregation are separate system-security concerns.

Protection layer	Primary objective	What it does not guarantee
Local data retention	Prevents routine central pooling of raw patient data	Does not hide information potentially encoded in model updates
Transport encryption	Protects updates in transit	Does not prevent an authorized server from inspecting a decrypted update

Secure aggregation	Hides an individual client update from direct server inspection	Does not provide a formal statistical privacy budget
Differential privacy	Bounds influence of a protected unit under an explicit privacy definition	Does not authenticate clients or stop poisoning attacks
Governance controls	Controls access, retention, auditing and approved use	Does not replace technical privacy mechanisms

Table 1. Layered privacy and security responsibilities.

3.2 Secure Aggregation

Let Δ_k denote the update generated by hospital k . In a pairwise-mask secure aggregation design, each client adds masks that cancel when participating client vectors are summed. The server receives masked contributions and recovers the aggregate but not an unmasked Δ_k under the assumed protocol conditions.

$$u_k = \Delta_k + \sum_j r_{kj} - \sum_j r_{jk} \quad (1)$$

$$\sum_k u_k = \sum_k \Delta_k \quad (2)$$

Because the pairwise masks cancel in the aggregate, the learned update is unchanged relative to ordinary aggregation, although cryptographic operations, key exchange, and dropout-resilient coordination add communication and latency overhead. This is why secure aggregation should be assessed using both predictive and systems metrics.

3.3 Differential Privacy

Differential privacy is incorporated as an optional statistical protection layer. The update is first clipped to limit sensitivity and then perturbed by Gaussian noise. The protected unit—patient, minibatch, or hospital update—must be explicitly defined before interpreting the privacy budget.

$$\bar{\Delta}_k = \Delta_k \cdot \min(1, C / \|\Delta_k\|_2) \quad (3)$$

$$\tilde{\Delta}_k = \bar{\Delta}_k + N(0, \sigma^2 C^2 I) \quad (4)$$

Here C is the clipping norm and σ is the noise multiplier. A privacy accountant tracks the cumulative (ϵ, δ) guarantee across training rounds. Smaller ϵ generally indicates stronger formal privacy under a fixed accounting framework, but stronger protection can reduce utility because clipping and noise suppress information contained in local updates. The source explicitly requires reporting ϵ , δ , clipping norm, sampling process, training steps, and accounting method instead of reporting ϵ alone.

3.4 Multimodal Explainable Artificial Intelligence

The explanation layer preserves the semantics of each modality rather than forcing all attributions into a single score. For chest radiographs, gradient-based localization such as Grad-CAM or attention rollout highlights influential image regions. For EHR and vital-sign variables, integrated gradients or SHAP-like feature attribution estimates positive and negative contributions. For clinical text, token-level attribution identifies phrases associated with the predicted disease while retaining negation and temporal context. For local clinical graphs, GAT coefficients identify neighboring representations that influenced graph refinement.

These outputs should be displayed as coordinated evidence: predicted probability, salient image region, top structured variables, important clinical phrases, and relevant graph relationships. Their numerical scales arise from different mathematical mechanisms and should not be interpreted as causally equivalent. Explanation quality is evaluated through localization plausibility, perturbation sensitivity, attribution stability, and clinical review rather than visual appeal alone.

3.5 Probability Calibration

Discrimination and calibration answer different questions. AUROC measures whether positive cases tend to receive higher scores than negative cases, whereas calibration measures whether reported probabilities correspond to empirical frequencies. Temperature scaling fits a scalar T on validation data and applies it to model logits before the softmax or sigmoid probability transformation.

$$p_cal(y=c|x) = \text{softmax}(z_c / T) \quad (5)$$

T is fixed before final test evaluation. Calibration is evaluated with expected calibration error (ECE), Brier score, negative log-likelihood, and reliability plots. The source also allows uncertainty indicators such as predictive entropy, deep ensembles, or Monte Carlo dropout when computationally feasible, with high uncertainty supporting a 'review required' state rather than an autonomous diagnosis.

4. Experimental Evaluation Design

The source evaluation assumes, for illustration, three participating hospitals and a five-class thoracic disease task containing Normal, Pneumonia, Tuberculosis, Pleural Effusion, and Cardiomegaly. FedAdam is used as the example federated optimizer, while the present paper varies privacy and trustworthiness settings. The final experimental study should preserve fixed patient-level splits and identical model architecture across privacy comparisons so that changes can be attributed to the privacy mechanism rather than to a change in data or model capacity.

Dimension	Measures
Predictive utility	Accuracy, macro F1, AUROC, AUPRC, sensitivity, specificity
Privacy	ϵ , δ , clipping norm, noise multiplier, protected unit, accountant
Secure aggregation	Round time, communication overhead, client dropout handling
Calibration	ECE, Brier score, NLL, reliability diagram
Explainability	Localization plausibility, attribution stability, perturbation tests, case review
Robustness	Site-wise metrics, worst-site performance, missing-modality behavior

Table 2. Evaluation dimensions for privacy-preserving trustworthy clinical AI.

5. Illustrative Results from the Source Thesis

The numerical results in this section are not implementation-verified measurements. They are the illustrative values supplied in the source thesis to demonstrate how the final experimental results should be

reported. They must be replaced by values generated by the actual implementation before journal or conference submission.

5.1 Differential-Privacy Utility Trade-Off

Privacy setting	ϵ	Accuracy (%)	Macro F1 (%)	Utility change vs. no DP
Strong DP	0.5	88.4	88.0	-6.8 pp
Strong-moderate DP	1	90.9	90.5	-4.3 pp
Moderate DP	2	92.8	92.5	-2.4 pp
Practical DP setting	4	94.1	93.9	-1.1 pp
Light DP	8	94.6	94.5	-0.6 pp
No DP	∞	95.2	95.1	Reference

Table 3. Illustrative privacy–utility trade-off reported in the source thesis.

The illustrative pattern is monotonic: stronger privacy corresponds to lower predictive utility, while performance approaches the no-DP reference as ϵ becomes less restrictive. The source explicitly cautions that these values do not constitute a recommended deployment epsilon. The operating point must be chosen according to the defined threat model, governance requirements, and acceptable loss in predictive performance.

5.2 Secure Aggregation and Operational Overhead

Configuration	Accuracy (%)	Mean round time (s)	Relative communication overhead	Raw patient data transferred
Plain federated aggregation	95.2	38.4	1.00×	No
Secure aggregation	95.2	43.1	1.12×	No
Secure aggregation + DP	94.6	45.7	1.19×	No

Table 4. Illustrative effect of privacy mechanisms on utility and systems overhead.

The source reports unchanged illustrative accuracy between ordinary federated aggregation and secure aggregation because masking changes update visibility rather than the reconstructed aggregate. The extra round time reflects cryptographic and protected-message overhead. Differential privacy introduces both additional overhead and an intentional utility reduction because the update values themselves are modified.

5.3 Calibration Results

Calibration metric	Before calibration	After calibration
Expected Calibration Error (ECE)	0.047	0.018
Brier score	0.082	0.061
Negative log-likelihood	0.284	0.221
Accuracy (%)	95.2	95.1

Table 5. Illustrative probability-calibration results.

The example shows a substantial reduction in ECE with negligible change in classification accuracy. This is the intended behavior of post-hoc temperature scaling: probability reliability improves while the underlying ranking and most predicted labels remain nearly unchanged. In a clinical decision-support setting, better calibration can support threshold selection, uncertainty flagging, and referral of ambiguous cases.

5.4 Explainability Analysis

Case	Predicted class	Confidence	Top multimodal evidence	Expected explanation
1	Pneumonia	0.96	CXR opacity; fever; low SpO ₂ ; cough text	Heatmap over affected lung region; fever and SpO ₂ highly weighted
2	Tuberculosis	0.93	Upper-zone CXR pattern; prolonged symptom text	Upper-lobe image focus; text tokens related to chronic symptoms
3	Pleural Effusion	0.91	Basal CXR opacity; reduced breath-sound context	Localization near fluid region with supporting clinical variables

Table 6. Illustrative multimodal XAI reporting format based on the source framework.

The purpose of this analysis is not to claim that the highlighted evidence is causal. Instead, explanations make the model inspectable and can expose implausible dependencies. A heatmap concentrated on image borders, a structured attribution dominated by an administrative variable, or a text explanation focused on post-outcome wording would indicate a problem in data preparation or model behavior.

6. Discussion

The central methodological result of this paper is that privacy, security, interpretability, and confidence reliability should be reported as separate but interacting properties. Keeping raw data inside a hospital reduces routine data movement, yet does not by itself prevent information leakage through model updates. Secure aggregation addresses direct visibility of individual updates but does not create a formal privacy budget. Differential privacy provides a formal mechanism, yet imposes a utility trade-off and does not solve authentication or poisoning. These distinctions prevent the common but inaccurate claim that a federated system is automatically private or secure.

The same separation is necessary for trustworthiness. Explainability does not validate a model, and attention or attribution values are not causal explanations. Calibration does not make an inaccurate model clinically useful, but it improves the interpretation of confidence when discrimination is otherwise acceptable. The value of the proposed framework lies in reporting these dimensions together, allowing a reviewer to see whether a gain in privacy comes with a measurable loss in utility, whether cryptographic protection creates operational cost, and whether high predictive performance is accompanied by reliable probabilities and plausible model evidence.

The illustrative results also show why privacy should be treated as a curve rather than a binary setting. The no-DP example reaches 95.2% accuracy, whereas progressively stronger privacy settings reduce accuracy. A light-DP illustration retains 94.6%, while strong DP at $\epsilon=0.5$ falls to 88.4%. This relationship is useful only when accompanied by the precise protected unit, clipping rule, noise multiplier, sampling assumptions, total training steps, and accounting method. Without those details, an epsilon value is not sufficient for scientific comparison.

7. Limitations and Future Work

The source framework is retrospective and does not constitute prospective clinical validation. Secure aggregation is analyzed under an honest-but-curious server assumption and does not provide full defense against malicious clients or Byzantine attacks. Differential-privacy values in the results section are illustrative rather than measured from a completed privacy accountant. XAI outputs require domain-expert assessment and stability testing, and calibrated probabilities may become miscalibrated under deployment-time distribution shift. Future work should therefore include adversarial robustness, personalized or adaptive federation, prospective multi-site validation, clinician-centered explanation studies, calibration monitoring after deployment, and experimentally measured privacy accounting under a clearly defined protected unit.

8. Conclusion

This paper isolates the privacy and trustworthiness contribution of the broader multimodal federated clinical intelligence framework. The proposed design keeps raw multimodal patient data local, protects individual model updates through secure aggregation, optionally adds differential privacy with explicit privacy accounting, calibrates prediction probabilities, and produces modality-specific explanations for image, structured, textual, and graph evidence. The framework treats privacy and interpretability as measurable engineering and evaluation dimensions rather than descriptive labels. Its main methodological contribution is the coordinated reporting of predictive utility, privacy budget, secure-aggregation overhead, calibration reliability, explanation quality, and site-level robustness. The illustrative values supplied by the source thesis demonstrate the expected reporting format but must be replaced with actual experimental results before publication.

References

- [1] K. Bonawitz et al., “Practical secure aggregation for privacy-preserving machine learning,” Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), pp. 1175–1191, 2017. doi:10.1145/3133956.3133982.
- [2] M. Abadi et al., “Deep learning with differential privacy,” Proc. ACM SIGSAC Conf. Computer and Communications Security (CCS), pp. 308–318, 2016.
- [3] R. R. Selvaraju et al., “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” Proc. IEEE Int. Conf. Computer Vision (ICCV), pp. 618–626, 2017.
- [4] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” Proc. 34th Int. Conf. Machine Learning (ICML), pp. 3319–3328, 2017.

- [5] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.
- [6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” *Proc. 34th Int. Conf. Machine Learning (ICML)*, pp. 1321–1330, 2017.
- [7] C. Rudin, “Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead,” *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019. doi:10.1038/s42256-019-0048-x.
- [8] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” *Proc. AISTATS*, pp. 1273–1282, 2017.
- [9] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, “Federated optimization in heterogeneous networks,” *Proc. Machine Learning and Systems*, vol. 2, pp. 429–450, 2020.
- [10] S. J. Reddi et al., “Adaptive federated optimization,” *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [11] J. Rieke et al., “Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture,” 2024, systematic review indexed in PubMed (PMID: 38340728).
- [12] Y. Wang, C. Yin, and P. Zhang, “Multimodal risk prediction with physiological signals, medical images and clinical notes,” *Heliyon*, vol. 10, no. 5, e26772, 2024. doi:10.1016/j.heliyon.2024.e26772.