INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 4, April 2026 | ISSN: 2456-4184 | IJNRD.ORG

SECURE HEALTHCARE DATA RETRIEVAL AND INTELLIGENT CLINICAL REPORT GENERATION USING RETRIEVAL-AUGMENTED GENERATION ARCHITECTURE

Mr. M. Nagarasan, Mr. J.V. Bharath Kumar, Ms. R. Raghavi, Mr. A. Ramana, Mr. M. Vigneshwara

Department of Computer Science and Engineering
INFO Institute of Engineering, Kovilpalayam, Coimbatore, India – 641107
{nagarasancs, bharathkumarjv28, raghavirajadurai8, ramanasbr98, vigneshwara4011}@gmail.com

***Abstract*—**The rapid advancement of digital technologies in healthcare has led to the widespread adoption of Electronic Health Records (EHRs), laboratory systems, and clinical data repositories. While these systems have improved data storage, efficient retrieval and secure access to patient information remain significant challenges. Traditional healthcare systems rely heavily on manual search and structured query mechanisms, which are time-consuming, fragmented, and prone to human error. Additionally, standalone Large Language Models (LLMs) used in healthcare applications often suffer from hallucination, lack of domain-specific grounding, and security vulnerabilities.

To address these challenges, this paper proposes a secure and intelligent healthcare data retrieval system based on Retrieval-Augmented Generation (RAG). The system integrates hybrid retrieval techniques combining semantic vector search and keyword-based retrieval to fetch relevant medical records. The retrieved data is then used by a Large Language Model to generate context-aware and factually grounded responses. To ensure data privacy and compliance with healthcare standards, the system incorporates Role-Based Access Control (RBAC), encryption mechanisms, and audit logging. Furthermore, the system enables automated clinical report generation, reducing manual workload and improving efficiency. The proposed framework enhances accuracy, security, and usability in healthcare data management systems.

Index Terms—Healthcare AI, Retrieval-Augmented Generation (RAG), Electronic Health Records (EHR), RBAC, Medical Report Generation, NLP

I. INTRODUCTION

The rapid evolution of digital technologies has significantly transformed the healthcare industry, leading to the widespread adoption of Electronic Health Records (EHRs), hospital information systems, and cloud-based medical platforms. Modern healthcare institutions generate vast volumes of data on a daily basis, including patient histories, laboratory reports, prescriptions, diagnostic images, and clinical notes. While the digitization of healthcare data has improved storage and accessibility, it has also introduced new challenges related to efficient data retrieval, integration, and security. Healthcare professionals often struggle to quickly access relevant patient information due to fragmented systems and inefficient search mechanisms, which can delay clinical decision-making and impact the quality of patient care.

Traditional healthcare data retrieval systems rely heavily on structured queries, keyword-based searches, and manual navigation through multiple interfaces. These approaches are not only time-consuming but also require technical expertise and fail to capture the contextual meaning of user queries. As a result, important medical information may be overlooked or retrieved inaccurately. Furthermore, healthcare data is often distributed across multiple subsystems such as laboratory management, radiology systems, and pharmacy databases, making it difficult to obtain a unified view of patient information. This lack of integration leads to inefficiencies and increases the workload on healthcare professionals.

In recent years, advancements in Artificial Intelligence (AI) and Natural Language Processing (NLP) have opened new possibilities for improving healthcare data retrieval. AI-powered systems can understand natural language queries, extract relevant information, and provide meaningful insights to users. Large Language Models (LLMs), in particular, have demonstrated strong capabilities in text generation, summarization, and conversational interfaces. However, standalone LLMs are not suitable for healthcare applications due to their tendency to generate hallucinated or unverified information. In critical domains such as healthcare, even minor inaccuracies can lead to serious consequences, making reliability and factual correctness essential.

In addition to data retrieval, healthcare systems also require efficient mechanisms for generating clinical documentation. Medical professionals spend a significant amount of time preparing reports such as discharge summaries, diagnostic interpretations, and treatment notes.

The proposed system aims to address the key challenges of healthcare data retrieval, including inefficiency, lack of integration, security concerns, and manual workload. By leveraging advanced AI techniques and secure data management practices, the system enhances the accessibility, accuracy, and reliability of healthcare information. Ultimately, this contributes to improved clinical decision-making, better patient outcomes, and more efficient healthcare services.

II. LITERATURE SURVEY

Several research studies have explored the application of Retrieval-Augmented Generation (RAG), Artificial Intelligence (AI), and Natural Language Processing (NLP) in healthcare data retrieval, clinical decision support, and medical report generation systems.

Paper 1: Enhancing Health Information Retrieval with RAG. Upadhyay and Viviani (2025) proposed a RAG-based framework that improves healthcare information retrieval by prioritizing topical relevance and factual accuracy. The system integrates dense retrieval with re-ranking techniques to enhance precision in medical data access.

Paper 2: Joint Medical LLM and Retrieval Training. Wang et al. (2024) introduced a joint training approach that combines medical Large Language Models with retrieval mechanisms to improve reasoning capabilities. The model is trained to retrieve relevant medical knowledge while simultaneously generating accurate responses.

Paper 3: Rationale-Guided Retrieval-Augmented Generation. Sohn et al. (2025) proposed a rationale-guided RAG framework for medical question answering. The system retrieves relevant documents along with explanatory rationales that guide the generation process.

Paper 4: Dual Retrieval and Ranking Medical LLM. Yang et al. (2025) developed a dual retrieval and ranking mechanism integrated with a medical Large Language Model. The system combines multiple retrieval strategies and applies ranking algorithms to select the most relevant medical information.

Paper 5: Explainable AI in Decision Making Systems. El-Enen et al. (2025) presented a comprehensive survey of RAG models in healthcare. The study analyzes various retrieval techniques, embedding strategies, evaluation metrics, and challenges in deploying RAG systems.

Paper 6: Privacy-Aware RAG for Telehealth Systems. Roberts and Zhao (2024) proposed a privacy-aware RAG framework designed for telehealth applications. The system incorporates privacy-preserving techniques to protect sensitive patient data during retrieval and generation.

Paper 7: RAG for Medical Image Report Generation. Kaur and Patel (2025) developed a RAG-based system for generating medical image reports. The system retrieves relevant clinical information and combines it with image analysis to generate structured reports.

Paper 8: Adaptive RAG Models for Multimodal Healthcare QA. Liu and Gupta (2024) proposed an adaptive RAG model that supports multimodal healthcare question answering using text, images, and structured data.

III. RESEARCH GAP

Despite the rapid development of AI-based healthcare systems, several limitations still exist in current solutions. Most traditional hospital management systems rely on manual data retrieval processes, which are inefficient and time-consuming.

Existing Retrieval-Augmented Generation models improve contextual understanding but often lack strong security mechanisms such as Role-Based Access Control (RBAC), encryption, and audit logging. This creates risks of unauthorized access and data breaches in healthcare environments

The proposed system addresses these gaps by integrating secure data access, hybrid retrieval mechanisms, and context-aware AI generation within a unified framework.

IV. PROPOSED FRAMEWORK

The proposed system introduces a secure and intelligent healthcare data retrieval framework using Retrieval-Augmented Generation (RAG).

The system is designed to provide accurate, context-aware, and secure access to patient information through natural language queries.

To ensure data security and privacy, the system incorporates Role-Based Access Control (RBAC), which restricts access to patient data based on user roles such as doctors, administrators, and patients. Encryption techniques are used to protect data during storage and transmission, while audit logging ensures accountability by recording all user interactions.

This reduces manual workload and improves efficiency in healthcare operations

Overall, the proposed system enhances data retrieval accuracy, ensures security, and supports intelligent clinical decision-making.

V. SYSTEM ARCHITECTURE

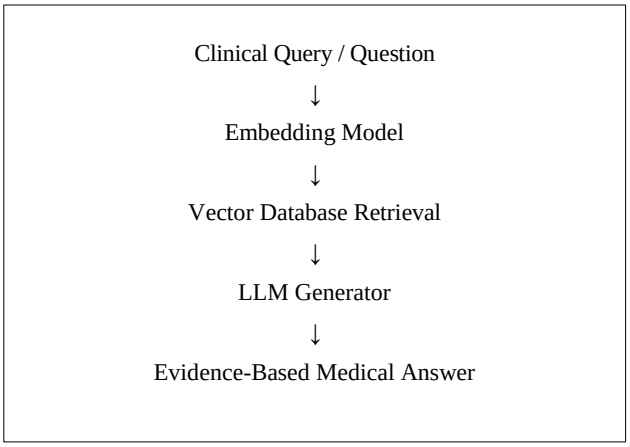

Figure 1. Architecture of RAG Workflow

VI. COMPARISON OF EXISTING PLATFORMS

Table 1. Comparison of Interview Platforms

Platform	Feature	Technology	Limitation
Upadhyay & Viviani (2025)	Health Information Retrieval	RAG + Dense Retrieval	Low Security (No RBAC, No Encryption)
Wang et al. (2024)	Medical Reasoning System	LLM + Retrieval Training	No Privacy Protection
Sohn et al. (2025)	Coding Medical QA with Explanation	Rationale-Guided RAG	Limited Scalability
Yang et al. (2025)	Dual Retrieval & Ranking	RAG + Ranking Algorithms	No Access Control
Roberts & Zhao (2024)	Telehealth Data Processing	Privacy-Aware RAG	Limited Features

VII. EXPERIMENTAL SETUP

The proposed system is implemented using Python with FastAPI for backend development and React for the frontend interface. The RAG model integrates a Large Language Model with hybrid retrieval techniques to evaluate performance based on accuracy, response time, and relevance of generated outputs.

VIII. SYSTEM MODULES

The system is divided into several modules to ensure efficient operation and maintainability. The healthcare data ingestion module collects patient data from various sources and ensures secure storage. The preprocessing module cleans and structures the data, extracting relevant medical information.

A. Healthcare Data Ingestion Module

The first stage of the system involves collecting healthcare data from multiple sources such as Electronic Health Records (EHRs), laboratory reports, prescriptions, and diagnostic systems. The module ensures that data from different formats and systems is securely stored in a centralized database. Data integrity and confidentiality are maintained during ingestion using secure data transfer protocols. This module forms the foundation for further processing and analysis within the system.

B. Data Preprocessing and Embedding Module

The collected healthcare data is processed to remove inconsistencies, handle missing values, and standardize formats. Important medical entities such as patient details, diagnosis, medications, and lab results are extracted and structured. The processed data is then converted into vector embeddings using machine learning models. These embeddings capture semantic relationships within the data, enabling efficient and accurate retrieval during query processing.

C. Hybrid Retrieval Module

The hybrid retrieval module is responsible for fetching relevant healthcare data based on user queries. It combines keyword-based search with semantic vector search to improve retrieval accuracy. Keyword search ensures exact matching, while semantic search captures contextual meaning. A ranking mechanism is applied to prioritize the most relevant results. This module plays a critical role in reducing irrelevant data retrieval and improving response quality.

D. Query Processing and RAG Module

The query processing module accepts user input in natural language and converts it into a format suitable for retrieval and generation. The Retrieval-Augmented Generation (RAG) module integrates the retrieved data with a Large Language Model to generate context-aware and medically accurate responses. By grounding the generated output in retrieved data, the system minimizes hallucination and ensures factual correctness.

E. Clinical Report Generation Module

The report generation module automatically creates structured clinical documents such as discharge summaries, diagnostic reports, and treatment notes. The module uses the output from the RAG system and organizes it into a standardized format suitable for healthcare professionals. This reduces manual effort, improves documentation accuracy, and enhances workflow efficiency in healthcare environments.

F. Security and Access Control Module

Security is a critical component of the system. This module implements Role-Based Access Control (RBAC) to restrict access to sensitive patient data based on user roles such as doctors, administrators, and patients.

G. Performance Evaluation Module

The performance evaluation module monitors the efficiency and accuracy of the system. It evaluates metrics such as retrieval accuracy, response time, precision, recall, and F1-score.

IX. MATHEMATICAL MODEL

The proposed system can be represented using a mathematical model that describes the relationship between healthcare data, user queries, and generated responses.

Let the healthcare data be represented as a set of documents:

$$D=(d1,d2,d3,\ldots,dn) \quad (1)$$

where each document d_i represents a medical record such as Electronic Health Records (EHRs), lab reports, prescriptions, or clinical notes.

The final response generated by the system is computed using Retrieval-Augmented Generation as:

$$\text{Response}=\text{LLM}(q,R) \quad (2)$$

where q represents the user query, R represents the set of retrieved relevant documents, and LLM denotes the Large Language Model used to generate context-aware and accurate responses.

X. DATASET DESCRIPTION

The dataset used in this system consists of Electronic Health Records, laboratory reports, prescriptions, and diagnostic data collected from healthcare systems. The data includes both structured and unstructured information, such as patient details, medical history, and clinical notes.

Before processing, the data undergoes preprocessing to remove inconsistencies, handle missing values, and standardize formats. Important clinical entities such as diagnosis codes, medications, and lab results are extracted and organized into structured formats.

Semantic embeddings are generated from the processed data to enable efficient retrieval using vector search techniques. The dataset is continuously updated to ensure that the system provides accurate and up-to-date information.

XI. EVALUATION METRICS

- The performance of the proposed system is evaluated using several metrics to ensure accuracy, efficiency, and reliability. Accuracy measures the correctness of retrieved and generated responses, while precision and recall evaluate the relevance of retrieved data.
- The F1-score provides a balanced measure of precision and recall. Response time is used to evaluate system efficiency, ensuring that queries are processed quickly. Additionally, system performance is assessed under different workloads to ensure scalability.
- These metrics help in analyzing the effectiveness of the system and identifying areas for improvement.

These metrics provide a comprehensive evaluation of system effectiveness.

XII. IMPLEMENTATION DETAILS

The system is implemented using Python as the primary programming language. Backend development is carried out using frameworks such as FastAPI or Flask, while the frontend interface is developed using React.

Healthcare data is stored in databases such as MongoDB or PostgreSQL, and semantic embeddings are managed using vector databases like FAISS. The RAG framework integrates retrieval mechanisms with a Large Language Model to generate responses.

Security is implemented using JWT-based authentication, RBAC, and encryption techniques. The system is deployed on a secure server environment to ensure scalability and accessibility.

The system is deployed on cloud infrastructure to ensure scalability and accessibility for multiple users.

XIII. ADVANTAGES OF PROPOSED SYSTEM

The proposed system provides several advantages compared to traditional interview preparation methods.

- Provides secure access to healthcare data using Role-Based Access Control (RBAC)
- Reduces hallucination in AI-generated responses using Retrieval-Augmented Generation (RAG).
- Ensures accurate and relevant data retrieval through hybrid search (keyword + semantic).
- Automates clinical report generation, reducing manual workload for healthcare professionals.
- Supports real-time query processing with faster response time.

XIV. LIMITATIONS

Despite its advantages, the system has certain limitations. The performance of the system depends on the quality and completeness of the dataset. Inaccurate or incomplete data may affect the quality of generated responses.

The system also requires significant computational resources for processing large datasets and running AI models. Additionally, real-time integration with hospital systems may require further development and testing.

XV. RESULTS AND DISCUSSION

The system was tested under various conditions to evaluate its performance and reliability. The results indicate that the hybrid retrieval mechanism improves the accuracy of data retrieval compared to traditional methods.

XVI. FUTURE WORK

Future enhancements can further improve the system's capabilities. Integration with real-time hospital systems will enable live data access and updates. Advanced AI models can be incorporated to improve accuracy and reduce errors.

XVII. CONCLUSION

The proposed Retrieval-Augmented Generation based healthcare system provides an effective solution for secure and intelligent medical data retrieval. By integrating hybrid retrieval techniques with AI-based generation, the system improves accuracy and efficiency in accessing patient information. The implementation of security mechanisms such as RBAC and encryption ensures data privacy and compliance with healthcare standards. The system also reduces manual workload through automated report generation.

REFERENCES

[1] R. Upadhyay and M. Viviani, "Enhancing Health Information Retrieval with RAG by Prioritizing Topical Relevance and Factual Accuracy," Discover Computing Education: Artificial Intelligence, Springer, 2025.

[2] W. Wang, Z. Yang, Z. Yao, and H. Yu, "Joint Medical LLM and Retrieval Training for Enhancing Reasoning," arXiv preprint, 2024.

[3] J. Sohn et al., "Rationale-Guided Retrieval-Augmented Generation for Medical Question Answering," in Proc. NAACL, 2025.

[4] Q. Yang et al., "Dual Retrieving and Ranking Medical Large Language Model with Retrieval Augmented Generation," Scientific Reports, Na ture, 2025.

[5] M. A. El-Enen, S. Saad, and T. Nazmy, "A Survey on Retrieval Augmented Generation Models for Healthcare Applications," Neural Computing and Applications, Springer, 2025.

[6] E. Roberts and A. Zhao, "Privacy-Aware RAG for Telehealth Data Interpretation," IEEE Access, vol. 12, pp. 45501–45515, 2024.

[7] M. Kaur and D. Patel, "RAG for Medical Image Report Generation," in Proc. IEEE International Conference on Healthcare Informatics, 2025.

[8] H. Liu and R. Gupta, "Adaptive RAG Models for Multimodal Healthcare QA," IEEE/ACM Transactions on Computational Biology and Bioinfor matics, 2024

[9] L. Chen and P. Kumar, "RAG-Enhanced Clinical Text Normalization," Neural Computing and Applications, Springer, 2023.

[10] M. Singh and I. Ahmed, "Hybrid Sparse–Dense RAG for Clinical Note Retrieval," in Lecture Notes in Computer Science, Springer, 2025.

[11] V. Karpukhin et al., "Dense Passage Retrieval for Open-Domain Question Answering," Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[12] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," Proceedings of NAACL-HLT, 2019.

[13] P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2020.

[14] A. Vaswani et al., "Attention Is All You Need," Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), 2017.

[15] J. Johnson, M. Douze, and H. J´ egou, "Billion-Scale Similarity Search with FAISS," Facebook AI Research Technical Report, 2017.

[16] C. Lin et al., "Clinical Information Extraction Using Natural Language Processing Techniques," Journal of Biomedical Informatics, 2021.

[17] Y. Wang et al., "Applications of Artificial Intelligence in Healthcare Data Retrieval Systems," Healthcare Informatics Research Journal, 2022.

[18] E. Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.

[19] A. Esteva et al., "A Guide to Deep Learning in Healthcare," Nature Medicine, vol. 25, pp. 24–29, 2019.

[20] A. Rajkomar et al., "Scalable and Accurate Deep Learning with Electronic Health Records," NPJ Digital Medicine, vol. 1, no. 18, 2018.

[21] R. Miotto et al., "Deep Learning for Healthcare: Review, Opportunities and Challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.

[22] B. Shickel et al., "Deep Learning in Electronic Health Records: A Systematic Review," Journal of Biomedical Informatics, vol. 83, pp. 168–185, 2018.

[23] A. E. W. Johnson et al., "MIMIC-III Clinical Database," Scientific Data, vol. 3, Article no. 160035, 2016.

[24] A. L. Beam and I. S. Kohane, "Big Data and Machine Learning in Healthcare," Journal of the American Medical Association (JAMA), vol. 319, no. 13, pp. 1317–1318, 2018.

[25] S. Vasudevan et al., "Secure Role-Based Access Control Models for Healthcare Information Systems," IEEE Access, vol. 9, pp. 45678–45689, 2021.

[26] T. Brown et al., "Language Models are Few-Shot Learners," Proceedings of the Conference on Neural Information Processing Systems (NeurIPS), vol. 33, pp. 1877–1901, 2020.

[27] I. Beltagy, M. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[28] K. Clark, M. Luong, Q. Le, and C. Manning, "ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators," Proceedings of the International Conference on Learning Representations (ICLR), 2020.

[29] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," arXiv preprint arXiv:1706.05098, 2017.

[30] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed., Pearson, 2022.