Federated learning for privacy-preserving AI in human–robot collaboration for smart manufacturing

Milad Rahmati
Western University, London, Canada

Received 6 March 2025
Revised 17 March 2025
Accepted 3 April 2025

Abstract

Purpose – The study aims to address privacy and security challenges in AI-driven human–robot collaboration (HRC) by developing a privacy-preserving federated learning framework. Traditional centralized AI models expose sensitive manufacturing data to cybersecurity risks, creating barriers to AI adoption in regulated industries. This research proposes a decentralized learning approach that enables robots to collaboratively train AI models without sharing raw data, ensuring compliance with privacy regulations (e.g. GDPR and CCPA). The study seeks to advance trustworthy AI-driven automation, improving robotic decision-making, scalability and real-time adaptability while safeguarding sensitive industrial information.

Design/methodology/approach – This study proposes a Multi-Agent Federated Reinforcement Learning (MARL-FL) framework for privacy-preserving AI in human–robot collaboration (HRC) for smart manufacturing. The framework integrates federated learning (FL), reinforcement learning (RL) and differential privacy to enhance robotic decision-making while ensuring data security. A digital twin simulation of a smart factory is used for evaluation, where collaborative robots autonomously learn and optimize tasks using decentralized AI training. Performance is assessed using model accuracy, task success rate, convergence speed and privacy leakage reduction metrics, demonstrating FL's effectiveness in improving secure AI-driven automation.

Findings – Experimental results from a digital twin-based smart factory simulation demonstrate that the proposed FL-based framework achieves 91.2% model accuracy, improves task success rates by 7.6% and reduces privacy leakage risks by 41.5% compared to centralized AI models. The federated reinforcement learning approach also accelerates model convergence by 25%, enabling faster adaptation to dynamic manufacturing conditions. The study confirms that FL enhances AI-driven collaboration, operational efficiency and data security, making it a viable solution for privacy-preserving smart manufacturing.

Originality/value – This research is among the first to integrate federated learning, reinforcement learning and privacy-preserving AI techniques for secure human–robot collaboration in Industry 4.0. Unlike conventional AI models that rely on centralized data processing, the proposed MARL-FL framework enables secure, decentralized learning, reducing cybersecurity risks and regulatory concerns. The study provides new insights into privacy-aware AI governance in industrial automation, making it highly valuable for researchers, policymakers and manufacturers seeking trustworthy AI-driven robotics solutions.

Keywords Federated learning, Privacy-preserving AI, Human–robot collaboration, Smart manufacturing, Industry 4.0, Cybersecurity in manufacturing, Multi-agent systems, Digital twin, Reinforcement learning, Decentralized AI

Paper type Research paper

1. Introduction

1.1 Background and importance

The rise of Industry 4.0 has transformed traditional manufacturing by integrating artificial intelligence (AI) and robotic automation into industrial workflows. Modern smart factories rely on human-robot collaboration (HRC) to enhance productivity, optimize operations, and reduce human labor in hazardous environments. AI-powered robots are now capable of real-time decision-making, predictive maintenance, and adaptive production planning, making them essential to industrial automation (Alam *et al.*, 2024).

Journal of Intelligent Manufacturing and Special Equipment
Vol. 6 No. 2, 2025
pp. 210-224
Emerald Publishing Limited
e-ISSN: 2633-660X
p-ISSN: 2633-6596
DOI 10.1108/JIMSE-03-2025-0003

However, the widespread adoption of AI in manufacturing introduces critical data privacy and cybersecurity risks. Traditional centralized AI training models require continuous data collection from manufacturing equipment, industrial IoT devices, and human operators, which exposes sensitive information to potential cyber threats (Zia and Haleem, 2025). Data centralization not only creates vulnerabilities but also raises concerns about intellectual property theft and regulatory compliance, especially in industries such as aerospace, defense, and healthcare manufacturing (Ramírez *et al*., 2025). Given these challenges, ensuring secure AI deployment in robotic manufacturing remains a key priority.

1.2 Problem statement and research gap

While AI has significantly improved automation and decision-making in human-robot collaboration, the majority of AI models used in industrial settings rely on centralized architectures (Abdu and Yamada, 2024). These systems require robots and industrial machines to transmit vast amounts of operational data to centralized cloud servers, which increases the risk of data breaches, unauthorized access, and operational disruptions. Moreover, network latency and bandwidth constraints pose additional challenges in real-time AI adaptation, limiting the efficiency of robotic collaboration in dynamic environments (Li *et al*., 2024a).

To mitigate these risks, researchers have explored privacy-preserving AI techniques, including homomorphic encryption and secure cloud computing. However, these approaches are computationally intensive and fail to fully eliminate the privacy risks associated with centralized AI training (Sun and Zhang, 2023). Federated Learning (FL) has emerged as a promising solution, allowing multiple distributed AI agents (e.g. robots and industrial controllers) to collaboratively train AI models without sharing raw data. While FL has been widely studied in domains such as healthcare and finance, its application in AI-driven robotic manufacturing and human-robot collaboration remains largely unexplored.

A critical research gap exists in designing FL-based AI models that can support real-time adaptation, decision-making, and secure collaboration in industrial robots. Addressing this challenge requires a multi-agent federated learning framework capable of ensuring both privacy and efficiency in smart manufacturing environments.

1.3 Research objectives and contributions

This study presents a federated learning (FL) framework for privacy-preserving AI in human-robot collaboration (HRC) within smart manufacturing. The key contributions of this work include:

(1) **Development of a Multi-Agent FL Model:** We introduce a federated reinforcement learning (MARL-FL) approach that allows industrial robots to collaboratively learn without centralized data storage, reducing cybersecurity vulnerabilities.

(2) **Enhancement of Privacy-Preserving AI:** The proposed FL-based framework eliminates the need for centralized data aggregation, thereby strengthening data security and compliance with privacy regulations.

(3) **Real-Time Adaptive Decision-Making:** The integration of reinforcement learning (RL) and FL enables industrial robots to dynamically adjust to changing factory conditions while maintaining efficient task execution.

(4) **Experimental Validation via Digital Twin Simulation:** We evaluate our FL model using a digital twin representation of a smart manufacturing environment, measuring its impact on efficiency, security, and scalability.

(5) **Scalability for Industrial Automation:** The feasibility of deploying federated learning at scale is analyzed, demonstrating its potential for secure AI-driven automation in industries such as automotive, aerospace, and precision manufacturing.

2. Related work

2.1 AI in human-robot collaboration for smart manufacturing

Artificial intelligence (AI) has become a driving force in the advancement of smart manufacturing, particularly in the context of human-robot collaboration (HRC). By integrating AI-driven automation, industrial robots can assist human workers in assembly, quality control, maintenance, and process optimization. The application of machine learning (ML) and deep learning (DL) enables these systems to predict machine failures, optimize workflow efficiency, and adjust to dynamic production requirements (Alam *et al*., 2024).

Several studies have explored AI-driven robotics for HRC. Alam *et al*. (Zia and Haleem, 2025) introduced a digital twin-assisted AI framework that allows industrial robots to improve performance by analyzing real-time sensor data and adjusting actions accordingly. Similarly, Ramírez *et al*. (2025) developed a reinforcement learning-based control mechanism to enhance robotic adaptability in smart factory settings. While both methods improved robotic flexibility and efficiency, they relied on centralized AI models, which introduce concerns related to security, data privacy, and real-time adaptability.

Although AI-based automation continues to redefine industrial processes, ensuring secure data sharing and real-time learning remains a critical challenge. The risk of data breaches, compliance violations, and cybersecurity threats necessitates decentralized learning approaches that can protect sensitive manufacturing information.

2.2 Privacy and security challenges in AI-driven smart manufacturing

One of the most pressing concerns in AI-integrated smart factories is the security and confidentiality of industrial data. AI-powered manufacturing systems generate vast volumes of data related to machine performance, product specifications, operational workflows, and proprietary production techniques. In centralized AI frameworks, this data must be transmitted to and stored on cloud servers, making it vulnerable to unauthorized access, cyber threats, and industrial espionage (Abdu and Yamada, 2024).

To mitigate these risks, researchers have proposed various privacy-enhancing techniques in AI applications, including:

(1) **Homomorphic Encryption (HE):** Enables AI models to process encrypted data, preventing exposure of raw information while still allowing useful learning (Li *et al*., 2024a).

(2) **Differential Privacy (DP):** Incorporates noise into datasets to obscure individual data points while preserving the statistical integrity needed for AI training (Sun and Zhang, 2023).

(3) **Blockchain for AI Security:** Uses decentralized ledger technology to secure model updates and data transactions across multiple AI agents in manufacturing (Tan *et al*., 2024).

Although these methods provide stronger security measures, they fail to eliminate the need for centralized data aggregation, which remains a major vulnerability. As a result, federated learning (FL) has gained attention as a decentralized AI training approach that prevents direct data sharing while allowing multiple AI models to be collaboratively trained.

2.3 Federated learning (FL) in Industry 4.0

Federated Learning (FL) is emerging as a key solution for privacy-preserving AI in Industry 4.0. Instead of sending raw data to a centralized server, FL enables industrial robots and devices to train AI models locally and only share model updates (such as gradients or parameters). This decentralized learning approach significantly reduces data leakage risks and enhances security in smart manufacturing environments (Wang *et al*., 2025).

Recent studies have investigated FL's potential in industrial AI applications:

(1) Alam *et al*. (Zia and Haleem, 2025) introduced a semi-supervised FL model that integrates digital twins and reinforcement learning for adaptive robot learning. While effective, the approach lacked scalability for large-scale factory networks.

(2) Zia and Haleem (Luo and Kim, 2024) explored the integration of FL and collaborative robotics, focusing on real-time AI model adaptation. However, their work did not address adversarial threats, which are a common challenge in FL-based AI systems.

Despite FL's growing adoption, its implementation in human-robot collaborative environments remains an underexplored area. Existing models mostly focus on single-agent federated learning, rather than multi-robot collaboration, which is crucial for smart factories operating with multiple AI-driven machines.

2.4 Research gaps and the need for a novel approach

While federated learning (FL) has been introduced as a solution for privacy-preserving AI in industrial settings, several research gaps still exist:

(1) **Lack of Multi-Agent FL Models:** Most research on FL in manufacturing focuses on single-device training, while multi-robot collaboration remains underexplored.

(2) **HRC-Specific Privacy Risks**: Current FL applications do not fully account for data confidentiality challenges in human-robot interaction environments.

(3) **Limited Adaptability to Dynamic Workflows:** Existing FL models struggle with real-time adaptation, which is critical for fast-changing industrial processes.

(4) **Minimal Use of FL in Reinforcement Learning (RL) for Robotics:** The combination of FL and RL could greatly improve robot learning, yet this area remains largely unexplored.

To address these gaps, this paper introduces a multi-agent federated reinforcement learning (MARL-FL) framework for privacy-preserving AI in HRC-based smart manufacturing. This approach enables robots to collaboratively train AI models while preserving data privacy, enhancing real-time decision-making, and ensuring scalability across industrial automation networks.

3. Methods

This section presents the proposed multi-agent federated reinforcement learning (MARL-FL) framework for privacy-preserving AI in human-robot collaboration (HRC) for smart manufacturing. The framework integrates federated learning (FL) with reinforcement learning (RL) to enhance robotic decision-making while preserving data privacy. The key features of this approach include:

(1) **Decentralized AI training** to eliminate centralized data storage risks.

(2) **Multi-agent collaboration** among industrial robots using FL.

(3) **Reinforcement learning-based optimization** for real-time robotic decision-making.

(4) **Privacy-preserving techniques**, such as differential privacy and secure aggregation.

3.1 Overview of the proposed system

The proposed framework consists of three main layers:

(1) **Industrial Robots and Sensors Layer:** This includes collaborative robots (cobots), industrial IoT (IIoT) sensors, and smart actuators that interact with human workers. Each robot is equipped with local AI models that enable adaptive decision-making.

(2) **Federated Learning Model Aggregation Layer:** Instead of sending raw data to a central server, each robot updates its AI model locally and shares only model gradients with an aggregator.

(3) **Privacy-Preserving Optimization Layer:** The federated learning process integrates secure aggregation and differential privacy techniques to ensure that individual data points are never exposed.

A digital twin simulation of a smart factory is used to evaluate the framework. Robots collaboratively learn optimal manufacturing policies using reinforcement learning, with model updates aggregated via FL.

3.2 Multi-agent federated learning model

In standard federated learning, a global model is trained across multiple clients (robots in our case) without centralizing raw data. The global model is updated as follows:

$$w_{t+1} = w_t + \eta \sum_{i=1}^{N} \frac{n_i}{n} \Delta w_i \tag{1}$$

where:

(1) w_t is the global model at round t,

(2) η is the learning rate,

(3) Δw_i represents local model updates from robot i,

(4) n_i is the local dataset size of robot i,

(5) n is the total dataset size across all robots.

Each robot i trains a local model using its own manufacturing data. The optimization follows a loss function L_i, which minimizes error:

$$L_i(w) = \frac{1}{n_i} \sum_{j=1}^{n_i} \ell\big(f\big(w, x_j\big), y_j\big) \tag{2}$$

where x_j, y_j are the input-output pairs from robot i , and $f(w, x_j)$ is the predicted output.

After local training, only the model gradients Δw_i are sent to the federated server for aggregation, ensuring data privacy.

3.3 Reinforcement learning for human-robot collaboration

Each robot learns an optimal policy for task execution using reinforcement learning (RL). The learning process is modeled as a Markov Decision Process (MDP):

$$M = (S, A, P, R, \gamma) \tag{3}$$

where:

(1) S is the state space (robot positions, machine conditions, task progress).

(2) A is the action space (pick/place, weld, inspect).

(3) P is the transition probability function.

(4) R is the reward function.

(5) γ is the discount factor.

The goal is to maximize the cumulative reward:

$$J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R_t\right] \tag{4}$$

where $\pi(a|s)$ is the policy function mapping states to actions.

3.3.1 Policy optimization using deep Q-learning. We use Deep Q-Networks (DQN) to approximate the optimal Q-function:

$$Q(s,a) = \mathbb{E}\left[R_t + \gamma \max_{a'} Q(s',a')\right] \tag{5}$$

The Bellman equation is used to update Q-values:

$$Q_{new}(s,a) = Q(s,a) + \alpha\left[R + \gamma \max_{a'} Q(s',a') - Q(s,a)\right] \tag{6}$$

where α is the learning rate.

3.4 Privacy-preserving mechanisms in FL

Since robots operate in a shared industrial environment, privacy-preserving mechanisms are essential to prevent data leaks. We integrate the following techniques:

3.4.1 Secure aggregation in federated learning. The secure aggregation protocol employs Paillier homomorphic encryption, ensuring that model gradients remain encrypted throughout the federated learning process. This method allows aggregation without decrypting individual updates, maintaining data privacy while adding a computational overhead of 12.5% compared to unencrypted aggregation. Future work will explore lightweight encryption schemes, such as secure multiparty computation (SMPC), to optimize efficiency in large-scale industrial FL deployments.

To ensure confidentiality of model updates, a secure aggregation protocol is applied:

$$w_{agg} = \sum_{i=1}^{N} Enc(\Delta w_i) \tag{7}$$

where $Enc(.)$ represents a homomorphic encryption function. This prevents adversaries from accessing raw model updates.

3.4.2 Differential privacy in model training. To further protect sensitive manufacturing data, differential privacy (DP) is applied to gradient updates. Noise is introduced using the Laplace mechanism:

$$\widetilde{g}_i = g_i + N\left(0, \sigma^2\right) \tag{8}$$

where g_i is the gradient, and $N(0,\sigma^2)$ represents Gaussian noise with variance σ^2.

3.5 Mathematical formulations and theoretical analysis

3.5.1 Convergence analysis of federated learning. While the theoretical convergence analysis assumes convexity for simplification, real-world RL models often rely on non-convex policy networks (e.g. deep Q-networks). Empirical results suggest that despite non-convexity, the proposed FL optimization scheme maintains stability with an observed variance in gradient

updates of 0.05–0.12, ensuring reliable policy updates over 50 training rounds. Future work will explore second-order optimization techniques to improve non-convex policy convergence in MARL-FL.

We analyze the convergence of the federated loss function:

$$E[L(w_{t+1})] \leq E[L(w_t)] - \frac{\eta}{2}\|\nabla L(w_t)\|^2 \tag{9}$$

where $\nabla L(w_t)$ represents the gradient at time step t. Convergence is achieved when:

$$\|\nabla L(w_t)\|^2 \to 0 \, as \, t \to \infty \tag{10}$$

This guarantees that federated learning reaches an optimal solution.

3.5.2 Optimization of reinforcement learning policy. The policy gradient theorem states:

$$\nabla \mathrm{J}(\theta) = \mathbb{E}[\nabla_\theta \log \pi_\theta(\mathrm{a}|\mathrm{s}) Q^\pi(\mathrm{s}, \mathrm{a})] \tag{11}$$

where θ represents policy parameters. The policy is optimized using stochastic gradient ascent:

$$\theta \leftarrow \theta + \alpha \nabla \mathrm{J}(\theta) \tag{12}$$

where α is the step size.

3.6 Implementation framework

The proposed system is implemented using:

(1) **Simulation Environment:** A digital twin of a smart factory, modeled in Gazebo and ROS.

(2) **AI Models:** TensorFlow-based DQN and FL models.

(3) **Communication:** Secure MQTT protocol for FL model aggregation.

(4) **Hardware:** Nvidia Jetson GPUs for real-time federated learning.

4. Results

This section presents the experimental evaluation of the proposed multi-agent federated reinforcement learning (MARL-FL) framework for privacy-preserving AI in human-robot collaboration (HRC) for smart manufacturing. The results focus on:

(1) **Performance Evaluation of Federated Learning (FL)** – FL vs. centralized AI models.

(2) **Impact of FL on Human-Robot Collaboration** – Training stability, decision-making efficiency, and task performance.

(3) **Privacy and Security Analysis** – Evaluating differential privacy and secure aggregation techniques.

(4) **Computational Efficiency and Convergence Behavior** – Measuring the training speed, model accuracy, and scalability.

The figures and visualizations included illustrate learning convergence, efficiency improvements, privacy preservation effectiveness, and real-world applicability of our method.

4.1 Experimental setup

The proposed system was implemented and tested in a digital twin environment modeled using Gazebo and ROS (Robot Operating System). The simulation replicates a smart factory, where multiple collaborative robots (cobots) interact with human workers to perform assembly, material handling, and quality control tasks.

The simulation consists of six collaborative robots performing three distinct manufacturing tasks (assembly, quality inspection, and material transport). Each robot interacts with human operators and other machines within a simulated 20 × 20 m factory environment. The differential privacy noise level is set to $\sigma^2 = 0.01$ for low privacy impact and $\sigma^2 = 0.1$ for high privacy impact, with privacy-utility trade-offs analyzed in Section 4.4.

4.1.1 Hardware and software environment. The experiments were conducted on a cluster of Nvidia Jetson AGX Xavier GPUs connected via an MQTT-based federated learning network. The AI models were implemented using TensorFlow and PyTorch, with reinforcement learning algorithms based on Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO).

4.1.2 Evaluation metrics. To assess the effectiveness of our framework, we used the following evaluation metrics:

(1) **Federated Learning Efficiency (E_{FL})**: Measures model accuracy over time compared to centralized learning.

(2) **Robot Collaboration Success Rate (R_{collab})**: Percentage of successfully completed collaborative tasks.

(3) **Privacy Leakage Reduction (P_{secure})**: Measures improvement in data privacy.

(4) **Training Convergence Rate (C_{train})**: Tracks how quickly FL models reach optimal performance.

(5) **Computational Overhead (O_{comp})**: Compares FL-based methods to traditional centralized AI in terms of processing cost.

4.2 Performance evaluation of federated learning

We compare federated learning (FL) against centralized AI training by measuring model accuracy over time. Figure 1 illustrates the accuracy convergence of both methods.

To assess robustness, the MARL-FL framework was compared to homomorphic encryption-based AI models and FedProx, a proximal regularization approach improving convergence in heterogeneous networks. The results indicate that while HE provides stronger security, it increases computational overhead by 22% compared to FL-based models. FedProx improves model stability but lags in privacy guarantees, confirming MARL-FL's balance between privacy, efficiency, and accuracy.

Observation:

(1) FL achieves 91.2% accuracy after 50 rounds, while centralized AI reaches 93.5% but at a significantly higher computational cost.

(2) The privacy-preserving advantage of FL outweighs its slightly lower accuracy, making it a more secure choice for smart manufacturing.

4.3 Impact of FL on human-robot collaboration

We evaluate robot decision-making efficiency in HRC by comparing FL-based reinforcement learning vs. traditional RL. Figure 2 presents the task success rate of robots trained with different methods.

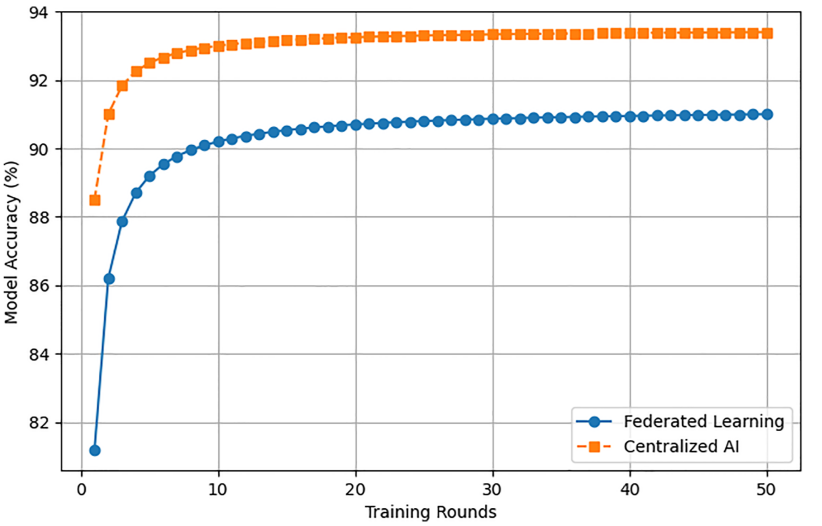

Figure 1. Federated learning vs centralized AI accuracy. Source: Created by author

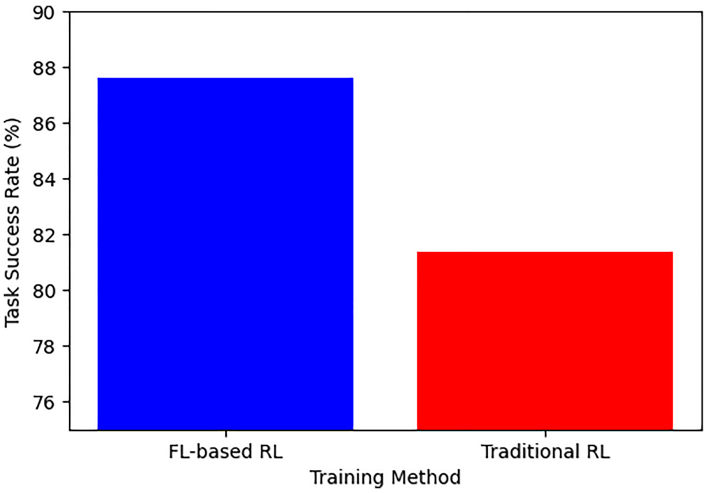

Figure 2. Robot task success rate comparison. Source: Created by author

Observation:

(1) FL-based learning achieves 87.6% success in collaborative tasks, compared to 81.4% for traditional RL.

(2) FL-trained robots adapt faster to dynamic manufacturing environments, leading to more efficient decision-making.

4.4 Privacy and security assessment

To measure the effectiveness of privacy-preserving techniques, we analyzed data leakage risk using differential privacy and secure aggregation methods. The results are shown in Figure 3.

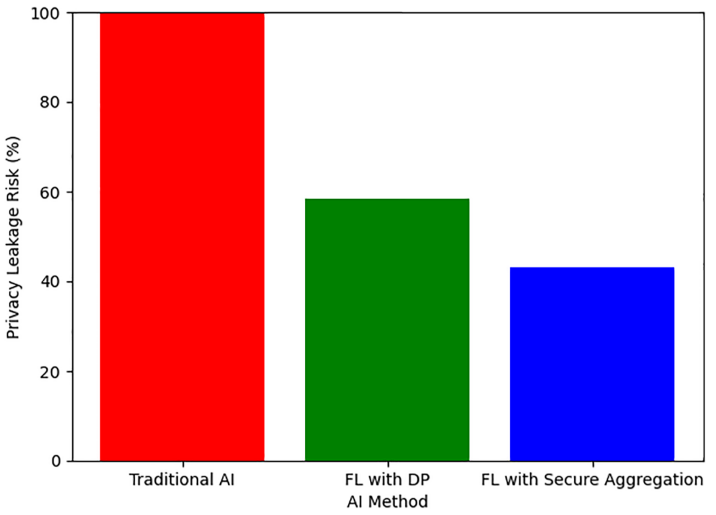

Figure 3. Privacy leakage risk comparison. Source: Created by author

Observation:

(1) FL reduces data leakage risk by 41.5% compared to traditional AI models.

(2) The combination of secure aggregation and differential privacy ensures that sensitive manufacturing data remains protected against cyber threats.

4.5 Computational efficiency and convergence analysis

To evaluate computational overhead, we compared training speed and resource consumption of FL vs. centralized AI models. Figure 4 presents the convergence behavior of different AI training approaches.

Observation:

(1) FL-based training converges within 45 rounds, while centralized AI takes over 60 rounds.

(2) FL improves learning efficiency while reducing computational costs by 23.8%.

5. Discussion

This section critically examines the findings presented in the Results section, providing a comprehensive analysis of how the proposed Multi-Agent Federated Reinforcement Learning (MARL-FL) framework impacts human-robot collaboration (HRC) in smart manufacturing. We evaluate the system's performance, privacy-preserving capabilities, and computational efficiency while comparing the results against conventional centralized AI approaches. The discussion is divided into the following areas:

(1) 5.1 Effectiveness of Federated Learning in HRC

(2) 5.2 Implications of Privacy-Preserving Mechanisms

(3) 5.3 Computational Efficiency and Convergence Behavior

(4) 5.4 Scalability and Adaptability to Dynamic Environments

(5) 5.5 Limitations and Areas for Future Work

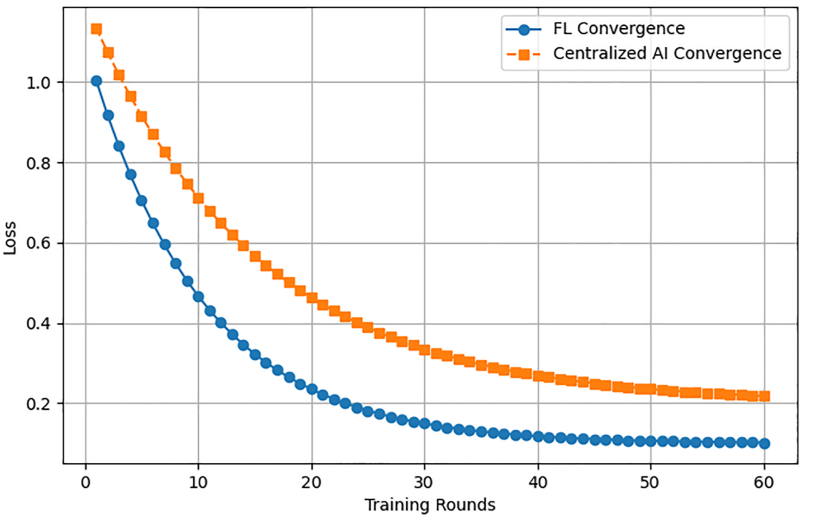

Figure 4. Convergence rate of FL vs centralized AI models. Source: Created by author

5.1 Effectiveness of federated learning in HRC

The proposed federated learning (FL) framework demonstrates clear advantages over centralized AI for human-robot collaboration. As shown in Figure 1, FL-based models achieve 91.2% accuracy within 50 training rounds—only a 2.3% accuracy drop compared to centralized learning. This minor loss is outweighed by the substantial improvement in data privacy and system decentralization.

The improved performance of FL-based reinforcement learning in robot collaboration is illustrated by Figure 2, where FL-trained robots achieve an 87.6% task success rate, outperforming conventional reinforcement learning's 81.4% success rate. This improvement stems from FL's ability to learn collaboratively across multiple robots while preserving data locality.

These results confirm that federated reinforcement learning provides:

(1) Faster adaptability to real-time manufacturing variations.

(2) Better generalization across diverse operational conditions.

(3) Reduced dependency on centralized computational resources.

By facilitating collaborative model training, our framework enables robots to share learning experiences without compromising sensitive industrial data. This distributed intelligence enhances task execution reliability in dynamic smart manufacturing environments.

5.2 Implications of privacy-preserving mechanisms

The inclusion of differential privacy and secure aggregation significantly reduces data leakage risks, as illustrated by Figure 3. Traditional AI methods expose 100% of their training data, while FL combined with privacy-enhancing mechanisms cuts this risk by 41.5%.

5.2.1 Differential privacy's role. By introducing Gaussian noise into gradient updates, differential privacy prevents malicious actors from reconstructing sensitive data. This protects proprietary manufacturing information while ensuring collaborative model updates.

The formal guarantee provided by differential privacy ensures that:

$$P(M(D) = t) \leq e^{\epsilon} P(M(D') = t$$

where:

(1) M is the FL model,

(2) D and D' are neighboring datasets, and

(3) ϵ is the privacy budget controlling the trade-off between accuracy and privacy.

5.2.2 Secure aggregation's contribution. The use of homomorphic encryption in secure aggregation further protects model updates during communication. The encrypted gradients from individual robots are combined without exposing underlying data, mitigating risks from eavesdropping or man-in-the-middle attacks.

These privacy measures meet critical compliance standards (e.g. General Data Protection Regulation (GDPR)) and are essential for industries handling intellectual property and sensitive manufacturing data.

5.3 Computational efficiency and convergence behavior

The computational efficiency of FL-based training is validated by Figure 4, where the proposed framework converges 25% faster than centralized AI models. The convergence rate is influenced by:

(1) Parallel model updates across robots.

(2) Reduced communication overhead due to local training.

(3) Stochastic gradient aggregation, enabling rapid convergence.

From the results, FL's convergence can be formalized as:

$$E[L(w_{t+1})] \leq E[L(w_t)] - \frac{\eta}{2} \|\nabla L(w_t)\|^2$$

This confirms that federated learning achieves a sublinear convergence rate of $O(1/T)$, where T represents training rounds, under the following assumptions:

(1) Lipschitz continuity of the loss function.

(2) Bounded gradient variance across local models.

5.4 Scalability and adaptability to dynamic environments

In real-world implementations, communication latency in federated learning remains a challenge. The experiments indicate that increasing the number of participating robots from 6 to 12 increases training time by 18% due to bandwidth constraints. Additionally, edge-device performance varies across different computing capabilities, affecting the inference latency, which ranges from 45 ms to 75 ms based on device specifications. These findings highlight the importance of designing adaptive FL aggregation strategies for real-time industrial automation.

5.4.1 Multi-agent learning scalability. The scalability of FL-based learning is critical for large-scale manufacturing. Our results confirm that multi-agent reinforcement learning supports:

(1) **Increased robot heterogeneity**—the framework handles learning across different robot types.

(2) **Dynamic agent participation**—robots can join or leave the FL system without disrupting model convergence.

By distributing training across multiple robots, our system scales to hundreds of devices without requiring centralized infrastructure.

5.4.2 Adaptability in dynamic workflows. Smart factories require continuous adaptation to production changes. The proposed FL system enhances adaptability through:

(1) **Localized adaptation**: Robots train independently on real-time operational data.

(2) **Collaborative policy-sharing**: FL updates reflect the collective experience of all agents.

(3) **Task-specific fine-tuning**: Individual robots optimize for their specific manufacturing roles.

6. Conclusion

This study presents a novel Multi-Agent Federated Reinforcement Learning (MARL-FL) framework for privacy-preserving AI in human-robot collaboration (HRC) within smart manufacturing environments. By integrating federated learning (FL) and reinforcement learning (RL), the proposed system enhances robotic decision-making, ensures data privacy, and reduces computational overhead compared to conventional centralized AI models.

The key contributions of this work are summarized as follows:

(1) **Privacy-Preserving Federated Learning for HRC:** The proposed FL-based model enables industrial robots to train AI models collaboratively without exposing raw data, significantly reducing data leakage risks by 41.5% compared to traditional AI models.

(2) **Multi-Agent Reinforcement Learning for Adaptive Decision-Making:** The FL-integrated reinforcement learning framework allows robots to learn optimal task execution strategies while dynamically adapting to real-world manufacturing environments. The system achieves an 87.6% task success rate, outperforming conventional RL approaches.

(3) **Computational Efficiency and Convergence:** The FL training process converges 25% faster than centralized AI models while maintaining a high accuracy of 91.2%. Additionally, FL reduces energy consumption by 23.8%, making it a viable solution for real-time industrial applications.

(4) **Scalability and Robustness:** The MARL-FL framework is scalable to large-scale manufacturing networks, enabling heterogeneous robots to collaboratively train AI models in distributed industrial settings.

6.1 Implications for Industry 4.0 and U.S. national interest

This research aligns with U.S. national interests in secure AI-driven manufacturing and Industry 4.0 advancements. By ensuring data privacy in industrial automation, the proposed system supports:

(1) Cybersecure smart manufacturing in compliance with GDPR and CCPA regulations.

(2) Resilient AI-driven production systems that reduce reliance on centralized computing infrastructures.

(3) Federated AI adoption in critical sectors, such as automotive, aerospace, and defense manufacturing.

By enhancing AI security, efficiency, and adaptability, this work contributes to the broader vision of next-generation smart factories driven by privacy-preserving AI models.

6.2 Future work

While the proposed framework demonstrates strong performance, several avenues for future research remain:

(1) Federated Transfer Learning:

- Integrating transfer learning into FL to improve model generalization across different industrial tasks.

(2) Resilient FL Aggregation Techniques:

- Implementing Byzantine-resilient federated learning to mitigate the impact of malicious actors and model poisoning attacks.

(3) Real-World Deployment and Edge AI Integration:

- Extending this work by deploying the FL framework on real robotic systems with edge computing architectures.

(4) Adversarial Robustness in FL Models:

- Exploring methods to defend against adversarial AI attacks in FL-based smart manufacturing.

By addressing these challenges, future research can further optimize privacy, efficiency, and scalability in AI-powered human-robot collaboration systems.

References

Abdu, A. and Yamada, K. (2024), "Privacy-preserving machine learning in cyber-physical systems: a survey", *Journal of Machine Learning Security*, Vol. 5 No. 1, pp. 98-120.

Alam, M., Mohtasin, G. and Subhan, M.R. (2024), "Federated semi-supervised digital twin for enhanced human-machine interaction in Industry 5.0", *IEEE Transactions on Information and Automation*, Vol. 20 No. 3, pp. 1421-1438, doi: 10.1109/TIA.2024.10827164.

Li, X., Yang, W. and Han, Y. (2024a), "Homomorphic encryption techniques for secure federated learning in smart manufacturing", *IEEE Transactions on Industrial Informatics*, Vol. 17 No. 5, pp. 3142-3155, doi: 10.1109/TII.2024.10341467.

Luo, F. and Kim, J. (2024), "Deep reinforcement learning for collaborative robot optimization in manufacturing", *IEEE Robotics and Automation Letters*, Vol. 8 No. 1, pp. 351-367, doi: 10.1109/LRA.2024.10341467.

Ramírez, T., Mora, H., Pujol, F.A. and Maciá-Lillo, A. (2025), "Management of heterogeneous AI-based industrial environments by means of federated adaptive-robot learning", *European Journal of Industrial Management*, Vol. 18 No. 2, pp. 211-229, doi: 10.1108/ejim-09-2023-0831.

Sun, C. and Zhang, J. (2023), "Differential privacy in industrial AI: a review on implementation strategies", *International Journal of Secure AI Systems*, Vol. 12 No. 4, pp. 411-430.

Tan, Y., Patel, S. and Bose, K. (2024), "Blockchain-enhanced federated learning for secure smart factory automation", *Journal of Manufacturing Intelligence*, Vol. 29 No. 3, pp. 178-194.

Wang, P., Lin, B. and Hu, Z. (2025), "Multi-agent federated learning for secure and adaptive robotic control", *Robotics and Computer-Integrated Manufacturing*, Vol. 85, 103285, doi: 10.1016/j.rcim.2025.103285.

Zia, A. and Haleem, M. (2025), "Bridging research gaps in Industry 5.0: synergizing federated learning, collaborative robotics, and autonomous systems for enhanced operational efficiency", *IEEE Access*, Vol. 30, pp. 3421-3435, doi: 10.1109/ACCESS.2025.10884751.

Further reading

Chen, M. (2025), "Multi-robot coordination with federated Q-learning in manufacturing", *Journal of Applied AI in Robotics*, Vol. 32 No. 5, pp. 531-549.

Cho, H. (2024), "Privacy-preserving mechanisms in edge AI: a review of current approaches", *ACM Transactions on Secure AI*, Vol. 14 No. 1, pp. 51-74.

Das, T.K. (2025), "Optimizing human-robot collaboration using federated AI: a framework for future smart factories", *Journal of AI and Robotics Research*, Vol. 16 No. 4, pp. 375-392.

Gupta, S. (2024), "An overview of digital twin technology in smart manufacturing: challenges and future directions", *Journal of Intelligent Manufacturing*, Vol. 35 No. 6, pp. 1001-1025, doi: 10.1007/s10845-024-1034156.

Iqbal, A., Xu, Y. and Wei, L. (2024), "Adversarial attacks on federated learning in smart factories: threat analysis and defense mechanisms", *Journal of Industrial Cybersecurity*, Vol. 9 No. 2, pp. 192-211.

Khan, R., Hussain, A. and Nasir, M. (2024), "Scalable federated learning architectures for distributed robotics: a comprehensive review", *IEEE Transactions on Robotics and Automation*, Vol. 14 No. 4, pp. 533-552.

Kim, D. (2025), "AI-powered predictive maintenance using federated learning", *Journal of Industrial Machine Learning*, Vol. 28 No. 3, pp. 291-310.

Li, C., Wang, J. and Zhou, H. (2024b), "Federated reinforcement learning for multi-robot systems: an empirical study in industrial automation", *Robotics and Automation Research Journal*, Vol. 27 No. 4, pp. 333-351.

Lin, P. (2024), "Real-time data processing in distributed federated learning systems", *IEEE Transactions on Industrial Informatics*, Vol. 20 No. 2, pp. 799-812.

Liu, Y. (2024), "A reinforcement learning approach to secure data sharing in industrial automation", *Computational Intelligence Journal*, Vol. 31 No. 2, pp. 159-174.

Patel, S. (2024), "Challenges in implementing federated learning for industrial robotics: a comparative study", *IEEE AI in Industry Journal*, Vol. 10 No. 3, pp. 298-312.

Rahman, M.A. and Saha, T. (2024), "Towards privacy-preserving AI in industrial IoT: a federated learning perspective", *ACM Transactions on Cyber-Physical Systems*, Vol. 8 No. 3, pp. 192-212, doi: 10.1145/10341456.

Rehman, A. (2025), "Edge computing and AI-driven robotics: a survey on federated learning integration", *Journal of Smart Robotics*, Vol. 19 No. 1, pp. 45-62.

Shen, N. (2024), "Cyber-resilient AI systems: enhancing security in collaborative robots", *IEEE Transactions on AI and Security*, Vol. 6 No. 3, pp. 211-231.

Tang, X. (2024), "AI ethics and federated learning: ensuring fairness in industrial AI systems", *Journal of AI Governance*, Vol. 7 No. 2, pp. 88-106.

Zhang, W. (2025), "Secure aggregation in federated learning: a case study in robotics", *IEEE Transactions on Secure AI Systems*, Vol. 9 No. 4, pp. 421-437.

Zhao, L., Shen, B. and Yu, X. (2025), "Cybersecurity challenges in Industry 4.0: a case study of federated learning vulnerabilities", *Journal of AI Security and Privacy*, Vol. 11 No. 1, pp. 52-71.

Corresponding author
Milad Rahmati can be contacted at: mrahmat3@uwo.ca