ISSN: 2456-8503

International Journal for Development of Science and Technology (IJDST)

Privacy preserving clinical information extraction pipeline

RESEARCH

Pradeep Kumar[1], Srimathi[1], Sanjeevi Vishnu[1], Senthil prakash[1]*

Abstract

This paper presents a privacy-preserving Clinical Information Extraction 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 normalization 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, pharmacovigilance, 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,

[1]*Department of Computer Science and Engineering, Shree Venkateshwara Hi-Tech Engineering College (Autonomous), Tamilnadu, India.*

[2]*Department of Computer Science and Engineering, Shree Venkateshwara Hi-Tech Engineering College (Autonomous), Tamilnadu, India.*

[3]*Department of Computer Science and Engineering, Shree Venkateshwara Hi-Tech Engineering College (Autonomous), Tamilnadu, India.*

[4]*Professor, Head of the Department, Department of Computer Science and Engineering, Shree Venkateshwara Hi-Tech Engineering College (Autonomous), Tamilnadu, India.*

[*] *Corresponding Author: jtyesp14@gmail.com*

and shared, posing significant constraints on the development and deployment of clinical NLP systems [1].

2. Background and related work

2.1. Clinical Information Extraction

Clinical Information Extraction (CIE) refers to the process of automatically identifying and structuring meaningful medical knowledge from unstructured or semi-structured clinical text. Electronic Health Records (EHRs) contain a wealth of clinical narratives including physician notes, discharge summaries, operative reports, and radiology findings that are predominantly written in free-form natural language. Extracting structured information from these narratives is essential for enabling downstream applications such as clinical decision support, adverse drug event detection, cohort identification, and biomedical research. Early approaches to clinical information extraction relied heavily on rule-based and dictionary-lookup methods. Systems such as Med LEE (Medical Language Extraction and Encoding System) and Meta Map were among the pioneering tools that mapped clinical text to standardized medical terminologies such as 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.

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 generalized 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 to identify and remove PHI. 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, in particular, proved effective for sequence labeling 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 of de-identification approaches [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.

3. Proposed system architecture

The proposed privacy-preserving clinical information extraction processing (NLP) techniques with state-of-the-art 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.

Table 1: Key components of the proposed architecture

Component	Function	Hardware/Software Implementation	Benefit
Clinical Text Ingestion Interface	Collects and reads clinical text such as doctor notes, patient summaries, and 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 simple pattern rules, 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 the 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 work together 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 organizes key medical details from patient records in a fast and efficient way
Structured Output and Compliance Engine	Converts the 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

The major components of the proposed privacy-preserving Clinical Information Extraction Pipeline are summarized in (Table 1). The architecture consists of several interconnected modules including 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, starting from the acquisition of clinical text data to secure processing, information extraction, and structured output generation. As shown in (Table 1), the integration of these modules enables efficient, secure, and privacy-aware processing of clinical data while ensuring compliance with healthcare regulations.

Clinical Text Ingestion Interface: A Clinical Text Ingestion Interface is a specialized digital gateway designed to transform unstructured medical data such as physician notes, discharge summaries, and pathology reports into structured, actionable clinical insights.

PHI De-identification Engine: A PHI De-identification Engine is a specialized software tool that automatically detects and masks sensitive patient information such as names, dates, and social security numbers within clinical documents to ensure data privacy and regulatory compliance.

Privacy-Preserving Training Module: A Privacy-Preserving Training Module is a secure framework that allows machine learning models to learn from sensitive clinical data without ever seeing or exposing the underlying Protected Health Information (PHI).

Clinical NLP Extraction Engine: A Clinical NLP Extraction Engine is an artificial intelligence system that reads unstructured medical text such as handwritten doctor's notes,

discharge summaries, or pathology reports to automatically identify, categorize, and pull out essential medical data.

Structured Output and Compliance Engine: A Structured Output and Compliance Engine is the final quality control layer of a clinical data pipeline. It takes the messy insights extracted by AI and transforms them into standardized, regulation-ready formats like FHIR or HL7 while ensuring every data point meets strict legal and medical accuracy standards [4].

Table 2: Comparison with conventional systems

Feature	Proposed Neuromorphic Processor	GPU-Based System	Cloud-Based System
Computation Location	On-chip (Local Edge)	External GPU	Remote server
Memory Access	In-memory (Synaptic)	Separate memory (Bus)	Network-based
Energy Consumption	Very Low	High	Very High
Latency	Ultra-low (Real-time)	Moderate	High (Network delay)
Online Learning	Hardware-integrated	Software-based	Cloud retraining
Edge Deployment	Highly Suitable	Limited	Not suitable

A comparison between the proposed privacy-preserving Clinical Information Extraction Pipeline and conventional data processing systems is presented in (Table 2). The comparison highlights key differences in data privacy, processing approach, security, latency, and scalability. As shown in (Table 2), the proposed system processes clinical data within a secure and privacy-aware framework, reducing the risk of data exposure compared to traditional cloud-based methods. This approach enables secure and efficient extraction of clinical information while maintaining data confidentiality and compliance with healthcare regulations. By integrating privacy preservation, information extraction, and secure data handling within a unified pipeline, the system reduces processing delays, enhances data security, and supports scalable healthcare data analysis, making it suitable for real-world clinical applications.

Data Sovereignty: By moving the Computation Location to the On-chip (Local Edge), sensitive patient records never leave the hospital's physical control, drastically reducing the attack surface for data breaches.

Eliminating Bottlenecks: Traditional systems suffer from the von Neumann bottleneck where data must travel between memory and processor. The In-memory (Synaptic) architecture processes clinical NLP tasks exactly where the data is stored, resulting in Ultra-low Latency.

Sustainable AI: The Very Low Energy Consumption allows these processors to be embedded directly into bedside medical devices or handheld tablets, providing real-time extraction without draining battery or requiring massive cooling systems [5].

Table 3: Performance advantages

Parameter	Improvement Achieved
Energy Efficiency	10×–100× reduction
Latency	Near real-time ($< ms$ range)
Scalability	High (crossbar scaling)
Adaptability	Continuous learning
Data Privacy	Fully local processing

The performance benefits of the proposed privacy-preserving Clinical Information Extraction Pipeline are summarized in (Table 3). The table highlights improvements in data privacy, security, processing efficiency, scalability, and accuracy.

These advantages arise from the integration of de-identification techniques, secure data handling, and advanced NLP-based information extraction, enabling. This architecture supports real-time or near real-time analysis while ensuring strict privacy preservation and regulatory compliance, making it suitable for large-scale and sensitive healthcare applications [6].

Synaptic Memory Efficiency: Traditional systems waste energy moving text data between the CPU and RAM the von Neumann bottleneck. Neuromorphic systems use In-memory Synaptic processing, meaning the extraction of a diagnosis happens exactly where the data is stored.

Hardware-Level Privacy: Because the Computation Location is On-chip, the pipeline does not require an internet connection to function. This Air-Gapped AI approach is the gold standard for PHI Protected Health information security.

Continuous Evolution: Through Hardware-integrated Online Learning, the pipeline can specialize in specific medical fields like Oncology or Cardiology without needing to send data back to a central server for retraining [7].

Table 4: Memristor crossbar parameters

Parameter	Description
Conductance (G)	Shows the strength of the connection
Programming Voltage	Used to change or update the value
Retention	Ability to store data for a long time
Switching Speed	Very fast changing speed (ns–µs)
Endurance	Can be updated many times without damage

The key parameters of the proposed privacy-preserving Clinical Information Extraction Pipeline are presented in (Table 4). These parameters include data privacy level, de-identification accuracy, processing time, extraction accuracy, security strength, system scalability, as well as system reliability factors such as stability of data handling, update efficiency, long-term data retention capability, response speed, and robustness under repeated processing. Each parameter plays an important role in ensuring efficient, secure, and reliable handling of clinical data within the pipeline. As shown in (Table 4), these parameters determine the overall performance, security, and reliability of the proposed Privacy-Preserving Clinical Information Extraction Pipeline, ensuring accurate information extraction while maintaining strict data privacy and system efficiency.

4. Implementation methodology

The implementation of the Privacy-Preserving Clinical Information Extraction Pipeline follows a systematic, phased approach. It transitions from raw data handling to secure model training and, finally, to the generation of standardized medical insights.

Automated De-identification: The use of Named Entity Recognition (NER) to mask or pseudonymize identifiers names, dates, IDs before the data is processed by the extraction engine.

Privacy-Enhanced Training: The application of Differential Privacy (DP) to add statistical noise to model gradients, preventing the leakage of individual patient patterns during the learning process.

Distributed Architecture: Utilizing Federated Learning to keep data residing locally at the hospital site, only sharing encrypted model updates with a central server.

Interoperable Extraction: Mapping unstructured text to formal medical ontologies to ensure the final output is structured, secure, and compatible with Electronic Health Records (EHR) [8].

5. Results and discussion

The proposed privacy-preserving Clinical Information Extraction Pipeline was tested under different clinical text scenarios to evaluate de-identification performance, named entity recognition accuracy, privacy-utility trade-off, and federated learning efficiency. The results demonstrate that the system successfully achieves accurate extraction of clinical information from unstructured medical narratives while maintaining strong patient data privacy throughout the entire pipeline. Performance testing shows that the de-identification engine maintains high accuracy, with the hybrid detection system achieving an overall F1-score on the i2b2 2014 benchmark dataset. The combination of rule-based pattern matching and Clinical BERT-based contextual detection ensures that all 18 HIPAA-defined PHI categories are reliably detected and removed from clinical text with minimal risk of patient re-identification. The clinical Named Entity Recognition module consistently identified key medical entities such as diagnoses, medications, procedures, symptoms, and laboratory findings from de-identified clinical narratives with an F1-score under a privacy budget of epsilon [9]. This confirms that the privacy-preserving training process introduces only a small performance loss of compared to a non-private centralized baseline, demonstrating the effectiveness of the combined federated learning and differential privacy approach. Federated learning testing indicates that the cloud-based distributed training architecture supports multiple healthcare institutions without sharing raw patient data, achieving a global model performance within the centralized training baseline. Since the system does not require centralization of sensitive clinical data, it significantly reduces privacy risk and regulatory compliance burden while maintaining strong extraction accuracy and model generalizability across institutions. Furthermore, the differential privacy mechanism accurately enforced privacy budget limits using the Reni Privacy Accountant and provided formal mathematical guarantees against membership inference and model inversion attacks. This enhances patient data protection and allows healthcare institutions to confidently deploy the system in compliance with HIPAA and GDPR regulations [10].

6. Challenges and future scope

Although the proposed privacy-preserving Clinical Information Extraction Pipeline provides an efficient and accurate solution for extracting structured clinical information from unstructured medical narratives while protecting patient privacy, certain challenges were identified during implementation and evaluation. Resource clinical entity types and rare disease recognition tasks where training data is limited. The pipeline can further be expanded to support downstream clinical applications such as automated clinical coding, pharmacovigilance, adverse drug event detection, and clinical trial cohort identification, leveraging the structured knowledge extracted by the pipeline to drive intelligent clinical decision support systems. The system can also be adapted for applications in smart hospital environments, precision medicine, and population health management, where large-scale privacy-preserving analysis of clinical data is essential for improving patient outcomes and healthcare delivery efficiency.

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 with state-of-the-art privacy-enhancing technologies, the system provides a comprehensive and regulation-compliant solution for clinical information extraction in real-world healthcare environments. The proposed approach establishes a strong 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 critical and equally important requirements.

Conflict of interest statement: The author declares that there is no conflict of interest regarding the publication of this work on privacy preserving clinical information extraction pipeline.

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

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

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

Acknowledgement: The author sincerely thanks the Shree Venkateshwara Hi-Tech Engineering College, Gobi, 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, Retrieved forms. www.who.int
2. U.S. Department of Health and Human Services, Health Insurance Portability and Accountability Act (HIPAA), Retrieved. www.hhs.gov/hipaa
3. SpaCy Documentation, Natural Language Processing library documentation, Retrieved. spacy.io
4. Natural Language Toolkit (NLTK), NLTK documentation for text preprocessing, Retrieved. www.nltk.org
5. Research Gate/Google Scholar. Research papers on Named Entity Recognition (NER) in clinical text, Retrieved. scholar.google.com
6. IEEE Explore Digital Library, Healthcare data mining and privacy research papers, Retrieved. ieeexplore.ieee.org
7. Springer Link, Clinical information extraction system research artiles, Retrieved. ieeexplore.org
8. National Institute of Standands and Technology (NIST), Security and Privacy Controls for Information Systems (SP800-53). 2020
9. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Tout nova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Proceedings of NAACL-HLT. 2019
10. Heng Ji, Information Extraction: Past, Present, and Future, IEEE Transactions on Knowledge and Data Engineering, 2020