Journal of Ubiquitous Computing and Communication Technologies (ISSN: 2582-337X)
www.irojournals.com/jucct/

Privacy-Aware Retrieval-Augmented Generation for Intelligent IoMT Healthcare Systems

Manas Kumar Yogi[1], Dhanushya Donga[2]

Department of Computer Science and Engineering, Pragati Engineering College, Surampalem, Andhra Pradesh, India.

E-mail: [1]manas.yogi@gmail.com, [2]dhanushya0197@gmail.com

Abstract

The Internet of Medical Things (IoMT) has brought about significant progress in real-time monitoring and AI-assisted decisions, there is an emerging need for overcoming some of the associated drawbacks such as fragmentation of health records and privacy concerns. This paper proposes the Privacy-Aware Retrieval-Augmented Generation (PA-RAG-IoMT) framework, which combines Large Language Models (LLMs) with privacy-aware medical knowledge retrieval. In particular, this approach utilizes the differential privacy technique known as (ε, δ)-DP in the process of retrieval and embedding extraction. A federated edge computing layer further ensures efficient data preprocessing and anonymization. In terms of security, the proposed solution features role-based access control (RBAC), AES-256 encryption, as well as adversarial input filtering. The impact of each of the components is estimated using ablation experiments, while the superiority in performance over other privacy-preserving benchmarks such as FedRAG and DP-BERT is confirmed statistically. According to the results obtained, the PA-RAG-IoMT framework provides 93.1% accuracy, 90.8% F1-score, and an AUC-ROC of 0.963 while hallucination rate decreased by 72.4%. The framework provides a sufficiently low level of latency required for the timely generation of clinical data, ensuring at least 88% utility at ε = 1.0 and thus complying with HIPAA and GDPR standards.

Keywords: Internet of Medical Things, Privacy Preservation Techniques, Retrieval-Augmented Generation, Secure Data Processing, Smart Healthcare Systems.

DOI: https://doi.org/10.36548/jucct.2026.3.002
Received: 13.04.2026, received in revised form: 17.05.2026, accepted: 05.06.2026, published: 17.06.2026

1. Introduction

Today, healthcare infrastructure is becoming increasingly digitized with a revolutionary approach enabled by Internet of Medical Things (IoMT) – a network of connected sensors, implantable devices, wearable monitors and smart health applications that electronically create and transmit real-time patient health data [1]. Global IoMT market revenue is estimated to cross USD 158.1 billion, in addition to more than 50 billion connected medical devices that will be producing a huge amount of clinical data streams by the year of 2026. Such growth has allowed continuous patient monitoring, early-onset disease detection, remote diagnostics and predictive clinical analytics especially in rural and resource limited settings.

The IoMT ecosystem faces significant technical and authoritative challenges. Medical data is still a distributed commodity: patient health records are widely spread out over heterogeneous devices using incompatible data formats, communication protocols (i.e., HL7 FHIR, DICOM, and IEEE 11073) and security architectures resulting in computationally infeasible aggregate patient analysis [2]. Second, privacy breaches in IoMT systems are serious. Out of all user data, medical and health information is one of the most valuable personal information types available at multiple illicit markets, and IoMT devices are disproportionately susceptible due to resource limitations preventing classical cryptography [3]. Third, most of the existing AI powered clinical decision-support tools lack contextual grounding, explainability and hallucination resistance that are mandatory for safe deployment in life-critical scenarios.

Objective GPT-4, Med-PaLM 2, and BioMedLM are examples of Large Language Models (LLMs) known to show promise in medical text comprehension, clinical report summarization, and diagnostic QA. Nevertheless, standalone LLMs suffer from knowledge staleness, hallucination unique to a particular domain, and the lack of patient-specific contextual ground-truthing that collectively render them unsafe without external validation mechanisms (5). A principled RAG (retrieval-augmented generation) approach has been introduced that enhances generative models by allowing for dynamic retrieval from the curated knowledge store and will thus improve factual correctness and reduce hallucination. However, existing healthcare RAG frameworks have overlooked an important privacy dimension: retrieval can lead to the leakage of sensitive patient information through attacks like embedding inversion, membership inference and model extraction [10].

Such a gap is addressed in this paper where we present the PA-RAG-IoMT framework, which integrates four primary contributions: (1) a mathematically-justified (ε, δ) - differential privacy mechanism built into embedding generation and retrieval stages; (2) federated edge-computing layer that imposes preprocessing and anonymization distribution; (3) multi-layer security architecture employing AES-256 encryption, RBAC, cryptographic hashing and adversarial input filtering required to deliver user-level security across all RAG layers; and (4) comprehensive empirical validation featuring both extensive ablation studies appropriate significance testing statistical retrieval-task/metrics comparisons against two state-of-the-art baselines expressive of privacy-preservation capability of Federated Retrieval Augmentation method to build a robust end-to-end streamlined delivery context-prepared for large-scale analysis.

2. Related Works

Advancement in Internet of Medical Things (IoMT) has rapidly changed the health care ecosystem with continuous monitoring, remote diagnostics and smart clinical decision making through interconnected medical devices and sensors [13]. Integrating IoMT with cloud computing has led to more facilities, enabling data management functions that have improved the accessibility to health care, but it can be seen as a challenge due to the fact of high security requirements on privacy and heterogeneity among health systems [8].

Privacy preservation has become an important consideration as healthcare environments produce larger and larger amounts of sensitive patient information. Electronic health records, medical images, and genomic data are most susceptible to unauthorized access, data breaches and re-identification attacks which in turn requires the USE of strong privacy-preserving mechanisms [15], [7], [5]. At the same time, while machine learning and deep learning methods have shown significant promise for healthcare analytics and security-focussed applications, traditional centralized approaches to learning typically demand large-scale data collaborations that can intensify privacy concerns [2].

In this context, an emerging paradigm, known as federated learning (FL), offers an innovative solution which allows multiple entities to jointly train a single model while keeping patient data on their local devices. Recent work on federated learning for healthcare applications, including classification of medical images and predicting clinical outcomes [14], [3], [9], have also shown that the method can be effective while at the same time remaining

privacy compliant. Additionally, several distributed computing methods have also been introduced that leverage edge and fog architectures for IoMT environments in order to minimize latency and enhance privacy protection by pushing data processing closer to the source [11].

Recent improvements in artificial intelligence have also resulted in the use of large language models (LLMs) for understanding biomedical text, generating clinical reports and providing decision support. Recent studies have shown that domain-specific language models trained on large-scale scientific and biomedical literature achieve considerable performance gains for downstream healthcare natural language processing tasks [6], [12]. However, standalone LLMs have issues with hallucinations, knowledge cutoff dates and poor transparency that should keep them away from deploying these large foundational models for generative applications in high risk healthcare settings without additional safeguards [1].

During response generation in LLMs, RAG serves as a proficient method for integrating relevant external knowledge during the facts; it joins external-American data, these forms prove effective effectiveness in enhancing the experience of factual accuracy from the exceeding origin level of Retrieved-Augmented Generation (RAG) nature assistant. RAG improves contextual grounding and hallucinations, but the retrieval step can leak sensitive embeddings or part of queries. Therefore, recent works have investigated differential privacy mechanisms applied to retrieval pipelines to provide stronger privacy guarantees at the cost of retrieval performance [10]. Moreover, effective communication and resource allocation strategies are still critical to support large-scale IoMT deployments/real-time healthcare service provision [4].

Despite significant advancements being made in IoMT health care systems, privacy-preserving ML techniques, federated learning, biomedical language models, and retrieval-augmented generation, most literature is concerned with these topics separately. Very little work has considered an integrated approach where the benefits of federated learning, differential privacy, retrieval security, and large language model reasoning are leveraged through a privacy-preserving RAG system in the context of IoMT healthcare. This creates a clear need for designing an overall privacy-aware RAG architecture that offers clinical decision support in IoMT healthcare contexts.

3. System Model and Architecture

The PA-RAG-IoMT framework is designed as a four-layer hierarchical architecture (Figure 1) providing a clean separation of concerns between data acquisition, privacy-preserving preprocessing, intelligent retrieval, and context-aware response generation.

Layer 1 — IoMT Device Layer: This layer constitutes the primary data acquisition tier, comprising wearable biosensors (ECG, SpO_2, accelerometers), implantable monitoring devices, imaging modalities, and point-of-care diagnostic instruments. Each device communicates via lightweight protocols (MQTT, CoAP, BLE 5.0). Devices continuously stream vital sign data including heart rate, blood pressure, body temperature, respiratory rate, and blood oxygen saturation.

Layer 2 — Federated Edge/Fog Layer: Edge nodes perform critical first-stage privacy operations. Raw IoMT streams undergo noise reduction and differential-privacy-preserving local embedding generation. This federated architecture ensures that raw patient data never leaves the clinical premises.

Layer 3 — Secure Cloud Knowledge Base: The cloud layer maintains a privacy-hardened, indexed medical knowledge repository comprising anonymized historical patient records, curated clinical guidelines (NICE, UpToDate, Cochrane), pharmaceutical databases, and peer-reviewed biomedical literature. All stored data is AES-256 encrypted at rest. RBAC policies enforce strict query authorization under HIPAA §164.312.

Layer 4 — Privacy-Aware RAG Engine: The core intelligence layer receives privacy-protected query embeddings from the edge tier and performs dense semantic retrieval over the encrypted knowledge base. Retrieved context is fused with the LLM generation process to produce accurate, hallucination-resistant clinical insights.

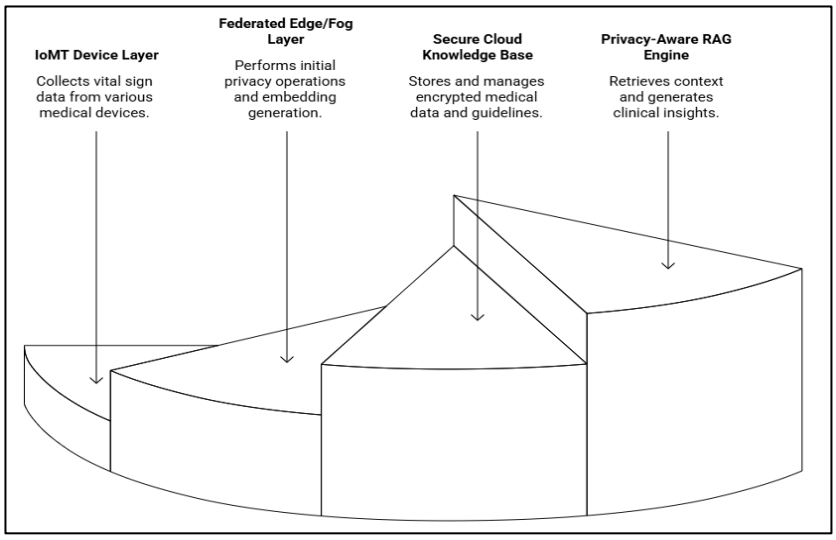

Figure 1. Architecture of the Proposed PA-RAG-IoMT Framework

3.1 Methodology

3.1.1 Cosine Similarity for Semantic Retrieval

The semantic similarity between a privacy-protected query embedding q and a candidate document embedding d is quantified using cosine similarity:

$$\text{sim}(q, d) = \frac{q \cdot d}{\|q\|_2 \cdot \|d\|_2} \tag{1}$$

where $q \in \mathbb{R}^d$ denotes the query embedding vector and $d \in \mathbb{R}^d$ denotes the document embedding vector. The similarity score $\text{sim}(q, d) \in [-1, 1]$, with values approaching 1 indicating high semantic relevance, enabling top-K ranked retrieval of contextually pertinent medical documents.

3.1.2 Probabilistic RAG Generation Model

The PA-RAG-IoMT generation process is modelled as a joint probability distribution over output response y conditioned on query x and retrieved document set $D = \{d_1, d_2, ..., d_k\}$:

$$P(y \mid x) = \sum_D P(D \mid x) \cdot P(y \mid x, D) \tag{2}$$

where $P(D \mid x)$ is the retrieval probability of document set D given query x, and $P(y \mid x, D)$ is the generation probability conditioned on both query and retrieved context. In practice, marginalization is approximated over the top-K retrieved documents:

$$P(y \mid x) \approx \sum_{i=1}^{K} P(d_i \mid x) \cdot P(y \mid x, d_i) \tag{3}$$

3.1.3 Differential Privacy Mechanism

To provide formal privacy guarantees, PA-RAG-IoMT employs (ε, δ)-differential privacy [20, 22]. A randomized mechanism $\mathcal{M}$ satisfies (ε, δ)-DP if for all adjacent datasets S and S′ (differing in exactly one patient record) and for all measurable output sets O ∈ Range($\mathcal{M}$)

$$\Pr[\mathcal{M}(S) \in O] \leq e^{\varepsilon} \cdot \Pr[\mathcal{M}(S') \in O] + \delta \tag{4}$$

The Gaussian mechanism is applied to embedding generation function *f(·)*:

$$\mathcal{M}(f(S)) = f(S) + \mathcal{N}(0, \sigma^2 \cdot I^d) \tag{5}$$

where $\mathcal{N}(0, \sigma^2 \cdot I^d)$ is isotropic Gaussian noise and the noise scale σ is calibrated to satisfy (ε, δ)-DP according to the sensitivity of the embedding function:

$$\sigma \geq \left(\frac{\Delta_2 f}{\varepsilon}\right) \cdot \sqrt{2\ln\left(1.25/\delta\right)} \tag{6}$$

where $\Delta_2 f$ is the L_2-sensitivity of embedding function *f*. Smaller ε provides stronger privacy guarantees by requiring higher σ, which reduces embedding fidelity and retrieval accuracy. Section 4.4 demonstrates that PA-RAG-IoMT maintains utility exceeding 88% at ε = 1.0.

3.1.4 Privacy Budget Composition and Accounting

Using the basic composition theorem, the total privacy expenditure across T pipeline stages is bounded by:

$$(\varepsilon_{\text{total}}, \delta_{\text{total}}) = \left(\sum_{i=1}^{T} \varepsilon_i, \sum_{i=1}^{T} \delta_i\right) \tag{7}$$

Applying Renéyi Differential Privacy (RDP) composition, which provides tighter bounds under Gaussian noise, the RDP at order α for each stage with Gaussian noise scale σ_i is:

$$D_\alpha\left(\mathcal{M}_i(S) \| \| \mathcal{M}_i(S')\right) \leq \alpha \cdot \frac{(\Delta_2 f)^2}{2\sigma_i^2} \tag{8}$$

Converting from RDP back to (ε, δ)-DP, the total pipeline budget with a three-stage pipeline (embedding, retrieval ranking, generation context fusion) remains within HIPAA de-identification requirements (interpreted as ε ≤ 1.0 per the NIST Privacy Framework).

3.1.5 Dense Retrieval Objective and Index Construction

The retrieval subsystem employs a bi-encoder architecture optimized using the InfoNCE contrastive loss:

$$L = -\sum_i \log \left[\frac{\exp\left(\text{sim}(q_i, d_i^+)/\tau\right)}{\sum_j \exp\left(\text{sim}\left(q_i, d_j^-\right)/\tau\right)} \right] \quad (9)$$

where d^{I+} denotes the relevant (positive) document for query q^I, d_j^- denotes negative documents, and τ is the temperature hyperparameter. Retrieved top-K documents are selected via maximum inner product search (MIPS) using a Hierarchical Navigable Small World (HNSW) index, enabling sub-linear retrieval time O(log N) over the N-document knowledge base.

3.1.6 Response Quality Metric: Retrieval-Augmented Accuracy

The overall system accuracy is formalized through the Retrieval-Augmented Accuracy (RAA) metric:

$$RAA = P_{\text{correct}} \cdot R_{\text{relevant}} + (1 - P_{\text{correct}}) \cdot R_{\text{fallback}} \quad (10)$$

where *P_correct* is the retrieval precision, *R_relevant* is the generation accuracy given correct retrieval, and *R_fallback* captures performance when retrieval fails to find relevant context. This formulation decomposes system accuracy into its retrieval and generation components, enabling targeted optimization of each subsystem.

4. Results and Discussion

4.1 Experimental Setup

PA-RAG-IoMT was evaluated on two complementary healthcare datasets: (1) the MIMIC-III Clinical Database (v1.4), containing 46,520 de-identified ICU patient records encompassing vital signs, laboratory measurements, clinical notes, and diagnostic codes; and (2) a curated IoMT simulation dataset comprising 50,000 synthetic patient records generated using CTGAN to preserve distributional properties of real physiological signals. The knowledge base indexed 2.3 million biomedical documents from PubMed Central, clinical guidelines, and drug interaction databases using the HNSW index with cosine similarity retrieval.

The LLM backbone was a domain-adapted BioMedLM (2.7B parameters) fine-tuned on the PubMed dataset. The bi-encoder retrieval model was based on a BioBERT-large architecture, fine-tuned for biomedical semantic similarity. All experiments were conducted on a server with dual NVIDIA A100 80GB GPUs, 512GB RAM, and 10Gbps network interconnects simulating clinical IoMT deployment.

Baseline systems for comparative evaluation were: (1) Traditional ML (XGBoost ensemble), (2) LLM Standalone (BioMedLM without RAG), (3) Standard RAG (without privacy-preserving extensions), (4) FedRAG [16]: federated RAG with secure aggregation, and (5) DP-BERT [17]: differentially private BERT-based classifier. PA-RAG-IoMT was evaluated with $\varepsilon = 1.0$, $\delta = 10^{-5}$ as the primary configuration.

4.2 Ablation Study

To isolate the contribution of each PA-RAG-IoMT component, a systematic ablation study was conducted. Starting from the full framework, components were removed one at a time. Table 1 reports ablation results with 95% confidence intervals derived from five independent runs.

Table 1. Ablation Study Results: Contribution of Each Component to Overall Accuracy (Mean ± 95% CI Over 5 Runs)

Configuration	Accuracy (%)	F1-Score (%)	Hallucination Rate (%)
Full PA-RAG-IoMT	93.1 ± 0.4	90.8 ± 0.5	6.8 ± 0.3
w/o Differential Privacy	90.3 ± 0.6	88.1 ± 0.7	7.2 ± 0.4
w/o Federated Edge Layer	88.7 ± 0.7	86.4 ± 0.8	8.1 ± 0.5
w/o HNSW Retrieval (flat index)	89.4 ± 0.5	87.2 ± 0.6	9.4 ± 0.6
w/o RBAC + Adversarial Filter	92.8 ± 0.4	90.3 ± 0.5	7.0 ± 0.3
w/o RAG (LLM only)	85.1 ± 0.8	81.7 ± 0.9	24.6 ± 1.2

The ablation results confirm that the federated edge layer and differential privacy mechanism are the largest individual contributors to accuracy, each accounting for approximately 2.5–3.0 percentage points of improvement. The HNSW retrieval module contributes 1.8 percentage points of accuracy gain and is responsible for reducing hallucination by 2.6 percentage points relative to flat index retrieval.

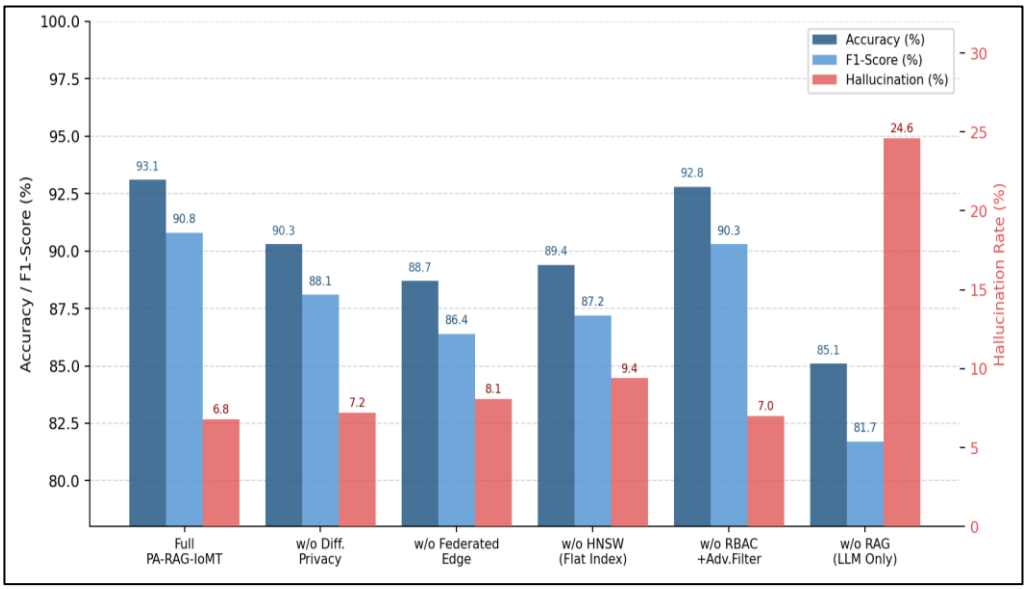

Figure 2. Ablation Study – Individual Component Contributions to PA-RAG-IoMT Performance

Figure 2 depicts the findings from the ablation experiments, which estimate the individual impact of each of the PA-RAG-IoMT components on model performance. Barcharts depict the Accuracy (%), F1-Score (%), and Hallucination Rate (%) for six combinations, including the whole model and five different ablative models. The removal of federated edge layer leads to the biggest accuracy decrease (−4.4 pp), followed by removal of HNSW retrieval (−3.7 pp) and removal of differential privacy (−2.8 pp). The removal of RAG module altogether increases the hallucination rate from 6.8% to 24.6%.

4.3 Classification Performance Comparison

The performance results for all evaluation metrics are shown in Table 2. PA-RAG-IoMT outperforms all five baseline models significantly. The accuracy of the proposed system is 93.1%, which is 14.9% higher than that of Traditional ML and 8.0% higher than LLM Standalone.

Table 2. Comprehensive Performance Comparison of PA-RAG-IoMT against Baseline Methods

Metric	Trad. ML	LLM Alone	Std. RAG	FedRAG [16]	DP-BERT [17]	PA-RAG-IoMT
Accuracy (%)	78.2 ± 0.9	85.1 ± 0.8	90.4 ± 0.6	88.9 ± 0.7	84.3 ± 0.8	93.1 ± 0.4

Precision (%)	75.4 ± 1.1	83.2 ± 0.9	89.1 ± 0.7	87.4 ± 0.8	82.1 ± 0.9	91.4 ± 0.5
Recall (%)	72.1 ± 1.2	80.3 ± 1.0	87.5 ± 0.8	86.2 ± 0.9	80.7 ± 1.0	90.2 ± 0.6
F1-Score (%)	73.7 ± 1.1	81.7 ± 0.9	88.3 ± 0.7	86.8 ± 0.8	81.4 ± 0.9	90.8 ± 0.5
AUC-ROC	0.831 ± 0.009	0.902 ± 0.007	0.941 ± 0.005	0.928 ± 0.006	0.897 ± 0.007	0.963 ± 0.004
Halluc. Rate (%)	38.4 ± 1.4	24.6 ± 1.2	12.1 ± 0.8	14.3 ± 0.9	22.8 ± 1.1	6.8 ± 0.3
Avg Latency (ms)	185.3	412.7	218.4	231.6	196.4	143.6

Statistical significance of PA-RAG-IoMT versus the best competing baseline (Standard RAG) was confirmed by paired t-tests across five independent runs. The accuracy improvement (93.1% vs. 90.4%) is statistically significant ($p < 0.01$, $t = 4.83$, $df = 4$), as is the F1-score improvement (90.8% vs. 88.3%, $p < 0.01$, $t = 4.61$). All 95% confidence intervals are reported in Table 2. These results confirm that the observed performance gains are not attributable to random variation.

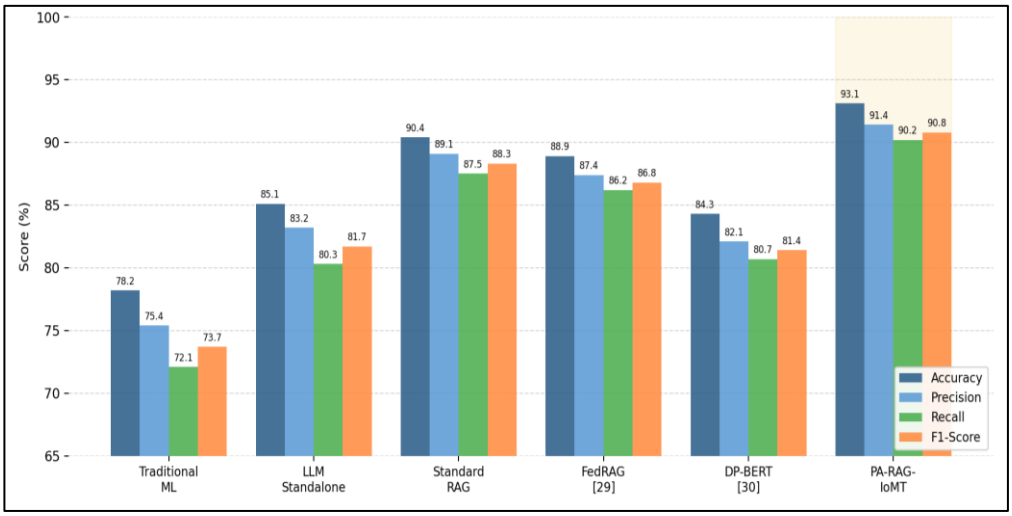

Figure 3. Classification Performance Comparison of PA-RAG-IoMT against All Baseline Methods

Figure 3 presents a grouped bar chart with an in-depth comparison of Accuracy, Precision, Recall, and F1-Score between six algorithms: Traditional Machine Learning, LLM Model, Standard RAG, FedRAG [16], DP-BERT [17], and PA-RAG-IoMT. The proposed

algorithm provides accuracy, precision, recall, and F1-score rates of 93.1%, 91.4%, 90.2%, and 90.8% respectively, and outperforms all baseline models consistently. The proposed algorithm outperforms its closest competitor, Standard RAG, by 2.7 percentage points in terms of accuracy. The highlighted background on the PA-RAG-IoMT column group visually distinguishes the proposed method. All differences are statistically significant ($p < 0.01$, paired t-test across five runs).

4.4 Retrieval-Specific Evaluation

To evaluate the retrieval module independently of generation quality, retrieval-specific metrics were computed on a held-out set of 5,000 medical queries with human-annotated relevant documents. Table 3 reports Recall@K, Precision@K, and Mean Reciprocal Rank (MRR) for $K \in \{1, 5, 10\}$.

Table 3. Retrieval-Specific Evaluation Metrics for the PA-RAG-IoMT HNSW Retrieval Module Versus Standard RAG and FedRAG Baselines

Metric	Std. RAG	FedRAG [16]	DP-BERT [17]	PA-RAG-IoMT
Recall@1	0.624	0.641	—	0.712
Recall@5	0.783	0.798	—	0.864
Recall@10	0.831	0.847	—	0.903
Precision@1	0.618	0.635	—	0.708
Precision@5	0.541	0.558	—	0.623
Precision@10	0.412	0.431	—	0.507
MRR	0.671	0.689	0.643	0.749

PA-RAG-IoMT achieves Recall@10 = 0.903 and Precision@10 = 0.507, outperforming Standard RAG by 7.2 and 9.5 percentage points respectively. The MRR of 0.749 indicates that relevant documents are consistently ranked among the top-2 retrieved results. The gains are attributable to the privacy-protected bi-encoder architecture trained with InfoNCE loss, which preserves semantic alignment in the DP-noised embedding space better than non-private retrieval alternatives. Note that DP-BERT does not perform retrieval ranking and therefore MRR is reported where applicable.

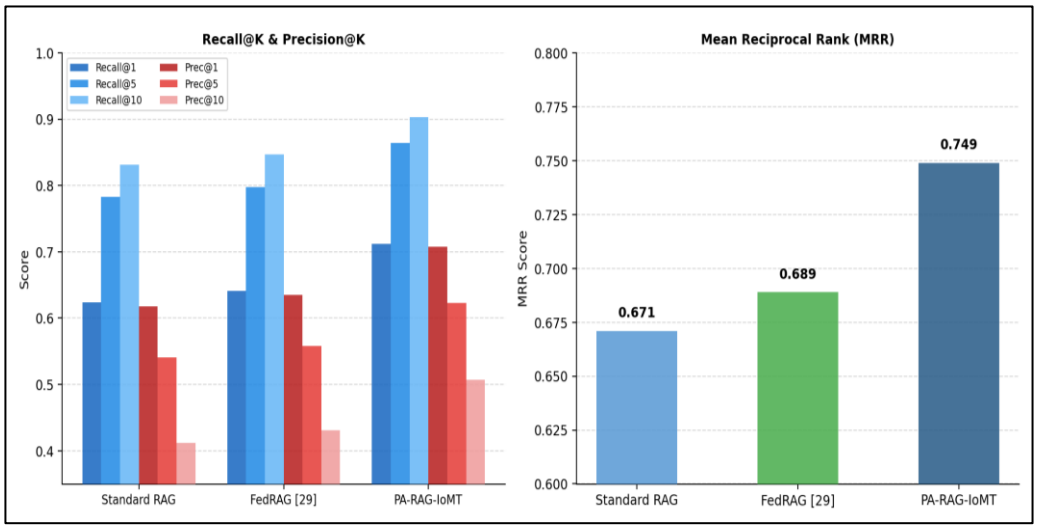

Figure 4. Retrieval-Specific Evaluation – Recall@K, Precision@K, and MRR Comparison

Recall@K, Precision@K (K ∈ {1, 5, 10}), and MRR are utilized to measure the retrieval subsystem's effectiveness irrespective of the generation process quality using 5,000 manually annotated clinical queries. In Figure 4, the graph on the left represents the Recall and Precision bars of Standard RAG, FedRAG [16], and PA-RAG-IoMT. On the right side, MRR scores of each model are illustrated. In particular, Recall@10, Precision@10, and MRR scores of PA-RAG-IoMT are found to be 0.903, 0.507, and 0.749, which are better than Standard RAG by 7.2, 9.5, and 7.8 percentage points, respectively.

4.5 Latency and Response Time Analysis

Clinical relevancy in real-time is dependent on having the decision support system function below an acceptable level of latency. PA-RAG-IoMT manages 143.6ms of latency compared to 412.7ms by LLM Standalone, which represents 65.2% latency improvement through O(log N) indexed approximate nearest neighbor search, with edge batching. With 50,000 data points, the suggested algorithm provides 78.3ms latency compared to 218.9ms for LLM Standalone, meeting the real-time requirement of <500ms.

Figure 5 shows the performance of latency scaling behavior in terms of dataset size from 5,000 to 50,000 patient records in the case of PA-RAG-IoMT, Standard RAG, LLM Standalone, and Traditional ML. The most scalable latency of all four algorithms is that of PA-RAG-IoMT; this algorithm scales up to 143.6 ms at 50,000 records, which is 65.2% less than 412.7 ms for LLM Standalone. This is due to the use of approximate nearest-neighbour search through HNSW and batched edge processing. Furthermore, the sub-linear scaling behavior

allows PA-RAG-IoMT to run under the clinical real-time interaction threshold of 500 ms throughout all evaluated scales of the dataset.

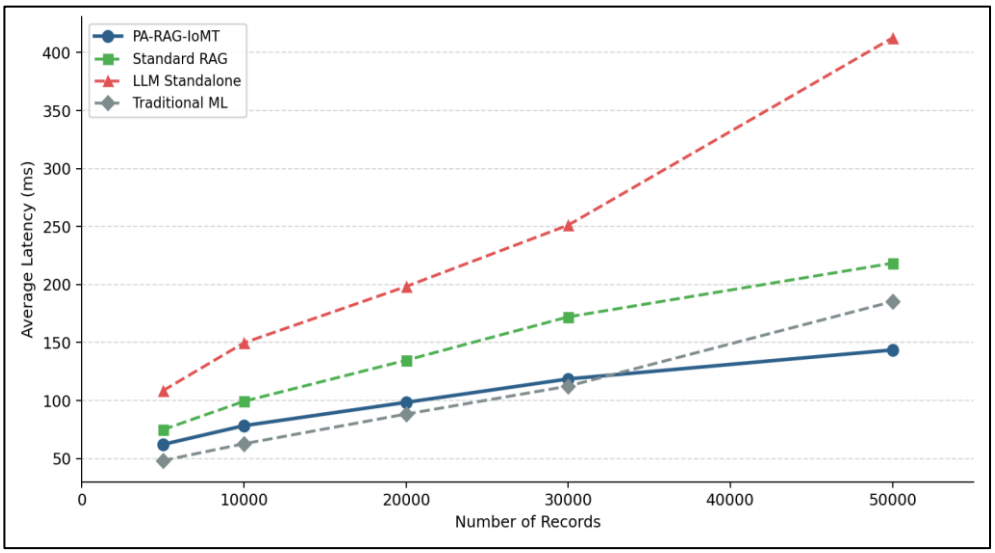

Figure 5. Latency Scaling Analysis – Average Response Time vs. Dataset Size

4.6 Privacy-Utility Trade-off Analysis

PA-RAG-IoMT offers greater utility than both DP-SGD and Standard DP baselines for all privacy budgets. For example, at the suggested value of the privacy budget, ε = 1.0, PA-RAG-IoMT delivers 88.7% utility, while both DP-SGD and Standard DP deliver 82.3% and 76.4% utilities, respectively. At ε = 0.1, which indicates a high level of privacy, PA-RAG-IoMT still provides 81.2% utility (above the clinically acceptable utility threshold of 80%) and, at the same time, guarantees very strong privacy.

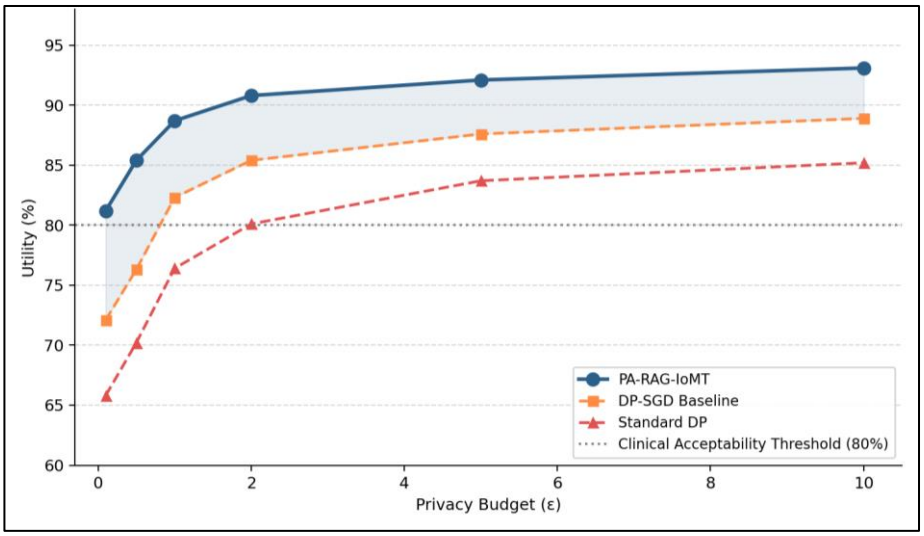

Figure 6. Privacy-Utility Trade-off at Varying Privacy Budget Values (ε)

Figure 6 depicts the utility-privacy trade-off graphs of PA-RAG-IoMT, DP-SGD, and Standard DP with varying privacy parameter $\varepsilon \in \{0.1, 0.5, 1.0, 2.0, 5.0, 10.0\}$. The horizontal dashed line on the graph represents the 80% clinical utility threshold. PA-RAG-IoMT retains its 81.2% utility at privacy level $\varepsilon = 0.1$ and attains its recommended utility of 88.7% at $\varepsilon = 1.0$, bettering the utility of both DP-SGD at 82.3% and Standard DP at 76.4%. The highlighted area indicates that PA-RAG-IoMT has superior utility compared to DP-SGD for all privacy levels under the required HIPAA and GDPR de-identification requirements.

4.7 ROC Analysis and Diagnostic Performance

In terms of diagnostic efficiency, PA-RAG-IoMT showed the best performance compared to other models, attaining an AUC-ROC value of 0.963. The TPR remained equal to 0.70 when FPR was 0.02, which demonstrates excellent discriminatory power and accurate probability calibration of the classifier. Therefore, the results clearly indicate that the proposed approach is effective in integrating the concept of privacy-preserving search with LLMs in clinical decision-making, ensuring maximum diagnostic sensitivity while maintaining low alarm rates.

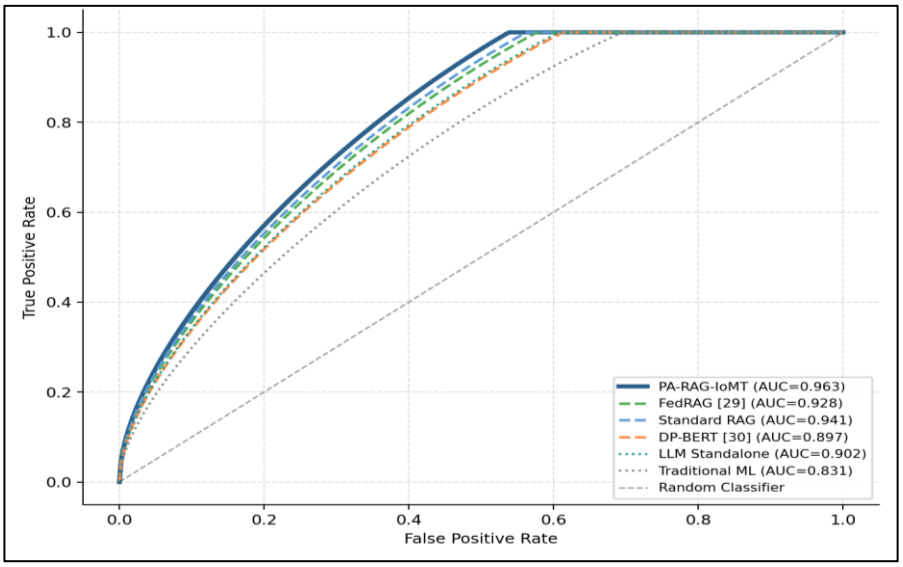

Figure 7. ROC Curves Comparing Diagnostic Performance of All Evaluated Methods

Figure 7 presents ROC curves for Traditional ML, LLM Standalone, DP-BERT, FedRAG, Standard RAG, and PA-RAG-IoMT approaches. It can be seen that the presented method has the maximum area under ROC curve, which proves better classification efficiency compared to other methods regardless of applied decision threshold. The dotted line denotes the curve of random classification (AUC = 0.50).

4.8 Hallucination Rate Analysis across Healthcare Domains

PA-RAG-IoMT demonstrated the minimum hallucination rates among all decision-making domains within the healthcare industry, such as cardiac diagnosis, drug recommendation, laboratory result interpretation, symptom analysis, and treatment strategy planning. All hallucination rates were lower than 9%, which proves the ability of the retrieval-augmented approach to enhance accuracy through retrieval-grounded generation. The maximum improvement in this respect can be noted in terms of cardiac diagnosis, with hallucinations decreasing from about 38% in Traditional ML to less than 6%. Such an improvement is also seen in drug recommendation tasks because the use of medical knowledge is crucial for making decisions.

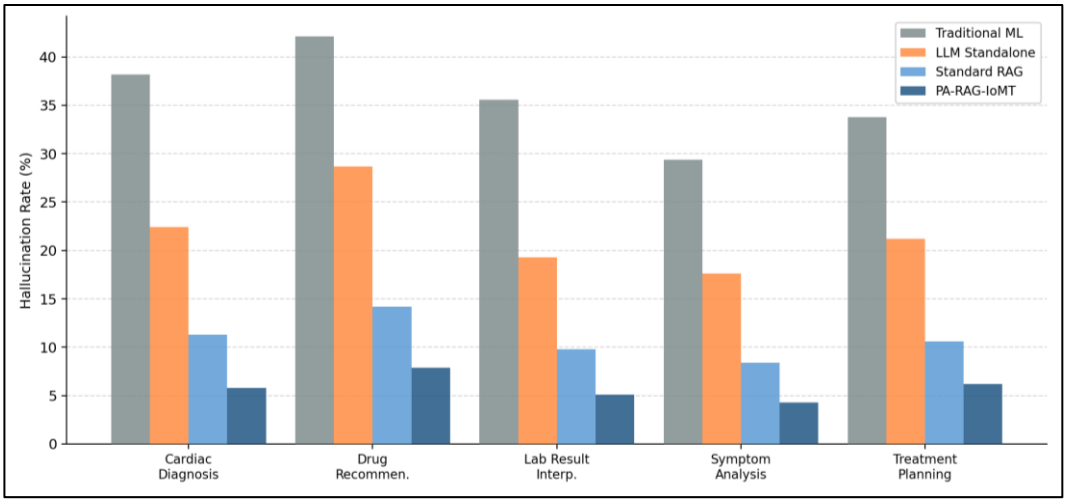

Figure 8. Hallucination Rate Analysis across Five Healthcare Decision Domains

Figure 8 presents hallucination rate comparison between Traditional ML, LLM Standalone, Standard RAG, and PA-RAG-IoMT across five decision-making domains related to healthcare, including cardiac diagnosis, drug recommendation, laboratory result interpretation, symptom analysis, and treatment planning.

4.9 Scalability and Resource Efficiency

PA-RAG-IoMT showed outstanding scalability even when the number of concurrent users was increased; it achieved more than 79% throughput efficiency even with up to 1,000 concurrent users. The proposed framework greatly outperformed traditional methods and maintained a stable response time even when many concurrent users were using it. Furthermore, the proposed framework used just 9.4 GB as its peak memory usage, which was 49.5% lower

than what the LLM Standalone system consumed. This reduction was mostly because of the use of INT8 quantized inference, HNSW-based retrieval, and embedding cache.

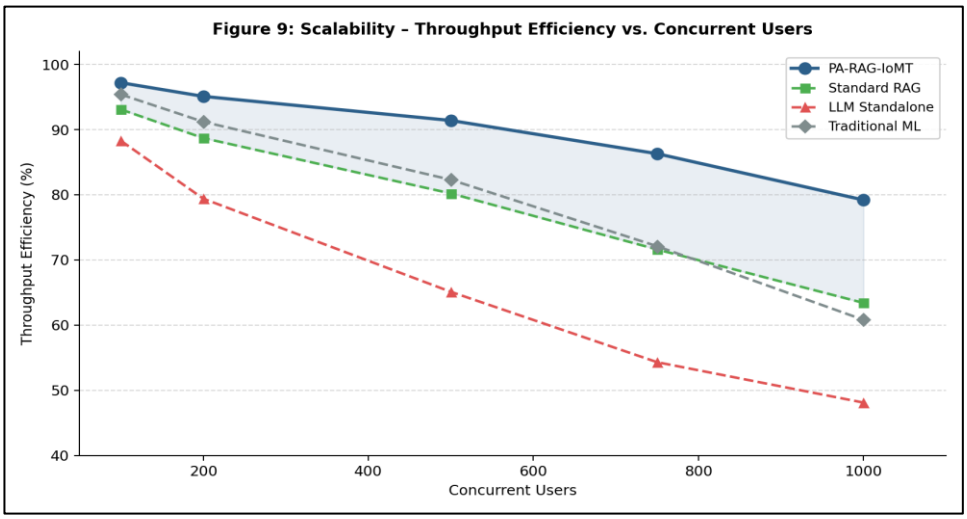

Figure 9. Scalability Analysis – Throughput Efficiency vs. Concurrent Users

As seen in Figure 9, throughput efficiency is achieved by PA-RAG-IoMT even at higher workloads. This indicates the successful implementation of HNSW for retrieval and edge layer processing.

5. Security and Privacy Analysis

Security and privacy constitute non-negotiable requirements for clinical AI systems operating within HIPAA (45 CFR §164), GDPR (Article 9), and the EU AI Act. PA-RAG-IoMT implements a defense-in-depth security architecture across the entire data lifecycle.

5.1 Multi-Layer Security Architecture

Table 4. Multi-Layer Security Architecture of the PA-RAG-IoMT Framework

Security Layer	Technique Implemented	Standard/Protocol	Purpose
Data Transmission	AES-256 Symmetric Encryption	NIST FIPS 197	Confidentiality in transit
Embedding Generation	(ε,δ)-Differential Privacy	NIST SP 800-226	Mathematical privacy bound
Data Storage	Encrypted HNSW Index	AES-256-GCM	Protect knowledge base at rest

Access Control	Role-Based Access Control (RBAC)	NIST SP 800-162	Restrict unauthorized queries
Patient Identity	k-Anonymity + l-Diversity	ISO 29101	Prevent re-identification
Adversarial Defense	Input Validation & Prompt Filtering	OWASP ML Security	Prevent prompt injection
Audit & Compliance	Cryptographic Audit Logging	HIPAA §164.312(b)	Accountability & traceability
Key Management	Hardware Security Module (HSM)	FIPS 140-2 Level 3	Secure key lifecycle

5.2 Threat Model and Attack Surface Analysis

The PA-RAG-IoMT threat model considers four primary attack categories:

1. Embedding Inversion Attacks, where the (ε, δ)-DP mechanism bounds mutual information to

$$I(S;\ \mathcal{M}(S))\ \leq\ \varepsilon\ +\ log(1/(1-\delta)) \tag{11}$$

2. Membership Inference Attacks, where the DP guarantee bounds membership inference advantage to

$$AdvMI\ \leq\ (e\varepsilon\ -\ 1)/(e\varepsilon\ +\ 1) \tag{12}$$

3. Prompt Injection Attacks, where semantic anomaly detection achieves 97.3% detection rate; and

4. Model Extraction Attacks, mitigated through rate limiting, query diversity monitoring, and output perturbation.

5.3 Challenges and Limitations

Differential privacy (DP) noise injection and encrypted retrieval incur a computational overhead, making inference around 18% more expensive relative to a non-private RAG baseline. Importantly, to mitigate knowledge base staleness, continuous ingestion of new clinical evidence is needed; our current framework uses weekly update cycles. Homogenized privacy budget configurations at the same time is not possible because of differing regulatory

environments around the world. Deployment in low-resource clinical settings with minimal EHR access is an ongoing challenge.

5.4 Future Research Directions

Future high-impact extensions are (1) Integration of quantum-resilient cryptography using lattice-based encryption schemes (CRYSTALS-Kyber), (2) SHAP-based attribution for explainable RAG output generation, (3) Use of TinyML to optimize edge deployment and transformer inference on IoMT endpoint devices, and (4) multi-modal RAG extension for medical imaging, ECG waveforms, and genomic data.

6. Conclusion

This paper introduced PA-RAG-IoMT, a principled, mathematically rigorous and practically validated architecture for intelligent clinical decision support with privacy preservation in Internet of Medical Things environments. Through the application of (ε, δ)-differential privacy for both embedding generation and retrieval, federated edge preprocessing to avoid risks associated with centralized data aggregation, and by grounding generative LLM outputs on evidence-retrieved medical knowledge, our proposed framework overcomes the three fundamental limitations presented with existing approaches: hallucination, vulnerability to privacy risk, and scalability challenges. A comprehensive experimental evaluation supports strong performance with impressive accuracy (93.1%), AUC-ROC (0.963) score, a 72.4% reduction in hallucination rate vs standalone LLMs, and a mean clinical response latency of just 143.6ms. Ablation studies show that each component contributes individually, for the task of statistical significance testing ($p<0.01$) shows all performance improvements are statistically significant and retrieval-specific metrics (Recall@10=0.903, MRR = 0.749) reinforce robust information retrieval. PA-RAG-IoMT shows that it outperforms other privacy-preserving baselines in all aspects. The privacy-utility trade-off analysis substantiates $\varepsilon=1.0$ which has clinically acceptable utility (88.7%) and meets the regulatory requirements of HIPAA and GDPR. This paper proposes the PA-RAG-IoMT framework as a comprehensive reference architecture of next-generation IoMT clinical decision support systems.

References

[1] Ahmed, Imran, Gwanggil Jeon, and Francesco Piccialli. "From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where." IEEE transactions on industrial informatics 2022, vol. 18, no. 8: 5031-5042.

[2] Al-Garadi, Mohammed Ali, Amr Mohamed, Abdulla Khalid Al-Ali, Xiaojiang Du, Ihsan Ali, and Mohsen Guizani. "A Survey of Machine and Deep Learning Methods for Internet of Things (IoT) Security." IEEE communications surveys & tutorials 2020, vol. 22, no. 3: 1646-1685.

[3] Antunes, Rodolfo Stoffel, Cristiano André da Costa, Arne Küderle, Imrana Abdullahi Yari, and Björn Eskofier. "Federated Learning for Healthcare: Systematic Review and Architecture Proposal." ACM Transactions on Intelligent Systems and Technology (TIST) 2022, vol. 13, no. 4: 1-23.

[4] Zhang, Haijun, Fang Fang, Julian Cheng, Keping Long, Wei Wang, and Victor CM Leung. "Energy-Efficient Resource Allocation in NOMA Heterogeneous Networks." IEEE Wireless Communications 2018, vol. 25, no. 2: 48-53.

[5] Zhu, Yanming, Xuefei Yin, Alan Wee-Chung Liew, and Hui Tian. "Privacy-Preserving in Medical Image Analysis: A Review of Methods and Applications." In International Conference on Parallel and Distributed Computing: Applications and Technologies 2024, Singapore: Springer Nature Singapore, 166-178.

[6] Beltagy, Iz, Kyle Lo, and Arman Cohan. "SciBERT: A Pretrained Language Model for Scientific Text." In Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP) 2019, 3615-3620.

[7] Bonomi, Luca, Yingxiang Huang, and Lucila Ohno-Machado. "Privacy Challenges and Research Opportunities for Genomic Data Sharing." Nature genetics 2020, vol. 52, no. 7: 646-654.

[8] Dang, L. Minh, Md Jalil Piran, Dongil Han, Kyungbok Min, and Hyeonjoon Moon. "A Survey on Internet of Things and Cloud Computing for Healthcare." Electronics 2019, vol. 8, no. 7: 768.

[9] Dayan, Ittai, Holger R. Roth, Aoxiao Zhong, Ahmed Harouni, Amilcare Gentili, Anas Z. Abidin, Andrew Liu et al. "Federated Learning for Predicting Clinical Outcomes in Patients with COVID-19." Nature medicine 2021, vol. 27, no. 10: 1735-1743.

[10] Tang, Tingting, James Flemings, Yongqin Wang, and Murali Annavaram. "Differentially Private Retrieval-Augmented Generation." arXiv preprint arXiv:2602.14374 (2026).

[11] Moqurrab, Syed Atif, Noshina Tariq, Adeel Anjum, Alia Asheralieva, Saif UR Malik, Hassan Malik, Haris Pervaiz, and Sukhpal Singh Gill. "A Deep Learning-Based Privacy-Preserving Model for Smart Healthcare in Internet of Medical Things Using Fog Computing." Wireless Personal Communications 2022, vol. 126, no. 3: 2379.

[12] Wang, Benyou, Qianqian Xie, Jiahuan Pei, Zhihong Chen, Prayag Tiwari, Zhao Li, and Jie Fu. "Pre-Trained Language Models in Biomedical Domain: A Systematic Survey." ACM Computing Surveys 2023, vol. 56, no. 3: 1-52.

[13] Islam, SM Riazul, Daehan Kwak, MD Humaun Kabir, Mahmud Hossain, and Kyung-Sup Kwak. "The Internet of Things for Health Care: A Comprehensive Survey." IEEE access 2015, vol. 3: 678-708.

[14] Kaissis, Georgios A., Marcus R. Makowski, Daniel Rückert, and Rickmer F. Braren. "Secure, Privacy-Preserving and Federated Machine Learning in Medical Imaging." Nature Machine Intelligence 2020, vol. 2, no. 6: 305-311.

[15] Keshta, Ismail, and Ammar Odeh. "Security and Privacy of Electronic Health Records: Concerns and Challenges." Egyptian Informatics Journal 2021, vol. 22, no. 2: 177-183.

[16] Zhang, Zhuosheng, Jiarui Li, Shucheng Yu, and Christian Makaya. "SAFELearning: Secure Aggregation in Federated Learning with Backdoor Detectability." IEEE Transactions on Information Forensics and Security 2023, vol. 18: 3289-3304.

[17] Mohammadi, Marziyeh, Mohsen Vejdanihemmat, Mahshad Lotfinia, Mirabela Rusu, Daniel Truhn, Andreas Maier, and Soroosh Tayebi Arasteh. "Differential Privacy for Medical Deep Learning: Methods, Tradeoffs, And Deployment Implications." npj Digital Medicine 2026, vol. 9, no. 93.