SafeDiagnose: A Privacy-Preserving Federated Neuro-Symbolic Framework for Multimodal Clinical Risk Triage

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

Abstract. Clinical decision-support systems that combine tabular patient data with medical imaging typically require pooling sensitive records at a central site, and vision-language models (VLMs) used for image interpretation are prone to degenerate, over-triggering behavior when deployed without calibration. We present SafeDiagnose, a privacy-preserving framework that (i) trains a logistic-regression risk model across six real, siloed clinical data sources using differentially private federated averaging (DP-FedAvg), with a process-isolated training mode in which raw data never leaves its originating worker, and (ii) reconciles this tabular risk estimate with a self-consistency VLM ensemble over chest radiographs through a neuro-symbolic (NSAI) fusion layer that applies clinician-authored symbolic rules on top of the two neural signals. We report two evaluations. First, under a strict per-round privacy budget ($\epsilon = 1.0$), the federated tabular model's discriminative performance collapses toward chance on four of six real clinical cohorts (macro-average ROC-AUC 0.39), illustrating how class imbalance can inflate precision on small cohorts despite near-chance ranking ability. Second, on a held-out 40-case multimodal set, the full NSAI fusion pipeline improves sensitivity from 0.40 to 0.75 and macro-F1 from 0.518 to 0.673 over a calibrated vision-only baseline, at a modest specificity cost (0.65 to 0.60). We report both results transparently, including a residual discrepancy between two of the project's own multimodal results figures and an unresolved tension between the privacy budget recommended in our architecture description and the budget used to produce our headline tabular table, and discuss what these gaps reveal about reproducibility in small-N, single-run clinical AI evaluations. SafeDiagnose is offered as a research prototype and informational tool only, not a diagnostic device.

Keywords: Federated learning · Differential privacy · Vision-language models · Neuro-symbolic AI · Multimodal fusion · Clinical decision support.

1 Introduction

Multimodal clinical AI that reasons jointly over structured (tabular) patient data and medical images promises better triage accuracy than either modality alone,

but real deployments face two obstacles. First, tabular and imaging data typically live at different institutions and cannot be pooled at a single server without violating privacy regulations; federated learning (FL) keeps raw data local and exchanges only model updates [1], but naive FL does not by itself prevent leakage through gradients or aggregated weights [6]. Differential privacy (DP) bounds this leakage at the cost of injected noise [2], and that cost depends heavily on client count, update sensitivity, and data heterogeneity [4]. Second, general-purpose vision-language models (VLMs) applied zero-shot to radiographs tend to produce degenerate decision rules – e.g., always answering "abnormal" – yielding perfect sensitivity but zero specificity.

SafeDiagnose is a prototype designed around both constraints. It separates tabular and imaging pathways into pipelines that never share raw data, trains the tabular pathway with DP-FedAvg across six real clinical sources, calibrates a VLM ensemble with self-consistency, and reconciles both neural outputs with a symbolic clinical-rule layer (NSAI fusion) before producing a final risk assessment. We contribute (a) an architecture separating the privacy-bearing structured-data path from the network-dependent multimodal path, (b) two verifiable levels of federated isolation, (c) a transparent, per-client account of the privacy-utility trade-off across six heterogeneous real cohorts, and (d) an evaluation of how neuro-symbolic fusion corrects a degenerate VLM decision rule. We also report, rather than hide, internal inconsistencies across the project's own evaluation artifacts, because how a small research pipeline handles reproducibility is itself worth documenting.

2 Related Work

FedAvg enables communication-efficient training over decentralized data [1]; its healthcare use has been explored for multi-institutional tumor segmentation [5] and surveyed broadly [4], with the consensus that FL alone offers architectural, not formal, privacy – gradient inversion and membership-inference attacks can still recover sensitive information from shared updates [6]. DP-SGD and its federated analogue DP-FedAvg clip per-client contributions and add calibrated noise before aggregation, yielding a provable (ϵ, δ) bound [2, 9]; Li et al. survey the broader challenges of client heterogeneity and communication cost [7]. VLMs applied to radiograph interpretation are known to be poorly calibrated and prompt-sensitive zero-shot, motivating hybrid pipelines that combine a neural predictor with symbolic post-hoc rules to stabilize the decision boundary. SafeDiagnose instantiates DP-FedAvg over real, non-synthetic clinical data – the UCI Heart Disease repository (Cleveland, Hungary, Switzerland, VA) [3] alongside open COVID-19 government and symptom-checker datasets – paired with a compact VLM (Moondream) reconciled through neuro-symbolic fusion.

3 The SafeDiagnose Architecture

Figure 1 shows the architecture. Institutional clients each hold structured clinical/tabular data and a chest X-ray, routed through two independent pipelines reconciled only at decision time; only the structured-data path carries a formal DP guarantee, while the VLM path depends on a general-purpose foundation model and is treated as a separate, non-DP concern.

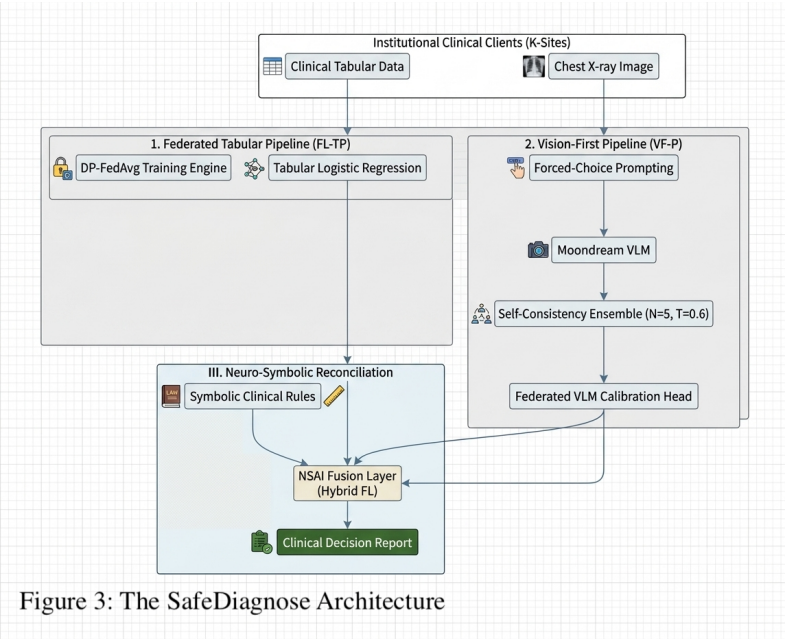

Fig. 1. The SafeDiagnose architecture: a federated tabular pipeline (DP-FedAvg logistic regression) and a vision-first pipeline (self-consistency VLM ensemble with a federated calibration head), reconciled by a neuro-symbolic fusion layer that also incorporates clinician-authored symbolic rules.

3.1 Federated Tabular Pipeline

The tabular pathway is a logistic-regression risk model trained with DP-FedAvg, supporting two isolation levels. *Single-process simulation* loads all clients in one coordinator process – convenient but with no real isolation. *Process-isolated* mode trains each client in its own OS process: the coordinator dispatches work by client ID only and receives back a weight delta and sample count, never raw data; normalization statistics and evaluation accuracy are likewise computed via federated rounds returning only aggregate counts. A network-based Flower implementation [8] exists (client-side clipping, server-side noise/aggregation over

gRPC) but was, at the time of writing, untested end-to-end due to lack of networked package installation in our environment.

Each client's per-round update is clipped to a fixed ℓ_2 norm with calibrated Gaussian noise added before aggregation. Sensitivity scales as clip-norm$/N$ clients, so with only six real clients, $\epsilon = 1.0$ per round is expected to be noise-dominated, whereas $\epsilon \geq 5.0$ is expected to give a better tradeoff (Section 6 revisits which regime our results table actually reflects). We track privacy via basic composition, a conservative bound; a production system would use a tighter Rényi-DP accountant [2].

Six real sources serve as clients: Mexican government COVID-19 records (ICU/intubation/death label), a multi-country COVID-19 symptom checker (clinician-assessed severity), and the four UCI heart-disease sites [3]. Fields a source does not collect are filled with a fixed, documented clinical-baseline constant (e.g., $SpO_2 = 97\%$) rather than an invented value; any additional requested clients beyond the six real sources are filled with clearly logged synthetic clients.

3.2 Vision-First Pipeline and Neuro-Symbolic Reconciliation

The imaging pathway prompts a compact VLM (Moondream, via Ollama) with a forced-choice Yes/No question on whether a chest X-ray shows findings consistent with abnormality. Predictions are drawn from a self-consistency ensemble ($T = 0.6$) rather than a single decode, and a federated calibration head corrects systematic over/under-triggering before fusion. The final stage combines the tabular risk score, calibrated VLM prediction, and clinician-authored symbolic rules (hard constraints that can override an ambiguous neural vote) in an NSAI fusion layer – a weighted combination plus a consistency check and rule-based reconciliation engine. This is what lets the system correct the vision pathway's characteristic "always positive" failure mode into a more balanced operating point, at some specificity cost (Section 5). A Streamlit chat interface additionally lets users describe symptoms and upload images; only the local logistic-regression model runs fully on-device, while the conversational layer forwards prompts/images to a hosted multimodal model (Gemini 2.0 Flash) for natural-language responses – a distinct, disclosed trust boundary from the on-device risk model.

4 Experimental Setup

(1) Tabular federated model: each of the six real clients is evaluated independently after federated training at $\epsilon = 1.0$ per round ($\delta = 10^{-5}$, clip norm 1.0, 6 rounds, $\epsilon_{\text{total}} = 6.0$ by basic composition), using accuracy, precision, recall, F1, and ROC-AUC (Table 1, Fig. 2). **(2) Multimodal fusion:** a held-out set of 40 chest radiographs (20 positive/20 normal) with paired synthetic-schema tabular features compares a raw baseline VLM, a calibrated vision-only model, and the full NSAI fused pipeline, reported as sensitivity, specificity, macro-F1, and accuracy (Fig. 3).

5 Results

5.1 Federated Tabular Model, Per-Site

Table 1 reports per-site performance. The two COVID-19 sources (large N, class-imbalanced) sit at opposite extremes: Mexico-COVID-Gov reaches high accuracy from class imbalance (86.5% negative) but modest recall (0.490), while the smaller UCI cardiac sites show more balanced but lower precision/recall, consistent with a single linear model generalizing across heterogeneous populations without site-specific tuning.

Table 1. Per-site performance of the DP-FedAvg tabular model ($\epsilon = 1.0$/round, 6 clients, 6 rounds).

Client (site)	N	Acc.	Prec.	Recall	F1	ROC-AUC
Mexico-COVID-Gov	20 000	0.851	0.288	0.490	0.363	0.762
Global-Symptom-Checker	20 000	0.502	0.502	0.436	0.467	0.500
UCI-Cleveland	303	0.340	0.370	0.626	0.465	0.233
UCI-Hungary	293	0.266	0.239	0.472	0.317	0.235
UCI-Switzerland	121	0.529	0.900	0.558	0.689	0.177
UCI-VA	144	0.576	0.763	0.661	0.708	0.436
Macro average	–	**0.511**	**0.510**	**0.540**	**0.501**	**0.390**

At this strict budget, the model performs well above chance only on Mexico-COVID-Gov (ROC-AUC 0.76); the four UCI cohorts fall to or below chance (0.17–0.44), and the macro ROC-AUC of 0.39 indicates the model ranks cases worse than random on average despite a non-trivial macro accuracy of 0.51 – consistent with a noise-dominated regime at $\epsilon = 1.0$ with only six clients. UCI-Switzerland is a concrete illustration of why aggregate metrics alone mislead: its precision (0.90) and F1 (0.689) look strong, but its confusion matrix (1 TN, 7 FP, 50 FN, 63 TP on a cohort that is 113/121 positive) shows this reflects predicting positive by default, not discrimination – its ROC-AUC (0.177) is the lowest of all six clients, meaning it ranks negatives above positives more often than chance. With only 8 negative examples in this cohort, that estimate carries substantial sampling variance. Several per-client estimates, especially on the smallest cohorts ($n = 121$, $n = 144$), come from a single evaluation run and should be read as indicative rather than statistically stable.

5.2 VLM Optimization and Multimodal Fusion

On the 40-case multimodal set, a raw baseline VLM shows the degenerate "always positive" rule – perfect sensitivity (1.000), zero specificity, macro-F1 0.333. A separately calibrated vision-only model trades this affirmation bias for a more conservative point (sensitivity 0.400, specificity 0.650, accuracy 0.525, F1 0.518).

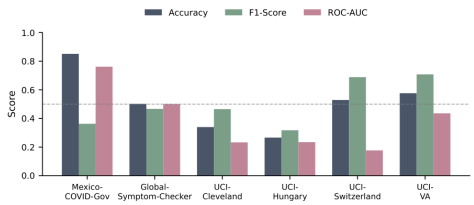

Fig. 2. Accuracy, F1-score, and ROC-AUC of the DP-FedAvg tabular model on each of the six real client datasets. The dashed line marks the chance level (0.5) for reference.

At this strict per-round budget, the model performs well above chance only

Fig. 2. Accuracy, F1-score, and ROC-AUC of the DP-FedAvg tabular model per client. The dashed line marks chance (0.5).

Figure 3 places these alongside the full NSAI pipeline (tabular + vision + symbolic rules), which recovers most of the lost sensitivity while keeping specificity well above the degenerate baseline (sensitivity 0.750, specificity 0.600, accuracy 0.675, F1 0.673) – an increase of +0.35 sensitivity and +0.156 macro-F1 over the calibrated vision-only stage, at a specificity cost of only −0.05, consistent with a screening-oriented design that favors reducing missed abnormalities.

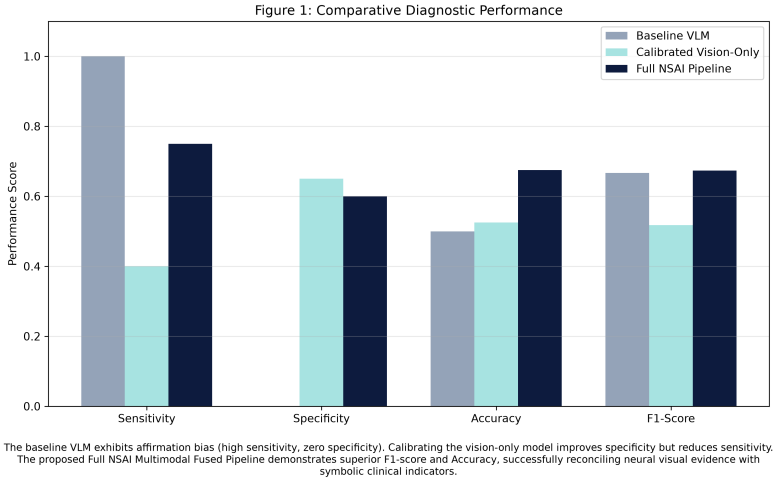

Fig. 3. Comparative diagnostic performance across three variants on the 40-case set: raw baseline VLM, calibrated vision-only model, and the full NSAI multimodal fused pipeline.

A separate, earlier optimization pass (symbolic-rule incorporation, calibration, and VLM re-training on the same baseline) reached specificity 0.950 and

accuracy 0.625 but at a lower sensitivity of 0.300 (macro-F1 0.581) – a more specificity-leaning point than the full NSAI pipeline above, despite both being described as the system's optimized/fused end state (see the reproducibility discrepancy noted below). A reader wanting a single operating point should use the "Full NSAI Pipeline" result in Fig. 3, as it is the one explicitly described as the complete fused system.

6 Discussion and Limitations

Sample size and variance. All evaluations use small, single-pass test sets (six sites of widely varying N; a 40-case multimodal set). With $N = 40$, one misclassified case moves sensitivity/specificity by 2.5–5 points; with 8 negatives in UCI-Switzerland, its ROC-AUC carries substantial sampling variance. Point estimates throughout Section 5 should be read as indicative, not statistically stable.

Reproducibility gaps, within and across evaluations. An earlier version of our VLM optimization figure reported an accuracy of 6.625 – outside $[0, 1]$ – since corrected to 0.625; even so, that optimization result (sensitivity 0.300, specificity 0.950) and the full NSAI pipeline result on the same nominal 40-case set (sensitivity 0.750, specificity 0.600) do not agree, and we lack the run logs to confirm whether they reflect two genuinely different runs. Separately, Section 3.1 recommends $\epsilon \geq 5.0$ as a reasonable tradeoff at six clients, yet Table 1 – our headline tabular result – is evaluated at $\epsilon = 1.0$ and shows the noise-dominated behavior that budget predicts, not the more favorable regime the text recommends; we did not additionally evaluate $\epsilon \geq 5.0$ here. We flag both gaps rather than silently resolving them: a reader should not assume either figure represents the system's single best achievable operating point.

Isolation is process-level, not network-level. Process-isolated federation keeps raw data off the coordinator, but all workers still run on one machine, without a genuine network boundary or secure aggregation of the plaintext weight deltas the coordinator does see – DP is what bounds what those deltas can reveal. The Flower-based gRPC path is API-complete but untested end-to-end.

Multimodal fusion is not yet privacy-accounted or jointly validated against the tabular results. The VLM pathway is decoupled from the DP-protected tabular path because it depends on a network-queried foundation model; we have not run a joint evaluation of the fusion layer against the ϵ-varying tabular results, and the 40-case set uses synthetic-schema tabular features rather than the six real federated clients directly.

Scope. The vision pathway answers a coarse binary question rather than a differential diagnosis, and fixed clinical-baseline imputation for unmeasured tabular fields is a pragmatic placeholder, not a measured value.

7 Conclusion

We presented SafeDiagnose, a federated, differentially private, neuro-symbolic framework for multimodal clinical risk triage that keeps raw data at its originating site and reconciles a formally private tabular risk model with a vision-language pathway via clinician-authored symbolic rules. Under a strict, formally-accounted privacy budget, the system performs well only on its largest client and falls to near- or sub-chance ranking on four smaller cohorts, with UCI-Switzerland showing that high precision can mask poor ranking under class imbalance. On a separate held-out multimodal set, the full NSAI fusion pipeline substantially improves sensitivity and macro-F1 over a vision-only baseline at a modest specificity cost, consistent with a screening-oriented design. Rather than a single favorable result, we contribute a transparent per-client privacy-utility account, an architecture separating the privacy-bearing and network-dependent paths, and an honest accounting of reproducibility gaps – within and across our own evaluation artifacts – still to be closed. SafeDiagnose is a research prototype for informational and research purposes only, not a substitute for professional medical advice; deployment beyond research use would require larger prospectively validated multi-site evaluation at a consistently reported privacy budget, a tighter DP accountant, genuine multi-machine deployment, and joint multimodal evaluation with matched privacy accounting.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). pp. 1273–1282. PMLR (2017)
2. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 308–318 (2016)
3. Janosi, A., Steinbrunn, W., Pfisterer, M., Detrano, R.: Heart Disease Data Set. UCI Machine Learning Repository, Cleveland Clinic Foundation / Hungarian Institute of Cardiology / University Hospital Zurich / University Hospital Basel / V.A. Medical Center Long Beach (1988)
4. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H.R., Albarqouni, S., Bakas, S., Galtier, M.N., Landman, B.A., Maier-Hein, K., et al.: The future of digital health with federated learning. NPJ Digital Medicine **3**, 119 (2020)
5. Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R.R., Bakas, S.: Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports **10**, 12598 (2020)
6. Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients – how easy is it to break privacy in federated learning? In: Advances in Neural Information Processing Systems (NeurIPS) **33**, 16937–16947 (2020)
7. Li, T., Sahu, A.K., Talwalkar, A., Smith, V.: Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine **37**(3), 50–60 (2020)

8. Beutel, D.J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., Sani, L., Kwing, H.L., Parcollet, T., Gusmão, P.P.d., Lane, N.D.: Flower: A friendly federated learning research framework. arXiv preprint arXiv:2007.14390 (2020)
9. Dwork, C., Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science **9**(3–4), 211–407 (2014)