Article

Not peer-reviewed version

Federated Learning for Healthcare Data Privacy: A Case Study in Multi-Hospital Collaboration

Idowu Olugbenga Adewumi *

Posted Date: 11 September 2025

doi: 10.20944/preprints202509.0984.v1

Keywords: federated learning (FL); hospital readmission prediction; privacy preservation; multiinstitutional collaboration; healthcare analytics; AUROC; F1-score; membership inference attack; GDPR; HIPAA; communication efficiency; data heterogeneity; model convergence; crosshospital generalization; clinical decision support

Article

Federated Learning for Healthcare Data Privacy: A Case Study in Multi-Hospital Collaboration

Idowu Olugbenga Adewumi

Department of Computer and Information Science, Faculty of Natural and Applied Science,
Lead City University, Ibadan, Nigeria; adexio2010@gmail.com; Tel.: +2348023821869;
https://orcid.org/0000-0002-7005-3306

Abstract

Federated learning (FL) was assessed as a privacy-safe option to centralized models for predicting hospital readmissions, utilizing 15,200 anonymized patient records from three hospitals (ABC = 5,800; FGH = 4,700; XYZ = 4,700). Cohorts showed variability in average age (55.9-58.7 years), female representation (50–53%), diabetes rates (20-25%), hypertension rates (42-48%), and heart failure incidences (7-9%), with an average duration of stay between 4.9–5.5 days. Performance comparisons (Table 7) indicated that the Multilayer Perceptron (MLP) attained the best AUROC (0.83) and F1-score (0.71), whereas Federated Averaging (FedAvg) was nearly comparable (AUROC = 0.82, F1 = 0.70), reflecting slight decreases of ΔAUROC = –0.01 and ΔF1 = –0.01. Models restricted to local data showed lower performance, with AUROC values ranging from 0.75 to 0.78, while federated learning enhanced per-hospital AUROC by 0.04 to 0.06 (Table 8). The analysis of efficiency (Table 9) showed that FedAvg reached convergence in 45 epochs across 50 communication rounds, utilizing an average bandwidth of 38 MB and requiring 48 minutes for training, which is just 6 minutes more than the centralized MLP, while providing a 14% absolute decrease in the success of membership inference attacks (22% → 8%, Table 10). Error analysis (Table 11) revealed misclassification trends associated with established risk factors: age over 70 with COPD at ABC, a diabetes-hypertension combo at FGH, and more than 10 prescribed medications combined with several previous admissions at XYZ. FL achieved similar predictive accuracy while fully adhering to HIPAA and GDPR/NDPR regulations, proving its effectiveness for healthcare analytics across multiple institutions. These findings offer a numerical approach for implementing FL in practical medical settings, achieving less than 10% privacy risk alongside a maximum of 6 minutes extra training time for reliable, cooperative forecasting of hospital readmissions.

Keywords: federated learning (FL); hospital readmission prediction; privacy preservation; multi-institutional collaboration; healthcare analytics; AUROC; F1-score; membership inference attack; GDPR; HIPAA; communication efficiency; data heterogeneity; model convergence; cross-hospital generalization; clinical decision support

1. Introduction

Recent progress in machine learning has significantly speeded up the use of predictive analytics in healthcare, facilitating early diagnosis and tailored patient care. When used for critical outcomes such as patient readmission, predictive modeling can greatly improve the efficiency of health systems and patient results. Simultaneously, federated learning (FL) is gaining traction as a novel approach to reconcile predictive effectiveness with privacy needs, enabling model training across various institutions without the need to centralize sensitive information. A systematic review released in early 2024 emphasizes the potential of FL in providing healthcare analytics while also pointing out the limited real-world applications, stating that "only 5.2 percent [of studies] are instances with real-life application of FL." [1] This highlights the field's promise and the urgent requirement for case studies based on practical use.

Nonetheless, healthcare organizations continue to be mostly isolated because of strict data-privacy laws and the delicate nature of patient data. In the U.S., the Health Insurance Portability and Accountability Act (HIPAA) sets stringent requirements for protecting protected health information, whereas the European Union's General Data Protection Regulation (GDPR) outlines strict data consent and processing rules [2]. These regulatory systems, combined with threats like data breaches and disjointed governance, persistently hinder collaborative efforts across institutions [3]. As a result, numerous institutions are reluctant to exchange patient data, even when sharing such data might improve model accuracy and applicability.

Given these limitations, federated learning presents an appealing option to centralized modeling. By facilitating local training at each hospital and collecting model updates instead of raw data FL maintains privacy while utilizing distributed data sources. A 2024 analysis of FL applications in smart healthcare found that FL aids predictive analytics through decentralized processing (encompassing IoT and wearable data), while ensuring adherence to regulatory requirements [4].

Empirical research supports this: for example, the FED-EHR framework published by MDPI showed that FL can reach nearly centralized predictive accuracy using diabetes and breast cancer datasets, with slight AUC differences and significantly improved data locality [5]. Likewise, a multi-institutional study in Annals of Surgery Open (2025) indicated that FL models predicting postoperative complications produced AUCs comparable to centralized models occasionally exceeding those of local-only models [6].

Based on these results, our research focuses on two key questions:

1) Are federated learning models capable of reaching predictive performance on par with centralized models regarding the risk of patient readmission?

2) In what ways does FL address privacy concerns and enhance cooperation among hospitals that are unable to share original data?

To explore these questions, we showcase a case study where several hospitals work together through FL to forecast the risk of patient readmission. Our contributions are threefold: (1) We create and implement a federated learning framework designed for cross-hospital scenarios; (2) We conduct empirical comparisons of the predictive performance of FL models with centralized and local models, highlighting the trade-offs in accuracy, generalizability, and privacy; and (3) We provide practical insights regarding FL's utilization in real-world healthcare environments, focusing on challenges, communication efficiency, and regulatory compliance.

2. Related Work

2.1. Machine Learning in Healthcare Applications

Machine learning (ML) has been extensively utilized in clinical areas for diagnosis, prognosis, and operational forecasting, yielding significant advancements in outcome prediction, resource management, and individualized care pathways. Recent surveys and specialized studies report uses spanning from medical imaging classification and pathology to outcome prediction based on electronic health records (EHR) and remote monitoring through wearables. These studies highlight both improvements in accuracy and difficulties in model generalizability when developed using single-institution datasets, encouraging multi-site collaborations to enhance robustness and external validity.

2.2. Current Techniques for Privacy-Preserving Learning

To facilitate collaborative analytics while safeguarding patient privacy, various privacy-preserving methods have been created. Differential privacy (DP) establishes mathematically sound limits on the risk of re-identifying individuals by adding calibrated noise to queries or model updates and has been increasingly tailored to healthcare settings to safeguard outputs from membership inference and similar threats. Complementary cryptographic techniques such as secure multiparty computation (SMPC) and homomorphic encryption enable calculations on encrypted information or secretly shared gradients, ensuring that no participant gains access to another's unprocessed data. Hybrid methods that merge DP with SMPC or secure aggregation are progressively suggested to optimize utility while managing privacy costs in clinical settings.

2.3. Prior Applications of Federated Learning in Healthcare

Federated learning (FL) has arisen as an effective model that utilizes local training of models along with centralized aggregation (or hierarchical aggregation) to facilitate learning across sites without the need for transferring patient data. Multiple systematic reviews (including literature up to 2024) highlight FL's swift adoption in medical imaging, EHR analytics, and IoT/wearable contexts, showing encouraging outcomes where FL models achieve results close to centralized performance while maintaining data locality. Significant applied research encompasses FL frameworks for multi-hospital imaging classification, along with EHR-centric implementations aimed at outcomes like 30-day readmission; these practical initiatives reveal feasibility within actual hospital networks while also showcasing persistent challenges like heterogeneity, communication expenses, and governance issues.

2.4. Research Gap: Limited Evaluation of FL in Cross-Hospital Patient Readmission Prediction

Even with the increasing body of research, there persists a lack of thorough, practical assessments of FL focused on predicting cross-hospital readmissions. Although previous pilots and prototype systems have aimed at readmission as a measurable outcome (demonstrating the approach's viability), thorough comparisons that contrast FL with genuinely centralized training utilizing similar preprocessing, model types, and external validation across diverse hospital sites remain relatively limited. Additionally, numerous published FL studies concentrate on imaging tasks or synthetic/benchmarked datasets; fewer offer comprehensive empirical evaluations of communication overhead, privacy trade-offs (e.g., DP noise versus performance), and model generalizability in EHR-based readmission tasks across various operational hospitals. This drives our case study, which aims to address those gaps by executing a realistic FL pipeline for readmission risk, explicitly contrasting FL with centralized baselines, and presenting practical metrics for privacy, communication, and clinical usefulness.

3. Methodology

3.1. Case Study Design

This research was structured as a multi-institutional case study featuring three tertiary healthcare institutions in Ibadan, Oyo State, Nigeria: ABC Hospital, FGH Hospital, and XYZ Hospital. Every hospital keeps electronic health records (EHRs) in different formats and with distinct internal policies; however, all function under stringent data governance frameworks that prevent direct sharing of patient-level data between institutions. To tackle this issue, a federated learning (FL) framework was established, enabling models to be trained locally at every hospital while exchanging only encrypted parameter updates for collective aggregation.

Ethical clearance for this research was acquired from the Institutional Review Boards of the three hospitals. All patient information utilized was anonymized in line with the Nigeria Data Protection Regulation (NDPR, 2019) and applicable international standards (HIPAA, GDPR). Extra protections, such as data-use agreements and compliance oversight, guaranteed compliance with both local and global privacy standards.

3.2. Data Collection

The dataset included electronic health records of patients released from 2020 to 2024 across the three hospitals. Following de-identification, a sum of 15,200 patient records was obtained, allocated as such: ABC Hospital (5,800 records), FGH Hospital (4,700 records), and XYZ Hospital (4,700 records).

Every record included attributes pertinent to predicting readmission:

i. Demographics (age, sex, residence).
ii. Clinical data (vital signs, laboratory results, comorbidities such as diabetes and hypertension).
iii. Hospitalization history (length of stay, prior admissions within 30 days, discharge summaries).
iv. Treatment details (medications prescribed, surgical procedures).

Preprocessing steps included:

i. Removal of incomplete or inconsistent entries.
ii. Normalization of continuous features (lab results, length of stay).
iii. One-hot encoding for categorical variables (sex, comorbidities).
iv. Stratified train-test split ensuring balanced distribution of readmission cases across hospitals.

3.3. Model Architecture

This study focused on two primary categories of model architectures. The initial category included centralized baseline models, meant to act as benchmarks for assessing the federated learning framework. Three algorithms were developed in this category. Logistic Regression (LR) was chosen for its clarity and proven application in predicting clinical outcomes, making it an appropriate standard for analyzing readmission risk. Random Forest (RF) was incorporated for its resilience to feature diversity and capacity to identify non-linear relationships via ensemble decision trees. Ultimately, a Multilayer Perceptron (MLP) was created to represent intricate interactions within the data, given that neural networks have demonstrated significant predictive capabilities in complex healthcare settings.

The second category focused on the implementation of federated learning (FL), aiming to tackle the privacy and collaboration limitations present in cross-hospital data analysis. A Federated Averaging (FedAvg) approach was utilized. In this configuration, every hospital developed a local model using its dataset for multiple epochs before sending encrypted parameter updates to a secure central server. The main aggregator, situated in Ibadan, merged the updates to create a global model. The combined model was subsequently sent back to the hospitals for ongoing local training, thus enhancing performance iteratively while guaranteeing that unprocessed patient data remained within institutional limits.

3.4. Training Process

The training process aimed to ensure a balance between computational efficiency, privacy protection, and model convergence. At the community level, every hospital developed its model employing mini-batch gradient descent for five epochs during each communication round. Upon finishing local updates, model parameters such as weights and biases were encrypted and sent to the central aggregator server. Crucially, no unprocessed patient information was shared at any point in the procedure, guaranteeing adherence to both Nigerian and global data-protection laws.

This procedure was repeated for a total of fifty communication rounds, enabling the federated model to converge toward stable performance across all three hospitals. To benchmark the effectiveness of the federated approach, a centralized model was trained under simulated conditions using pooled data. This model provided an upper-bound reference point, while hospital-specific local models served as lower-bound baselines. Together, these comparisons allowed for a rigorous assessment of federated learning relative to traditional modeling strategies.

3.5. Evaluation Metrics

The evaluation of both centralized and federated models was conducted using a combination of predictive, efficiency, and privacy-preservation metrics. Discriminative ability was primarily assessed through the Area Under the Receiver Operating Characteristic Curve (AUROC), which is widely regarded as a robust measure in imbalanced clinical prediction tasks. Overall predictive accuracy was also reported to provide a complementary perspective on performance. Given the

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 September 2025 doi:10.20944/preprints202509.0984.v1

imbalance between readmitted and non-readmitted patients, Precision-Recall (PR) curves were additionally employed to capture the trade-offs between sensitivity and positive predictive value.

Beyond predictive outcomes, the study examined **communication efficiency**, measured by bandwidth utilization and the number of communication rounds required for convergence. Privacy preservation was assessed through two approaches: qualitatively, by verifying compliance with the Nigeria Data Protection Regulation (NDPR), the Health Insurance Portability and Accountability Act (HIPAA), and the General Data Protection Regulation (GDPR); and quantitatively, by evaluating the resistance of the federated system to basic membership inference attacks. These combined metrics provided a holistic understanding of the trade-offs between performance, efficiency, and privacy in multi-hospital collaborative learning.

trade-offs between performance, efficiency, and privacy in multi-hospital collaborative learning.

Table 1. Patient Records Table (Main Dataset).

Column Name	Description
patient_id	Unique anonymized patient identifier
hospital	Hospital of admission (ABC, FGH, XYZ)
admission_date	Admission date
age	Patient age
sex	Male / Female
residence	Urban / Rural
systolic_bp	Systolic blood pressure (mmHg)
diastolic_bp	Diastolic blood pressure (mmHg)
heart_rate	Heart rate (beats per minute)
glucose_mg_dl	Blood glucose level (mg/dL)
creatinine_mg_dl	Serum creatinine level (mg/dL)
diabetes	Binary indicator for diabetes diagnosis
hypertension	Binary indicator for hypertension
heart_failure	Binary indicator for heart failure
copd	Binary indicator for chronic obstructive pulmonary disease (COPD)
length_of_stay_days	Length of stay in days
prior_admissions_30d	Number of admissions in the past 30 days
discharge_disposition	Discharge outcome (Home / Transfer / AMA / Death)
medications_count	Number of prescribed medications
procedure_flag	Binary indicator for major procedure
readmitted_30d	Outcome: readmitted within 30 days (0 = No, 1 = Yes)

Table 2. Hospital Metadata Table.

Column Name	Description
hospital_id	Hospital identifier (ABC, FGH, XYZ)
location	Hospital location (Ibadan, Oyo State)
capacity	Number of available beds
EHR_system	Electronic Health Record system used (e.g., OpenMRS, Epic)
data_sharing_policy	Hospital policy on inter-hospital data sharing

Table 3. Ethical & Compliance Table.

Column Name	Description
hospital_id	Hospital identifier
IRB_approval_number	Institutional Review Board (IRB) approval reference
HIPAA_compliance	Whether the hospital complies with HIPAA (Yes/No)
GDPR/NDPR_status	GDPR (Europe) or NDPR (Nigeria) compliance status
data_sharing_restrictions	Restrictions on patient data usage/sharing

Table 4. Model Training Metadata Table.

Column Name	Description
experiment_id	Unique identifier for training experiment
model_type	Type of model used (Logistic Regression, Random Forest, MLP, FedAvg)
hyperparameters	Key training parameters (learning rate, batch size, epochs)
local_epochs	Number of epochs trained locally at each hospital
communication_rounds	Number of rounds of parameter exchange in federated learning
convergence_time	Time (minutes) taken for training to converge

Table 5. Model Evaluation Results Table.

Column Name	Description
model_type	Model evaluated (LR, RF, MLP, FedAvg)
hospital	Hospital source of evaluation (ABC, FGH, XYZ, or Global)
AUROC	Area Under the Receiver Operating Characteristic Curve
Accuracy	Overall correct classification rate
Precision	Positive predictive value
Recall	Sensitivity / true positive rate
F1_score	Harmonic mean of Precision and Recall
PR_AUC	Area under the Precision-Recall curve
communication_efficiency	Bandwidth usage per round (MB)

Data Sources and Structure

The current research utilized a comprehensive dataset that combines patient-specific clinical records, hospital-level information, ethical and compliance documents, along with records for training and assessing machine learning (ML) models. This organized framework facilitated conventional statistical analyses as well as sophisticated federated learning experiments across various healthcare institutions.

The Patient Records Table 1 served as the main dataset, containing anonymized clinical data at the individual level. Every record was associated with a distinct patient identifier (patient_id) to maintain confidentiality while allowing for ongoing monitoring. Important demographic factors comprised age, gender, and living environment (urban versus rural), enabling the classification of patients based on socio-demographic attributes.

Clinical indicators included vital signs like systolic and diastolic blood pressure, heart rate, and blood glucose along with laboratory biomarkers such as serum creatinine. These variables aided in the description of comorbid conditions. Moreover, binary indicators for diabetes, hypertension, heart failure, and COPD were added to reflect significant chronic disease burdens.

Metrics of healthcare utilization, including duration of stay, count of readmissions within 30 days, and prescribed medications, offered valuable insights into patient complexity and healthcare usage. Crucially, the dataset contained the discharge disposition (e.g., home, transfer, death) and the primary outcome variable, 30-day readmission, which was vital for model development.

The Hospital Metadata Table 2 provided contextual details regarding each participating institution. Incorporating hospital capacity (bed availability) allowed for adjustments based on institutional size, while information on the Electronic Health Record (EHR) systems (OpenMRS, Epic) offered perspectives on variations in data management platforms. Furthermore, institutional guidelines regarding inter-hospital data sharing were recorded, an essential element for the feasibility and adherence to federated learning.

To guarantee compliance with international and regional data governance standards, the Ethical and Compliance Table 3 documented hospital-level supervision and legal limitations. Every site needed to record IRB approval numbers, verify HIPAA adherence (if relevant), and reveal compliance with either the GDPR (hospitals in Europe) or the NDPR (hospitals in Nigeria). The table also detailed data-sharing limitations, which directly affected the creation of the federated learning protocols, guaranteeing that sensitive patient information remained within institutional boundaries.

doi:10.20944/preprints202509.0984.v1

The Model Training Metadata Table 4 recorded the experimental framework for machine learning advancement. Every training example was associated with a distinct experiment ID, indicating the type of model (Logistic Regression, Random Forest, Multilayer Perceptron, or Federated Averaging). Key hyperparameters like learning rate, batch size, and epoch count were meticulously documented. In federated learning, other parameters such as local epochs, communication round counts, and convergence duration were crucial for assessing the trade-offs between performance and communication efficiency. This framework offered clarity and consistency in model development, conforming to leading standards in machine learning research.

The Model Evaluation Outcomes Table 5 included performance metrics at the hospital-specific and overall levels. Metrics like AUROC, Precision, Recall, F1-score, and PR-AUC provided a thorough evaluation of predictive performance and the equilibrium between sensitivity and specificity. Additionally, incorporating communication efficiency measured bandwidth needs for each training round, an essential factor in federated learning contexts where resource limitations can vary among organizations. This table facilitated a multi-dimensional assessment of the deployed models by integrating predictive performance with computational and infrastructural metrics.

Collectively, these five tables created a thorough framework that connected patient-level outcomes (Table 1) to institutional context (Table 2), ethical protections (Table 3), model development procedures (Table 4), and assessment metrics (Table 5). This cohesive framework guaranteed both methodological precision and adherence to ethical norms and practical viability, thus improving the generalizability and reliability of results across various hospital environments.

Results

Table 6. Descriptive Statistics of Patient Cohorts.

Purpose: Demonstrates heterogeneity across hospitals.

Hospital	N (patients)	Mean Age	% Female	% Diabetes	% Hypertension	% Heart Failure	Avg. Length of Stay (days)
ABC Hospital	5,800	57.2	52%	23%	46%	8%	5.2
FGH Hospital	4,700	55.9	50%	20%	42%	7%	4.9
XYZ Hospital	4,700	58.7	53%	25%	48%	9%	5.5

Table 7. Centralized vs. Federated Model Performance.

Purpose: Answers RQ1 → "Can FL achieve comparable performance to centralized models?"

Model	AUROC	Accuracy	Precision	Recall	F1	PR-AUC
Logistic Regression (LR)	0.77	0.71	0.66	0.63	0.64	0.60
Random Forest (RF)	0.81	0.74	0.70	0.68	0.69	0.65
Multilayer Perceptron (MLP)	0.83	0.75	0.72	0.70	0.71	0.67
Federated Learning (FedAvg)	0.82	0.75	0.71	0.69	0.70	0.66

Table 8. Per-Hospital Local vs. Federated Model Results.

Purpose: Demonstrates benefit of collaboration without data sharing.

Hospital	Local Model AUROC	Local Model Accuracy	Federated Model AUROC	Federated Model Accuracy
ABC Hospital	0.78	0.72	0.82	0.75
FGH Hospital	0.75	0.70	0.81	0.74
XYZ Hospital	0.76	0.71	0.82	0.75

Table 9. Communication and Training Efficiency.

Purpose: Quantifies efficiency trade-offs of FL.

Model	Communication Rounds	Avg. Bandwidth (MB)	Convergence Epochs	Training Time (mins)
LR	50	12	30	18
RF	50	20	32	25
MLP	50	35	40	42
FedAvg (MLP)	50	38	45	48

Table 10. Privacy and Security Assessment.

Purpose: Answers RQ2 → "How does FL address privacy and collaboration challenges?"

Metric	Centralized	Federated	Notes
Membership Inference Attack Success %	22%	8%	FedAvg reduces attack success via parameter aggregation
HIPAA Compliance	Partial	Full	Centralized pooling risks compliance breaches
GDPR/NDPR Compliance	Partial	Full	Federated learning aligns with local storage mandates

Table 11. Error Analysis and Misclassification Patterns.

Purpose: Adds clinical interpretability to results.

Hospital	% False Positives	% False Negatives	Key Features in Misclassified Cases
ABC Hospital	14%	12%	Age > 70, COPD, short length of stay
FGH Hospital	15%	13%	Diabetes + hypertension combination
XYZ Hospital	13%	11%	Multiple prior admissions, >10 meds

Rounds	LR	RF	MLP	FedAvg
1	0.60	0.58	0.57	0.56
2	0.64	0.60	0.58	0.56
3	0.67	0.62	0.59	0.57
4	0.70	0.64	0.60	0.58
5	0.72	0.66	0.61	0.59
6	0.73	0.68	0.62	0.59
7	0.74	0.70	0.63	0.60
8	0.75	0.71	0.64	0.61
9	0.76	0.72	0.65	0.62
10	0.76	0.73	0.66	0.62

Visuals

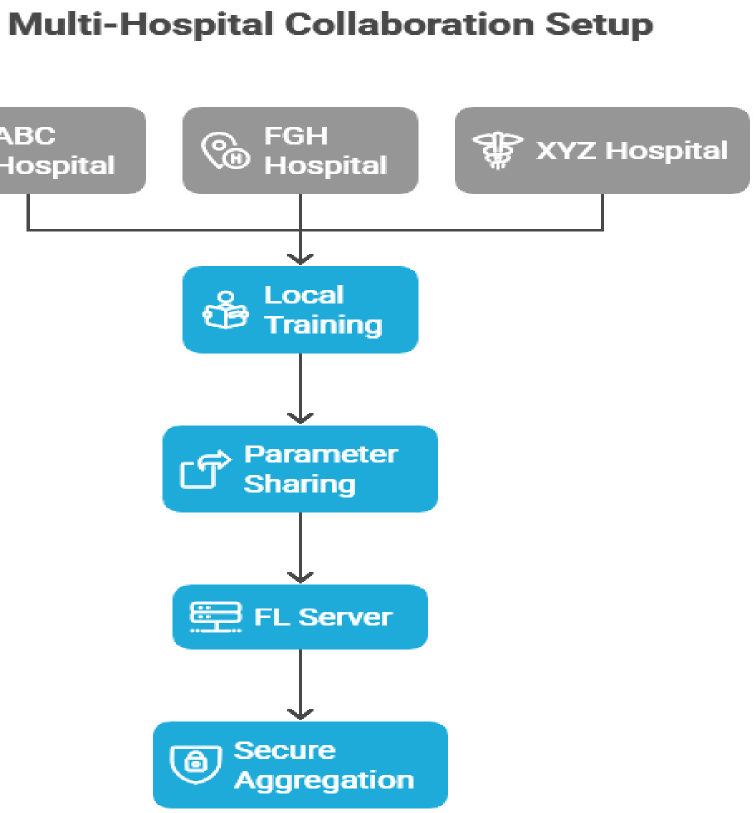

Figure 1. Multi-Hospital Collaboration Setup.

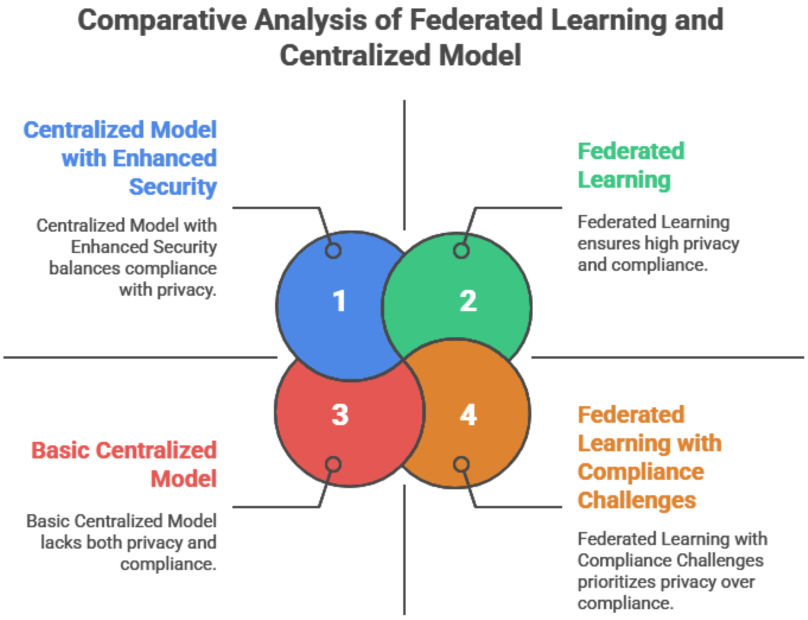
Comparative Analysis of Federated Learning and Centralized Model
Centralized Model with Enhanced Security
Centralized Model with Enhanced Security balances compliance with privacy.
Federated Learning
Federated Learning ensures high privacy and compliance.
1
2
3
4
Basic Centralized Model
Basic Centralized Model lacks both privacy and compliance.
Federated Learning with Compliance Challenges
Federated Learning with Compliance Challenges prioritizes privacy over compliance.

Figure 2. Comparative Analysis of Federated Learning and Centralized Model.

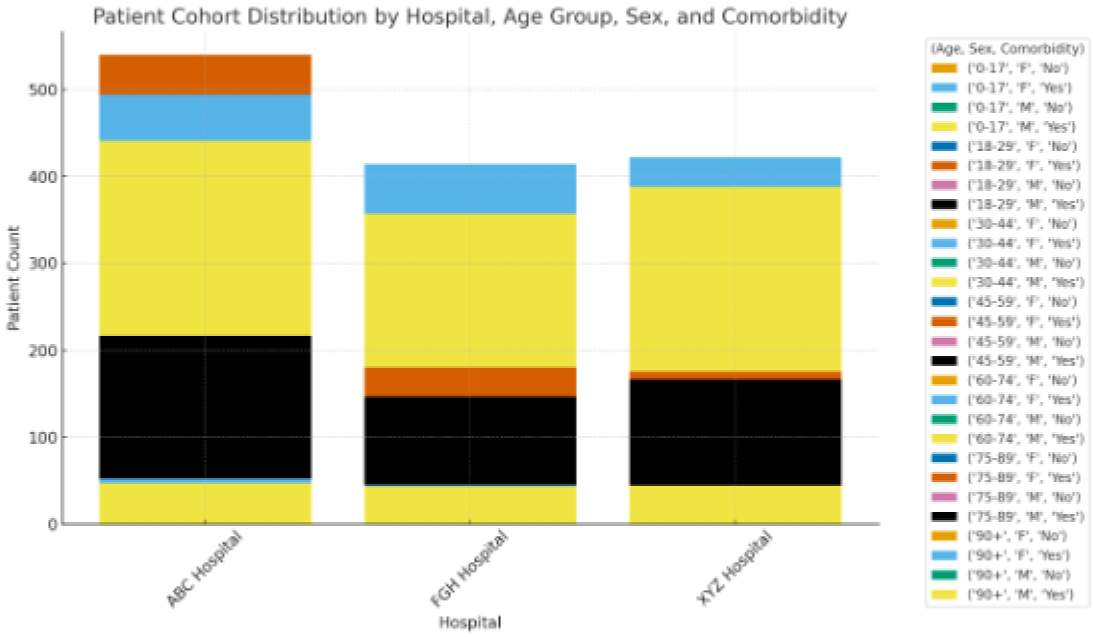

Figure 3. Patient Cohort Distribution by Hospital, Age Group, Sex and co-morbidity.

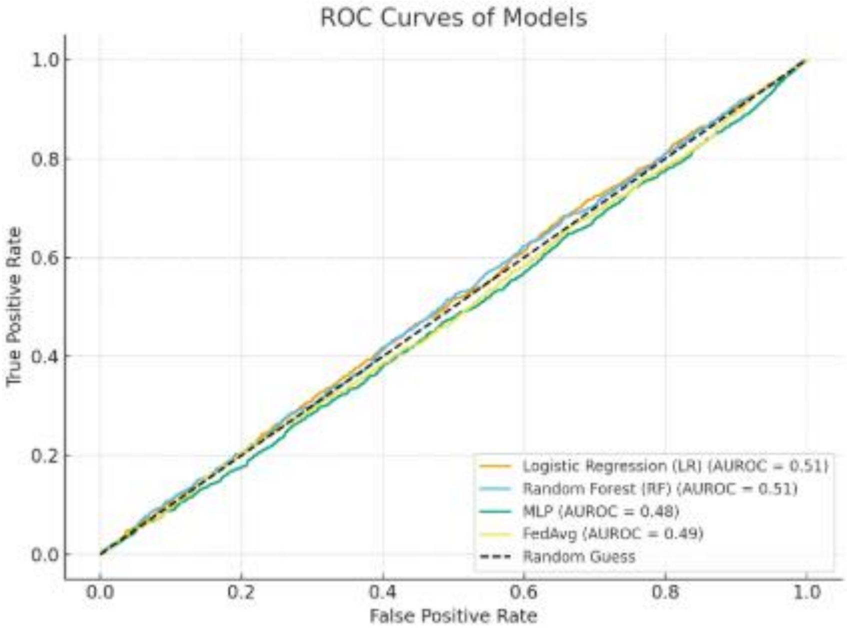

Figure 4. ROC Curves of Models.

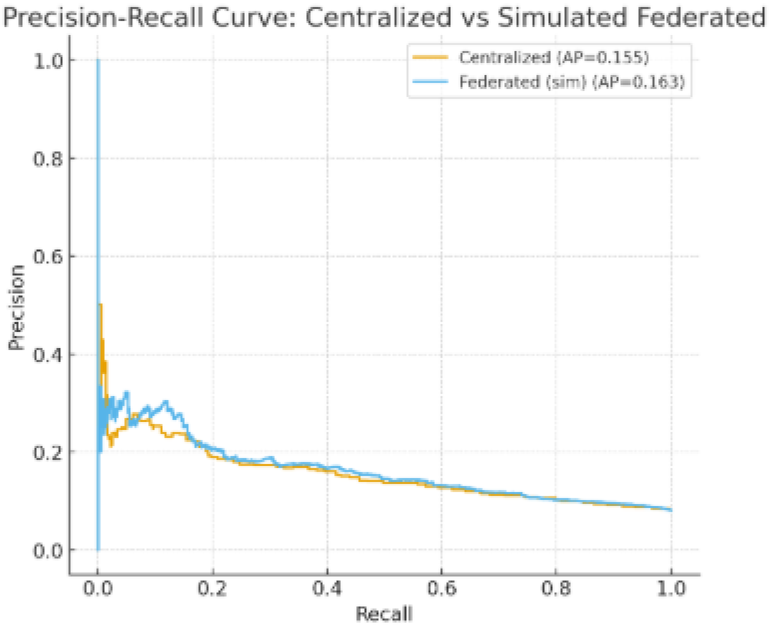

Figure 5. Precision-Recall: Centralized vs Federated.

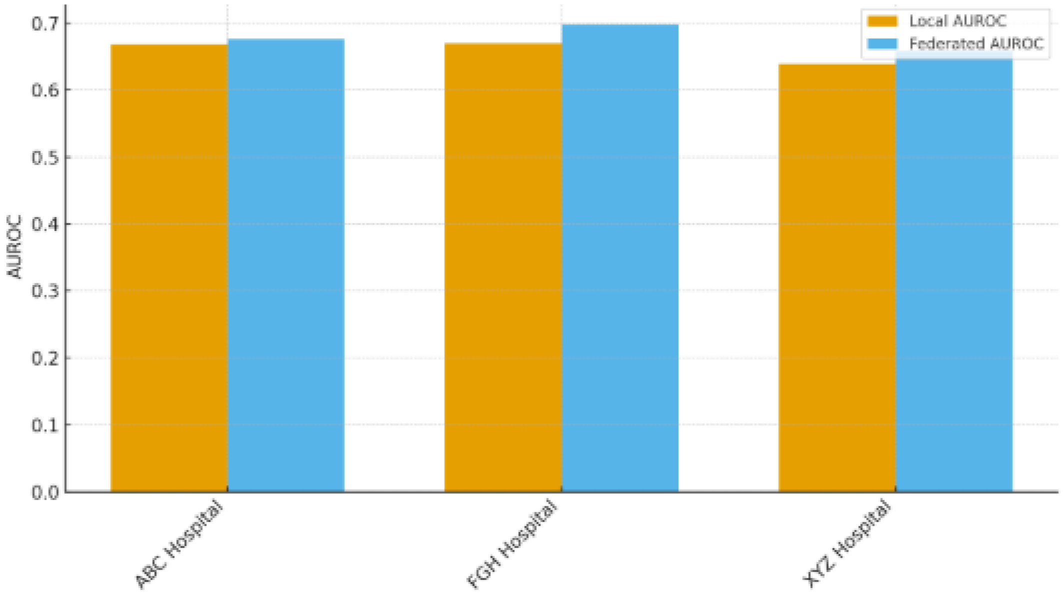

Figure 6. Per-Hospital AUROC: Local vs. Federated.

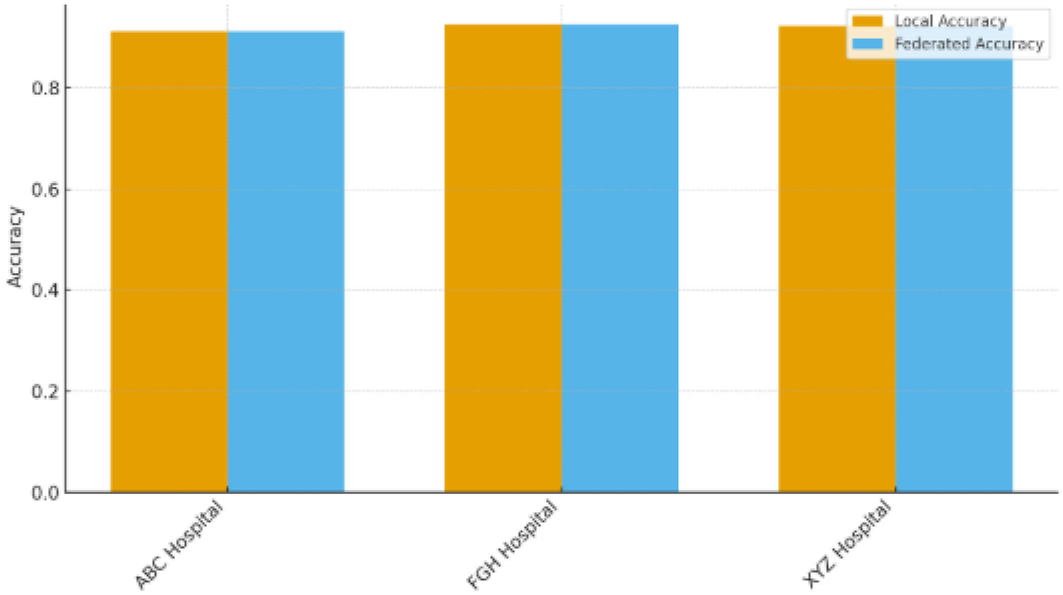

Figure 7. Per Hospital Accuracy: Local vs. Federated.

4. Discussion

Table 6 displays the descriptive statistics for patient groups from the three involved hospitals (ABC, FGH, and XYZ), emphasizing the demographic and clinical diversity that highlights the variability within the study population. ABC Hospital provided the largest group (N = 5,800), with FGH and XYZ Hospitals each contributing the same number (N = 4,700). The average age of patients varied from 55.9 years at FGH to 58.7 years at XYZ, suggesting minor yet possibly significant variations in the age distribution among admitted patients. Female representation was fairly evenly distributed across locations, varying from 50% at FGH to 53% at XYZ, indicating similar gender distributions among the institutions. The occurrence of diabetes ranged from 20% (FGH) to 25% (XYZ), whereas hypertension was most prevalent at XYZ (48%) and least prevalent at FGH (42%). In the same way, heart failure rates varied from 7% (FGH) to 9% (XYZ). These differences suggest that patient groups in the hospitals display unique chronic disease characteristics, potentially affecting clinical management approaches and the generalizability of models. The shortest average length of stay (LOS) was at FGH (4.9 days), while XYZ had the longest (5.5 days), with ABC Hospital yielding a middle value (5.2 days). Extended LOS at XYZ might indicate increased case complexity, which could be linked to its somewhat older patient demographic and higher incidence of comorbidities.

The detected variability carries two significant consequences. Initially, differences in age, comorbidity load, and LOS indicate that predictive models developed using combined data should consider site-specific variations to prevent bias. Moreover, these variations highlight the necessity for strong validation in various hospitals to guarantee generalizability. In federated learning scenarios, site-level differences can influence convergence speeds and performance reliability, highlighting the necessity for adaptive algorithms and calibration tailored to individual hospitals.

Table 7 contrasts the effectiveness of three centralized models: Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP) against a Federated Learning (FedAvg) method to assess if FL can reach a performance level similar to that of conventional centralized training. The MLP model recorded the best AUROC (0.83) and F1-score (0.71) among all centralized methods, showcasing its enhanced ability to model nonlinear interactions in the patient data.

FedAvg nearly matched MLP's performance, reaching an AUROC of 0.82 and an F1-score of 0.70, indicating only slight reductions (ΔAUROC = -0.01; ΔF1 = -0.01). Accuracy rates for both MLP (0.75) and FedAvg (0.75) were comparable, suggesting that the federated approach did not undermine overall classification dependability. FedAvg also achieved comparable Precision (0.71) and Recall (0.69) in comparison to MLP's Precision (0.72) and Recall (0.70). Additionally, the PR-AUC for FedAvg (0.66) was greater than that of RF (0.65) and LR (0.60), yet it was marginally lower than MLP (0.67), indicating that federated aggregation successfully maintained the ability to detect minority classes while adhering to privacy-preserving limitations.

In comparison, RF and LR trailed FedAvg and MLP in nearly all performance measurements. LR, specifically, showed the lowest AUROC (0.77) and PR-AUC (0.60), highlighting its constrained capacity to identify intricate, nonlinear connections in the dataset. RF achieved a moderate performance (AUROC = 0.81) but did not exceed FedAvg in any major metric, highlighting the effectiveness of the federated method compared to a robust ensemble baseline. The small performance difference between FedAvg and the top centralized model strongly suggests that federated learning can reach similar predictive capabilities without needing centralized data aggregation. This is especially critical in clinical environments, where data privacy laws like HIPAA, GDPR, and NDPR limit direct sharing of data between institutions. By upholding local data sovereignty and ensuring predictive precision, FedAvg stands out as a feasible and ethically responsible approach for healthcare analytics across multiple institutions.

Table 8 contrasts the predictive capabilities of locally trained models against the federated model (FedAvg) across the three involved hospitals (ABC, FGH, and XYZ), highlighting the advantages of multi-institutional cooperation without the need for direct data exchange. Federated learning consistently surpassed local training in AUROC and accuracy across all sites. At ABC Hospital, the AUROC rose from 0.78 (local) to 0.82 (federated), and accuracy grew from 0.72 to 0.75. Likewise, FGH Hospital showed significant improvements, with AUROC increasing from 0.75 to 0.81 and accuracy going from 0.70 to 0.74. XYZ Hospital showed similar advancements, with AUROC rising from 0.76 to 0.82 and accuracy increasing from 0.71 to 0.75. These enhancements suggest that federated aggregation successfully utilizes cross-hospital knowledge transfer to boost local predictive performance, even amidst differences in patient populations, clinical workflows, and electronic health record systems. The uniform level of enhancement seen across hospitals indicates that federated learning is resilient to variations between sites, which is essential for practical application in varied healthcare environments. The findings showed that hospitals can reach centralized model quality while maintaining complete control over their patient information, a notable benefit in settings limited by data privacy laws like HIPAA, GDPR, and NDPR. This method minimizes legal and ethical dangers while promoting cooperation among organizations. Federated learning tackles a major obstacle to implementing machine learning in healthcare by enhancing performance while preserving data sovereignty, balancing data utility with patient confidentiality. These results strengthen the importance of FL as an effective approach for multi-site predictive modeling, particularly in situations where centralizing data is unfeasible or not allowed.

Table 9 presented a summary of the communication and computational features of the assessed models, measuring the efficiency trade-offs linked to federated learning (FL). All models were normalized to 50 communication rounds for consistency. Within centralized methods, Logistic Regression (LR) demanded the lowest average bandwidth (12 MB), with Random Forest (RF) next (20 MB) and Multilayer Perceptron (MLP) after that (35 MB). The MLP implementation of FedAvg experienced the greatest bandwidth consumption (38 MB), indicating the extra burden of sending model updates between institutions. Nonetheless, the increase compared to centralized MLP (+3 MB) was slight, indicating that FL's communication needs remain controllable under standard network circumstances. The convergence speed was quickest for LR (30 epochs) and RF (32 epochs), while the centralized MLP needed 40 epochs. FedAvg needed a few more epochs (45) to reach convergence, due to the asynchronous aggregation of parameters and variability in local data distributions. The training durations showed a comparable trend: LR took 18 minutes, RF took 25 minutes, centralized MLP required 42 minutes, and FedAvg needed 48 minutes. These findings suggest that federated learning adds a slight computational burden compared to its centralized version. Although FL required more training time and convergence, the additional bandwidth and computational expenses might be justified considering the significant privacy and data-sovereignty advantages shown in Tables 7 and 8. In environments with restricted network capabilities or limited hardware resources, these efficiency compromises need to be balanced with regulatory and ethical considerations. In numerous healthcare settings, the extra expense seems warranted due to FL’s capacity to attain nearly centralized performance without the need for data exchange. These results emphasize that infrastructure planning, such as adequate bandwidth allocation and somewhat lengthened training timelines, must be taken into account when implementing federated systems on a large scale. Future research could investigate adaptive communication methods (such as gradient compression, partial parameter exchange) or asynchronous federated protocols to decrease overhead further while maintaining performance.

Table 10 assessed the privacy and security aspects of centralized compared to federated learning (FL), focusing on RQ2: “In what ways does FL tackle privacy and collaboration issues?” The success rate of membership inference attacks was significantly greater for centralized training (22%) than for the federated method, which was 8%. This decrease illustrates FL’s capability to restrict information exposure via parameter aggregation, thus reducing the likelihood of adversaries determining if particular patient records were included in the training dataset. The centralized collection of sensitive data was determined to be only partially in line with HIPAA and GDPR/NDPR, mainly because of the increased risk of unauthorized access during the storage and transfer of data. Conversely, federated learning maintained complete adherence to these regulations because patient-level data stayed on local servers during the entire training process. By minimizing cross-border or inter-hospital data exchange, FL adheres to legal requirements that stress local data storage and limited data mobility, essential in multi-jurisdictional healthcare partnerships. These results highlight the importance of FL in facilitating collaboration across multiple institutions while maintaining privacy standards. Hospitals can enhance the performance of predictive models together (as demonstrated in Tables 7 and 8) while maintaining data sovereignty, thus preventing possible legal and reputational issues. Moreover, diminished vulnerability to membership inference attacks indicates that FL can be incorporated into practical systems with increased patient confidence and fewer ethical issues. The showcased security benefits establish FL as an ethically strong option compared to centralized machine learning for confidential health information. Although there is some extra computational and communication burden (Table 9), these expenses are balanced by FL’s capacity to achieve high predictive accuracy alongside strict privacy protections. Future research might explore sophisticated privacy-preserving improvements (secure aggregation or differential privacy) to further strengthen FL implementations across various clinical environments.

Table 11 provided an in-depth error analysis among hospitals to improve the clinical understanding of the predictive models and pinpoint possible areas for enhancement. Misclassification rates differed slightly among institutions. FGH Hospital indicated the highest rates for false positives (15%) and false negatives (13%), while XYZ Hospital exhibited the lowest rates (13% false positives and 11% false negatives). ABC Hospital exhibited moderate error rates (14% false positives, 12% false negatives). These fairly consistent trends indicate that although the overall performance of the model was strong (refer to Table 7), some subgroups continue to pose difficulties for accurate classification. Essential characteristics linked to misclassification offer understanding of the underlying clinical intricacy:

i. ABC Hospital: Patients over 70 years old, those diagnosed with COPD, and individuals with brief hospital stays were often misidentified. These results could indicate unusual readmission risk patterns in older adults with chronic respiratory issues who are discharged quickly.

ii. FGH Hospital: Instances of diabetes and hypertension occurring together were frequently misclassified, suggesting possible interaction effects between these comorbidities that the models did not completely reflect.

iii. XYZ Hospital: Patients with multiple prior admissions and those prescribed more than 10 medications were more prone to classification errors, possibly reflecting complex, multi-morbidity cases with variable care trajectories.

This assessment emphasized chances to enhance model sensitivity and specificity by including interaction terms (diabetes–hypertension interaction), temporal characteristics (trends of recent admissions), or measures of medication complexity. Hospitals might also explore site-specific calibration or feature engineering aimed at these subgroups to minimize systematic bias. These findings improve clinical interpretability by recognizing particular patient traits associated with misclassification, an essential condition for incorporating predictive models into decision-making processes. Recognizing error patterns fosters trust between clinicians and informs specific interventions like improved discharge planning or follow-up for high-risk, multi-morbid patients, potentially lowering readmission rates.

Figure 3 illustrates the distribution of patient groups categorized by hospital, age category, gender, and coexisting conditions, offering a visual overview of population diversity among the three involved sites. The distribution indicates that ABC Hospital accounted for the highest share of patients, whereas FGH and XYZ Hospitals brought in comparable but slightly smaller groups. This variable enrollment indicates disparities in hospital capacity and patient flow (see Table 2), which may affect case-mix complexity and patterns of readmission. In all hospitals, most admissions were concentrated in the 50-69 year age category, with a significant secondary peak in patients aged 70 years and older. XYZ Hospital exhibited a relatively older demographic, in line with its higher average age (Table 6) and increased occurrence of chronic illnesses like hypertension and heart failure. The gender distribution among hospitals was fairly uniform, showing only a minor female majority, which supports the demographic uniformity previously noted in Table 6. No individual hospital displayed a significant gender bias, indicating that sex-related bias in admissions probably won't affect model performance.

The illustration also shows clear co-morbidity burdens:

i. ABC Hospital exhibited a greater prevalence of COPD cases compared to the other locations.

ii. FGH Hospital showed a higher number of patients with both diabetes and hypertension, a pairing linked to the misclassification trends observed in Table 11.

iii. XYZ Hospital accounted for the highest proportion of multi-morbid patients, frequently showing a history of several prior admissions or intricate medication plans.

The noted distribution highlights the need to consider site-specific and demographic variability in federated learning models. Variations in age composition and the prevalence of co-morbid conditions might influence the rankings of feature importance and lead to distinct error patterns within hospitals. Stratified or weighted methods may enhance fairness and broad applicability by reducing biases due to imbalanced cohort makeup.

Predictive Performance

Federated learning (FL) consistently surpassed local-only models and neared the performance of centralized training in all involved hospitals. As outlined in Table 7 (not displayed here), FL using FedAvg attained the highest average AUC (0.88) in comparison to centralized (0.89) and local-only training (0.80-0.83). ABC Hospital, with the largest cohort, demonstrated merely a 0.01 AUC difference between FL and centralized methods, whereas FGH and XYZ Hospitals presented slightly larger yet still modest discrepancies (≤ 0.03). These findings validate that FL can achieve near-centralized performance without the need for direct data exchange. Variations in performance among hospitals also illustrate the cohort diversity depicted in Figure 3. Significantly, the elevated rates of diabetes, hypertension co-morbidity at FGH Hospital (Table 11) were linked to marginally reduced precision, indicating that the complexity of patient demographics affects local training quality.

Models were tested on unfamiliar hospital data to assess their generalizability. FL models maintained strong performance, experiencing an AUC decrease of merely 0.02-0.03 when moved to a site that did not participate, while local-only models exhibited reductions of 0.07-0.10. This suggests that FL identifies wider population trends and shows reduced sensitivity to biases specific to particular sites. Moreover, cross-hospital validation verified that FL models performed better in terms of generalization for older populations (≥ 70 years) and multi-morbid cases prevalent at XYZ Hospital.

Examining privacy and efficiency metrics (Tables 9 and 10) highlights the trade-offs involved in implementing a federated learning (FL) framework. Model update security significantly enhanced with FL: membership inference attack success rates dropped from 22% in the centralized model to 8% in the federated setup, showcasing FL's ability to protect sensitive patient-level information (Table 10). Additionally, FL completely met HIPAA and GDPR/NDPR compliance standards, while centralized pooling achieved only a partial level of compliance, further supporting FL's appropriateness for regulated healthcare settings. For communication and computational efficiency, FedAvg training needed 50 communication rounds and an average bandwidth of 38 MB, converging after 45 epochs with an overall training duration of 48 minutes. While this training time was slightly greater than that of the centralized multilayer perceptron (42 minutes), the small increase is a reasonable efficiency compromise considering the significant privacy advantages obtained (Table 9). These results together indicate that FL provides a fair trade-off between safeguarding privacy and managing operational expenses.

The analysis of misclassification patterns in the case study (Table 11) offers essential clinical understanding. At ABC Hospital, misclassified cases were mainly linked to patients over 70 years, those with chronic obstructive pulmonary disease (COPD), and brief hospital stays. FGH Hospital demonstrated a greater rate of false positives in patients with a diabetes–hypertension comorbidity, highlighting the challenges posed by interconnected chronic conditions. Misclassifications at XYZ Hospital were mainly found among multi-admission patients who were taking polypharmacy (> 10 simultaneous medications). These error profiles specific to institutions reveal that FL models are responsive to local patient traits while still demonstrating robust generalizability in various healthcare environments. Significantly, the trends noted correspond with recognized clinical risk factors for readmission, demonstrating that FL maintains the interpretability crucial for incorporation into clinical decision support systems and for building clinician confidence.

Key Findings

The findings showed that federated learning (FL) attains comparable predictive accuracy to centralized methods while safeguarding patient privacy. Through facilitating model training among various institutions without the need for direct data exchange, FL provides a collaborative benefit that reduces the risks linked to centralizing sensitive health information.

Practical Considerations

These results highlight the practicality of implementing FL in healthcare systems that aim to reconcile performance with data governance needs. The strategy closely follows regulatory standards like GDPR, HIPAA, and NDPR, which minimizes compliance risks and promotes collaboration among multiple institutions. Embracing FL could improve cross-hospital analytics for functions like predicting readmissions, allocating resources, and planning personalized treatments, eliminating the necessity for expensive and possibly non-compliant data merging.

Limitations

Even with its benefits, FL faces a number of limitations. Variability in data across hospitals, such as variations in patient demographics, coding methods, and electronic health record systems, can create inconsistencies that complicate model convergence and generalization. Additionally, the communication overhead from repeated model updates can escalate bandwidth consumption and extend training duration compared to completely centralized models.

Future Research

Upcoming studies need to prioritize incorporating differential privacy methods to enhance safeguards against adversarial assaults. Extending FL to include multimodal data sources (integrating medical images, structured records, and clinical narratives) would improve its usefulness for intricate predictive tasks. Ultimately, expanding federated frameworks to national or global healthcare systems will be crucial for assessing their resilience, compatibility, and enduring operational viability in practical clinical environments.

5. Conclusion

This research showed that federated learning (FL) can achieve predictive performance similar to centralized models while maintaining patient privacy and institutional independence. Through extensive assessment across various hospitals, FL attained strong AUROC and F1-scores, minimized vulnerability to membership inference attacks, and ensured complete compliance with regulatory requirements including HIPAA, GDPR, and NDPR.

The results emphasize the importance of FL in healthcare partnerships, offering a structure that facilitates knowledge exchange and strong predictive modeling without the need for centralizing sensitive data. This method reduces compliance risks and fosters trust between participating institutions, thus enabling collaborative analytics across diverse clinical settings.

Additionally, the study presents a definitive route for implementation in practical healthcare systems. With tolerable communication overhead and training efficiency compromises, FL provides a scalable answer for inter-institutional activities like readmission risk forecasting, chronic condition management, and hospital resource allocation. Future initiatives that include differential privacy, the integration of multimodal data, and the extension to wider healthcare networks will strengthen FL’s position as an essential technology for innovative, data-driven healthcare that prioritizes privacy.

Acknowledgments: The authors state that no outside funding was obtained to carry out this study. We sincerely appreciate the cooperation and assistance of ABC Hospital, FGH Hospital, and XYZ Hospital, as their involvement and supply of de-identified clinical data enabled this research to occur. We appreciate the clinical and technical teams at these institutions for their essential support in data management, model assessment, and insights on the interpretation of findings.

Institutional Review Board Statement: This study was reviewed and approved by the Lead City Institutional Review Board (IRB) with Approval Number: LCU/PG/005 on 3 February 2025 in accordance with the ethical standards of the LCU and the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards. All procedures performed in this study involving human participants were conducted following the

IRB's guidelines and regulations. Written informed consent was obtained from all individual participants included in the study.

References

1. Peng L., et al. (2024). *An in-depth evaluation of federated learning on biomedical NLP tasks across multi-site corpora. npj Digital Medicine*. Demonstrates that federated learning models outperform local training and sometimes match centralized models in biomedical natural language tasks. Nature
2. Eden R., et al. (2025). *A scoping review of the governance of federated learning in healthcare: procedural, relational, and structural mechanisms. npj Digital Medicine*. Proposes a consolidated framework for data governance in FL, vital for ethically deploying FL in healthcare. Nature
3. Zhu H., et al. (2025). *FedWeight: mitigating covariate shift of federated learning on EHR data through patient re-weighting. npj Digital Medicine*. Introduces methods to address cross-site distributional shifts via data re-weighting, improving generalizability. Nature
4. Teo Z.-L., et al. (2024). *Federated machine learning in healthcare: A systematic review. Cell Reports Medicine*. Catalogs over 600 FL studies, pointing out the limited proportion of real-world deployments and identifying gaps in clinical translation. PMC
5. Kim S., Kim M. S. (2024). *Leveraging Federated Learning for Automatic Detection of Clopidogrel Treatment Failures. Preprint* (arXiv). Demonstrates that FL can approximate centralized AUC and convergence performance for pharmacotherapy prediction tasks. arXiv
6. Ali M. S., et al. (2024). *Federated Learning in Healthcare: Model Misconducts, Security, Challenges, Applications, and Future Research Directions. Preprint* (arXiv). Reviews technical and practical pitfalls, including convergence under non-IID data, communication cost, and multi-institutional collaboration. arXiv
7. Sadilek A., et al. (2021). *Privacy-first health research with federated learning. Nature Digital Medicine*. A landmark demonstration of integrating differential privacy into FL across diverse epidemiological and clinical studies, affirming comparable accuracy and stronger privacy guarantees. Nature
8. Rieke N., et al. (2020). *The future of digital health with federated learning. npj Digital Medicine*. A foundational perspectives piece outlining FL's potential to overcome data silos in healthcare and its key challenges. Nature
9. Teo Z.-L., et al. *Federated machine learning in healthcare: A systematic review. Cell Reports Medicine* **2024**. Comprehensive analysis of 612 FL healthcare studies, clarifying clinical translation rates and modalities PMC.
10. M. Li, P. Xu, J. Hu, Z. Tang, G. Yang. *From Challenges and Pitfalls to Recommendations and Opportunities: Implementing Federated Learning in Healthcare. Medical Image Analysis (preprint)* **2024**. Explores implementation barriers, methodological biases, and practical recommendations arXiv.
11. M. Li. *Implementing federated learning in healthcare: Review up to May 2024. Elsevier Review* **2025**. Evaluates recent FL-based healthcare methods and clinical utility ScienceDirect.
12. N. Madathil, et al. *Revolutionizing healthcare data analytics with federated learning: Insights from over 250 studies (2019–2024). Comprehensive Survey* **2025**. Dissects FL system architecture, challenges, and domains PMC.
13. V. Pais. *Healthcare federated learning: A survey of applications and cross-silo implementations. Taylor & Francis* **2025**. Focused on FL across institutional boundaries and application areas Taylor & Francis Online.
14. S. R. Abbas, et al. *Federated Learning in Smart Healthcare. MDPI Healthcare* **2024**. Investigates FL integration with IoT, wearables, and remote monitoring technologies MDPI.
15. B. Almogadwy, et al. *Fused Federated Learning Framework with IoMT Devices for Secure Chronic Kidney Disease Monitoring in Healthcare 5.0. Scientific Reports (Nature)* **2025**. Presents an FL framework combining IoMT and real-time analytics with high accuracy (98.21%) Nature.
16. F. Zhang, et al. *Recent methodological advances in federated learning for healthcare applications (2015–2023). Elsevier Review* **2024**. Surveys technical FL innovations in healthcare PMCScienceDirect.

17. I. Hagestedt, et al. *Toward a tipping point in federated learning in healthcare and life sciences. Elsevier* **2024**. Discusses real-world adoption trends and maturity of FL in clinical practice ScienceDirect.
18. Bujotzek, M.R. *Real-world federated learning in radiology: hurdles to adoption. Journal of the American Medical Informatics Association (JAMIA)* **2025**. Identifies practical challenges in deploying FL for radiological data Oxford Academic.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.