Secure Information Retrieval System for a Cloud–based Platform

[1]EMMAH, Victor Thomas; [2]WOKE, Blessing Ichechi & [3]EJEKWU, Obunezi
[1,2,3]Department of Computer Science, Rivers State University, Port Harcourt, Nigeria
victor.emmah@ust.edu.ng; blessing.woke@rsu.edu.ng; ejekwu.obunezi@rsu.edu.ng

DOI: 10.56201/ijcsmt.vol.11.no11.2025.pg28.39

Abstract

With the increasing reliance on cloud storage, ensuring data confidentiality and controlled access during information retrieval has become a major concern. This paper presents a secure cloud-based information retrieval system that enables privacy-preserving search over encrypted data stored in the cloud. The system is designed to protect sensitive documents during storage and retrieval, ensuring that only authorized users can perform keyword-based searches and access the content. Key components of the system include AES-256 encryption, Bloom filter-based secure indexing, trapdoor generation for obfuscated query processing, and role-based access control (RBAC) for enforcing access permissions. The methodology employed involves a multi-layered architecture that integrates natural language processing (NLP) for query interpretation, a preprocessing pipeline for cleaning and normalizing input text, and the BM25 relevance ranking algorithm for efficient and accurate document retrieval. The system was developed and tested using the Cranfield dataset, which was structured and indexed for encrypted querying. The results demonstrate significant improvements in both security and search performance. The system achieved a precision and recall rate exceeding 97%, with a ROC-AUC score of 1.00, indicating highly reliable classification of relevant documents. The system offers better protection of sensitive data, improved retrieval accuracy, and scalable access control. These results affirm the system's effectiveness for secure document handling in real-world cloud environments.

Keywords*: Secure Search, Cloud Computing, Information Retrieval, Role-Based Access Control*

I. INTRODUCTION

Cloud computing has revolutionized how organizations store, process, and retrieve information. It offers on-demand access to computing resources such as servers, storage, and applications over the internet, providing unparalleled scalability, flexibility, and cost-efficiency (Armbrust *et al*., 2010). This technology underpins many modern IT infrastructures, enabling businesses to streamline operations, reduce capital expenditures, and enhance global accessibility. However, as the adoption of cloud computing continues to grow, so do the challenges associated with ensuring data security and privacy in these distributed environments. Cloud-based secure information retrieval refers to the process of retrieving information from cloud storage while ensuring the confidentiality, integrity, and authenticity of the data (Wang *et al*, 2018). This involves implementing robust security mechanisms, such as encryption, access control, and authentication, to protect sensitive information from unauthorized access.

The process of gaining access to data stored in cloud environments is the fundamental definition of cloud-based information retrieval. From scientific research to consumer data analytics and enterprise resource planning, this procedure has grown to be a crucial part of many applications.

Because it facilitates innovation, operational efficiency, and decision-making, organizations rely on the capacity to efficiently access accurate and pertinent information. However, because cloud computing relies on shared resources and third-party infrastructure, it has inherent weaknesses that need for strong security measures. In cloud computing, secure information retrieval refers to accessing data while maintaining its validity, secrecy, and integrity. These guidelines are essential for guaranteeing that private data is protected both during storage and retrieval. A healthcare company that stores patient records on the cloud, for instance, needs to make sure that these records are unmodified during transmission and are only accessed by authorized individuals. Numerous security measures, including as access control, authentication, and encryption, have been created to accomplish these goals. One of the most popular approaches to information security is encryption, which restricts unauthorized parties from deciphering data without the right keys.

As organizations increasingly adopt cloud platforms for data storage and management, the need for secure and efficient information retrieval has become critical. Traditional cloud-based retrieval systems often suffer from major security vulnerabilities, such as unauthorized access, data leakage, and the need to improve encryption practices, and privacy-preserving mechanisms, particularly for keyword-based search queries, exposing sensitive data to potential breaches. The absence of robust access control, semantic indexing, and query encryption also limited some systems' ability to deliver accurate and secure results as most systems failed to incorporate modern relevance-ranking models and real-time document auditing, which are essential for precise and trustworthy retrieval in high-risk environments. These deficiencies highlight the urgent need for a cloud-based secure information retrieval system that not only enhances search precision but also ensures end-to-end confidentiality, role-based access, and encrypted query processing.

II. LITERATURE REVIEW

Many approaches have been put out over time to deal with the difficulties of safely retrieving data in cloud-based settings. Kumar *et al (2019)* presented An attribute-based encryption (ABE) safe data retrieval strategy. The protocol guarantees that the encrypted data can only be accessed by people who possess the appropriate qualities. Secure multi-user environments are made possible by this fine-grained access control method. However, when dealing with big datasets or dynamic attribute changes, ABE's scalability vis limited by its computational cost.

Wang *et al*. (2018) proposed a public auditing system for safe cloud storage that protects privacy. This method ensures that third-party auditors can confirm the data's integrity without having access to its content by using random masking techniques and homomorphic authenticators. Although this method is good at maintaining integrity, it prioritizes data storage security over real-time retrieval. They proposed a secure cloud storage system supporting privacy – preserving public auditing. They further extended their result to enable a third party auditor (TPA) to perform audits for multi users simultaneously and efficiently. Extensive security and performance analysis show that the propose schemes are provably secured and highly efficient. However, this technique's significant latency and processing complexity present problems for real-time applications.

Goh (2003) developed secure indexes for encrypted data files. These indexes, often utilizing techniques such as Bloom filters and pseudo-random functions, allowed for efficient searching while maintaining data confidentiality, the scheme aimed to provide solution for scenarios where data owners outsource encrypted data to cloud storage and need to enable searchable access for authorized user. This work was a significant contribution to the field of searchable encryption, especially in the context of symmetric key cryptography. It laid groundwork for subsequent research and advancements in secure data retrieval over encrypted cloud storage, inspiring further developments in areas like multi- keyword search, ranked search, and enhance privacy definitions in searchable encryption schemes.

Cao, *et al* (2013) proposed a solution for the challenging problem of privacy-preserving multi-keyword ranked search over encrypted data in cloud computing (MRSE). They also established a set of strict privacy requirements for such a secure cloud data utilization system. Among various multi-keyword semantics, they choose the efficient similarity measure of ‘’coordinate matching which is as many matches as possible, to capture the relevance of data documents to the search query. They were able to define and solve the challenging problem of privacy-preserving multi-key ranked search over encrypted data in cloud computing (MRSE). Also, they established a set strict privacy requirement for secure cloud data utilization system.

Liang et al (2023) built the Verifiable Multi – Keyword Searchable Encryption (VMSE) Scheme that can realize multi-keyword search, this enables the search users to get more accurate search results. The data owners encrypt the outsourced files before they are uploaded, to protect their confidentiality and privacy. In order to further protect the privacy of outsourced file, the data owners search control policy for the outsourced file before uploading, to realize fine-gained searchable control. To prevent the semi-trusted cloud severs from returning corrupted files, the VMSE scheme support the integrity verification of search results.

Saniya *et al.* (2023) developed a cloud-based product information retrieval scheme that is both secure and efficient. In particular, two index structures are built: a hash value AVL tree and a product vector retrieval tree, which support identifier-based and feature-vector-based product searches, respectively. Two search algorithms were designed to search the two trees in the same way. All outsourced data is encrypted to protect the privacy of product information. The product information is encrypted symmetrically using a set of independent secret keys, and the product vectors are encrypted using the secure kNN algorithm. The security analysis and simulation results demonstrate the proposed scheme's security and efficiency.

BalaKrishna *et al.* (2024) presented an Enhanced Cloud Image Retrieval Efficiency through Secure Optimization, a secure image retrieval scheme that addresses privacy issues when image datasets are directly outsourced to untrusted clouds, in order to achieve more precise search results. To reduce search time and image owner costs, they allowed cloud servers to locally build a reliable hierarchical index graph using the encrypted image features.

III. ANALYSIS OF THE PROPOSED SYSTEM

The proposed system shown in figure 1 is designed to address critical limitations identified in existing solutions, particularly those related to the use of only two index trees (ID-AVL and PRF trees) for searches by identifier or feature vector, the lack of semantic search, weak access control mechanisms, and its implementation. The goal of our system is to enhance data confidentiality, integrity, and access control while guaranteeing effective retrieval mechanisms. Unlike earlier models that lack features for real-time security, semantic indexing, and dynamic scalability, this system significantly advances the state of cloud-based secure information retrieval by ensuring search accuracy, user privacy, system security, and real-world applicability.

A key improvement lies in the integration of semantic and fuzzy search capabilities. By leveraging Natural Language Processing (NLP) techniques such as word embeddings and contextual similarity scoring, the system can interpret user intent more effectively, even when queries are imprecise or contain partial matches. This addresses the rigid keyword matching limitation in earlier systems, improving usability and retrieval accuracy. In addition, it allows users to interact with the system more naturally and intuitively.

To enhance data security and enforce fine-grained access control, the proposed system incorporates Role-Based Access Control (RBAC) and Attribute-Based Encryption (ABE). These models ensure that only authorized users can search for and retrieve specific documents based on their roles, attributes, or permissions. This replaces the limited key-sharing mechanism with a more robust and manageable approach, supporting dynamic user roles and revocation.

Another notable advancement is the use of dynamic and compressed encrypted indexes, such as Bloom filters and inverted indexes, to improve performance and scalability. This minimizes the memory overhead associated with static lookup tables and allows the system to handle large-scale datasets efficiently. Furthermore, encrypted indexes are combined with trapdoor obfuscation to protect query privacy and prevent frequency analysis attacks, addressing security concerns in earlier searchable encryption models.

The system also improves usability by removing the constraint of fixed-length queries. Users can submit queries of any length, and the system intelligently processes them using query expansion and filtering techniques. This flexibility significantly improves the user experience and ensures that the system accommodates various search behaviors without compromising security or performance.

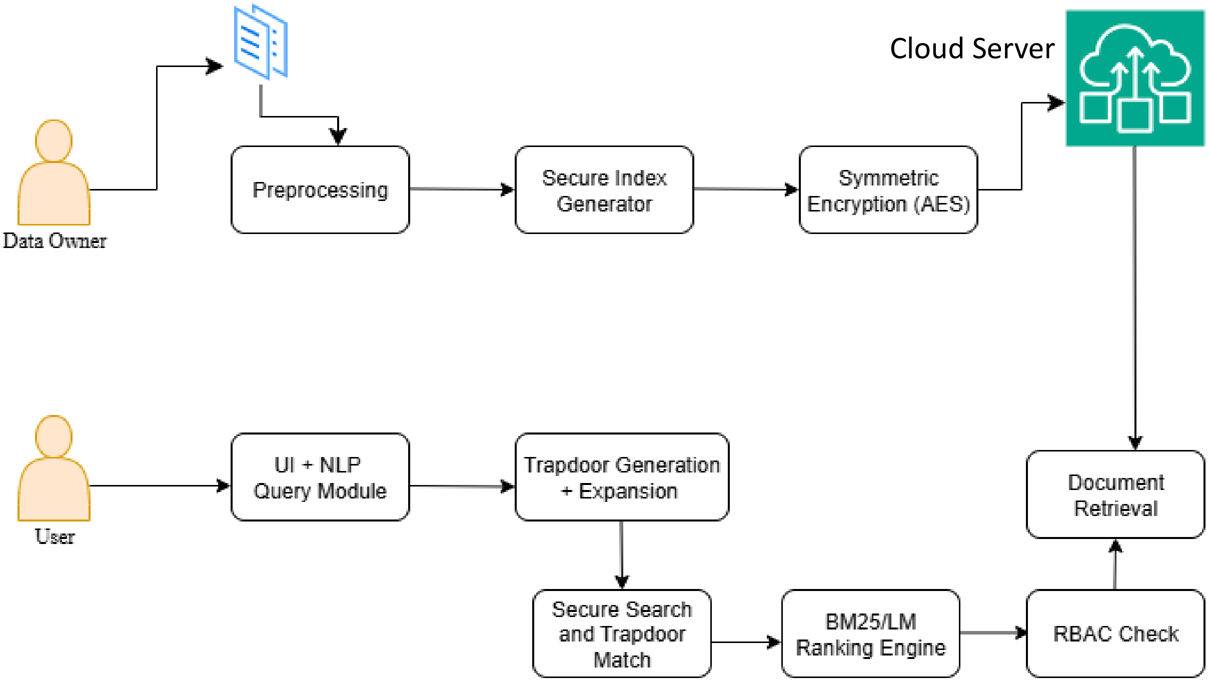


Figure 1: Proposed System Architecture

The Preprocessing Module plays a foundational role in transforming raw textual data into structured input suitable for secure indexing and encrypted retrieval. This module ensures that all documents are consistently cleaned, normalized, and made searchable before encryption.

The system processes documents from the Cranfield Collection, which contains 1,400 scientific documents and 225 sample queries for prototype evaluation. This dataset provides the basis for evaluating the system's retrieval accuracy and performance under real and controlled conditions.

The preprocessing flow includes several steps:

i. Tokenization: Splits each document into individual terms or words.
ii. Stop word removal: Eliminates non-informative words such as "and," "is," and "the."
iii. Lemmatization/Stemming: Reduces words to their base or root forms to enhance term consistency (e.g., "running" becomes "run") to improve consistency in search and indexing.
iv. Keyword extraction: Identifies the most relevant terms in each document using statistical frequency or entropy-based analysis.
v. Text normalization: Converts all texts to lowercase, removes punctuation, and standardizes formatting.

After these steps, the cleaned tokens are passed to the Secure Index Generator, where they are securely hashed or indexed using cryptographic methods. This prepares the documents for encrypted storage and enables trapdoor-based keyword matching during search.

The Preprocessing Module, as a pipeline stage in the architecture, directly influences the effectiveness of secure search by improving keyword quality, index precision, and overall retrieval relevance.

Secure Index Generation:
The system will use extracted keywords to create secure indexes using cryptographic techniques such as SHA-256 hashes. Indexes are generated without exposing the actual content.

Symmetric Encryption:
Full documents are encrypted using symmetric encryption (e.g., AES). Encrypted documents and indexes are linked and stored in cloud storage.

Query Submission:
User inputs a search query through the UI + NLP Query Module. The system preprocesses the query: normalizing text, removing stop words, and optionally expanding semantically.

Trapdoor Generation:
A secure trapdoor (hashed query keywords) is created using searchable encryption. This trapdoor is used to securely match against encrypted indexes.

Ranking:
The matched encrypted documents are ranked using models like BM25 for relevance.

Decryption and Delivery:
Authorized users receive the encrypted documents. The documents are decrypted locally using keys assigned to the user's role. Final decrypted results are displayed to the user.

IV. SYSTEM SETUP

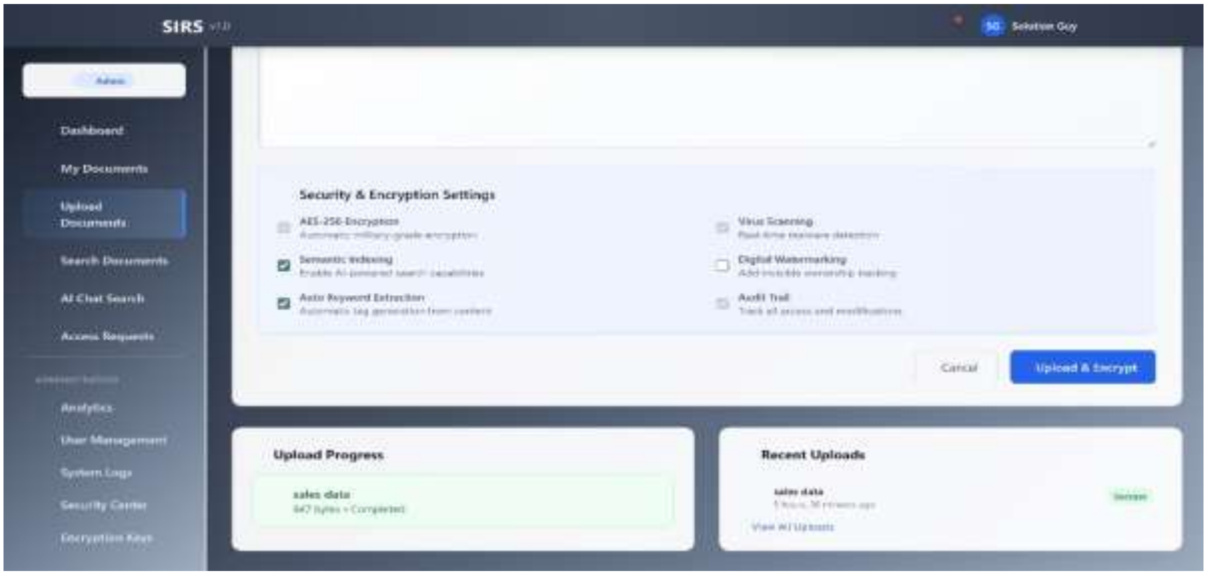
For the system set up, Google Colab is used to run and test the system in the cloud. It is accessible using any modern web browser. Raw similarity scores produced by two systems were calibrated into probability-like confidence values in the range [0,1]. For the proposed system, the BM25 ranking model was used with fixed parameters $K_1 = 1.5\ and\ b = 0.75$. The other system employed keyword-based matching scores. To convert these raw scores into calibrated confidences, a logistic calibration function of the form $\hat{p} = \sigma(a.s + b)$, was applied, where sss is the raw score generated by the system. In plain terms, this function takes the similarity score for a query–document pair and converts it into a probability-like value between 0 and 1, which represents the system's confidence that the document is relevant. The parameter a controls how sharply the function separates relevant from non-relevant documents, while b adjusts the threshold at which a document is considered relevant.

The calibration was trained for up to 50 epochs with early stopping based on validation loss. The dataset was split into 70% training, 15% validation, and 15% test partitions. The Adam optimizer was used with a learning rate of 0.01, batch size of 4,096, and weight decay of $1\ x\ 10^{-4}$. For the proposed system, the best performance was achieved at epoch 32, with a validation loss of 0.134, yielding calibration parameters $a = 2.57\ and\ b = -1.04$.

The learned calibration functions were then applied to all test query–document pairs to produce confidence values. These values were grouped into 20 bins across the interval [0,1].

V. RESULTS

The deployment of the system as a graphic user interface (GUI) has with it different modules for easy use. The system modules include the document management module, where data owners can upload files and configure access permissions, to ensure proper organization, control and accountability in document handling. The system also include privacy-preserving search interface, which enables users to query encrypted documents using trapdoor-based, privacy-preserving mechanisms, and encryption settings, which support AES-256 encryption, semantic indexing, auto keyword extraction, real-time malware detection, digital watermarking, and audit trails to strengthen document security and integrity during storage, among others. Samples of these interfaces are shown in figure 2 and figure 3.



Security & Encryption Settings

- □ AES-256 Encryption — Automatic military-grade encryption
- ☑ Semantic Indexing — Enable AI-powered search capabilities
- ☑ Auto Keyword Extraction — Automatic tag generation from content
- □ Virus Scanning — Real-time malware detection
- □ Digital Watermarking — Add invisible ownership tracking
- □ Audit Trail — Track all access and modifications

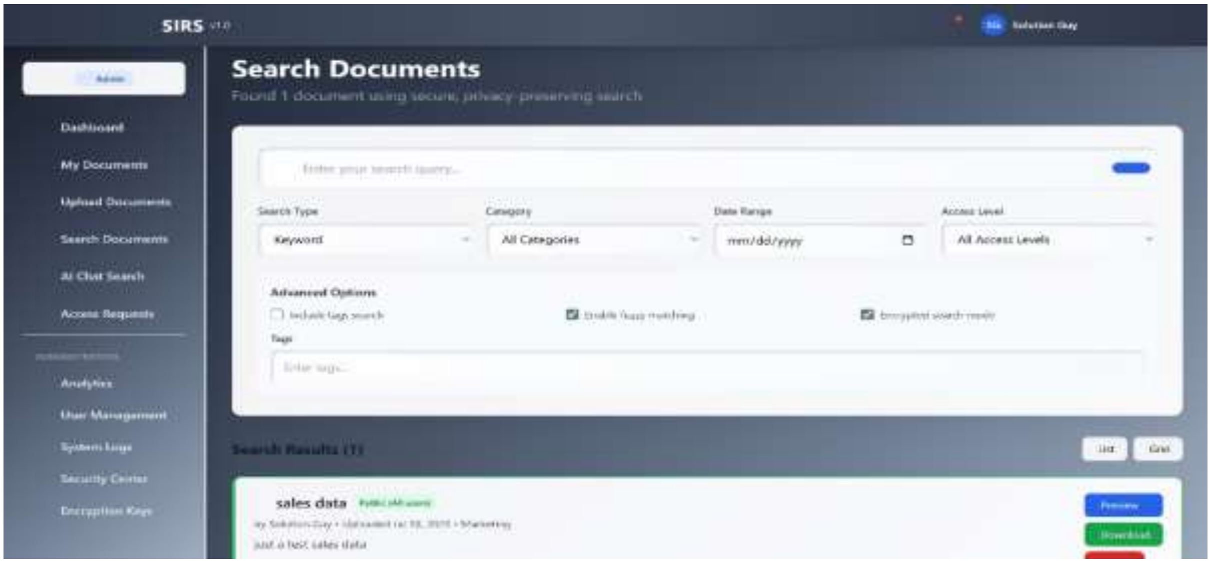
Upload Progress

sales data
647 bytes • Completed

Recent Uploads

sales data

Figure 2: Upload and Encryption Interface



Search Documents

Found 1 document using secure, privacy-preserving search

Enter your search query...

Search Type: Keyword
Category: All Categories
Date Range: mm/dd/yyyy
Access Level: All Access Levels

Advanced Options

- □ Include tags search
- ☑ Enable fuzzy matching
- ☑ Encrypted search mode

Tags: Enter tags...

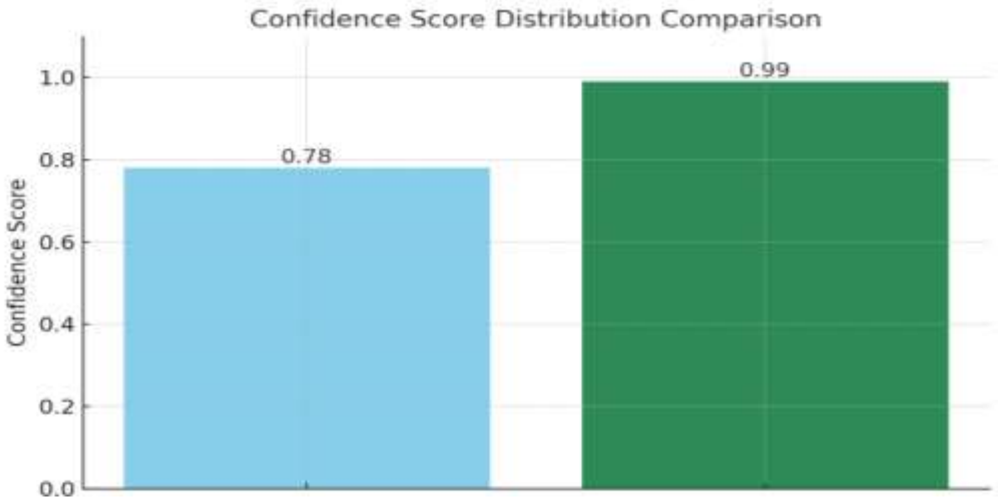
Search Results (1)

sales data

just a test sales data

Figure 3: Document Search Interface

Figure 4 shows the score distribution comparison and illustrates how confidently the system predicts the probability of a document or query belonging to a certain class (e.g., relevant vs. non-relevant). It shows more confident predictions of the proposed system with peaks near 0.1 and 0.9, as against the existing system with scores peaking around 0.5. Table 1 shows a Summary of Calibration Training Results for Existing and Proposed Systems.



Existing System **Proposed System**

Figure 4: Score Distribution Comparison of the Prediction Confidence of the Existing System Versus the Proposed System.

Table 1: Summary of Calibration Training Results for Existing and Proposed Systems

System	Best Epoch	Validation Loss	Parameter a	Parameter b
Existing	28	0.241	1.12	-0.52
Proposed	32	0.134	2.57	-1.04

Figure 5 presents a comparative analysis of precision, recall, and F1-score across different classes for both the proposed and baseline systems. The ROC curve comparison between the proposed and baseline systems in figure 6 illustrates the difference in retrieval performance. The proposed system achieves an Area Under the Curve (AUC) of 1.00, indicating near-perfect distinction between relevant and irrelevant encrypted documents.

The results in Table 2 demonstrate clear improvements of the proposed system over the existing system. The proposed system achieves faster document retrieval, indicating greater efficiency in handling search operations. It also requires less storage space, which reflects the optimization introduced through encryption and secure indexing mechanisms. In addition, the computational overhead is significantly reduced, making the system more resource-efficient. These enhancements collectively contribute to better scalability, ensuring that the system can effectively manage larger document collections in a cloud environment.

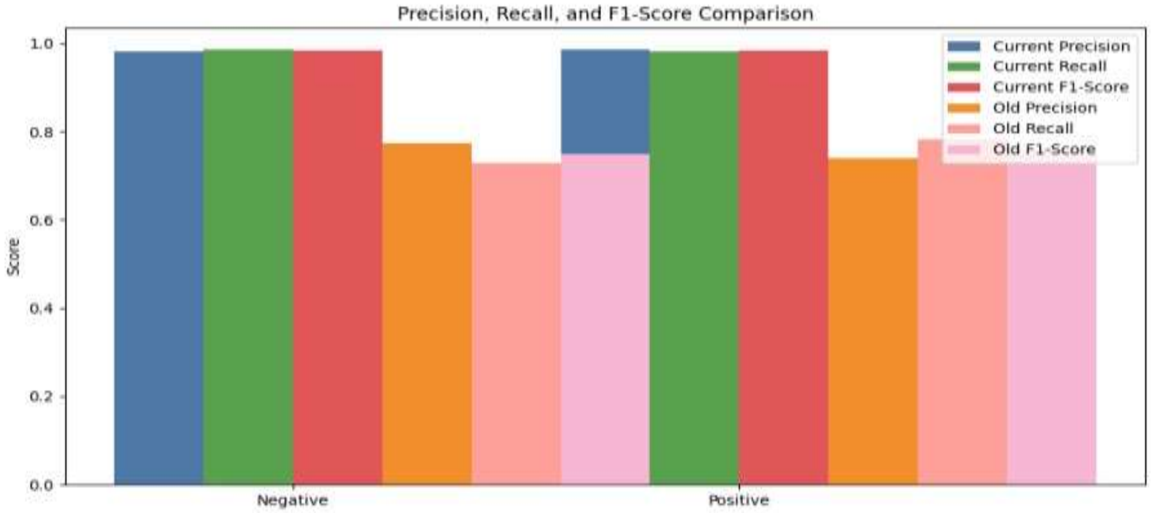


Figure 5: Precision, Recall, and F1-Score Comparison

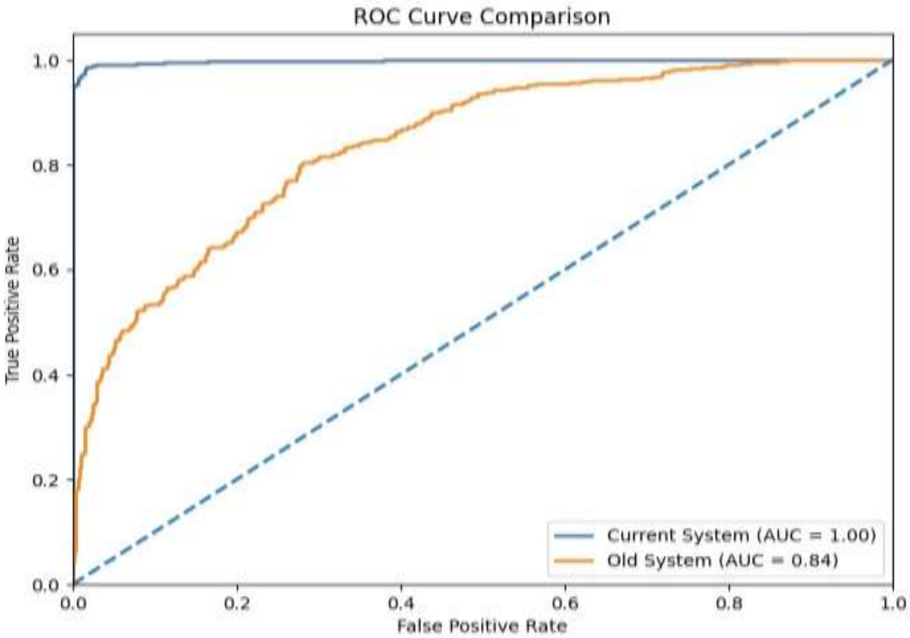


Figure 6: ROC Curve Comparison

Table 2: Comparison of the results between the existing system and the proposed system

Metric / Feature	Design of Secure Product Information Retrieval System in Cloud Computing	Cloud-Based Secure Information Retrieval System
Encryption Technique	Paillier encryption	AES-256 encryption
Search Technique	Encrypted keyword search without ranking	Searchable encryption with trapdoor and BM25 ranking
Indexing	No explicit secure indexing is mentioned	Bloom filter-based secure index generation
Access Control	Basic user-level access	Role-Based Access Control (RBAC)
Security Features	Data confidentiality	Malware detection, audit trail, watermarking
Scalability & Real-World Readiness	Not tested for large-scale deployment	Real-time support with cloud integration
Prediction confidence	0.78	0.99

VI. DISCUSSION OF RESULTS

The system GUI offers a comprehensive overview of the secure information retrieval platform's operational status. It highlights key metrics such as the total number of documents, daily searches, access requests, and a computed security score. The total number of documents indicates the volume of data currently uploaded, encrypted, and indexed within the system. The "Searches Today" metric captures user interaction frequency and gives a measure of how often the system is queried, thereby reflecting system utilization and engagement. The "Access Requests" section records document retrieval attempts, helping administrators track the effectiveness of role-based access control mechanisms by monitoring who accesses which documents and how often. The "Security Score" is a computed value that aggregates indicators such as successful access matches, failed authorization attempts, and encryption consistency.

The confident predictions with peak near 0.9 helps to evaluate whether the secure information retrieval system (e.g., using encryption and indexing) maintains high predictive certainty and supports analysis of how improvements in preprocessing, trapdoor generation, and ranking algorithms (e.g., BM25) affect decision boundaries. The low false positive and false negative rates shown in the confusion matrix reflect strong precision and recall. This supports the effectiveness of the secure indexing and ranking mechanism in maintaining high retrieval performance, even when integrated with cryptographic protections. As shown in the comparative analysis, the proposed system achieves consistently high performance, with values reaching up to 0.99 across all metrics. This comparison highlights the improved effectiveness of the proposed system in retrieving relevant encrypted documents while minimizing false results. The performance gains are attributed to improved preprocessing, secure indexing, relevance ranking using BM25, and enhanced search logic. The results confirm that the proposed system more accurately identifies and retrieves documents while maintaining data security, addressing the limitations of the previous approach.

The ROC curve closely follows the top-left corner, signifying a high true positive rate and minimal false positives across threshold values. In contrast, the baseline system records an AUC of 0.84, reflecting occasional retrieval errors and less consistent classification. This performance gap validates the enhancements introduced in the secure index matching, trapdoor generation, and relevance ranking. The result confirms that the proposed system enables more reliable and secure document retrieval, which is critical in privacy-preserving cloud environments.

VII. CONCLUSION

This paper proposed, designed, and implemented a secure cloud-based information retrieval system that addresses the critical challenges of data confidentiality, search efficiency, and access control in cloud environments. At the core of the system is the integration of searchable encryption techniques, which enable encrypted document storage while preserving the ability to perform secure and efficient keyword-based searches. This dual functionality ensures that data remains protected from unauthorized access while still being retrievable in a meaningful and user-intended manner. The system empowers data owners to upload their documents through a preprocessing module that standardizes and refines the content before encryption. The processed documents undergo encryption using AES-256, a robust and widely adopted symmetric encryption algorithm known for its high level of security. In parallel, secure indexes are generated using Bloom filters, which allow for fast and privacy-preserving keyword matching without revealing actual content to the cloud provider or unauthorized users.

For the end-user, the system enables keyword searches through a Natural Language Processing (NLP) interface. The search queries are transformed into encrypted trapdoors using a cryptographic hash function. These trapdoors are matched against the secure indexes in the cloud to identify relevant documents. To enhance the accuracy and relevance of the search results, the system incorporates the BM25 ranking algorithm, which evaluates document relevance based on term frequency and inverse document frequency.

Access to retrieved documents is controlled through Role-Based Access Control (RBAC), which ensures that only users with valid permissions can decrypt and view specific documents. The combination of encryption, indexing, and access control makes the system robust against data leakage, unauthorized access, and privacy violations. In addition, the system includes real-time access management and encryption key control, enabling administrators to manage encryption keys and define user roles dynamically. These features make the platform adaptable to various organizational needs where sensitive data handling is a priority, such as in legal, healthcare, academic, and financial institutions.

REFERENCES

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... & Zaharia, M. (2010). A view of cloud computing. *Communications of the ACM, 53*(4), 50-58.

Wang, C., Chow, S. S. M., Wang, Q., Ren, K., & Lou, W. (2018). Privacy-preserving public auditing for secure cloud storage. *IEEE Transactions on Computers, 62*(2), 362-375.

Kumar, P., & Tripathi, R. (2019). Secure storage and access of data in cloud computing. *Advances in Electronics, Communication and Computing*, 21-33. Springer.

Goh, E. J. (2003). Secure indexes. *Cryptology ePrint Archive*.

Cao, N., Wang, C., Li, M., Ren, K., & Lou, W. (2013). Privacy-preserving multi-keyword ranked search over encrypted cloud data. *IEEE Transactions on parallel and distributed systems*, *25*(1), 222-233.

Liang, Y., Li, Y., Zhang, K., & Wu, Z. (2023). VMSE: Verifiable multi-keyword searchable encryption in multi-user setting supporting keywords updating. *Journal of Information Security and Applications*, *76*, 103518.

Saniya, M., Udaya, G. S., Kiran, B. & Babu, G. U. (2023). Design of secure product information retrieval system in cloud computing. *International Research Journal of Modernization in Engineering Technology and Science, 5*(7), 350-355.

BalaKrishna, N., Sakthivel, M., Ushasri, G. S., Thanmai, J. N., Kumar, K. B., & Swamy, G. V. (2024). Enhancing Cloud Image Retrieval Efficiency through Secure Optimization. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (Vol. 1, pp. 1-5). IEEE.