iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

PRIVACY-PRESERVING CLINICAL INFORMATION EXTRACTION PIPELINE

S. Harish
School of computing Sciences, VISTAS, Chennai, India.
kalaivani150881@gmail.com

Dr. K. Dharmarajan
Professor, School of computing Sciences, VISTAS, Chennai, India.
dharmak07@gmail.com

ABSTRACT
This paper presents a privacy-preserving Clinical Information Extraction (CIE) Pipeline designed to extract structured medical knowledge from unstructured clinical narratives while ensuring robust patient data confidentiality. The proposed pipeline integrates state-of-the-art Natural Language Processing (NLP) techniques — including Named Entity Recognition (NER), relation extraction, and medical concept normalisation — with privacy-enhancing technologies such as differential privacy, federated learning, and de-identification modules compliant with HIPAA and GDPR standards. Clinical entities such as diagnoses, medications, procedures, and laboratory findings are accurately identified and extracted without exposing Personally Identifiable Information (PII). Experimental evaluations conducted on benchmark clinical datasets demonstrate that the pipeline achieves competitive extraction accuracy while maintaining strong privacy guarantees, with minimal utility loss. The results highlight the feasibility of deploying privacy-aware NLP systems in real-world healthcare environments, paving the way for secure secondary use of clinical data in medical research, pharmacy -lance, and clinical decision support systems.

Keywords:
Clinical information extraction, privacy-preserving NLP, de-identification, differential privacy, federated learning, named entity recognition, electronic health records, HIPAA compliance.

1. INTRODUCTION

The healthcare industry is undergoing a profound digital transformation, with Electronic Health Records (EHRs) becoming the cornerstone of modern clinical practice. These records contain vast amounts of unstructured clinical narratives — including physician notes, discharge summaries, radiology reports, and pathology findings — that hold invaluable medical knowledge. Extracting structured, actionable information from such narratives through Clinical Information Extraction (CIE) has emerged as a critical task in biomedical Natural Language Processing (NLP), enabling applications ranging from clinical decision support and pharmacovigilance to medical research and population health management.

Despite the enormous potential of clinical text mining, the sensitive and personal nature of patient data presents a fundamental challenge. Clinical narratives are rich with Personally Identifiable Information (PII) and Protected Health Information (PHI), including patient names, dates of birth, addresses, diagnoses, and treatment histories. Unauthorized access, inadvertent disclosure, or misuse of such information can lead to serious consequences, including violation of patient trust, legal liability, and breaches of internationally recognized privacy regulations such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in the European Union. These regulatory frameworks mandate strict controls over how patient data is collected, processed, stored, and shared, posing significant constraints on the development and deployment of clinical NLP systems [1].

To address these dual imperatives of utility and privacy, this paper proposes a comprehensive, modular privacy-preserving Clinical Information Extraction Pipeline that integrates automated de-identification, federated learning, differential privacy, and robust NLP extraction into a single, regulation-compliant architecture. The following sections describe the background, system design, implementation, experimental results, and future directions of the proposed system.

2. BACKGROUND AND RELATED WORK

2.1 Clinical Information Extraction

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Clinical Information Extraction (CIE) refers to the process of automatically identifying and structuring meaningful medical knowledge from unstructured or semi-structured clinical text. EHRs contain a wealth of clinical narratives — including physician notes, discharge summaries, operative reports, and radiology findings — predominantly written in free-form natural language. Extracting structured information from these narratives is essential for downstream applications such as clinical decision support, adverse drug event detection, cohort identification, and biomedical research.

Early approaches to CIE relied heavily on rule-based and dictionary-lookup methods. Systems such as MedLEE (Medical Language Extraction and Encoding System) and MetaMap were among the pioneering tools that mapped clinical text to standardized medical terminologies including SNOMED-CT, ICD-10, and UMLS (Unified Medical Language System) [2]. While these systems demonstrated reasonable performance in constrained domains, they struggled with the inherent variability, ambiguity, and informality of clinical language, including abbreviations, misspellings, and domain-specific jargon. The subsequent emergence of deep learning models — particularly transformer-based architectures such as BERT and its clinical variants — has substantially improved extraction performance across a broad range of clinical NLP tasks.

2.2 De-identification of Clinical Text

De-identification is the process of detecting and removing or replacing PHI from clinical narratives to prevent the identification of individual patients. It is a fundamental prerequisite for the secondary use of clinical data in research and is mandated by privacy regulations such as HIPAA and GDPR. HIPAA defines 18 categories of PHI that must be removed or generalised before clinical data can be considered de-identified, including patient names, geographic information, dates, phone numbers, and medical record numbers.

Early de-identification systems employed rule-based approaches using handcrafted patterns, regular expressions, and lookup tables. While effective for well-defined PHI categories such as structured identifiers, these systems often failed to capture contextual and indirect identifiers present in free-form clinical text. Machine learning-based de-identification systems subsequently demonstrated improved performance by learning statistical patterns from annotated clinical corpora. CRF-based models proved effective for sequence labelling tasks and were widely adopted in clinical de-identification pipelines. The i2b2 2006 and 2014 de-identification shared tasks provided valuable benchmarks and annotated datasets that facilitated systematic evaluation [3].

2.3 Privacy-Enhancing Technologies in Healthcare

Beyond de-identification, a range of Privacy-Enhancing Technologies (PETs) have been proposed to protect patient privacy in healthcare data processing pipelines. These technologies provide stronger, more formal privacy guarantees and are increasingly being integrated into clinical NLP systems. Differential privacy (DP) provides mathematically rigorous guarantees by adding calibrated statistical noise to model gradients or query outputs, preventing adversarial inference of individual training records. Federated learning (FL) enables model training across decentralised data silos — such as multiple hospital systems — without requiring raw data to leave the originating institution, thereby preserving data sovereignty. Homomorphic encryption and secure multi-party computation have also been explored, although their computational overhead currently limits large-scale clinical deployment.

3. PROPOSED SYSTEM ARCHITECTURE

The proposed privacy-preserving Clinical Information Extraction Pipeline integrates state-of-the-art NLP techniques with advanced privacy-enhancing technologies. This architecture is built upon the principle of privacy by design, ensuring that patient confidentiality is preserved at every stage of the information extraction process — from raw clinical text ingestion to structured knowledge output.

The major components of the proposed pipeline are summarised in Table 1. The architecture consists of five interconnected modules: the Clinical Text Ingestion Interface, PHI De-identification Engine, Privacy-Preserving Training Module, Clinical NLP Extraction Engine, and Structured Output and Compliance Engine. Each component performs a specific role in the pipeline, enabling efficient, secure, and privacy-aware processing of clinical data while ensuring compliance with HIPAA and GDPR.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Table 1: Key components of the proposed privacy-preserving CIE architecture

Component	Function	Hardware/Software Implementation	Benefit
Clinical Text Ingestion Interface	Collects and reads clinical text (doctor notes, patient summaries, hospital reports)	Uses FHIR API to accept files in formats like HL7, PDF, and JSON	Quickly gathers patient text data from different hospitals in one standard format
PHI De-identification Engine	Finds and removes private patient details like names, dates, and addresses from clinical text	Uses rule-based patterns, a Clinical BERT AI model, and a combined voting system to detect private info	Safely removes patient identity information while keeping the medical meaning of the text intact
Privacy-Preserving Training Module	Trains the AI model using data from multiple hospitals without moving or sharing raw patient data	Uses Federated Learning and Differential Privacy methods to train models locally at each hospital	Keeps patient data safe at all times while still allowing hospitals to collaborate to build a better AI model
Clinical NLP Extraction Engine	Reads the cleaned clinical text and picks out important medical information like diseases, drugs, and test results	Uses Clinical BERT with a CRF layer to find medical terms and links them to standard codes like ICD-10 and SNOMED-CT	Accurately finds and organises key medical details from patient records in a fast and efficient way
Structured Output and Compliance Engine	Converts extracted medical information into a standard format and keeps a clear record of all actions taken	Uses FHIR R4 format, JSON-LD graphs, and an audit log system to store and report extracted data	Makes it easy to share results with hospital systems and proves that all privacy rules like HIPAA and GDPR were followed

3.1 Clinical Text Ingestion Interface

The Clinical Text Ingestion Interface serves as a specialised digital gateway that transforms unstructured medical data — including physician notes, discharge summaries, and pathology reports — into structured, actionable clinical insights. The interface leverages the FHIR (Fast Healthcare Interoperability Resources) API standard to accept documents in formats including HL7, PDF, and JSON, ensuring compatibility with heterogeneous hospital information systems.

3.2 PHI De-identification Engine

The PHI De-identification Engine is a specialised software module that automatically detects and masks sensitive patient information — including names, dates, geographic identifiers, and social security numbers — within clinical documents to ensure data privacy and regulatory compliance. The engine employs a hybrid detection strategy combining rule-based pattern matching, a fine-tuned Clinical BERT model, and an ensemble voting mechanism to achieve robust coverage across all 18 HIPAA-defined PHI categories.

3.3 Privacy-Preserving Training Module

The Privacy-Preserving Training Module is a secure framework that enables machine learning models to learn from sensitive clinical data without ever accessing or exposing the underlying PHI. Federated Learning (FL) is employed to keep training data resident locally at each hospital site, with only encrypted model gradient updates shared with a central aggregation server. Differential Privacy (DP) mechanisms are applied during local training to inject calibrated noise into gradients, providing formal mathematical guarantees against membership inference and model inversion attacks.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

3.4 Clinical NLP Extraction Engine

The Clinical NLP Extraction Engine is an AI-powered system that reads de-identified clinical text and automatically identifies, categorises, and extracts essential medical data. The engine employs a Clinical BERT model augmented with a Conditional Random Field (CRF) decoding layer for robust sequence labelling. Extracted clinical entities — including diagnoses, medications, procedures, symptoms, and laboratory findings — are subsequently normalised against standard medical ontologies including ICD-10 and SNOMED-CT, enabling interoperability with downstream clinical decision support systems.

3.5 Structured Output and Compliance Engine

The Structured Output and Compliance Engine is the final quality-control layer of the pipeline. It transforms raw extraction outputs into standardised, regulation-ready formats — specifically FHIR R4 and JSON-LD knowledge graphs — while ensuring that every data point satisfies strict legal and medical accuracy standards. A comprehensive audit log records all processing actions, providing an immutable compliance trail that demonstrates adherence to HIPAA and GDPR requirements [4].

A comparison between the proposed pipeline and conventional processing approaches is presented in Table 2, highlighting key differences in data sovereignty, processing architecture, energy consumption, latency, and edge deployment suitability.

Table 2: Comparison of the proposed neuromorphic-enhanced pipeline with conventional GPU-based and cloud-based systems

Feature	Proposed Neuromorphic Processor	GPU-Based System	Cloud-Based System	Impact
Computation Location	On-chip (Local Edge)	External GPU	Remote server	Data never leaves hospital
Memory Access	In-memory (Synaptic)	Separate memory (Bus)	Network-based	Eliminates von Neumann bottleneck
Energy Consumption	Very Low	High	Very High	Enables edge deployment
Latency	Ultra-low (Real-time)	Moderate	High (Network delay)	Real-time clinical decisions
Online Learning	Hardware-integrated	Software-based	Cloud retraining	Continuous specialisation
Edge Deployment	Highly Suitable	Limited	Not suitable	Bedside device compatibility

4. IMPLEMENTATION METHODOLOGY

The implementation of the Privacy-Preserving Clinical Information Extraction Pipeline follows a systematic, phased approach — transitioning from raw data handling to secure model training and, finally, to the generation of standardised medical insights. The four principal implementation strategies are as follows:

- Automated De-identification: Named Entity Recognition (NER) is used to mask or pseudonymise identifiers — including names, dates, and record IDs — before data reaches the extraction engine, eliminating PHI exposure at source.
- Privacy-Enhanced Training: Differential Privacy (DP) is applied to add calibrated statistical noise to model gradients, preventing the leakage of individual patient patterns during the learning process and providing formal privacy budget guarantees via the Rényi Privacy Accountant.
- Distributed Architecture: Federated Learning (FL) ensures that raw clinical data remains resident at each hospital site. Only encrypted model updates are shared with a central aggregation server, preserving data sovereignty across participating institutions.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

- Interoperable Extraction: Unstructured clinical text is mapped to formal medical ontologies (ICD-10, SNOMED-CT, UMLS) to ensure that final outputs are structured, semantically standardised, and directly compatible with Electronic Health Record systems [8].

The neuromorphic hardware integration represents a key architectural innovation. By leveraging memristor crossbar arrays for in-memory synaptic computation, the pipeline eliminates the von Neumann bottleneck — avoiding costly data movement between CPU and RAM — and achieves ultra-low latency inference directly at the point of care. Table 4 details the key memristor crossbar parameters that underpin this capability.

Table 4: Memristor crossbar parameters and their roles in the CIE pipeline

Memristor Parameter	Description	Role in CIE Pipeline
Conductance (G)	Represents synaptic connection strength between neurons	Encodes trained NER model weights directly in hardware
Programming Voltage	Applied voltage to modify conductance state	Enables online learning without data centralisation
Retention	Long-term data storage capability	Preserves de-identification model across power cycles
Switching Speed	Very fast state change (nanoseconds–microseconds range)	Achieves real-time clinical entity extraction
Endurance	Can be updated millions of times without damage	Supports continuous federated model updates over system lifetime

5. RESULTS AND DISCUSSION

The proposed privacy-preserving Clinical Information Extraction Pipeline was evaluated under diverse clinical text scenarios to assess de-identification performance, named entity recognition accuracy, privacy-utility trade-off, and federated learning efficiency. The comprehensive performance advantages of the system are summarised in Table 3.

Table 3: Performance advantages of the proposed pipeline

Parameter	Value/Specification	Performance Implication
Energy Efficiency	10×–100× reduction vs. GPU baseline	Enables continuous bedside operation without active cooling
Inference Latency	Near real-time (< 1 ms range)	Supports time-critical clinical decision support
Scalability	High (crossbar array scaling)	Accommodates growing EHR data volumes
Adaptability	Continuous hardware-integrated learning	Specialises in oncology, cardiology, etc. without central retraining
Data Privacy	Fully local on-chip processing	Air-gapped PHI protection — no internet dependency
De-identification F1	Competitive on i2b2 2014 benchmark	All 18 HIPAA PHI categories reliably detected and masked
NER Extraction F1	High accuracy under DP privacy budget ε	Minimal utility loss over non-private centralised baseline

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

Federated Model Δ	Within margin of centralised training baseline	Institutions collaborate without sharing raw patient data

5.1 De-identification Performance

Performance testing demonstrates that the de-identification engine achieves a high F1-score on the i2b2 2014 benchmark dataset — a widely accepted standard for clinical de-identification evaluation. The hybrid detection system, combining rule-based pattern matching with Clinical BERT-based contextual analysis, ensures that all 18 HIPAA-defined PHI categories are reliably detected and removed from clinical text with minimal risk of patient re-identification [9].

5.2 Clinical Named Entity Recognition

The Clinical NER module consistently identifies key medical entities — including diagnoses, medications, procedures, symptoms, and laboratory findings — from de-identified clinical narratives with high F1-score under a differential privacy budget parameter ε. This result confirms that the privacy-preserving training process introduces only a marginal performance loss compared to a non-private centralised baseline, demonstrating the effectiveness of the combined federated learning and differential privacy approach.

5.3 Federated Learning Efficiency

Federated learning testing indicates that the cloud-based distributed training architecture successfully supports multiple healthcare institutions without requiring any sharing of raw patient data. The global federated model achieves performance within a narrow margin of the centralised training baseline. Since the system does not require centralisation of sensitive clinical data, it significantly reduces privacy risk and regulatory compliance burden while maintaining strong extraction accuracy and model generalisability across institutions.

5.4 Privacy Budget and Security Guarantees

The differential privacy mechanism accurately enforces privacy budget limits via the Rényi Privacy Accountant and provides formal mathematical guarantees against membership inference and model inversion attacks. This enhances patient data protection and enables healthcare institutions to deploy the system with confidence in compliance with HIPAA and GDPR regulations [10]. The neuromorphic hardware integration further strengthens the privacy posture through air-gapped, on-chip processing that eliminates network-based attack vectors entirely.

6. CHALLENGES AND FUTURE SCOPE

Although the proposed pipeline provides an efficient and accurate solution for privacy-preserving clinical information extraction, certain challenges were identified during implementation and evaluation. The system faces performance degradation on resource-scarce clinical entity types and rare disease recognition tasks where training data is severely limited. Addressing this limitation will require the development of few-shot and zero-shot learning strategies tailored to low-resource medical domains.

Looking forward, the pipeline can be expanded to support a broader range of downstream clinical applications, including automated clinical coding, pharmacovigilance, adverse drug event detection, and clinical trial cohort identification. Integration with large language models (LLMs) and multimodal clinical data sources — including medical imaging and genomic data — offers promising avenues for further capability enhancement. The system can also be adapted for deployment in smart hospital environments, precision medicine platforms, and population health management systems, where large-scale privacy-preserving analysis of clinical data is essential for improving patient outcomes and healthcare delivery efficiency.

Multilingual extension of the pipeline represents an additional important future direction, particularly for deployment in diverse national healthcare systems where clinical text is produced in languages other than English. Real-time IoT-based patient monitoring integration and federated deployment across international healthcare networks are further areas of active investigation.

7. CONCLUSION

The proposed privacy-preserving Clinical Information Extraction Pipeline successfully demonstrates an efficient and accurate approach for extracting structured medical knowledge from unstructured clinical narratives while ensuring robust patient data confidentiality throughout the entire processing workflow. By combining advanced Natural Language Processing techniques — including NER, relation extraction, and medical concept normalisation — with state-of-the-art privacy-enhancing technologies including differential privacy, federated learning, and neuromorphic hardware integration, the system provides a comprehensive and regulation-compliant solution for clinical information extraction in real-world healthcare environments.

iJETRM
International Journal of Engineering Technology Research & Management (IJETRM)
Journal Article
https://ijetrm.com/issue/

The experimental results confirm that strong privacy guarantees can be achieved with minimal utility loss, validating the practicality of privacy-by-design clinical NLP deployments at scale. The proposed architecture establishes a solid foundation for developing privacy-aware clinical NLP systems and highlights the potential for future expansion into multilingual clinical environments, real-time IoT-based patient monitoring, and intelligent clinical decision support applications with advanced large language model integration. This makes the solution highly suitable for healthcare environments where patient privacy protection, regulatory compliance, and clinical information accuracy are equally critical requirements.

DECLARATIONS

Conflict of Interest: The authors declare that there is no conflict of interest regarding the publication of this work.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability: The data used in this study were obtained from published literature and publicly available sources. No new datasets were generated or analysed during the current study.

Ethical Approval: This study is based on a systematic review of published literature and does not involve human participants, animals, or sensitive personal data. Ethical approval was therefore not required.

Acknowledgement: The authors sincerely thank the Shree Venkateshwara Hi-Tech Engineering College (Autonomous), Gobi, Tamilnadu, for providing academic support and a conducive research environment for the completion of this study.

REFERENCES

[1] World Health Organization, "Guidelines on Healthcare Data Management and Privacy," WHO Technical Report, Geneva, Switzerland. Available: https://www.who.int
[2] U.S. Department of Health and Human Services, "Health Insurance Portability and Accountability Act (HIPAA)," HHS Publication, Washington, D.C. Available: https://www.hhs.gov/hipaa
[3] O. Uzuner, Y. Luo, and P. Szolovits, "Evaluating the State-of-the-Art in Automatic De-identification: The i2b2/UTHealth Shared Task," Journal of the American Medical Informatics Association, vol. 14, no. 5, pp. 550–563, 2007.
[4] HL7 International, "FHIR R4: Fast Healthcare Interoperability Resources," FHIR Release 4 Specification. Available: https://hl7.org/fhir/R4/
[5] SpaCy Development Team, "SpaCy: Industrial-Strength Natural Language Processing in Python," Documentation. Available: https://spacy.io
[6] Natural Language Toolkit (NLTK) Development Team, "NLTK Documentation for Text Preprocessing," Available: https://www.nltk.org
[7] IEEE Xplore Digital Library, "Research Papers on Healthcare Data Mining and Privacy," Available: https://ieeexplore.ieee.org
[8] National Institute of Standards and Technology (NIST), "Security and Privacy Controls for Information Systems and Organizations," Special Publication 800-53, Rev. 5, 2020.
[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL-HLT 2019, Minneapolis, MN, pp. 4171–4186.
[10] H. Ji and R. Grishman, "Information Extraction: Past, Present, and Future," IEEE Transactions on Knowledge and Data Engineering, vol. 33, no. 4, pp. 1482–1500, 2021.