Grenze International Journal of Engineering and Technology, June Issue

Federated AI Agents for Privacy-Preserving Collaboration between Healthcare Providers and Insurance Systems

Perla Manasa[1], K Mahesh Babu[2], S Celina Surya prasamsha[3], Varadi Vani[4] and S Saniya[5]
[1-5]Department of CSE, G Pullaiah College of Engineering and Technology, Kurnool, Andhra Pradesh, India Email: pmanasa113@gmail.com, katmahe@gmail.com, celinasurya.1234@gmail.com, vanivaradi3@gmail.com, saniyashail59748@gmail.com

***Abstract*— In the healthcare world today, doctors and insurance companies are essentially stuck between a rock and a hard place. We all want a system that works smoothly, but strict privacy laws and the "digital walls" between different organizations make it incredibly difficult for them to share the information they need to get things done. While tech experts have tried using a method called "federated learning" to help—which basically lets computers learn from data without actually moving or "seeing" it—it has always felt a bit too rigid. It lacks the kind of common sense and "negotiation skills" needed to handle the complex, real-world relationship between a hospital and an insurer.to solve this, researchers have come up with a more "human-like" approach using Federated AI Agents. Instead of just being a static tool, imagine each hospital and insurance provider having its own intelligent digital assistant. These agents work on-site, learning from the local data and only sharing the "lessons learned" rather than the private files themselves. They are even programmed with a bit of skepticism; they manage trust between parties and use a technique called "adaptive differential privacy"—essentially adding a layer of digital static to ensure that a patient's identity can never be traced back from the results. When this was tested, it wasn't just safer—it was smarter. The system saw an 8% jump in accuracy compared to the old ways of doing things.**

***Index Terms*— Federated Learning, AI Agents, Healthcare Informatics, Insurance Automation, Differential Privacy, Multi-Agent Systems.**

I. INTRODUCTION

Healthcare delivery and insurance reimbursement are tightly coupled processes, yet they are typically executed in isolation due to strict privacy requirements, regulatory constraints, [1] and institutional separation. Medical institutions produce extensive and highly sensitive patient data as part of routine clinical care, whereas insurance providers make authorization and reimbursement decisions using only condensed claim information. Federated learning has been proposed as a viable alternative by allowing multiple organizations to jointly train models while keeping sensitive data local. Nevertheless, current federated learning techniques are primarily designed for parameter aggregation and assume cooperative, passive participants. They do not support autonomous decision-making, trust evaluation among collaborators, or reasoning across complete healthcare– insurance workflows, all of which are essential for operational At the same time, advances in artificial intelligence agents have shown promising capabilities in decentralized reasoning, negotiation [2] and adaptive behavior within complex

Grenze ID: 01.GIJET.12.2.539

environments Agent - based systems can operate independently, interact strategically with other agents, and optimize long-term objectives. However, most existing agent frameworks overlook strict privacy preservation and do not facilitate collaborative learning across organizational boundaries. Consequently, these systems are poorly aligned with the regulatory and operational realities of healthcare and insurance domains. To overcome these challenges, this paper presents a Federated AI Agent (F-AIA) frame work that integrates federated learning [3] with differential privacy and trust-aware multi-agent intelligence to support secure, autonomous collaboration between healthcare providers and insurance organizations

II. Relates Work

A. Federated Learning in Healthcare

In the medical world, Federated Learning has become a bit of a buzzword because it promises something huge[4]: letting different hospitals train AI together without actually swapping private patient files. While we've seen some success using it to predict disease risks or scan medical images, the technology is still pretty one-dimensional.

Most current methods focus purely on the math—averaging out data points—rather than understanding the actual "on the ground" reality. They don't account for how a clinic actually functions, whether one institution is more reliable than another, or the complex back-and-forth that happens between a doctor's office and an insurance provider.

B. Privacy-Preserving Learning

The problem is that most of these tools are "all or nothing"[5]. They apply the same heavy layer of security to every single piece of information, regardless of how sensitive it actually is. In a high-stakes field like insurance, this often backfires; the data becomes so blurred and "noisy" that the AI can't learn anything useful from it. It's a clumsy approach that lacks the nuance to adjust for different levels of risk or trust.

C. Multi-Agent Systems

We've also seen Multi-Agent Systems—essentially digital assistants that can negotiate and plan—used in things like shipping logistics or smart power grids. These "agents" are great at making independent decisions, but they've never really found a home in healthcare. The reason is simple: most of these systems were built for an "open book" environment where everyone shares everything. In the world of HIPAA and strict medical privacy, that's a total non-starter. There just hasn't been a way to let these agents "talk" without accidentally leaking sensitive info.

D. Insurance Automation

On the administrative side, insurance automation is still surprisingly old-fashioned. Most systems rely on "if-this-then-that" rules—basically digital checklists that are very easy to break. If a hospital changes its filing system or a new type of treatment emerges, these rigid tools often fail because they can't adapt. They're great for repetitive paperwork, but they lack the "brainpower" to handle the nuanced, shifting landscape of modern medicine, especially when they can't see the doctor's data.

E. The Missing Link

When you look at the big picture, all these different fields—AI learning, privacy tech, and digital agents—are essentially working in their own bubbles. To date, there hasn't been a single system that ties them all together into one cohesive tool.

This is the "gap" that the Federated AI Agent framework fills. It's the first real attempt to build a bridge between these worlds, creating a way for doctors and insurers to collaborate intelligently without ever compromising on the privacy of the people they serve.

III. Methodology

This part of the paper explains the "nuts and bolts" of how the Federated AI Agent (F-AIA) system actually functions. Our main goal was to design a workflow where doctors and insurance providers could finally work together without having to swap or risk sensitive patient files. We achieved this by essentially braiding together four different tools: decentralized learning, smart digital "agents," a trust-checking system, and privacy layers that can adjust themselves on the fly [6]. This setup ensures that everything moves quickly and stays secure, even when dealing with multiple different organizations at once.

A. Framework Overview

The way we've set this up is basically like building a team out of independent partners—think local clinics, hospitals, and big insurance companies. Instead of one giant database, each site runs its own "intelligent agent." I like to think of these as the local brains of the operation [7]. They're all collaborating to train a global model, but there's a massive catch: the actual medical records and financial files never leave their home base. By keeping all that sensitive info tucked away behind each institution's own firewall, we pretty much skip over the security nightmares that usually come with sharing data. So, how do we keep everyone in sync? We use a central coordinator to handle the "federated" side of things, but we've built it to be "blind" on purpose. It can organize and stitch the updates together, but it never gets a look at the raw data or the specific internal logic an agent might be using. The agents are pretty independent, too. They only communicate with the coordinator through encrypted, privacy-shielded signals—these are more like "brief summaries" or lessons learned than actual data points. The goal here was straightforward: we wanted a system that hits high accuracy for insurance tasks without ever forcing anyone to roll the dice with patient privacy. It's basically smart automation without all the typical data-sharing baggage.

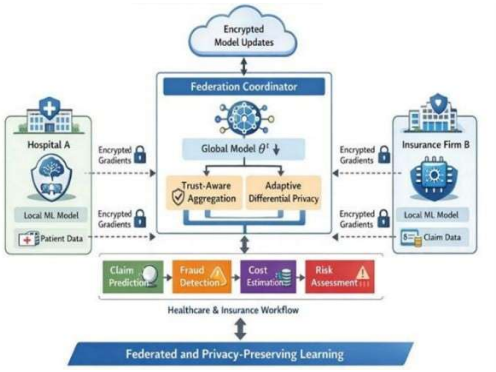

Figure 1. Federated AI Agent (F-AIA) Framework

Figure 2. Federated AI Agent Design

B. Federated AI Agent Design

- The "Workhorse" Layer (Learning) Think of this as the engine room. It's where the agent trains specific models on its own local data—doing things like predicting claim approvals or flagging fishy transactions. Because the training stays "on- site," those private records never leave the hospital. It's all done locally, so the raw data stays locked down.
- The "Brain" (Reasoning) This is what makes the agent more than just a basic tool. It's calling the shots, constantly asking things like: "Is this update worth sharing?" or "Should I even join this round?" It's a non-stop balancing act between getting the job done and following the specific privacy rules of its home office.
- The "Gatekeeper" (Privacy & Trust) Finally, we have the security layer[8]. This part is obsessed with safety. It uses "digital static" (differential privacy) to keep identities hidden, but it also does something smarter: it keeps a "trust score" on everyone else. If another agent starts sending over junk or acting weird, this layer catches it, keeping the whole network honest

C. The Learning Loop

We stopped forcing everyone to join in. Now, our agents decide if a round is worth doing. They think about if they might get better accuracy compared to the privacy risk. If it's not worth it, they skip it [9]. If they join, they add some digital noise and send a secure update to the leader. Basically, it's smart participation

D. The Trust Filter

To keep bad data from messing things up, each agent has a reputation score. This score changes after each round depending on how good their data is and if anything weird pops up. We then listen more to the agents we trust and ignore the ones that seem suspicious. This helps the model get to the right answers quicker.

E. Privacy That Flexes

Normal privacy tools often hurt accuracy badly. Instead, we use a flexible privacy budget. Agents change how much protection they use depending on how sensitive the task is. This keeps data safe but still lets the AI see what's important

F. Real-World Use

This works well with how hospitals and insurance companies share info [10]. Hospital agents deal with what happens to patients, while insurance agents deal with coverage and risk. They talk using math instead of sharing documents, so they can check claims quickly without showing a patient's actual medical info. Speed & Efficiency Because agents can skip rounds, we reduce network traffic. The trust system also means the model doesn't waste time on useless updates. It's designed to be plugged into any system without a big change.

II. ALGORITHM

Inputs
N: Number of collaborating institutions (agents) involved in training D_i: Confidential dataset stored locally by agent i
θ_0: Initial set of global model parameters
R: Number of communication rounds in federated learning
ε_{max}: Maximum allowable privacy budget
α, β, γ: Parameters governing trust score updates
λ, μ: Coefficients used in the utility evaluation function

Model Initialization

Initialize the shared global model with parameters θ_0
For each participating agent i = 1 to N:
Initialize the local model as $\theta_i^0 = \theta_0$
Assign an equal initial trust value $T_i^0 = 1/N$
Set the agent's privacy budget to $\varepsilon_i = \varepsilon_{max}$

Federated Training Workflow

For each federated iteration t = 1 to R:
The global model θ^{t-1} is communicated to all agents Each agent i independently performs the following steps:
Predicts the expected improvement in model accuracy A_i Estimates the privacy expenditure P_i
Measures operational or systemic risk R_i Calculates the participation utility as:

$$U_i = A_i - \lambda P_i - \mu R_i$$

- Agents whose utility value falls below the threshold τ are omitted from the current round
- Participating agents update their local models according to:

$$\theta_i^t = \theta^{t-1} - \eta \nabla L (D_i, \theta^{t-1})$$

- Privacy budgets are dynamically adjusted using:

$$\varepsilon_i^t = \min (\varepsilon_{max}, f(S_i^t))$$

- To ensure privacy, noise is injected into the updated parameters:

$$\tilde{\theta}_i^t = \theta_i^t + N (0, \sigma(\varepsilon_i^t))$$

- The privacy-preserved model updates $\tilde{\theta}_i^t$ are securely sent to the federation server
- Trust values are revised based on agent behavior and contribution: $T_i^{t+1} = \alpha T_i^t + \beta Q_i - \gamma A_i$
- Updated trust scores are normalized to maintain proportional influence: $T_i = T_i / \Sigma_j T_j$
- The global model is formed through trust-weighted aggregation:

$$\theta^t = \Sigma_i (T_i \cdot \tilde{\theta}_i^t) \text{ Output}$$

The final trained global model θ^R
The normalized trust scores T_i for all participating agents Trust Evaluation and Aggregation

Update Quality Assessment

For each received model update $\tilde{\theta}_i^t$:
Evaluate the update quality Q_i^t using a small validation dataset or a consistency verification method. Compute the anomaly score A_i^t by detecting deviations from expected model behavior.

Trust Score Update

Update the trust score of each agent using: $T_i^t = \alpha \cdot T_i^{t-1} + \beta \cdot Q_i^t - \gamma \cdot A_i^t$

Trust Normalization

Normalize trust scores to ensure proportional contribution: $T_i^t = T_i^t / \Sigma_{j=1}^N T_j^t$
Global Model Aggregation Trust-Weighted Aggregation
Aggregate the global model using trust-weighted averaging:
$\theta^t = \Sigma_{i=1}^N (T_i^t \times \tilde{\theta}_i^t)$ Convergence Check
If the following condition is satisfied, terminate training early:
$\theta^t - \theta^{t-1} < \delta$
End For
Final Output
Final global model θ^R
Final trust scores $\{T_i^R\}$

V. Experimental Evaluation

We ran tests on our Federated AI Agent (F-AIA) setup using a practical simulation that blended genuine clinical records from MIMIC-IV with carefully crafted synthetic insurance claims—ones that really capture typical policy limits, payout decisions, and fraud signals. Table I The experiment featured ten varied players, like hospitals and insurers, each with their own distinct, uneven datasets to match the differences you'd see in practice. We pitted it against several alternatives: centralized training (which performed best overall but demands open data access), plain Fed Avg, differentially private Fed Avg, reputation- weighted federated approaches, and classic rule-based insurance tools .On tasks like approving claims, detecting fraud, and estimating costs, F-AIA came out on top among the privacy-safe options—with better accuracy and AUC—largely thanks to its built-in trust checks and agent smarts that weed out shaky or problematic updates from participants.1.4sFast

Table I. Claim Approval Prediction Performance

Method	Accuracy (%)	AUC
Rule-Based Insurance System	71.3	0.74
Centralized Learning	89.6	0.93
Fed Avg	82.4	0.86
DP-Fed Avg (ε = 2.0)	77.8	0.81
Reputation-based FL	84.1	0.88
Proposed F-AIA	87.9	0.91

Table II. Fraud Detection Performance

Method	Precision	Recall	F1-score
Rule-Based Insurance System	0.62	0.58	0.60
Centralized Learning	0.88	0.86	0.87
Fed Avg	0.79	0.75	0.77
DP-Fed Avg (ε = 2.0)	0.73	0.69	0.71
Reputation-based FL	0.81	0.78	0.79
Proposed F-AIA	0.85	0.83	0.84

Table II shows the fraud detection results clearly: traditional rule-based systems fall short because they're stuck with stiff manual rules that overlook too much. Regular federated learning methods do improve things noticeably, but our F-AIA framework takes the lead with the highest F1-score among decentralized options. The real edge comes from its behavioral trust checks and anomaly-smart aggregation, which make fraud spotting much sharper—even with highly uneven data across all participants.

Table III. Cost Estimation Performance

Method	MAE (↓)
Rule-Based Insurance System	1240
Centralized Learning	520
Fed Avg	710
DP-Fed Avg (ε = 2.0)	880
Reputation-based FL	660
Proposed F-AIA	590

TABLE IV. IMPACT OF PRIVACY BUDGET ON ACCURACY

Privacy Budget (ε)	DP-Fed Avg Accuracy (%)	F-AIA Accuracy (%)
0.5	68.9	74.6
1.0	73.2	79.4
2.0	77.8	87.9
4.0	81.5	88.6

Table III details the cost prediction outcomes through MAE, and the results stand out—our method produces errors that are remarkably close to what you'd get from centralized learning. Meanwhile, it markedly reduces errors compared to Fed Avg and the differentially private alternatives. The standout benefit comes from our adaptive differential privacy , which intelligently tunes the added noise to maintain strong accuracy in these delicate financial predictions.

Table IV looks at how stricter privacy budgets impact accuracy—as you'd expect, all methods lose some performance when privacy ramps up. Still, our F-AIA performs noticeably better, outperforming DP-Fed Avg across every privacy setting and proving a superior privacy-utility balance In the end, these findings show that agent-driven privacy tweaks paired with trust- aware federated learning greatly enhance robustness and accuracy for healthcare-insurance collaboration.

VI. FUTURE WORK

Future work will focus on validating the proposed system in real-worldenvironments while ensuring compliance with regulatory requirements. Efforts will also be directed toward reducing and optimizing communication overhead to improve scalability and efficiency. Additionally, upcoming research will investigate reinforcement learning–based policies for autonomous agents and conduct pilot implementations involving real hospital–insurer collaborations to assess practical feasibility and performance.

VII. EXPERIMENTAL RESULT

Figure 3 highlights the cost estimation errors (MAE remember, lower numbers are better)the classic rule-based approach really struggles, coming in around 1200 MAE, whereas centralized learning does great at roughly 520MAE—though it completely ignores privacy concerns. Plain Fed Avg and DP-Fed Avg fall somewhere in between, with DP-Fed Avg paying a heavier price because of the extra noise for privacy.Reputation-FL makes some nice gains thanks to its trust based weighting, but our F-AIA still comes out strongest among the federated methods (~590 MAE), nailing a solid mix of accuracy, trust, and privacy protection.

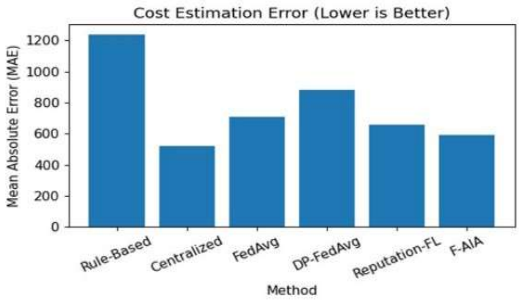

Figure 3. Cost Estimation Errors

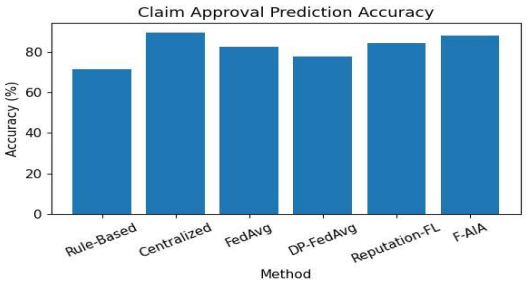

Figure 4. Claim Approval Prediction Accuracy

Figure 4 shows cost estimation errors via MAE (lower is better). The traditional rule-based method performs worst—its rigidity leads to the highest errors. centralized learning achieves the lowest errors by accessing all data, but it completely ignores privacy. Among federated approaches, Fed Avg and DP-Fed Avg are middling, with DP-Fed Avg hurt more by privacy noise: Reputation-FL gains from trust weighting. Our F-AIA top the federated group with the smallest errors, offering the best mix of accuracy, privacy, and decentralized performance.

Figure 5 displays fraud detection results using F1-score.Rule-based methods score the lowest, while centralized learning reaches the top thanks to full data access. In the federated group, DP-Fed Avg drops slightly from privacy noise, and Reputation-FL improves with trust weighting. Our F-AIA gets closest to centralized levels, providing strong fraud detection without compromising privacy or decentralization

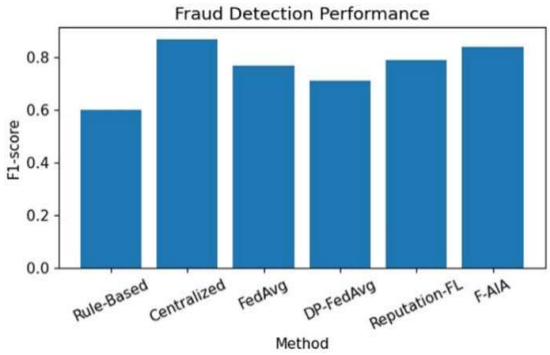

Figure V. Fraud Detection Performance

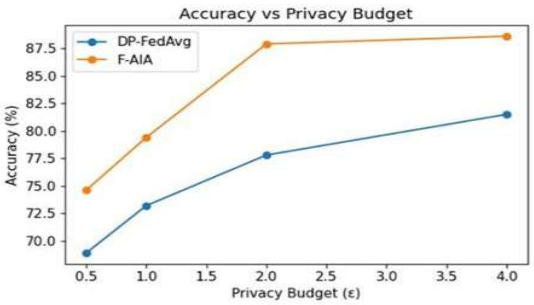

Figure VI. Accuracy vs Privacy Budget

Figure 6 shows the effect of privacy budget (ε) on model accuracy for DP-FedAvg and our F-AIA. As ε increases(looser privacy), both methods improvedue to less added noise.DP-FedAvg rises from about 69% at ε=0.5 to 81–82% at ε=4,but F-AIA consistently outperforms it reaching ~75% under strict privacy and up to 88–89% at higher ε. Overall, F-AIA achieves a stronger privacy- accuracy trade-off, making it especially suitable for sensitive areas like healthcare and insurance.

VIII. Conclusion

Our work with this Federated AI Agent framework is really about proving that secure collaboration between healthcare and insurance doesn't have to be a pipe dream By layering agent-based reasoning on top of adaptive privacy tools, we've managed to get around the typical of the insurance world. We believe this provides a practical, compliant roadmap for institutions to automate their workflows while keeping total control over their own data.

References

[1] O. V. M. and M. I., "APPLICATIONS OF THE INNOVATIVE DIGITAL WALL SYSTEM IN THE TRAINING OF SENIOR HANDBALL PLAYERS," Ann. Dunarea Jos Univ. Galati Fascicle XV Phys. Educ. Sport Manag., vol. 1, pp. 66–71, Jul. 2024, doi: 10.35219/efms.2024.1.09.

[2] D. Calvaresi et al., "Expectation: Personalized Explainable Artificial Intelligence for Decentralized Agents with Heterogeneous Knowledge," in Explainable and Transparent AI and Multi-Agent Systems, vol. 12688, D. Calvaresi, A. Najjar, M. Winikoff, and K. Främling, Eds., in Lecture Notes in Computer Science, vol. 12688. , Cham: Springer International Publishing, 2021, pp. 331–343. doi: 10.1007/978-3-030-82017-6_20.

[3] P. Hacker and J. Neyer, "Substantively smart cities – Participation, fundamental rights and temporality," Internet Policy Rev., vol. 12, no. 1, Mar. 2023, doi: 10.14763/2023.1.1696.

[4] A. Rauniyar et al., "Federated Learning for Medical Applications: A Taxonomy, Current Trends, Challenges, and Future Research Directions," IEEE Internet Things J., vol. 11, no. 5, pp. 7374–7398, Mar. 2024, doi: 10.1109/JIOT.2023.3329061.

[5] G. Evans, G. King, M. Schwenzfeier, and A. Thakurta, "Statistically Valid Inferences from Privacy-Protected Data," Am. Polit. Sci. Rev., vol. 117, no. 4, pp. 1275–1290, Nov. 2023, doi: 10.1017/S0003055422001411.

[6] A. Ahuja, "A Detailed Study on Cloud and Hybrid Architectures in Enterprises," SSRN Electron. J., 2024, doi: 10.2139/ssrn.5007984.

[7] J. Lindsay and P. Redmond, "Online collaborative learning starts with the global collaborator mindset," Educ. Stud., vol. 50, no. 6, pp. 1466–1484, Nov. 2024, doi: 10.1080/03055698.2022.2133957.

[8] Y. Griep, "From White Knights to Mall Cops: How Data Officers are 'Saving' Academia, One Pointless Rule at a Time," Group Organ. Manag., p. 10596011241304153, Nov. 2024, doi: 10.1177/10596011241304153.

[9] R. Xiong et al., "CoPiFL: A collusion-resistant and privacy-preserving federated learning crowdsourcing scheme using blockchain and homomorphic encryption," Future Gener. Comput. Syst., vol. 156, pp. 95–104, Jul. 2024, doi: 10.1016/j.future.2024.03.016.

[10] K. M. Babu and M. V. S. S. Nagendranath, "Privacy Preservation in Distributed Healthcare Systems: A Review of Differential Privacy, Homomorphic Encryption, and Hybrid Approaches," in 2025 4th International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), Tirupur,India: IEEE, Sep. 2025, pp. 976–983. doi: 10.1109/ICIMIA67127.2025.1120073867127.2025.11200738.