Centralized vs. Federated Learning in Healthcare: Addressing Data Privacy, Communication Efficiency, and Model Robustness

Yemi T. Fatokun, Arome J. Gabriel, Olufunso D. Alowolodu, Adebayo O. Adetunmbi
Osun State University, Osogbo
Federal University of Technology, Akure
Nigeria

Abstract

Healthcare data is inherently sensitive, making privacy and security crucial when developing AI models for medical applications. Centralized learning, the traditional approach, involves collecting data in a single location, which raises privacy concerns and increases data transfer costs. Conversely, federated learning allows decentralized model training, keeping data on local devices to preserve privacy. This paper provides a comparative analysis of centralized and federated learning in the healthcare domain, focusing on three critical aspects: data privacy, communication efficiency, and model robustness. We highlight the trade-offs between centralized data processing and decentralized approaches in terms of protecting patient data, reducing communication overhead, and maintaining model performance across diverse, non-IID datasets. The paper also discusses current limitations and proposes potential strategies to enhance model robustness and privacy in federated learning, aiming to guide future research in privacy-preserving AI for healthcare.

1. Introduction

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that focuses on creating statistical models and algorithms that let computers learn from, make choices about, and forecast based on data [1],[2]. Traditionally, researchers have to source vast amounts of data and move this data from different locations to a single location for learning. This generates lots of concern about the security and privacy of the data collated as some policies prevent data from leaving the data owners [3]. In Healthcare institutions, the medical data generated by each healthcare center is either stored locally on-premises or globally in the cloud [4]. More often a center might consider using hybrid measures [5]. Due to the volume of the medical data, its complexity, the running cost of holding data, privacy issues, and policies, a lot of medical institutions prefer to hold onto their medical data for privacy concerns [6]. In 2006, Google came up with Federated Learning (FL) [7],[8] which will allow healthcare institutions to collaborate to be able to formulate models by training data without revealing the privacy of the data [9]. Federated learning works by training a central model across decentralized devices or servers. Instead of moving all data to a central location, the model is trained locally on each device, and only the model updates are shared. Hence, Federated learning aims at training a machine learning algorithm [10].

2. Centralized Learning

Centralized Learning refers to the traditional machine learning approach where data is collected from various sources and centralized in a single location [11] as shown in Figure 1, typically on a central server or in a data warehouse, where the machine learning model is trained. This approach contrasts with federated learning, where the data remains decentralized, and only the model parameters

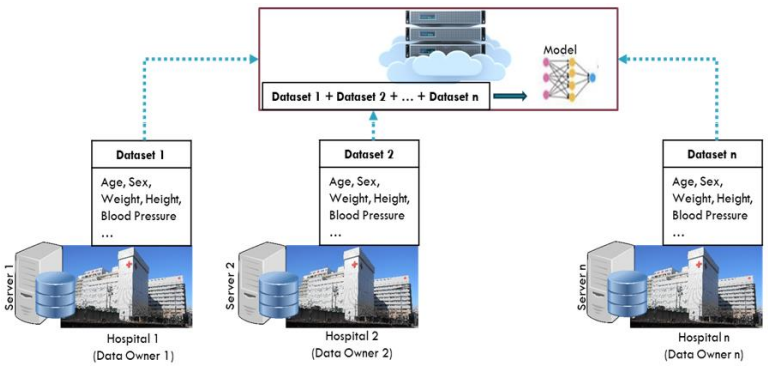

Figure 1. Overview of Centralized Learning

 | DOI: 10.20533/ijehis.2046.3332.2024.0037

are shared [12].

Key Characteristics of Centralized Learning

The key characteristics of Centralized Learning are but not limited to:

- Data Centralization: All data is gathered and stored in a central location (e.g., a cloud server or data center) where the machine learning model is trained. Data is collected from various devices, sources, or nodes and transferred to a central server [13],[14],[15],[16].
- Single Training Location: The model is trained on the centralized server using the entire dataset. Training usually happens at once, or in predefined batches, but in one central place [17],[18].
- Raw Data Transmission: In centralized learning, raw data from different sources is transmitted to the central location. This can include images, text, sensor data, and other formats [19].
- High Computation Requirements: The central server must have sufficient computational resources (e.g., CPU, GPU, memory) to handle the entire dataset and perform training efficiently.
- Model Update Process: Once the model is trained, it is deployed to the users or systems that require it. When the model is updated, data must be re-aggregated and retrained on the central server.

Workflow of Centralized Learning

The model developer collects data from users or sources:

- Sends and stores all the data in a centralized server.
- Train the model on this entire dataset.
- And deploy the model for inference or predictions.

Strengths of Centralized Learning

Most of the papers reviewed concluded that the strengths of CL lie in the following:

- Large-scale data availability: All data is collected in one place, making it easier to train high-performance models, particularly for intricate jobs like picture identification, natural language processing, and more.
- Efficient model training: Training the model over all datasets allows for easy optimization and fine-tuning without concerns about distributed computation.
- Comprehensive evaluation: Having access to the entire dataset enables better evaluation of the model's performance across different data types and distributions.

Challenges of Centralized Learning

- Data privacy concerns: Collecting all data in one place raises significant privacy risks, especially for sensitive data (e.g., healthcare, financial, or personal data). Compliance with privacy laws (like GDPR) can hinder the transmission of raw data in healthcare.
- Data transfer costs: Moving large amounts of data from edge devices to a central server can be costly and time-consuming, particularly when dealing with IoT devices or geographically dispersed users.
- Single point of failure: If the central server experiences downtime or is compromised, the entire system is affected. This introduces security and robustness concerns.
- Bias towards dominant data sources: Since all data is treated equally in centralized learning, dominant data sources (e.g., large contributors) can skew the model towards patterns found in their data, potentially neglecting underrepresented or minority data.

3. Federated Learning

Federated learning is a decentralized approach to model training, where data remains on the local devices (e.g., smartphones, IoT devices), and only model updates (like gradients or weights) are shared with a central server as depicted in Figure 2. The central server aggregates these updates to improve the global model without accessing the actual data.

Workflow

A global model is initialized and sent to the local devices:

- Each device trains the model locally on its own data using techniques like Stochastic Gradient Descent (SGD).
- Local models send their updates (not data) to the central server.
- The central server aggregates the updates (e.g., using Federated Averaging) and updates the global model.
- The updated global model is sent back to the devices for the next round of training.

 | DOI: 10.20533/ijehis.2046.3332.2024.0037

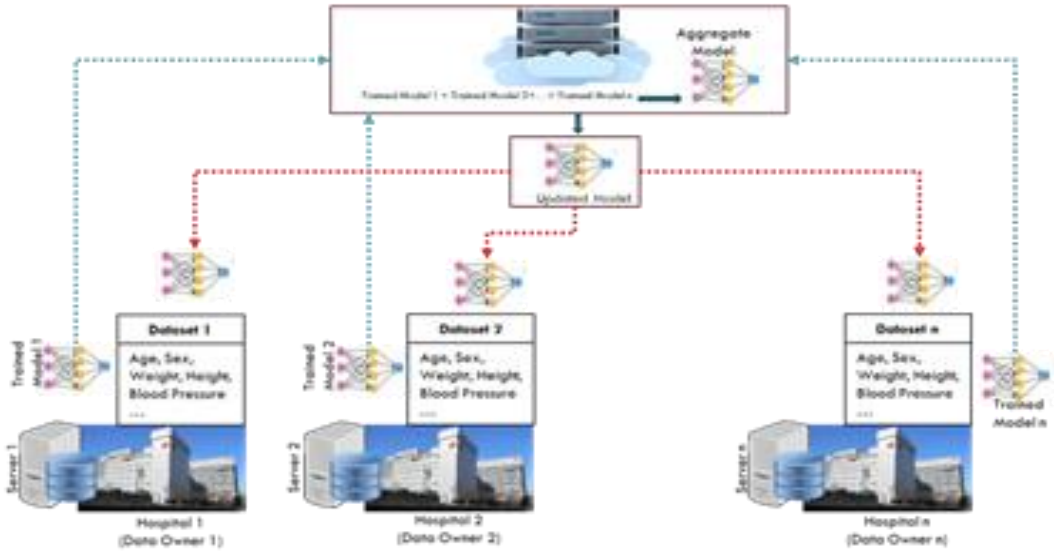

Figure 2. Overview of Federated Learning

Strengths of Federated Learning

- Data privacy: Since data never leaves the devices, FL reduces privacy risks significantly, allowing compliance with privacy regulations such as GDPR and HIPAA.
- Reduced communication costs: FL minimizes the need to transfer large volumes of data across networks, focusing instead on sharing smaller model updates.
- Personalization: Each device can train a model that is more personalized to its local data. While the global model serves all users, local models can improve based on unique characteristics of individual datasets.
- Scalability: FL allows learning across massive networks of devices (e.g., smartphones, IoT devices) without requiring all data to be centralized, enabling scaling across large numbers of devices.

Challenges in Federated Learning

- Data heterogeneity: Different devices may have non-IID (non-independent and identically distributed) data, meaning the data on one device may have a different distribution than on another. This can make training challenging because models must generalize across diverse datasets [20].
- Communication overhead: While FL reduces the need to transfer data, it can be expensive for devices and the central server to communicate model updates on a regular basis, especially in networks with limited bandwidth or unreliable connections.
- Device variability: Devices participating in FL may have varying computational power, battery life, and network connectivity, leading to imbalances in the training process (e.g., slower devices may delay the aggregation process).
- Security risks: Though data remains local, FL is vulnerable to attacks like model poisoning or backdoor attacks, where malicious devices send faulty updates to compromise the global model [21].
- Convergence and stability Due to decentralized data and device variability, FL often requires more sophisticated algorithms to ensure the model converges effectively. Balancing local updates and global model aggregation requires careful tuning of hyperparameters.

4. Algorithms for generating Models

Supervised Learning

Machine learning algorithms are diverse and can be broadly categorized into supervised, unsupervised, reinforcement, and semi-supervised learning [22]. Here's an outline of commonly used algorithms under these categories:

- Supervised Learning Algorithms: In supervised learning, the model learns from labelled data.
- Linear Algorithms: Linear Regression: Predicts continuous values based on a linear relationship.
- Logistic Regression: Used for binary classification problems.

 | DOI: 10.20533/ijehis.2046.3332.2024.0037

Non-Linear Algorithms

- Decision Trees: Builds a tree structure for decision-making based on input features.
- Random Forest: An ensemble of decision trees for improved accuracy.
- Support Vector Machines (SVM): This algorithm finds a hyperplane to separate classes in feature space.
- Neural Networks: This mimics the structure of the human brain to learn complex relationships.
- Gradient Boosting: XGBoost and LightGBM combine weak learners iteratively for robust predictions.

Unsupervised Learning Algorithms

In unsupervised learning, the model learns patterns from unlabelled data:

Clustering Algorithms

- K-Means: Partitions data into *k*. k clusters based on feature similarity.
- Hierarchical Clustering: Builds nested clusters by merging or splitting clusters iteratively.
- Density-based spatial *clustering* of applications with noise (*DBSCAN*) *clustering* method (DBSCAN): Groups data points into dense regions and labels outliers.

Dimensionality Reduction Algorithms

- Principal Component Analysis (PCA): Reduces dimensionality by projecting data onto principal components.
- t-SNE (t-Distributed Stochastic Neighbor Embedding): Visualizes high-dimensional data in 2D or 3D space.
- Autoencoders: Neural networks for compressing and reconstructing data.

Reinforcement Learning Algorithms

Reinforcement Learning involves learning by interacting with an environment to maximize cumulative reward:

- Q-Learning: A value-based method that learns the action-value function.
- Deep Q-Networks (DQN): Combines Q-learning with deep neural networks.
- Policy Gradient Methods: Directly optimizes the policy function (e.g., REINFORCE).
- Proximal Policy Optimization (PPO): Balances exploration and exploitation for robust learning.

Semi-Supervised Learning Algorithms

Semi-supervised learning uses both labelled and unlabelled data:

.

Self-Training: Iteratively labels the unlabelled data using predictions from the model itself.

Graph-Based Methods: Use graph structures to propagate label information.

Algorithm Selection Criteria

- Data Size: Algorithms like decision trees and SVMs perform well on smaller datasets, while neural networks excel with large data.
- Problem Type: Classification, regression, clustering, or dimensionality reduction.
- Data Structure: Structured (tabular) vs. unstructured (images, text).
- Computation Resources: Neural networks require more computational power compared to simpler models.

5. Future Research Issues

Privacy and Security

- Differential privacy: Ensuring that individual data points cannot be reconstructed from model updates is crucial. Techniques like differential privacy need to be better integrated into FL pipelines without compromising model performance.
- Secure aggregation: Developing encryption techniques that allow secure aggregation of model updates without compromising efficiency or adding too much computational overhead.
- Robustness to adversarial attacks: FL is vulnerable to malicious devices. Techniques like anomaly detection, robust aggregation algorithms, and Byzantine fault tolerance are critical areas for further research [23],[24],[25].

Efficient Communication

- Compression techniques: Compressing model updates (e.g., using quantization or sparsification) to reduce the communication burden without

 | DOI: 10.20533/ijehis.2046.3332.2024.0037

losing valuable information is an ongoing research area.

- Asynchronous communication: Current FL implementations often rely on synchronous aggregation, where all devices update at the same time. Asynchronous methods could improve efficiency by allowing updates from devices as they become available, without waiting for slower devices.

Personalization and Heterogeneity

- Personalized federated learning: Allowing models to better adapt to individual devices' data without sacrificing global performance. One approach is to combine local and global models, enabling better generalization and personalization.
- Handling non-IID data: Addressing data heterogeneity remains one of the most significant challenges in FL. Developing new algorithms that can effectively deal with non-IID data distributions is essential for improving performance [26],[27].

Scalability and Resource Constraints

- Scalability: FL needs to be scaled to hundreds of millions or billions of devices. Efficient techniques for selecting devices and managing their participation in training (e.g., based on device capabilities and data quality) are vital.
- Resource constraints: Many FL participants may have limited battery life, storage, and computational power. Algorithms must be lightweight to avoid overloading devices, especially in IoT or mobile environments.

Convergence and Model Stability

- Federated optimization algorithms: More research is needed to design optimization algorithms that ensure faster and more stable convergence in federated settings [28].
- Adaptive learning rates: Techniques like Adam and RMSProp need to be adapted to federated settings to improve convergence speed while managing communication costs.

6. Conclusion

Both centralized and federated learning have their strengths and limitations. Centralized learning is efficient and powerful when data can be centralized, but it faces increasing challenges related to privacy, data transfer costs, and security. Federated learning offers a decentralized alternative, preserving data privacy and enabling scalability, but it comes with significant challenges in communication efficiency, security, data heterogeneity, and convergence. Future research in federated learning will need to address these challenges to fully realize its potential in privacy-sensitive and distributed environments.

7. References

[1] Koneti C., Gonesh, D., Gonesh, S., Hasi, S., Samik, A., Manjul, S. (2023). The Impact of Artificial Intelligence and Machine Learning in Digital Marketing Strategies. European Economics Letters. 13. 982-992. DOI:10.52783/eel.v13i3.393.

[2] Mahesh Batta. (2019). Machine Learning Algorithms - A Review. International Journal of Science and Research (IJSR). 9. DOI:10.21275/ART20203995.

[3] Saravanan K., Jose A., Srinivasan R., Kavitha R., Suresh S. (2023). Handbook on Federated Learning. Advances, Applications and Opportunities. Boca Raton, CRC Press DOI:10.1201/9781003384854.

[4] Alexandru Adriana, Alexandru Cristina, Coardos Dora, Tudora Eleonora. (2016). Healthcare, Big Data and Cloud Computing. WSEAS Transactions on Computer Research. 4. 123-131.

[5] Quale John. (2017). Onsite vs. offsite. DOI:10.4324/9781315743332-9.

[6] Neame RLB. (2014). Privacy protection for personal health information and shared care records. Inform Prime Care. 21(2):84–91. http://dx.doi.org/10.14236/jhi.v21i2.55 (Access Date: 24th October, 2024).

[7] Bharati, S., Mondal, M., Rubaiyat P., Prajoy, Prasath, S. (2022). Federated learning: Applications, Challenges, and Future Directions. DOI:10.48550/arXiv.2205.09513.

[8] Kishor, K. (2022). Communication-Efficient Federated Learning. DOI:10.1007/978-3-030-85559-8_9.

[9] Ekolle, Z., Ochiai, H., Kohno, R. (2023). A Collaborative Machine Learning Model and its Application to the Security of Heterogeneous Medical Data in an IoT Network. DOI:10.36227/techrxiv.22714864.

[10] Ahmed, A., Diego O., Valentina E. (2024). Artificial Intelligence using Federated Learning. Fundamentals, Challenges, and Applications. Boca Raton, DOI:10.1201/9781003482000.

[11] Mari, F., Supianto, A., Bachtiar, F. (2023). Comparison of Federated and Centralized Learning for Image Classification. PIKSEL Penelitian Ilmu Komputer Sistem Embedded and Logic. 11. 393-400. DOI:10.33558/piksel.

[12] Gabrielli, E., Pica, G., Tolomei, G. (2023). A Survey on Decentralized Federated Learning. DOI:10.48550/arXiv.2308.04604.

[13] Velu, C., Madnick, S., Van A., Marshall. (2013). Centralizing Data Management with Considerations of Uncertainty and Information-Based Flexibility. Journal of Management Information Systems. 30. 179-212. DOI:10.2753/MIS0742-1222300307.

[14] Dandan, H., Yajuan, Z., Junfeng, L., Chen, L., Mo, X., Zhihai, S. (2017). Research on Centralized Data-Sharing Model Based on Master Data Management. MATEC Web of Conferences. 139. 00195. DOI:10.1051/matecconf/201713900195.

[15] Kristiina Sulankivi. (2004). Benefits of Centralized Digital Information Management in Multipartner Projects. ITcon Vol. 9 (2004); Sulankivi, pg. 35-63. https://www.itcon.org/papers/2004_3.content.08950.pdf (Access Date: 15 June 2025).

[16] Technical References-IDC Technologies. Advantages of Centralized Management Systems https://www.idc-online.com/technical_references/pdfs/civil_engineering/Advantages_of_Centralized_Management_Systems.pdf (Access Date: 19 June 2025).

[17] Chen Hua-Ching, Shih An, Li Chang, Tsung-Han Feng, hsuan-Ming Chen, Yun-Chien. (2024). Hybrid Centralized Training and Decentralized Execution Reinforcement Learning in Multi-Agent Path-Finding Simulations. Applied Sciences. 14. 3960. DOI:10.3390/app14103960.

[18] Elbir, A., Coleri, S. (2021). A Family of Hybrid Federated and Centralized Learning Architectures in Machine Learning. DOI:10.48550/arXiv.2105.03288.

[19] Syed R., Zeeshan A., Arifa Z., Seung Won L. (2024). Federated Learning in Smart Healthcare: A Comprehensive review on Privacy, Security, and Predictive Analytics with IoT Integration. DOI:10.3390/healthcare12242587.

[20] Zhenkun S., Pi Liu, Xiaoping L., Zhitao M., Jianqi Z., Qinhong W., Jibin Sun, Hongwu M., Yanhe Ma. (2022). Data-Driven Synthetic Cell Factories Development for Industrial Biomanufacturing. BioDesign Research. DOI:10.34133/2022/9898461.

[21] Hu Kai, Gong Sheng, Zhang Qi, Seng Chaowen, Xia Min, Jiang Shanshan. (2024). An overview of implementing security and privacy in federated learning. Artificial Intelligence Review. 57. DOI:10.1007/s10462-024-10846-8.

[22] Jahid Hasan (2023). Security and Privacy Issues in Federated Learning. https://arxiv.org/pdf/2307.12181 (Access Date: 8 June 2025).

[23] Kandati, D., Reddy S., Anusha. (2023). Security and Privacy in Federated Learning: A survey. Trends in Computer Science and Information Technology. 8. 029-037. DOI:10.17352/tcsit.000066.

[24] Sachin D. N., Basava, A., Hegde, S. Abhijit, C. Ambesange, S. (2024). FedCure: A Heterogeneity-Aware Personalized Federated Learning Framework for Intelligent Healthcare Applications in IoMT Environments. IEEE Access. PP. 1-1. DOI:10.1109/ACCESS.2024.3357514.

[25] Riedel, P., Schick, L., Von Schwerin, R. Comparative Analysis of Open-source Federated Learning Frameworks - A Literature-based Survey and Review. Int. J. Mach. Learn. and Cyber. 15, 5257–5278 (2024). DOI: 10.1007/s 13042-024-02234-z.

[26] Tong, Q., Liang, G., Bi, J. (2022). Calibrating the Adaptive Learning Rate to Improve Convergence of ADAM. Neurocomputing. 481. DOI:10.1016/j.neucom.2022.01.014.

[27] Fatokun, Y., Gabriel, A., Alowolodu, O. (2023). Secure Electronic Medical Records Using Lattice Cryptography and Fog Computing Techniques. Applied and Computational Engineering. 2. 111-118. DOI:10.54254/2755-2721/2/20220609.

[28] Wang, R. (2024). The Experiment of Federated Learning Algorithm. Applied and Computational Engineering. 39. 145-159. DOI:10.54254/2755-2721/39/20230592.

