

*Research Article***C-PPDSS: A Comprehensive Patient Personal Data Sovereignty System for Real-Time PII Redaction in Healthcare LLM Workflows**

**Edoghogho Olaye^a , Williams O. Aigbe^b , David Onwuka^{a,*} , Daniel Obuh^a ,
Amarachi A. Madukwe^a**

^a University of Benin, Edo State, Nigeria

^b Chemnitz University of Technology, Germany and Fraunhofer-Institute for Electronic Nano Systems (ENAS) Chemnitz, Germany

ARTICLE INFO

Article history:

Received 23 May 2026

Accepted 28 August 2026

Keywords:

C-PPDSS,
Healthcare AI,
Large Language Models,
Patient Data Privacy,
PII Redaction,
PIIRANHA

ABSTRACT

Health surveillance in modern healthcare increasingly relies on AI, raising concerns about the exposure of patients' identifiable information (PII) in automated workflows. This paper presents the Comprehensive Patient Personal Data Sovereignty System (CPPDSS), a privacy-preserving framework designed to enforce real-time PII redaction during outbound interactions with large language models. C-PPDSS employs a fine-tuned Custom PIIRANHA model, based on Microsoft's DeBERTa-v2 architecture, to intercept and sanitize HTTP POST requests containing sensitive healthcare data before they reach external LLM services. Evaluated on a held-out test set of 20,927 annotated examples, Custom PIIRANHA achieved 99.51% token-level accuracy and a micro-averaged F1 score of 95.45%, outperforming a fine-tuned Distilled BERT baseline across all major metrics. Quantitative benchmarking demonstrated a mean inference latency of 229.66 ms per request, a model memory footprint of 305.7 MB, and a sustained throughput of 4.35 requests per second on CPU-only hardware, confirming suitability for real-time deployment on modest intranet infrastructure. By securing patient information at the network level, C-PPDSS bridges a gap left by previous static-document-focused systems, enabling safe AI integration in healthcare workflows. Its architecture ensures reproducibility and feasibility for deployment in hospital intranets or standard commodity hardware setups.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. INTRODUCTION

From 24 May 2018, the General Data Protection Regulation (GDPR) of the European Union aims to protect individuals' privacy and data security which have a link to the Union's territory or market. This was demonstrated by the European Union's implementation of the comprehensive Data Protection Regulation. It mandates all Data Protection Authorities (DPAs) with the task to ensure companies clearly outline user agreements and forbids them from misleading or pressuring users into surrendering their privacy rights, and in the case of failure to adhere, they impose sanctions [1]. Individuals' data privacy and security become a concern that must be addressed, especially in this era of artificial intelligence (AI) [2], [3]. The adoption of AI in the healthcare sector

has rapidly improved disease prediction and diagnosis, treatment optimization, and overall health management. AI-powered health surveillance systems promise to enhance healthcare delivery efficiency, save lives, and reduce costs [4], [5].

Despite these opportunities and unprecedented capabilities, AI adoption in healthcare has also raised profound challenges around data privacy, equity, and accountability. As the healthcare sector increasingly relies on AI to monitor, predict, and diagnose patients' health, the absence of robust frameworks for digital sovereignty risks exposing sensitive health data to cross-border flows that may override national regulations and compromise privacy [6]. By "digital sovereignty," we follow [7] in treating the concept not merely as a descriptive label for

* Corresponding author. E-mail address: david.onwuka@eng.uniben.edu
DOI: 10.58190/ijamec.2026.178

actors exerting control over digital technologies, but as a normative framework centered on authority. This emphasizes that sovereignty in the digital domain requires legitimate control that is assessed through participation, accountability, consent and the ability to serve public needs and protect rights. Such an approach is particularly relevant to healthcare adopting AI-powered systems, where control must also be evaluated against standards of fairness and trust.

Framing the issue in this way raises critical questions: Who ultimately controls these systems? How are data collected, extracted, loaded, and transformed before, during, and after analysis?

Researchers have explored the challenges of data protection and privacy that result from the integration of AI in healthcare, highlighting risks of data breaches, algorithmic bias, and erosion of fundamental patient rights [8], [9]. To keep data localized, federated learning approaches have been proposed, yet these often lack end-to-end governance models that ensure true data sovereignty [10], [11].

In light of these challenges, this work seeks to address the following research questions:

How can sensitive health data be intercepted and protected before transmission to external AI or LLM systems?

Can privacy-preserving data redaction be performed in real time within healthcare workflows without disrupting clinical efficiency?

What governance and architectural frameworks can ensure that data sovereignty is preserved at both the institutional and national levels in AI-powered health surveillance systems?

By investigating these questions, this paper argues that the design and governance of AI-powered health technologies must prioritize innovation in ethical frameworks and controls that reconcile competing priorities of privacy, sovereignty, and equity to avoid eroding fundamental rights.

2. LITERATURE REVIEW

The integration of AI in healthcare raises urgent concerns around data privacy and sovereignty. While AI enhances diagnostic accuracy and treatment planning, it poses significant risks to patient confidentiality and data security [12]. Differential privacy and federated learning emerge as potential solutions to preserve patient privacy while utilizing healthcare data effectively [9]. The need for robust governance and risk management strategies is emphasized, with proposals like the 'Integrated Security and Ethics Model' addressing the dynamic nature of AI in healthcare [13]. Recent research has explored various approaches to ensure data security and sovereignty in healthcare surveillance. Patient-centric platforms using

proxy re-encryption allow patients to control access to their health records, enhancing data sovereignty [14]. Edge of Things (EoT) computing frameworks employing fully homomorphic encryption offer secure, real-time processing of IoT-generated patient data while preserving privacy [15]. Existing security measures include encryption, access controls, and secure authentication mechanisms, with a focus on compliance with healthcare acts like HIPAA [16]. Emerging technologies such as blockchain and artificial intelligence are being explored to fortify data security in healthcare [17], [18], [19]. [20] proposed a model combining Fuzzy C-means Clustering and Support Vector Machines for improved cloud security in surveillance, achieving 94.4% accuracy. [21] reviewed AI-driven cybersecurity in e-health systems, emphasizing AI's role in continuous monitoring, threat detection, and response. They also discussed future research directions, including quantum-resistant encryption and privacy-preserving AI techniques. [22] examined various AI and ML methods for healthcare system security, such as anomaly detection and risk prediction. The study evaluated different algorithms, including neural networks and decision trees, using metrics like accuracy and precision to assess their effectiveness in protecting patient data and IT networks from cyber threats. [6] developed a system that utilized an optical character recognition powered by AI to obfuscate personal identifiable patient's data capture before sending to the machine learning models for training. These papers collectively underscore the potential of AI and ML in bolstering security measures for sensitive data in healthcare and surveillance contexts. However, the protection of patient data before integration into large language models (LLMs) has received little scholarly attention. Our work involves the development of a Comprehensive Patient Personal Data Sovereignty System (C-PPDSS) in order to ensure on-premise personally identifiable information interception. Specifically, when healthcare practitioners use large language models (LLMs), our system ensures that the input received by the LLMs contains no personally identifiable patient information. Unlike approaches that focus primarily on data protection during the training stage, our work addresses the usage stage, ensuring that sensitive patient data are intercepted and obfuscated before being transmitted to the model for inference.

3. METHODOLOGY

The Comprehensive Patient Personal Data Sovereignty System (C-PPDSS) was designed to ensure that no personally identifiable information (PII) leaves the local environment when healthcare professionals interact with external large language models (LLMs). The system detects and filters personally identifiable information locally before data leaves the hospital, ensuring

compliance with regulations and preserving digital sovereignty. Relying on external personally identifiable information processing services was rejected due to risks of data misuse. Instead of using large, general-purpose LLMs like Mistral, the team developed a lightweight, fine-tuned model tailored to efficiently identify and replace healthcare-related personally identifiable information with placeholders.

3.1. System Design

The system was architected as a modular pipeline consisting of three primary components:

Request interception

Personally identifiable information filtration

Request transformation

The overall architecture of the Comprehensive Patient Personal Data Sovereignty System (C-PPDSS) is shown in Figure 1, which provides a high-level view of how these components interact. The architecture illustrates the flow of data from outbound requests, through interception and PII redaction, to transformation and forwarding to external large language models (LLMs). This diagram serves as a reference map for the subsections that follow, where each module of the pipeline is described in detail.

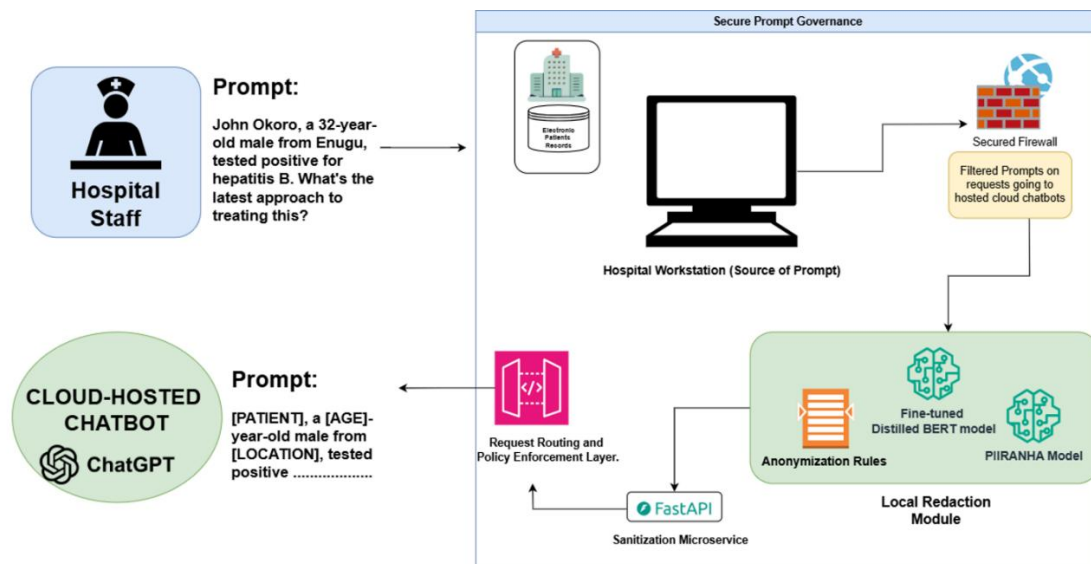


Figure 1. Overview of the C-PPDSS Architecture and Data Flow Pipeline

3.1.1. Request Interception

A system-level monitoring mechanism was implemented to detect outbound HTTP POST requests directed at OpenAI's ChatGPT service. Using browser developer tools, we first inspected the structure of these requests, confirming that user prompts were transmitted as JSON payloads within HTTPS POST calls to `api.openai.com/v1/chat/completions` as shown in Figures 2 and 3. To automate this process, we deployed a local HTTPS proxy to capture traffic at the network boundary. This enabled us to log, parse, and analyze the outbound request body in real time, providing visibility into exactly what information—potentially including PII—was leaving the hospital environment. The captured request structure formed the basis for the subsequent redaction module, which intercepts and replaces sensitive fields before forwarding the sanitized query to the LLM.

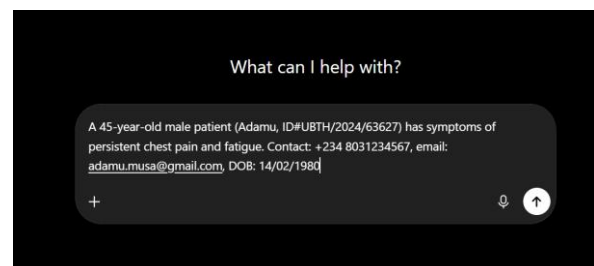


Figure 2. Sample ChatGPT web interface input containing potential patient PII

```

{
  "timestamp": "2026-08-11T13:52:02.715",
  "request": {
    "url": "https://chatgpt.com/backend-api/conversation",
    "headers": {
      ...
    },
    "content_type": "application/json",
    "body": {
      "action": "next",
      "messages": [
        {
          "id": "c3134bff-bf6b-4e02-bac1-84beae1e07e",
          "author": {
            "role": "user"
          },
          "create_time": 1746194852.685,
          "content": {
            "content_type": "text",
            "parts": [
              "A 45-year-old male patient (Adamu, ID#BIRTH/2024/63627) has symptoms of persistent chest pain. contacts: +234 8031234567, email: adamu.musa@gmail.com, DOB: 14/02/1980."
            ]
          },
          "metadata": {
            "selected_github_repos": [],
            "selected_all_github_repos": false,
            "serialization_metadata": {
              "custom_symbol_offsets": []
            },
            "dictation": false
          }
        ]
      }
    }
  }
}

```

Figure 3. System Log Entry showing a Captured HTTP POST Request containing a User Prompt

3.1.2. PII Filtration

Both candidate models were fine-tuned on the AI4Privacy PII-Masking-200k dataset (AI4Privacy, 2022), a publicly available Hugging Face dataset containing over 200,000 synthetically generated sentences annotated with personally identifiable information, including names, addresses, phone numbers, and identification codes. All data used across both fine-tuning procedures was either publicly sourced or synthetically generated, with no real patient data collected or processed at any stage.

The dataset was partitioned using a fixed random seed (seed = 42) into three non-overlapping splits, as shown in Table 1.

Table 1. Dataset Split Configuration

Split	Examples	Proportion
Training	167,416	80%
Validation	20,927	10%
Test	20,927	10%
Total	209,270	100%

The test set contains 56 distinct PII label types. Of these, 14 categories, GIVENNAME, SURNAME, EMAIL, TELEPHONENUM, ACCOUNTNUM, SOCIALNUM, PASSWORD, USERNAME, STREET, CITY, ZIPCODE, BUILDINGNUM, CREDITCARDNUMBER, and DATEOFBIRTH, fall within PIIRANHA's detection taxonomy and were used for evaluation scoring in Section 4. The remaining 42 label types (e.g., BITCOINADDRESS, USERAGENT, JOBTYP) lie outside PIIRANHA's classification scope and were excluded from scoring accordingly.

Both models were fine-tuned on a single personal computer to emphasize reproducibility on accessible, non-specialized hardware. The specifications of the shared fine-tuning environment are provided in Table 2.

Table 2. Specifications for Model Fine-tuning

Component	Specification	Purpose
CPU	Intel Core i5 (4 cores)	Parallel processing
GPU	—	Accelerated fine-tuning (not available)
RAM	8GB DDR4	Handling large datasets
Storage	256GB	Dataset and model checkpoint storage
Framework	PyTorch 2.13, Hugging Face Transformers	Model architecture and training

3.1.3. Fine-tuned Distilled BERT-Based Model

A distilled BERT model was fine-tuned on the dataset and split described above. To adapt the dataset to healthcare-specific contexts, additional annotations were curated by our team to capture domain-specific identifiers, including prescription codes and medical record numbers. Training required 78 hours of compute time under the environment specified in Table 2, following the configuration in Table 3.

Table 3. Distilled BERT-Based Model Fine-Tuning Configuration

Parameter	Value
Base model	distilbert-base-uncased (66M parameters)
Tokenizer	BERT WordPiece tokenizer
Maximum sequence length	512 tokens
Number of epochs	4
Batch size	16
Gradient accumulation steps	1 (effective batch size = 16)
Learning rate	3×10^{-5}
Learning rate scheduler	Linear decay (no warm-up)
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$)
Random seed	42
Training framework	PyTorch 2, Hugging Face Transformers
Training duration	Approximately 78 hours (CPU-only)

The model demonstrated high precision in detecting standard PII categories, such as email addresses, phone numbers, and medical record numbers. It exhibited signs of overfitting, however, suggesting that further optimization or larger compute resources would be required for production-grade deployment.

3.1.4. PIIRANHA Model with Region-Specific Enhancements

To address the limitations of the fine-tuned distilled BERT model, we adopted PIIRANHA, a model built on Microsoft's DeBERTa (Decoding-enhanced BERT with disentangled attention) [23]. PIIRANHA was further fine-tuned on a supplementary, region-specific synthetic dataset constructed to reflect Nigerian healthcare contexts, described below.

3.1.5. Region-Specific Dataset Construction

A supplementary dataset of 900 examples was generated programmatically using a template-based slot-

filling approach; no real patient data was collected or used. The templates represented common clinical scenarios, including hospital admissions, referral letters, and discharge summaries. These templates were populated with randomly sampled values from a curated list of 312 common Nigerian given names and 287 surnames, Nigerian telephone number formats (+234 XXXXXXXXXX), hospital medical record number formats following regional conventions (e.g., UBTH/YYYY/NNNN), and Nigerian street addresses and city names. Entity annotations were generated automatically from each template's slot positions, requiring no manual labelling. Generated entities were mapped to PIIRANHA's label taxonomy: patient names to GIVENNAME and SURNAME, hospital record numbers to ACCOUNTNUM, telephone numbers to TELEPHONENUM, and email addresses to EMAIL. The dataset was split into 720 training, 90 validation, and 90 test examples.

3.1.6. Fine-Tuning Configuration

PIIRANHA was fine-tuned on a combination of a sampled subset of the AI4Privacy training split and the region-specific dataset described above, under the configuration in Table 4.

Table 4. Custom PIIRANHA Fine-Tuning Configuration

Parameter	Value
Base model	iiiorg/piiranha-v1-detect-personal-information (DeBERTa-v2, ~0.3B parameters)
Tokenizer	DeBERTa-v2 SentencePiece tokenizer
Maximum sequence length	512 tokens
Number of epochs	3
Per-device batch size	1
Gradient accumulation steps	8 (effective batch size = 8)
Learning rate	2×10^{-5}
Learning rate scheduler	Linear decay with warm-up
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$)
Random seed	42
Training framework	PyTorch 2.13, Hugging Face Transformers
Training duration	Approximately 8 hours (CPU-only)

Both models were evaluated using standard classification metrics including precision, recall, and F1-score, on the withheld test set described in Table 1. These metrics were chosen due to their relevance in assessing both the accuracy and completeness of PII detection.

3.1.7. Request Transformation

After PII detection, the system applies a transformation stage that replaces sensitive entities with structured placeholders (e.g., replacing a patient's name with "[PATIENT]" or a medical record number with "[RECORD_ID]"). This substitution is performed at the token level to preserve the syntactic and semantic integrity of the request while ensuring that no identifiable information leaves the local environment.

To maintain consistency, identical entities within a

request are replaced with the same placeholder token. The transformed request is then reconstructed into its original JSON schema and forwarded to the target LLM endpoint. This design ensures that the LLM receives a valid, well-structured request without knowledge of the underlying sensitive attributes.

The response from the LLM is routed back through the system and delivered to the user without modification, creating the illusion of direct interaction while ensuring that all outbound communications remain free of personally identifiable information.

3.2. Model Evaluation and Deployment

Both models were evaluated on a held-out test set to determine their effectiveness in redacting personally identifiable information (PII). The evaluation process employed four key performance metrics commonly used in natural language processing tasks: Accuracy, Precision, Recall, and F1 Score. These metrics provided a balanced assessment of each model's ability to correctly identify and redact PII while minimizing false positives and false negatives.

To ensure fairness and reproducibility, the same test set was applied to both models. This enabled a direct comparison of their strengths and weaknesses in the healthcare-specific domain. Following evaluation, the deployment decision was guided by the superior performance of the Custom PIIRANHA model across accuracy, precision, recall, and F1 score. The deployment pipeline consisted of:

Containerization of the model into a lightweight runtime environment.

Serving the model as a RESTful API using FastAPI, selected for its ability to support low-latency, high-performance machine learning inference.

Integration of the API as an internal microservice within the request interception pipeline, enabling real-time PII redaction before prompts reach external large language models (LLMs).

4. RESULTS

4.1. Model Evaluation and Comparison

The Comprehensive Patient Personal Data Sovereignty System (C-PPDSS) was evaluated using two machine learning models: a fine-tuned Distilled BERT and a custom PIIRANHA model (based on Microsoft's DeBERTa). Both models were tested on the same held-out dataset to ensure fairness and reproducibility.

Table 5. Model Evaluation of the Fine-Tuned Distilled BERT and the Custom PIIRANHA

Metric	Fine-Tuned Distilled BERT	Custom PIIRANHA
Accuracy	94.9%	99.51%
Precision	90.85%	94.93%
Recall	91.36%	95.98%
F1 Score	92.60%	95.45%

The Custom PIIRANHA model outperformed the fine-tuned Distilled BERT across all major evaluation metrics, with a 4.61% increase in accuracy and a 2.85-point improvement in F1 score. These gains reflect a better balance between precision and recall, and motivated the choice of PIIRANHA as the model carried forward into deployment.

To understand where this improvement came from, we examined PIIRANHA's performance at the level of individual PII categories.

Table 6. Per-Category Token-Level Performance of the Custom PIIRANHA Model

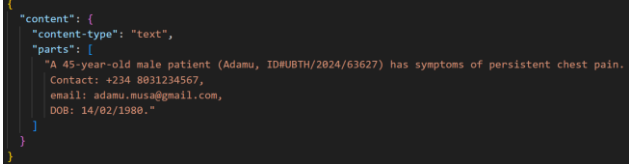
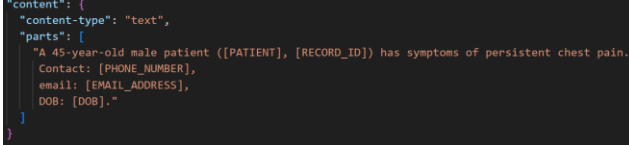
PII Category	Precision	Recall	F1-Score	Support (Tokens)
EMAIL	0.9960	1.0000	0.9980	1,245
SOCIALNUM	0.9920	1.0000	0.9960	374
TELEPHONENUM	1.0000	0.9898	0.9949	488
USERNAME	0.9912	0.9978	0.9945	449
PASSWORD	0.9885	1.0000	0.9942	1,118
STREET	0.9774	0.9774	0.9774	265
ACCOUNTNUM (MRN)	0.9468	0.9944	0.9700	179
ZIPCODE	0.9356	0.9561	0.9458	228
CITY	0.9447	0.9447	0.9447	217
GIVENNAME	0.9156	0.9632	0.9388	653
SURNAME	0.9050	0.9241	0.9144	237
CREDITCARDNUMBER	0.8714	0.9590	0.9131	537
BUILDINGNUM	0.8824	0.9000	0.8911	100
DATEOFBIRTH	0.7104	0.6068	0.6545	384
Micro-Average	0.9493	0.9598	0.9545	5,644
Macro-Average	0.9262	0.9442	0.9377	5,644
Accuracy (All Tokens)	—	—	0.9951	112,063

Structured alphanumeric PII categories including EMAIL, TELEPHONENUM, SOCIALNUM, PASSWORD, and USERNAME, achieved F1-scores of 0.994 or higher, reflecting the model's accuracy on entities with well-defined lexical patterns. ACCOUNTNUM, used

as a proxy for hospital medical record numbers, achieved an F1-score of 0.9700 with a recall of 0.9944, confirming reliable detection of healthcare-specific patient identifiers — the category most directly relevant to clinical deployment.

4.2. Real-Time Redaction Verification

Following this evaluation, the PIIRANHA model was deployed in the final system. It was containerized and served via a RESTful API using FastAPI, providing low-latency inference suitable for real-time request interception. Integrated as an internal microservice, the model successfully redacted sensitive information from outgoing HTTP POST requests before they reached external LLMs, confirming that the system could operate in real-time healthcare-specific use cases without introducing noticeable delays.


Figure 4. System log entry showing unredacted PII in an outbound request before C-PPDSS processing

Figure 5. System log entry showing redacted placeholders in an outbound request after C-PPDSS processing

4.3. System Performance Benchmarks

With redaction accuracy and functional correctness established, the deployed C-PPDSS microservice was evaluated for real-time operational suitability using 200 HTTP POST requests containing synthetic clinical prompts (mean prompt length: 154 characters; mean PII entities per prompt: 3.8), benchmarked on the hardware specification described in Table 2.

Table 7. C-PPDSS System Performance Benchmark Results

Metric	Measured Value
Model load time	2.50 s
Model RAM footprint	305.7 MB
Total process memory (runtime)	1,116.7 MB (~1.1 GB)
Mean inference latency	229.66 ms
Median latency (p50)	217.00 ms
95th percentile latency (p95)	325.69 ms
99th percentile latency (p99)	344.23 ms
Min / Max latency	153.63 ms / 368.76 ms
Throughput	4.35 requests/s
CPU utilization (mean)	196.6% (~2 of 4 cores)

The mean inference latency was 229.66 ms per request, with the 95th and 99th percentile latencies remaining below 350 ms. Given that typical LLM responses take between 2 and 10 seconds, the redaction process adds only a small fraction of the overall response time and is unlikely to be perceptible to end users.

Resource utilization stayed within the capabilities of the deployment hardware. The PIIRANHA model occupied 305.7 MB of memory after loading, while total runtime process memory averaged approximately 1.1 GB — well below the available 8 GB of system memory. Average CPU utilization was 196.6%, corresponding to roughly two fully utilized cores under psutil per-core accounting, indicating efficient multi-threaded execution without saturating the host system. Together, these measurements support low-latency operation and low resource consumption for the deployed C-PPDSS.

5. DISCUSSION

5.1. Interpretation of Model Performance

The evaluation of the two models — fine-tuned Distilled BERT and the custom PIIRANHA — points to clear differences in their suitability for healthcare-specific PII redaction. While Distilled BERT achieved a strong accuracy of 94.9% and an F1 score of 92.60%, PIIRANHA outperformed it, reaching an accuracy of 99.51% and an F1 score of 95.45%. The improved recall (95.98%) of the Custom PIIRANHA carries particular weight in this setting: healthcare applications demand a conservative approach, where missing a sensitive identifier can pose serious risks to patient privacy.

5.2. Answering the Research Questions

These findings map directly onto the three research questions posed at the outset of this study. First, the local HTTPS interception mechanism demonstrated that sensitive healthcare information can be identified and replaced before transmission to external LLM services. Second, the measured mean inference latency of 229.66 ms confirms that privacy-preserving redaction can be

performed in real time without introducing noticeable delays during clinical AI interactions. Third, by ensuring that all PII detection and substitution occur within the local deployment environment before requests are forwarded, the proposed architecture supports institutional data sovereignty while still allowing healthcare practitioners to benefit from external AI services.

5.3. Positioning Relative to Existing Work

Compared to existing systems like the original Patient Personal Data Sovereignty System (PPDSS) [6], which focused on static document sanitization, C-PPDSS introduces a paradigm shift by enforcing privacy at the network level. By intercepting outbound HTTP POST requests in real time, the system closes a gap that prior work left largely unaddressed: conversational prompts sent to large language model (LLM) APIs. This matters directly for interactive clinical workflows such as drafting medical notes, referral letters, or clinical summaries, where the sensitive content is generated on the fly rather than stored in a document

Table 8. Systematic Comparison of Privacy-Preservation Approaches for Healthcare AI

Approach	Example Work	Scope	Real-Time	Network-Level	LLM Usage Stage
Static document masking	Olaye et al. [6]	Documents	No	No	Pre-training only
NER-based redaction	C-PPDSS	Conversational	Yes	Yes	Usage/Inference
Federated learning	Anunita et al. [10]	Model training	No	No	Training only
Homomorphic encryption	Alabdu latif et al. [15]	IoT data	Partial	No	Pre-processing
Proxy re-encryption	Rodrigues et al. [14]	Health records	No	No	Storage/Access
Blockchain-based	Bhaskar et al. [17]	Data logging	No	No	Audit trail only

As Table 8 shows, existing privacy-preserving approaches concentrate on document sanitization, secure storage, encrypted computation, or model training. C-PPDSS instead operates at the inference stage, intercepting conversational requests before they leave the local environment. This lets sensitive information be removed at the network boundary without requiring any changes to the external LLM service itself.

5.4. Deployment Feasibility

The deployment architecture reinforces the practicality of this approach. The containerized FastAPI microservice combined low latency with minimal resource consumption, making the system viable on modest hardware or within hospital intranet environments.

Training and deployment were both shown to be feasible on a personal computer setup, without reliance on high-performance clusters or cloud-based infrastructure, which is a meaningful factor for reproducibility and for adoption in resource-constrained healthcare settings.

5.5. Error Analysis

Per-category evaluation (Table 6) shows the model performing consistently across most structured PII categories. EMAIL, TELEPHONENUM, PASSWORD, USERNAME, and SOCIALNUM all achieved F1-scores above 0.99, confirming reliable detection of identifiers with well-defined lexical patterns. ACCOUNTNUM, used as a proxy for hospital medical record numbers, likewise maintained a high recall of 0.9944 following region-specific fine-tuning.

The lowest performance was observed for DATEOFBIRTH ($F1 = 0.6545$). Most false negatives traced back to differences in date formatting: numeric, textual, and ISO representations, were all present in the evaluation data, and because these formats tokenize differently, detection accuracy varied across them. Incorporating date normalization before inference is expected to close much of this gap in future versions of the system.

GIVENNAME and SURNAME also scored lower than the structured identifiers, since name detection depends more on contextual understanding than on fixed lexical patterns. During deployment, this limitation was mitigated through word-level boundary expansion: when any subword token within a name is flagged as PII, the complete word is replaced, preventing partial disclosure of patient names.

5.6. Effect of Region-Specific Fine-Tuning

Including the supplementary Nigerian healthcare dataset improved the model's ability to recognize identifiers common to local clinical practice including hospital medical record number formats, Nigerian telephone prefixes, and common Nigerian personal names among them. The high recall obtained for ACCOUNTNUM suggests that adapting the model with region-specific examples makes it better suited to healthcare deployments where local identifier formats diverge from those represented in general-purpose datasets.

5.7. Multilingual Applicability

The base PIIRANHA model (iiiorg/piiranha-v1-detect-personal-information) supports six languages: English, Spanish, French, German, Italian, and Dutch. The current C-PPDSS evaluation is limited to English-language clinical prompts. Extension to multilingual healthcare settings such as Francophone African health systems, would require language-specific synthetic training data and evaluation sets, which is identified as future work.

5.8. Limitations

Several limitations remain. First, the current implementation is limited to text-based HTTP POST payloads and does not extend to multimodal data such as scanned medical documents, voice recordings, or medical images. Second, although both the AI4Privacy dataset and the supplementary Nigerian dataset are synthetically generated, real-world deployments may surface identifiers, abbreviations, or code-switching patterns that were not represented during training. Third, the system currently targets OpenAI-compatible HTTPS POST requests and does not yet support communication channels such as WebSocket or gRPC. Finally, the underlying DeBERTa architecture processes a maximum of 512 tokens per request, requiring long clinical narratives to be segmented before inference, a constraint that may increase processing time for lengthy documents.

Even with these limitations, the results suggest that lightweight, domain-specific models like PIIRANHA can play a meaningful role in enabling secure AI adoption in healthcare. By balancing accuracy with efficiency, the system offers a practical blueprint for integrating privacy-preserving safeguards into everyday clinical AI workflows without sacrificing usability.

6. CONCLUSION

This work presented the Comprehensive Patient Personal Data Sovereignty System (C-PPDSS), a privacy-preserving framework designed to safeguard sensitive information in healthcare interactions with large language models. By focusing on real-time interception of outbound HTTP POST requests and integrating lightweight yet accurate machine learning models, the system demonstrated a practical approach to enforcing data sovereignty at the point of data generation.

The evaluation of two candidate models highlighted the superiority of the custom PIIRANHA model, which achieved a 99.51% accuracy and an F1 score of 95.45%, outperforming the fine-tuned Distilled BERT baseline. Its favourable balance of recall and resource efficiency makes it particularly well-suited for healthcare contexts, where both privacy and responsiveness are critical.

Beyond its technical contributions, the system addresses a significant gap left by prior privacy-preservation efforts that concentrated on static documents while overlooking conversational AI use cases. By embedding privacy safeguards within the interaction pipeline, C-PPDSS demonstrates how healthcare institutions can confidently integrate AI tools into clinical workflows without compromising patient trust or regulatory compliance.

Future work will explore extending the system to multimodal data, adapting to encrypted communication protocols, and expanding the training corpus to include more diverse healthcare identifiers. These steps will

further strengthen its robustness and scalability.

In conclusion, C-PPDSS establishes a foundation for secure, real-time PII redaction in healthcare AI applications. Its combination of accuracy, efficiency, and architectural simplicity positions it as a viable solution for protecting sensitive data in an era of rapidly expanding AI adoption.

CRedit Authorship Contribution Statement

Edoghogho Olaye: Conceptualization, Methodology, Software. Williams O. Aigbe: Validation, Investigation, Writing. David Onwuka: Software, Formal Analysis, Writing. Daniel Obuh: Software, Data Curation. Amarachi A. Madukwe: Review & Editing.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Funding

This work was supported by the Tertiary Education Trust Fund National Research Fund grant (TETFund Ref: NRF2023/CC/STI/00031/VOL.1).

Ethical Standards

All data used in this study was either publicly available (AI4Privacy dataset, licensed under Creative Commons) or synthetically generated. No real patient data was collected or used. The study complies with all applicable ethical standards.

Data Availability

The AI4Privacy PII-Masking-200k dataset used in this study is publicly available through Hugging Face. The synthetic region-specific dataset and supporting materials generated for this study are available from the corresponding author upon reasonable request.

Acknowledgments

The authors thank TETFund and the Internet Integrity Initiative for the open-source release of the PIIRANHA model (iiiorg/piiranha-v1-detect-personal-information).

References

- [1] J. P. Albrecht, "How the GDPR Will Change the World," *Eur. Data Prot. Law Rev.*, vol. 2, no. 3, pp. 287–289, 2016, doi: 10.21552/EDPL/2016/3/4.
- [2] V. Rathod, S. Nabavirazavi, S. Zad, and S. S. Iyengar, "Privacy and Security Challenges in Large Language Models," in *2025 IEEE 15th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, Jan. 2025, pp. 00746–00752. doi: 10.1109/CCWC62904.2025.10903912.
- [3] G. Sperli, "Exploring Privacy and Security Risks in LLMs: Data Leakage, Prompt Injection, and Membership Inference," in *2025 International Symposium on Multimedia (ISM)*, Naples, Italy: IEEE, Dec. 2025, pp. 5–12. doi: 10.1109/ISM66958.2025.00011.
- [4] H. Alami, P. Lehoux, C. Papoutsis, S. E. Shaw, R. Fleet, and J.-P. Fortin, "Understanding the integration of artificial intelligence in healthcare organisations and systems through the NASSS framework: a qualitative study in a leading Canadian academic centre," *BMC Health Serv. Res.*, vol. 24, no. 1, p. 701, Jun. 2024, doi: 10.1186/s12913-024-11112-x.
- [5] P. Goktas and A. Grzybowski, "Shaping the Future of Healthcare: Ethical Clinical Challenges and Pathways to Trustworthy AI," *J. Clin. Med.*, vol. 14, no. 5, p. 1605, Feb. 2025, doi: 10.3390/jcm14051605.
- [6] E. Olaye, I. Omafovbe, W. O. Aigbe, and D. Obuh, "Personalized Governance Strategy for Patient Data in a Digital One Health Surveillance System," *Procedia Comput. Sci.*, vol. 254, pp. 20–29, 2025, doi: 10.1016/j.procs.2025.02.060.
- [7] H. Roberts, "Digital sovereignty and artificial intelligence: a normative approach," *Ethics Inf. Technol.*, vol. 26, no. 4, p. 70, Dec. 2024, doi: 10.1007/s10676-024-09810-5.
- [8] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC Med. Ethics*, vol. 22, no. 1, p. 122, Dec. 2021, doi: 10.1186/s12910-021-00687-3.
- [9] S. M. Williamson and V. Prybutok, "Balancing Privacy and Progress: A Review of Privacy Challenges, Systemic Oversight, and Patient Perceptions in AI-Driven Healthcare," *Appl. Sci.*, vol. 14, no. 2, p. 675, Jan. 2024, doi: 10.3390/app14020675.
- [10] K. P. Anunita, S. R. Devisri, and C. Arumugam, "Federated Learning: Countering Label Flipping Attacks in Retinal OCT," *Procedia Comput. Sci.*, vol. 254, pp. 58–67, 2025, doi: 10.1016/j.procs.2025.02.064.
- [11] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, "Ethical and regulatory challenges of AI technologies in healthcare: A narrative review," *Heliyon*, vol. 10, no. 4, p. e26297, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26297.
- [12] P. L. Chong *et al.*, "Integrating artificial intelligence in healthcare: applications, challenges, and future directions," *Future Sci. OA*, vol. 11, no. 1, p. 2527505, Dec. 2025, doi: 10.1080/20565623.2025.2527505.
- [13] S. Gawankar, S. Nair, V. Pawar, A. Vhatkar, and P. Chavan, "Patient Privacy and Data Security in the Era of AI-Driven Healthcare," in *2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India: IEEE, Aug. 2024, pp. 1–6. doi: 10.1109/ICCUBEA61740.2024.10775004.
- [14] B. Rodrigues, I. Amorim, I. Costa, and A. Mendes, "Patient-centric health data sovereignty: an approach using Proxy re-encryption," 2023, *arXiv*. doi: 10.48550/ARXIV.2307.01175.
- [15] A. Alabdulatif, I. Khalil, X. Yi, and M. Guizani, "Secure Edge of Things for Smart Healthcare Surveillance Framework," *IEEE Access*, vol. 7, pp. 31010–31021, 2019, doi: 10.1109/ACCESS.2019.2899323.
- [16] S. Chandra, S. Ray, and R. T. Goswami, "Big Data Security in Healthcare: Survey on Frameworks and Algorithms," in *2017 IEEE 7th International Advance Computing Conference (IACC)*, Hyderabad, India: IEEE, Jan. 2017, pp. 89–94. doi: 10.1109/IACC.2017.0033.
- [17] T. Bhaskar, M.N. Narsaiah, and M. Ravikanth, "Central Medical Centre Healthcare Data Security with Lightweight Blockchain Model in IoT Sensor Environment," *JSIHS*, vol. 1, no. 1, pp. 15–26, Dec. 2023, doi: 10.69996/jsihs.2023002.
- [18] G. P. Tobia *et al.*, "Privacy in Italian Clinical Reports: A NLP-Based Anonymization Approach," in *2025 IEEE 13th International Conference on Healthcare Informatics (ICHI)*, Rende, Italy: IEEE, Jun. 2025, pp. 630–635. doi: 10.1109/ICHI64645.2025.00077.
- [19] L. Mainetti and A. Elia, "Detecting Personally Identifiable Information Through Natural Language Processing: A Step Forward," *Appl. Syst. Innov.*, vol. 8, no. 2, p. 55, Apr. 2025, doi: 10.3390/asi8020055.
- [20] L. R. Kumar, C. Ashokkumar, P. S. Pandey, S. K. Kannaiah, B. J., and M. I. T. Hussan, "Security Enhancement in Surveillance Cloud Using Machine Learning Techniques," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 3s, pp. 46–55, Mar. 2023, doi: 10.17762/ijrtcc.v11i3s.6154.
- [21] M. Nankya, A. Mugisa, Y. Usman, A. Upadhyay, and R. Chataut, "Security and Privacy in E-Health Systems: A Review of AI and Machine Learning Techniques," *IEEE Access*, vol. 12, pp. 148796–148816, 2024, doi: 10.1109/ACCESS.2024.148796.

- 10.1109/ACCESS.2024.3469215.
- [22] N. V. D. P. Raju, "Design and Implementation of Artificial Intelligence and Machine Learning Techniques for Security of Healthcare Systems," in *2024 Asian Conference on Intelligent Technologies (ACOIT)*, KOLAR, India: IEEE, Sep. 2024, pp. 1–7. doi: 10.1109/ACOIT62457.2024.10941611.
- [23] J. Ganesh and A. Bansal, "Transformer-based Automatic Mapping of Clinical Notes to Specific Clinical Concepts," in *2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC)*, Torino, Italy: IEEE, Jun. 2023, pp. 558–563. doi: 10.1109/COMPSAC57700.2023.00080.