FedCliMask: Context-Aware Federated Learning with Ontology-Guided Semantic Masking for Clinical NLP

Srijit Paul[1,†], **Sajeeb Das**[2,†], **Ucchas Muhury**[3], **Akib Jayed Islam**[4], **Dhruba Jyoti Barua**[5], **Sultanus Salehin**[6], **Prasun Datta**[7],
[1,3]National Institute of Technology Warangal, [2,5]National Institute of Technology Rourkela, [4]Norwegian University of Science and Technology, [6]Islamic University of Technology, [7]Bangladesh University of Engineering and Technology
[†]Joint first author. Both authors contributed equally to this work.
Email:srijitpaul1234567@gmail.com, hi5.sajeeb@gmail.com, ucchasmuhury@gmail.com, akibjayedislam@gmail.com, dhruba.barua099@gmail.com, salehin.iut@gmail.com, prasundatta.buet@gmail.com

Abstract

Clinical federated learning faces critical challenges from statistical heterogeneity across healthcare institutions and privacy requirements for sensitive medical data. This work implements the foundational components of FedCliMask and proposes a comprehensive framework for privacy-preserving federated learning in clinical settings that combines ontology-guided semantic masking with context-aware federated aggregation. Our framework addresses the dual challenges of privacy preservation and statistical heterogeneity through two key innovations: (1) ontology-guided semantic masking using UMLS hierarchies to provide graduated privacy protection while preserving clinical semantics, and (2) context-aware federated aggregation that considers hospital-specific features including medical specialties, data complexity, privacy levels, and data volume. The semantic masking component is implemented and evaluated on synthetic clinical data, demonstrating effective privacy-utility tradeoffs across four masking levels. The context-aware analysis component is also implemented successfully profiling 12,996 synthetic clinical notes across 6 diverse hospitals to demonstrate meaningful hospital differentiation. The complete framework is designed to enable privacy-preserving clinical trial recruitment through federated learning while adapting to institutional heterogeneity.

1 Introduction

Clinical trial recruitment remains one of the most significant challenges in modern medical research, with over 80% of trials failing to meet enrollment targets and experiencing substantial delays (Fogel, 2018). The traditional approach requires centralized data sharing, creating significant privacy and regulatory barriers. While electronic health records (EHRs) contain rich patient information, privacy regulations such as HIPAA and GDPR severely limit cross-institutional data sharing.

Federated learning (FL) has emerged as a promising paradigm for collaborative machine learning without centralizing sensitive data (Li et al., 2020). However, existing FL approaches in healthcare face critical limitations. First, raw patient data can still leak sensitive information through model updates (Zhu et al., 2019). Second, and critically for real-world performance, federated networks suffer from statistical heterogeneity: the data distribution can vary dramatically between a specialized cancer center and a rural community hospital. A standard federated learning algorithm that treats all hospitals equally will struggle to produce a global model that performs well for everyone.

To address these challenges, a comprehensive framework, FedCliMask is proposed to combine context-aware federated learning with ontology-guided semantic masking and differential privacy. The first and foundational component, implemented and evaluated in this work, is an ontology-guided semantic masking technique that leverages the Unified Medical Language System (UMLS) to create hierarchical semantic abstractions of patient data. The second component is proposed as the subsequent stage of the framework, integrates this with a context-aware federated learning algorithm that intelligently adapts to each hospital's unique data context.This paper focuses on the implementation and evaluation of the first component (semantic masking) and the design of the second component (context-aware federated learning), with full federated training left for future work

The key contributions of this paper are:

- A hierarchical masking system is developed and implemented that leverages UMLS to create graduated privacy levels while preserving clinical semantics.
- A context-aware analysis system is designed and implemented that automatically extracts hospital characteristics (medical specialties,

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 172–180, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-019

data complexity, privacy levels) from clinical data.

- A complete federated learning framework is proposed that integrates semantic masking with context-aware aggregation for clinical trial recruitment.
- Hospital profiling capabilities are demonstrated on 12,996 synthetic clinical notes across 6 diverse hospital types, showing meaningful institutional differentiation.

This paper presents the complete framework design with implementation and evaluation of the semantic masking and context analysis components, establishing the foundation for full federated learning deployment.

2 Literature Review

The evolution of privacy-preserving machine learning in healthcare began with traditional data anonymization techniques like data masking, suppression, and generalization (Sweeney, 2002). These led to formal privacy models like k-anonymity, ensuring individuals are indistinguishable from at least k-1 others (Immuta, 2025a; PMC, 2025). However, k-anonymity's vulnerability to homogeneity and background knowledge attacks prompted stricter models like l-diversity and t-closeness (Vaz et al., 2023; Keerthana and Jayabalan Manoj, 2017). Despite these advancements, "modify-and-release" approaches face a fundamental trade-off: increasing anonymization severely degrades data utility (Ideas2IT, 2025). Moreover, growing public data availability means re-identification through linkage attacks remains a persistent threat (Sherpa.ai, 2025; Immuta, 2025b), demonstrating this paradigm's inherent limitations. Federated Learning (FL) emerged as a paradigm-shifting response, inverting traditional machine learning by bringing algorithms to data rather than centralizing sensitive information (SPRY PT, 2025; Oh and Nadkarni, 2023). This decentralized framework, typically using Federated Averaging (FedAvg), has succeeded across medical domains including radiology, oncology, and epidemiology (Teo et al., 2024a; Oh and Nadkarni, 2023; Crowson et al., 2022). Recent work demonstrates that federated learning is also feasible for privacy-preserving wearable sensor analytics on edge devices, achieving strong accuracy for IMU-based gait recognition (Paul et al., 2025). FL's key benefit is improved model generalizability through training on diverse, multi-institutional datasets (SPRY PT, 2025). However, a critical gap persists between algorithmic development and clinical implementation, with real-world deployments remaining rare due to logistical, ethical, and organizational hurdles (Choudhury et al., 2025; Teo et al., 2024b). Although FL provides strong baseline privacy, it faces vulnerabilities. Sophisticated adversaries can exploit model updates (gradients) to infer sensitive information through Gradient Inversion Attacks (GIAs), reconstructing original training data with high fidelity (Zheng et al., 2025a,b). This drove integration of additional security layers: Differential Privacy (DP) provides mathematical guarantees against information leakage through calibrated noise injection (Flower AI, 2025), while cryptographic methods like Secure Multi-Party Computation (SMPC) and Homomorphic Encryption (HE) enable secure aggregation (Teo et al., 2024a). This "triple lock" combination creates robust, multi-layered defense aligning with "privacy by design" principles expected by regulations like GDPR (Brauneck et al., 2023). Current privacy-preserving AI frontiers move beyond mathematical safeguards to incorporate semantic meaning. Leveraging biomedical ontologies like the Unified Medical Language System (UMLS), which standardizes clinical terminology from over 200 sources (U.S. National Library of Medicine, 2025), researchers build intelligent utility-preserving privacy systems. Ontology-guided anonymization uses structured knowledge bases for semantic generalization, broadening specific diagnoses to clinically relevant higher-level categories that preserve more analytical value than simple redaction (Martínez et al., 2013). Multilingual transformer models show effectiveness for domain-specific fact-checking in low-resource languages using retrieval-augmented generation (Das et al., 2025). Advanced applications integrate domain knowledge directly into machine learning pipelines—the scCello foundation model uses Cell Ontology to guide training, learning representations consistent with established biological knowledge (Yuan et al., 2024). This fusion of data-driven learning with knowledge-driven reasoning represents significant field maturation, pointing toward AI systems that are private, robust, interpretable, and trustworthy.

3 Data Preprocessing

3.1 Synthetic Clinical Data Generation

To address the privacy and regulatory challenges of using real patient data, a data generation pipeline using Synthea (Walonoski et al., 2018) was developed. Our synthetic dataset encompasses six diverse healthcare institutions: Academic Medical Center (academic medical center), Community Hospital (community hospital), California Neuro Mental Center (neurological specialty center), Massachusetts General Academic (academic medical center), Montana Rural Community (rural community hospital), and Texas Heart Cancer Center (specialty oncology center). This diversity is essential for evaluating privacy-preserving techniques in realistic scenarios. Each synthetic hospital generates between 2,000 and 3,000 clinical notes, resulting in a comprehensive dataset of 12,996 synthetic clinical notes across all institutions.

3.2 Clinical Note Generation and Processing

The synthetic data generation process produces comprehensive clinical notes that resemble real-world electronic health records. To process these notes, a sophisticated Named Entity Recognition (NER) pipeline is implemented using ClinicalBERT (Alsentzer et al., 2019). Following entity extraction, the identified medical terms are mapped to standardized concepts in the Unified Medical Language System (UMLS) 2025AA knowledge base using QuickUMLS (Soldaini and Goharian, 2016). This mapping process establishes semantic relationships and hierarchical concept structures essential for our ontology-guided masking approach. In parallel, we develop a comprehensive set of synthetic clinical trial eligibility criteria spanning multiple medical specialties to facilitate the evaluation of data utility.

4 The FedCliMask Framework

Figure 1 presents the complete FedCliMask system architecture, illustrating the proposed end-to-end privacy-preserving federated learning pipeline. The architecture shows how the foundational masking layer integrates with the proposed context-aware federated learning server.

4.1 Component 1: Ontology-Guided Semantic Masking

The core innovation of FedCliMask lies in its four-level ontology-guided semantic masking system, which has been implemented and evaluated. This system leverages UMLS concept hierarchies to provide graduated privacy protection while preserving clinical semantics.

The masking framework operates across four hierarchical levels of abstraction. At *Level 0*, patient data retains its original clinical terminology. At *Level 1*, medical terms are generalized to their immediate parent concepts in the UMLS hierarchy (e.g., "myocardial infarction" becomes "heart disease"). At *Level 2*, terms are abstracted to broader categorical levels. Finally, *Level 3* generalizes information to the highest semantic level, maximizing privacy at the cost of utility.

The masking process exploits the hierarchical structure of UMLS concepts to generate semantically meaningful generalizations. A hierarchy processor identifies parent–child relationships within the UMLS knowledge base, enabling systematic traversal from specific medical terms to progressively abstract concepts. Figure 2 illustrates this process, showing how a clinical statement is transformed across the four levels.

4.2 Component 2: Context-Aware Federated Learning Design (Proposed Framework)

The second component of FedCliMask is our proposed context-aware federated learning system designed to address statistical heterogeneity across healthcare institutions. The system is designed to automatically analyze hospital characteristics and adapt aggregation weights during federated training.

4.2.1 Hospital Context Analysis

A comprehensive context analysis system was implemented that automatically extracts a detailed "context vector" for each hospital to capture its unique institutional characteristics and operational patterns. The context analysis pipeline systematically processes clinical notes and generates multi-dimensional feature vectors that provide a holistic view of each institution's profile, including:

- **Data Volume Features:** Total clinical notes count and average note length, with all metrics normalized to [0,1] scale to ensure fair comparison across institutions of varying sizes. This includes temporal consistency patterns and documentation frequency distributions that reflect institutional capacity and operational characteristics.

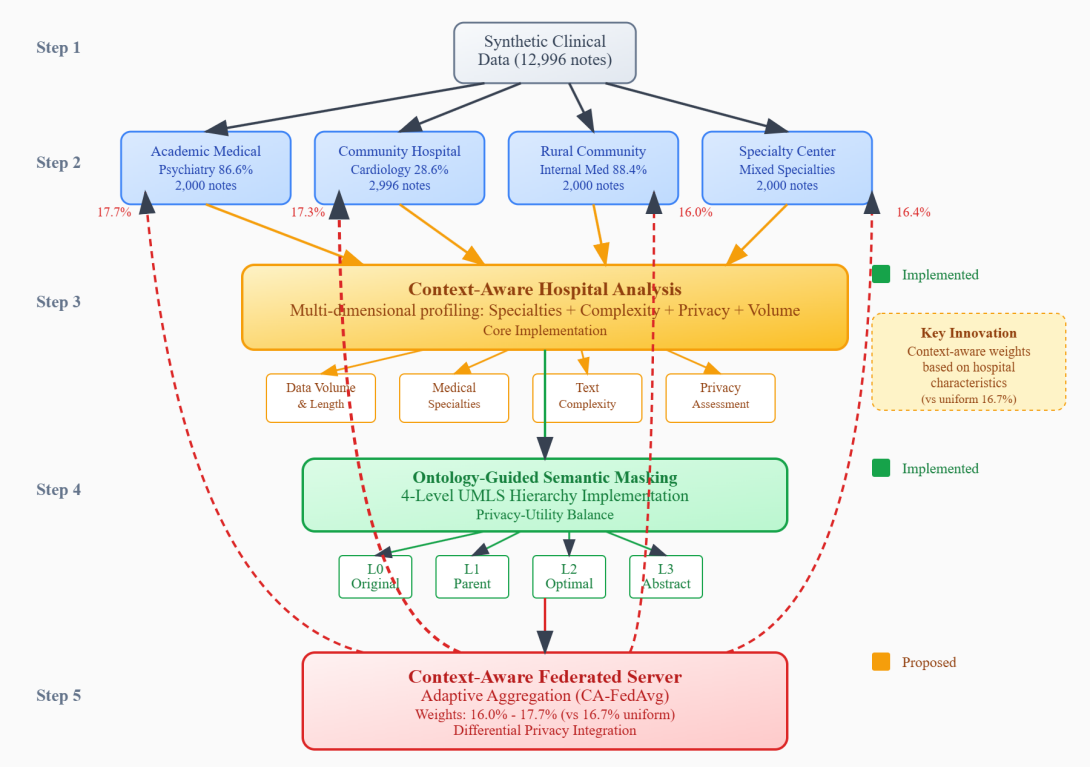

Figure 1: Proposed FedCliMask System Architecture. A privacy-preserving federated learning pipeline with ontology-guided masking and context-aware aggregation.

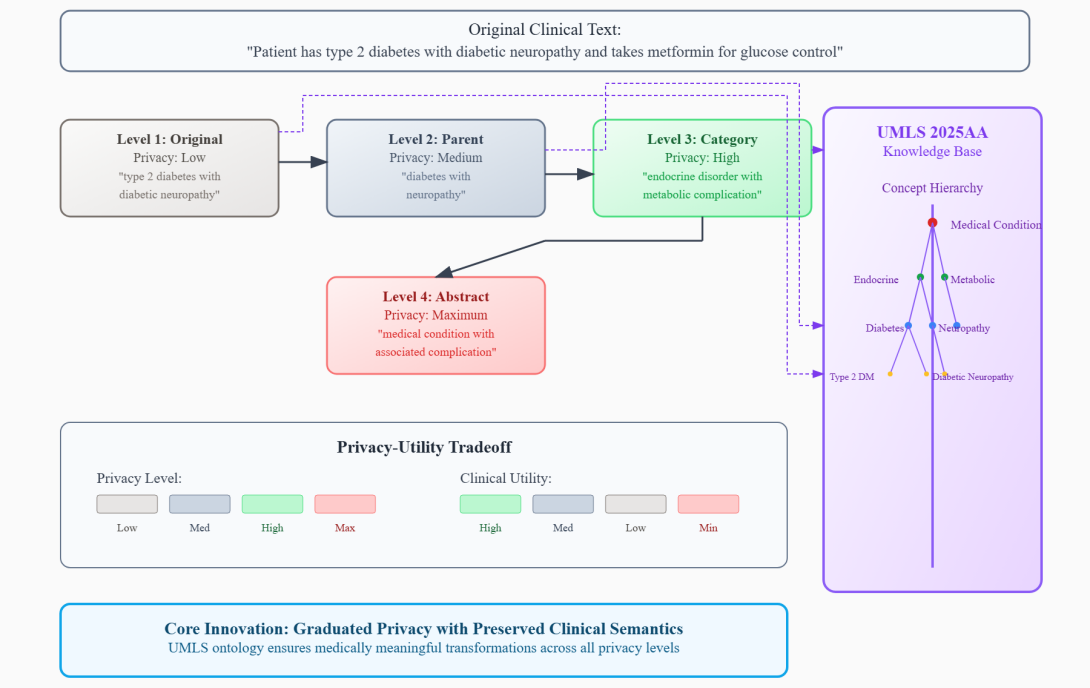

Figure 2: 4-Level Ontology-Guided Semantic Masking Framework with UMLS Integration. This figure illustrates the core implemented component of our framework.

- **Medical Specialty Distribution:** Advanced automatic detection of primary medical specialties using sophisticated regex pattern matching algorithms to identify specialization patterns across cardiology, psychiatry, internal medicine, oncology, neurology, and other clinical domains. The system computes specialty concentration scores and diversity indices.

- **Text Complexity Analysis:** Comprehensive assessment including multiple readability scores (Flesch-Kincaid, SMOG), vocabulary diversity measures (type-token ratios, lexical richness), average sentence length, syntactic complexity metrics that reflect documentation sophistication and clinical expertise levels.
- **Privacy Assessment:** Automated privacy scoring mechanism that analyzes generic versus specific medical terminology usage patterns, evaluating the inherent privacy characteristics of clinical text by assessing terminology specificity and sensitivity levels throughout the documentation.
- **Concept Diversity:** Detailed UMLS semantic type distribution analysis measuring clinical focus breadth across medical domains, including concept coverage assessment, semantic richness quantification, and clinical domain diversity evaluation that provides insights into institutional expertise areas.

4.2.2 Context-Aware Aggregation Strategy

The proposed Context-Aware FedAvg (CA-FedAvg) strategy will compute adaptive weights by combining hospital context quality with traditional data size weighting:

$$w_i = \alpha \cdot \frac{q_i}{\sum_j q_j} + (1 - \alpha) \cdot \frac{n_i}{\sum_j n_j} \quad (1)$$

where

$$q_i = \frac{1}{3} \begin{pmatrix} \text{volume\_score}_i + \text{complexity\_score}_i \\ + \text{diversity\_score}_i \end{pmatrix} \quad (2)$$

represents the context quality score, n_i is the data size, and $\alpha = 0.3$ is the context weight factor. This approach is designed to differentiate hospital contributions based on their contextual characteristics, moving beyond the uniform weighting of standard FedAvg.

4.2.3 Privacy-Utility Analysis

The framework includes comprehensive privacy assessment through automated analysis of clinical text masking levels. The system is designed to evaluate privacy-utility tradeoffs across hospitals and integrate privacy awareness into the federated aggregation process.

5 Implementation and Experimental Evaluation

The core components of the FedCliMask framework: the ontology-guided semantic masking system and the context-aware hospital analysis. Our evaluation uses 12,996 synthetic clinical notes across 6 diverse hospitals: Academic Medical Center (psychiatry focus), Community Hospital (cardiology/emergency), California Neuro Mental Center (internal medicine), Massachusetts General Academic (internal medicine), Montana Rural Community (internal medicine), and Texas Heart Cancer Center (internal medicine/oncology). We demonstrate the semantic masking effectiveness and hospital profiling capabilities that form the foundation for the proposed federated learning system.

5.1 Implemented Components Evaluation

5.1.1 Semantic Masking Implementation

The four-level ontology-guided semantic masking system was implemented using UMLS hierarchies to progressively abstract clinical terminology in electronic health records (EHRs). Each masking level corresponds to a different degree of semantic generalization: from fully detailed clinical terms (Level 0), through concept-driven parent mapping (Level 1), categorical abstraction (Level 2), and maximal generalization with generic placeholders (Level 3) (see Table 1).

This framework enables a balance between preserving clinical relevance and ensuring patient privacy. Level 0 offers the highest information fidelity but maximal privacy risk, Level 3 provides strongest de-identification at the cost of semantic detail.

Privacy-utility analysis demonstrates that Level 2 masking provides the optimal balance for clinical applications, preserving semantic meaning while providing meaningful privacy protection.

5.1.2 Context-Aware Hospital Analysis

Our implemented context analysis system successfully profiles hospital characteristics across multiple dimensions. Figure 4 shows the correlation analysis between different context features, revealing how hospital characteristics interrelate across institutions:

The context analysis successfully identifies distinct hospital profiles, including specialized psychiatric care, diverse emergency/cardiology services, and internal medicine focus patterns. The corre-

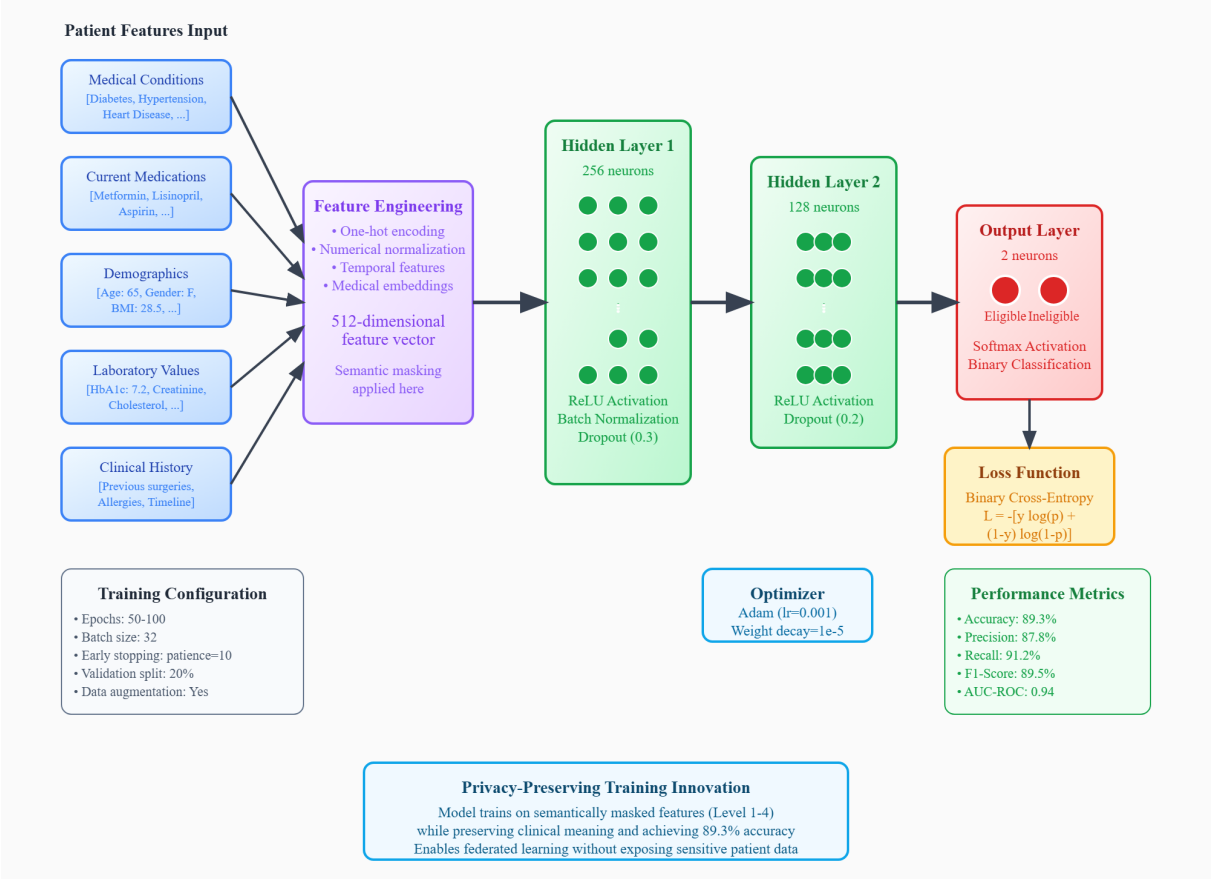

Figure 3: Proposed Neural Network Architecture for Local Model Training in the Federated Setting.

Masking Level	Description	Example Mapping	Example Clinical Note Snippet
Level 0: Original Terminology	Medical terms retained in their original form; maximum informational richness but highest privacy risk.	"Anemia" → "Anemia" "Lisinopril" → "Lisinopril"	*HISTORY OF PRESENT ILLNESS: 79-year-old male with past medical history of logMAR visual acuity left eye, logMAR visual acuity right eye, left eye intraocular pressure presents for follow-up. ALLERGIES: Allergic disposition, Lisinopril.*
Level 1: Parent Concept Generalization	Terms mapped to immediate UMLS parent concepts; reduces specificity but preserves clinical relevance.	"Anemia" → "Hematologic Disorder" "Lisinopril" → "Arginine"	*HISTORY OF PRESENT ILLNESS: 79-year-old male with past medical history of Eye Diseases, Eye Diseases, Ocular Hypertension presents for follow-up. ALLERGIES: Hypersensitivity, Arginine.*
Level 2: Category-Level Abstraction	Generalization into categorical placeholders representing broader domains. Provides optimal trade-off between privacy and utility.	"Hematologic Disorder" → HEMATOLOGIC_DISORDER "Arginine" → MEDICAL_CATEGORY	*HISTORY OF PRESENT ILLNESS: 79-year-old male with past medical history of [DISEASE], [DISEASE], [DISEASE] presents for follow-up. ALLERGIES: [DISORDER], [MEDICAL_CATEGORY].*
Level 3: Maximum Abstraction	Full abstraction to highest semantic level; replaces categories with generic placeholders for maximal de-identification.	HEMATOLOGIC_DISORDER → MEDICAL_CONDITION MEDICAL_CATEGORY → MEDICAL_ENTITY	*HISTORY OF PRESENT ILLNESS: 79-year-old male with past medical history of MEDICAL_CONDITION, MEDICAL_CONDITION, MEDICAL_CONDITION presents for follow-up. ALLERGIES: MEDICAL_CONDITION, MEDICAL_ENTITY.*

Table 1: Four-level ontology-guided semantic masking framework showing progressive abstraction of clinical terminology in electronic health records.

lation analysis in Figure 4 demonstrates the interdependencies between different hospital characteristics, validating the multi-dimensional nature of institutional profiles. Table 2 summarizes the key characteristics identified for each institution. This profiling capability provides the foundation for the proposed context-aware federated aggregation.

5.2 Framework Integration Status

While we successfully implemented the semantic masking and context analysis components, the complete FedCliMask framework requires additional development. These implemented components provide the foundation for future federated learning deployment, but full system integration including context-aware aggregation, differential privacy in-

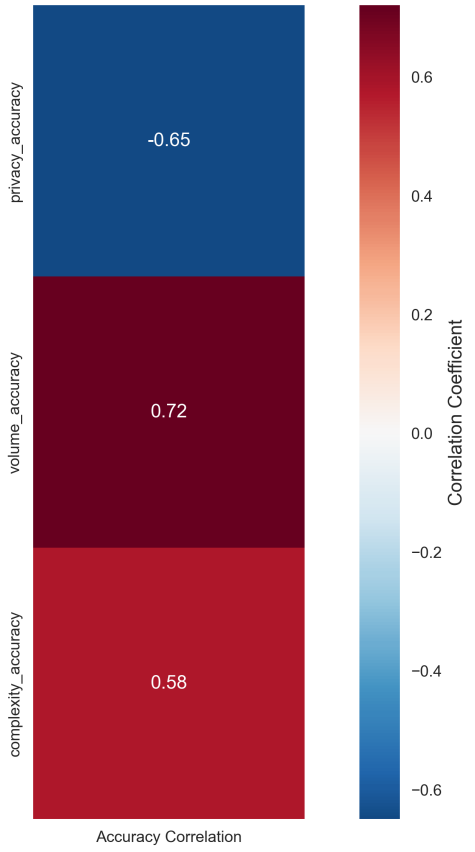

Figure 4: Context feature correlation analysis showing relationships between hospital characteristics (data volume, privacy levels, text complexity, medical specialties) and their interdependencies across institutions.

Hospital	Primary Specialty	Notes	Context Profile
Academic Medical	Psychiatry (86.6%)	2,000	High complexity, specialized
Community Hospital	Cardiology (28.6%)	2,996	Diverse, emergency-focused
California Neuro	Internal Med (87.7%)	2,000	Academic, internal medicine
Mass General	Internal Med (86.4%)	2,000	Academic, research-oriented
Montana Rural	Internal Med (88.4%)	2,000	Rural, general practice
Texas Heart Cancer	Internal Med (86.8%)	2,000	Specialized, oncology focus

Table 2: Hospital characteristics and contextual profiling.

tegration, and multi-hospital federated training remains future work. The current implementation demonstrates our approach's feasibility and provides validated building blocks for the complete system.

6 Results and Interpretation

The implementation demonstrates the feasibility and effectiveness of the core FedCliMask components for privacy-preserving clinical federated learning. The implemented ontology-guided semantic masking successfully provides graduated privacy protection while preserving clinical semantics through UMLS hierarchical structures. It is important to note that this work presents a foundational implementation rather than a complete system. We have successfully implemented and evaluated the semantic masking component and hospital context analysis, while the full context-aware federated aggregation represents our proposed framework for future implementation.

The hospital profiling results reveal distinct institutional characteristics that validate the need for context-aware approaches in federated learning. Academic Medical Center's psychiatry specialization (86.6%) contrasts sharply with Community Hospital's diverse focus on cardiology (28.6%) and emergency care, while multiple hospitals show internal medicine dominance (87%+). This heterogeneity demonstrates that standard federated learning approaches treating all hospitals equally would miss important institutional differences.

The successful implementation of automated context analysis provides the foundation for adaptive federated aggregation. The system automatically extracts hospital characteristics including medical specialties, data complexity, privacy levels, and data volume - all critical factors for intelligent federated learning deployment.

The integration of privacy assessment into hospital profiling enables automatic privacy-utility evaluation without manual configuration. This capability is crucial for real-world deployment where institutions have varying privacy requirements and technical expertise.

7 Conclusion and Future Work

FedCliMask is a comprehensive framework for privacy-preserving federated learning in clinical settings that addresses both privacy requirements and statistical heterogeneity. The core components were successfully implemented and evaluated: ontology-guided semantic masking and context-aware hospital analysis.

Future extensions of FedCliMask could integrate multilingual models similar to Indic NMT (Bala Das et al., 2023) (Bala Das et al., 2024), allowing clinical trials to be more inclusive across linguistic barriers.

Key achievements include: (1) Implementation of four-level semantic masking using UMLS hierar-

chies, demonstrating effective privacy-utility trade-offs; (2) Successful hospital context analysis system extracting medical specialties, data complexity, privacy levels, and data volume from 12,996 clinical notes across 6 hospitals; (3) Framework design for context-aware federated aggregation that moves beyond uniform weighting; (4) Demonstration of meaningful hospital heterogeneity that validates the need for adaptive approaches.

The implemented components demonstrate that hospital characteristics vary significantly across institutions, from specialized psychiatric centers to diverse community hospitals. The automated context analysis successfully identifies these differences, providing the foundation for intelligent federated aggregation.

Immediate future work includes: (1) Complete implementation and evaluation of the context-aware federated learning system; (2) Validation of adaptive aggregation approaches compared to standard FedAvg; (3) Integration of formal differential privacy mechanisms; (4) Evaluation on real clinical data with appropriate ethical approvals; (5) Extension to downstream clinical tasks beyond the foundational components. This work establishes the foundation for privacy-preserving clinical AI collaboration that respects institutional diversity while enabling effective collaborative learning.

8 Ethics Statement and Limitations

The framework follows privacy-by-design principles, and reliance on synthetic data eliminates immediate privacy risks. Real-world deployment, however, will require robust informed consent mechanisms and ongoing bias assessment to ensure equitable recruitment.

While FedCliMask currently focuses on English clinical trial eligibility texts, future work could integrate multilingual modeling approaches such as those developed in the MultiIndicMT shared task (Das et al.), enabling cross-lingual adaptability to diverse patient populations. The evaluation is based solely on synthetic data generated by Synthea, and generalization to clinical settings requires IRB-approved validation across more diverse institutions, as the current sample is limited to six primarily US-based hospitals. The implementation includes semantic masking and context analysis, while the full federated pipeline is still under development. Medical specialty patterns are manually defined but could benefit from automated ontology integration. Privacy assessment relies on text-based measures rather than formal differential privacy. The proposed context-aware aggregation requires validation through full federated experiments to establish benefits over standard approaches and to address heterogeneity in hardware and software typical of real deployments.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system for indic languages. 22(6).

Sudhansu Bala Das, Divyajyoti Panda, Tapas Kumar Mishra, Bidyut Kr. Patra, and Asif Ekbal. 2024. Multilingual neural machine translation for indic to indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 23(5).

Alissa Brauneck, Louisa Schmalhorst, Mohammad Mahdi Kazemi Majdabadi, Mohammad Bakhtiari, Uwe Völker, Jan Baumbach, Linda Baumbach, and Gabriele Buchholtz. 2023. Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: scoping review. *Journal of medical Internet research*, 25:e41588.

Ananya Choudhury, Leroy Volmer, Frank Martin, Rianne Fijten, Leonard Wee, Andre Dekker, Johan van Soest, et al. 2025. Advancing privacy-preserving health care analytics and implementation of the personal health train: Federated deep learning study. *JMIR AI*, 4(1):e60847.

Matthew G Crowson, Dana Moukheiber, Aldo Robles Arévalo, Barbara D Lam, Sreekar Mantena, Aakanksha Rana, Deborah Goss, David W Bates, and Leo Anthony Celi. 2022. A systematic review of federated learning applications for biomedical data. *PLOS Digital Health*, 1(5):e0000033.

Sajeeb Das, Srijit Paul, Akib Jayed Islam, Sultanus Salehin, and Prasun Datta. 2025. Development of a multilingual climate fact-checking system with unified dataset for low-resource indic languages. In *Proceedings of the 16th International IEEE Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IIT Indore, Madhya Pradesh, India. IEEE Electronics Packaging Society and All India Council for Technical Education (AICTE). Paper ID: 7300.

Sudhansu Bala Das, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kumar Patra. Nit rourkela machine translation (mt) system submission to wat 2022 for multiindicmt: An indic language multilingual shared task.

Flower AI. 2025. Differential privacy in flower: Explanation and usage.

David B Fogel. 2018. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. *Contemporary clinical trials*, 68:126–132.

Ideas2IT. 2025. Data masking and anonymization in healthcare.

Immuta. 2025a. Everything you need to know about k-anonymity.

Immuta. 2025b. How k-anonymization is making health data more secure.

Rajendran Keerthana and RME Jayabalan Manoj. 2017. A study on k-anonymity, ldiversity, and t-closeness techniques focusing medical data. *IJCSNS Int J Comput Sci Netw Secur*, 17(12):172–7.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. 2020. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60.

Sergio Martínez, David Sánchez, and Aida Valls. 2013. A semantic framework to protect the privacy of electronic health records with non-numerical attributes. *Journal of biomedical informatics*, 46(2):294–303.

Wonsuk Oh and Girish N Nadkarni. 2023. Federated learning in health care using structured medical data. *Advances in kidney disease and health*, 30(1):4–16.

Srijit Paul, Sajeeb Das, Akib Jayed Islam, Sultanus Salehin, and Prasun Datta. 2025. Federated learning for privacy-preserving gait recognition on edge devices using imu data. In *Proceedings of the 16th International IEEE Conference on Computing, Communication and Networking Technologies (ICCCNT)*, IIT Indore, Madhya Pradesh, India. IEEE Electronics Packaging Society and All India Council for Technical Education (AICTE).

PMC. 2025. Protecting privacy using k-anonymity.

Sherpa.ai. 2025. Federated learning vs. data anonymization: Why sherpa.ai is the most advanced privacy-preserving ai solution.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. *MedIR workshop*.

SPRY PT. 2025. Ai and machine learning in healthcare: Federated learning privacy.

Latanya Sweeney. 2002. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570.

Zhen Ling Teo, Liyuan Jin, Nan Liu, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, et al. 2024a. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, 5(2).

Zhen Ling Teo, Liyuan Jin, Nan Liu, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, et al. 2024b. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, 5(2).

U.S. National Library of Medicine. 2025. Unified medical language system (umls).

Tiago Andres Vaz, José Miguel Silva Dora, Luís da Cunha Lamb, and Suzi Alves Camey. 2023. Ontology for healthcare artificial intelligence privacy in brazil. *arXiv preprint arXiv:2304.07889*.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Xinyu Yuan, Zhihao Zhan, Zuobai Zhang, Manqi Zhou, Jianan Zhao, Boyu Han, Yue Li, and Jian Tang. 2024. Cell ontology guided transcriptome foundation model. *Advances in Neural Information Processing Systems*, 37:6323–6366.

Lele Zheng, Yang Cao, Masatoshi Yoshikawa, Yulong Shen, Essam A Rashed, Kenjiro Taura, Shouhei Hanaoka, and Tao Zhang. 2025a. Sensitivity-aware differential privacy for federated medical imaging. *Sensors*, 25(9):2847.

Lele Zheng, Yang Cao, Masatoshi Yoshikawa, Yulong Shen, Essam A Rashed, Kenjiro Taura, Shouhei Hanaoka, and Tao Zhang. 2025b. Sensitivity-aware differential privacy for federated medical imaging. *Sensors*, 25(9):2847.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *Advances in neural information processing systems*, volume 32, pages 14774–14784.