

AI-Enhanced Privacy-Preserving EMR Search using Natural Language Query Translation and Intelligent Revocation

Uday Kiran .M ¹, Eshwar .P ², Sampath .S ³, Saketh .P⁴

¹²³⁴Department of Information Technology, TKR College of Engineering and Technology, Telangana, India

Abstract - The rapid digitization of healthcare has led to a significant increase in Electronic Medical Records (EMRs), which contain highly sensitive patient information such as diagnoses, prescriptions, lab reports, and treatment history. While digital storage improves accessibility and clinical efficiency, it introduces major privacy and security challenges, especially when searching medical records stored in encrypted form. Traditional EMR systems either store records in plaintext or require full decryption during search, which increases the risk of data exposure. Moreover, conventional keyword-based retrieval lacks semantic understanding, resulting in low accuracy and poor relevance. This paper proposes an AI-enhanced privacy-preserving EMR search system that enables secure and intelligent retrieval of encrypted medical data. The system integrates Fernet AES encryption for EMR confidentiality, SHA-256 hashed keyword indexing for secure keyword lookup, and Sentence Transformer embeddings for semantic search. A hybrid query engine combines hashed keyword matching with embedding-based similarity ranking to improve retrieval accuracy. The framework also enforces strong security through Role-Based Access Control (RBAC), patient-ID restricted access, OTP-based authentication, and dynamic access revocation through key invalidation. A Stream lit-based dashboard provides role-specific interfaces for Admin, Doctor, Nurse, and Patient. Experimental results show that the proposed approach supports natural-language EMR search while preserving confidentiality, preventing unauthorized access, and ensuring clinically relevant results. The system demonstrates that secure searchable EMR platforms can achieve both high privacy and high usability for modern healthcare environments.

Key words : Electronic Medical Records, Privacy-Preserving Search, Searchable Encryption, Semantic Search, NLP, Sentence Transformers, RBAC, Fernet AES, SHA-256, Access Revocation.

1. INTRODUCTION

The rapid growth of digital healthcare has significantly increased the adoption of Electronic Medical Records (EMRs) across hospitals, clinics, and diagnostic centres. EMRs contain highly sensitive patient data such as medical history, diagnoses, prescriptions, laboratory reports, and treatment plans. While digitization improves clinical efficiency and accessibility, it also introduces major challenges related to data privacy, security, and controlled

access. Healthcare data breaches have become a serious concern due to unauthorized access, insider misuse, and weak access revocation mechanisms in conventional EMR systems.

1.1 Need for Secure EMR Storage

Traditional EMR platforms often rely on centralized database storage with basic encryption and authentication. However, these systems are vulnerable because encrypted records are usually decrypted during retrieval and search operations, which increases exposure risks. Searchable encryption has been proposed as an effective approach to allow retrieval without revealing plaintext medical content [1]. Therefore, strong cryptographic mechanisms are required to ensure confidentiality and tamper resistance.

1.2 Challenges in Searching Encrypted Medical Records

Performing search operations over encrypted EMRs is technically difficult because encryption hides the semantic meaning and structure of the stored data. Most existing systems require decryption before performing keyword search, which defeats the purpose of encryption. Privacy-preserving search techniques attempt to solve this issue by allowing search functionality while maintaining confidentiality [1]. However, keyword-only search still produces low accuracy and high false positives in clinical settings.

1.3 Role of NLP and Semantic Search

Natural Language Processing (NLP) has become a powerful tool for enabling intelligent retrieval of medical documents. Semantic search techniques allow the system to understand query intent and retrieve clinically relevant records even when the exact keyword is not present. Studies have shown that semantic retrieval using NLP-based embeddings significantly improves EMR search relevance [2]. Sentence Transformer models are widely used for generating semantic embeddings for text-based similarity matching [5].

1.4 Importance of Fine-Grained Access Control

In healthcare, not all users should access all records. Role-Based Access Control (RBAC) is a widely accepted

model for enforcing fine-grained permissions based on user roles such as doctor, nurse, and patient [3]. However, many systems fail to implement effective revocation, meaning users may retain access even after their role changes. Dynamic revocation is essential to prevent unauthorized decryption after privilege removal.

1.5 Motivation and Contribution of This Work

To address these limitations, this project proposes an AI-Enhanced Privacy-Preserving EMR Search System that integrates encryption, privacy-preserving indexing, semantic retrieval, and dynamic access control. The key contributions of the proposed system include: Securing EMRs using Fernet AES encryption [4] Supporting privacy-preserving keyword search through SHA-256 hashed indexing Enabling semantic EMR retrieval using Sentence Transformer embeddings [5] Implementing hybrid query processing (keyword + semantic) for higher accuracy Enforcing RBAC-based access control with dynamic revocation [3] Ensuring patient-ID restricted record access and OTP-based authentication

2. PROPOSED SYSTEM

The proposed system is a secure and intelligent Electronic Medical Records (EMR) search platform that enables privacy-preserving storage and retrieval of sensitive healthcare data. The system combines encryption, hashed keyword indexing, semantic search, hybrid query processing, and role-based access control to ensure that EMRs remain confidential while still allowing efficient and meaningful search. The framework is designed to support natural language queries and enforce strict authorization policies, making it suitable for modern healthcare environments.

2.1 EMR Encryption and Secure Storage

All EMR records are encrypted before storage using Fernet AES encryption to ensure confidentiality and integrity. This prevents unauthorized users from viewing plaintext medical data even if the database is compromised. Encrypted EMRs are stored as ciphertext, ensuring secure long-term storage.

2.2 SHA-256 Hashed Keyword Indexing

To enable privacy-preserving keyword search, the system extracts important medical terms from each EMR and converts them into SHA-256 hash values. These hashes are stored in an index table. During search, the query keywords are also hashed and matched against the stored hashes, allowing secure keyword lookup without revealing plaintext terms.

2.3 Semantic Search Using Sentence Transformer Embeddings

To improve retrieval accuracy beyond keyword matching, the system uses Sentence Transformer embeddings to represent EMRs and queries as semantic vectors. This enables context-aware search, allowing the system to retrieve relevant records even when different terms are used (e.g., “tension” matching “hypertension”). Semantic similarity scoring is used to rank EMRs based on meaning.

2.4 Hybrid Search and Dynamic Access Control

The system uses a hybrid search mechanism that combines hashed keyword matching and semantic ranking to generate accurate and clinically relevant results. Role-Based Access Control (RBAC) ensures that users can only access records permitted by their role. Dynamic access revocation is implemented through key invalidation and permission updates, ensuring immediate removal of decryption access when roles change. Patient-ID restrictions and OTP-based authentication further strengthen security.

3. IMPLEMENTATION DETAILS

The proposed system is implemented as a modular privacy-preserving EMR search platform that integrates cryptography, semantic NLP, hybrid retrieval, and secure access control. The implementation is designed to ensure that EMRs remain encrypted during storage and retrieval while still supporting accurate natural language search. The system follows a layered approach consisting of encryption, indexing, semantic embedding generation, hybrid search processing, and role-based access enforcement.

3.1 Data Encryption and Secure Storage Module

In this module, every EMR is encrypted before being stored in the database. Fernet AES encryption is used to provide both confidentiality and integrity. A unique Data Encryption Key (DEK) is generated for EMR encryption, ensuring that records remain protected even if the storage server is compromised. The encrypted EMRs are stored in SQLite as ciphertext blobs. This module guarantees that no plaintext medical information is stored directly in the database.

3.2 Hashed Keyword Indexing Module

To support privacy-preserving keyword search, the system extracts important keywords from each EMR (such as diseases, symptoms, medications, and lab test names). Each extracted keyword is converted into an irreversible SHA-256 hash. These hashed values are stored in a separate keyword-index table linked to the EMR record ID. During search, query keywords are hashed using the same SHA-256 algorithm and compared with the stored hashes. This

enables keyword-based retrieval without exposing any plaintext medical terms.

3.3 Semantic Embedding Generation Module

To enable context-aware search, semantic embeddings are generated for each EMR using a Sentence Transformer model. The embedding represents the overall meaning of the EMR content in a numerical vector format. Similarly, when a user enters a natural language query, the query is converted into an embedding vector. Semantic similarity measures are then used to identify EMRs that match the intent of the query, even when exact keywords do not match. This module significantly improves retrieval accuracy for clinical queries.

3.4 Hybrid Search, RBAC, and Access Revocation Module

This module performs the main retrieval and security enforcement tasks. The final results are merged and ranked to produce accurate and clinically relevant EMR retrieval. Role-Based Access Control (RBAC) is enforced to restrict EMR access based on user roles such as Admin, Doctor, Nurse, and Patient. Patient-ID restricted access ensures that patients can only view their own records. Dynamic access revocation is implemented by invalidating encryption keys or removing wrapped DEKs when user permissions are revoked, ensuring immediate loss of decryption capability. OTP-based authentication strengthens login security, and audit logs store all search and access activities for accountability.

3.5 System Architecture

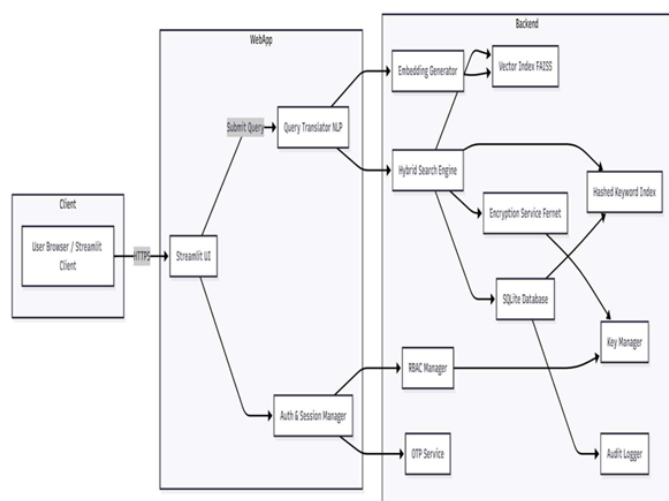


Fig. 1: Architecture of Encrypted EMR Storage and Hybrid Search Framework

Fig. 1 illustrates the overall architecture of the proposed AI-enhanced privacy-preserving EMR search system. The architecture is divided into three major layers: Client Layer, Web Application Layer, and Backend Layer. The client

accesses the system through a secure HTTPS connection using a browser-based Streamlit interface.

The WebApp layer contains the Streamlit UI, where users such as Admin, Doctor, Nurse, and Patient interact with the system. User queries are submitted through the interface and forwarded to the NLP Query Translator, which converts natural language queries into both keyword tokens and semantic representations. The WebApp also includes an Authentication and Session Manager, responsible for secure login and session handling.

The Backend layer performs all privacy-preserving operations. The Embedding Generator produces Sentence Transformer embeddings for EMRs and user queries, which are stored and searched using a FAISS Vector Index. In parallel, the system maintains a Hashed Keyword Index using SHA-256 to support secure keyword lookup without revealing plaintext medical terms. The Hybrid Search Engine combines results from both semantic similarity search and hashed keyword matching to generate accurate ranked outputs.

For confidentiality, EMR records are encrypted using the Fernet Encryption Service, and encrypted records are stored in an SQLite database. The Key Manager handles encryption key generation, storage, and key invalidation for dynamic access revocation. Access permissions are enforced by the RBAC Manager, ensuring only authorized users can decrypt and view EMRs. Additionally, the OTP Service provides strong authentication, and an Audit Logger records all search and access activities for accountability and compliance.

4. RESULTS AND PERFORMANCE ANALYSIS

The proposed AI-enhanced privacy-preserving EMR search system was successfully implemented and evaluated to verify its security, usability, and retrieval effectiveness. The evaluation focused on encrypted EMR storage, hybrid search accuracy, semantic relevance, and access control enforcement. The system was tested using sample EMR datasets containing patient history, diagnoses, prescriptions, lab values, and clinical notes. The results confirm that the proposed system provides meaningful search performance while ensuring strict privacy preservation.

4.1 Secure Storage and Encryption Validation

The first performance outcome verified that all EMRs are securely stored in encrypted form using Fernet AES encryption. No plaintext EMR content was stored in the SQLite database. Even when database entries were inspected directly, only ciphertext was visible. This confirms that the proposed system ensures confidentiality at rest and prevents direct exposure of sensitive medical information.

4.2 Keyword Indexing and Privacy-Preserving Search Results

The SHA-256 hashed keyword indexing mechanism successfully supported secure keyword search without storing plaintext medical terms. When users searched using medical keywords such as “diabetes,” “creatinine,” or “hypertension,” the system generated query hashes and matched them against stored hashed indexes. The results showed correct record retrieval while maintaining privacy. Since hashing is irreversible, the keyword index did not reveal sensitive terms to unauthorized observers.

4.3 Semantic Search Accuracy and Query Understanding

The semantic retrieval module demonstrated strong performance in interpreting clinical queries. Sentence Transformer embeddings enabled the system to retrieve relevant EMRs even when users entered indirect or alternate terms. For example, searching for “tension” successfully retrieved EMRs containing “hypertension,” and searching for “sugar” returned diabetes-related records. This shows that semantic embeddings improve search relevance and reduce dependence on exact keyword matching.

4.4 Hybrid Search Performance and Relevance Ranking

The hybrid search engine combined hashed keyword matching and semantic similarity scoring to generate ranked results. Compared to keyword-only search, hybrid retrieval reduced false positives and improved clinical relevance. The system returned accurate results for complex queries such as “patients with diabetes and abnormal creatinine,” by matching both structured medical terms and semantic meaning. The hybrid method produced more meaningful retrieval outcomes than using either keyword or semantic search alone.

4.5 Access Control Enforcement and Revocation Testing

Role-Based Access Control (RBAC) was tested across Admin, Doctor, Nurse, and Patient roles. Unauthorized access attempts were successfully blocked. Patient-ID restricted access ensured that patients could only view their own records. Dynamic access revocation was validated by removing user privileges and invalidating keys, after which previously authorized users could no longer decrypt EMRs. This confirms that the proposed system supports real-time access revocation and prevents privilege misuse.

4.6 System Usability and Interface Validation

The Streamlit dashboard provided a user-friendly interface for all roles. Doctors were able to upload EMRs securely,

nurses could update vitals, and patients could view only their authorized records. OTP-based authentication improved login security and ensured strong identity verification. Audit logging successfully captured search and access events, supporting accountability and compliance requirements.

5. CONCLUSION

This project successfully presents an AI-enhanced privacy-preserving Electronic Medical Records (EMR) search system that ensures secure storage, controlled access, and intelligent retrieval of sensitive healthcare data. Unlike traditional EMR platforms, the proposed system enables encrypted EMRs to be stored and searched without exposing plaintext information during indexing or retrieval.

The system integrates Fernet AES encryption for confidentiality, SHA-256 hashed keyword indexing for privacy-preserving keyword search, and Sentence Transformer embeddings for semantic understanding of natural language queries. A hybrid search mechanism combining keyword matching and semantic similarity ranking improves retrieval accuracy and reduces false positives, making the system suitable for clinical environments.

In addition, the implementation of Role-Based Access Control (RBAC), patient-ID restricted access, OTP-based authentication, audit logging, and dynamic access revocation ensures that only authorized users can decrypt and access medical records, even after role changes. Overall, the proposed framework demonstrates that strong privacy protection and meaningful EMR usability can be achieved simultaneously, offering a secure and scalable foundation for modern healthcare data management systems.

6. FUTURE WORK

In future, the proposed system can be enhanced by integrating advanced clinical transformer models and domain-specific NLP techniques to improve medical context understanding and reduce retrieval errors. Explainable AI (XAI) features can also be added to justify why specific EMRs are retrieved, improving transparency and trust for clinicians. Deploying the system on cloud infrastructure would improve scalability for large hospitals and high-volume datasets, while enabling multi-branch access with better performance. Additionally, stronger privacy-preserving computation methods such as homomorphic encryption or secure multiparty computation can be explored to further protect data during query processing. A user feedback mechanism can be introduced to allow doctors and nurses to rate results, enabling continuous improvement in retrieval accuracy. Finally, voice-based or chatbot-driven EMR search can be implemented to support real-time clinical workflows and increase system usability.

ACKNOWLEDGEMENT

There are many individuals who contributed directly or indirectly to the successful completion of this project, and we take this opportunity to express our sincere gratitude to all of them.

We are extremely thankful and indebted to our supervisor, Mrs. B. Rajani, Assistant Professor, Department of Information Technology, TKR College of Engineering and Technology, for her constant guidance, encouragement, and moral support throughout the project.

We also extend our sincere thanks to Dr. N. Satyanarayana, Head of the Department (I/c), Department of Information Technology, TKR College of Engineering and Technology, for his continuous encouragement and support during the course of this work.

Our heartfelt gratitude is extended to Dr. D. V. Ravi Shankar, Principal, TKR College of Engineering and Technology, for his timely support and valuable suggestions throughout the project period.

Finally, we would like to thank all the faculty and staff of the Department of Information Technology, as well as our parents and friends, for their cooperation, encouragement, and support in successfully completing this project.

REFERENCES

- [1] Y. Li, H. Zhu, and M. Yu, "Privacy-Preserving Searchable Encryption for Healthcare Data," *IEEE Transactions on Information Forensics and Security*, 2020.
- [2] R. Johnson and A. Patel, "Semantic Retrieval of Electronic Health Records Using NLP Techniques," *Journal of Biomedical Informatics*, 2021.
- [3] R. Sandhu and D. Ferraiolo, "Role-Based Access Control Models," *IEEE Computer*, vol. 33, no. 2, pp. 38–47, 2000.
- [4] Cryptography.io Team, "Cryptography Library Documentation (Fernet AES)," 2024.
- [5] Sentence Transformers Team, "Sentence Transformers Documentation," 2024.