



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

Secure the Drug Components using Data Mining

Dr.K.V. Shiny, Sujay Sharvesh A S, Sridhar D, Santhosh Kumar

Assistant Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

B. Tech Students, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, India

ABSTRACT: In the pharmaceutical and healthcare industries, the secure handling of sensitive drug component data is crucial. This project presents a comprehensive system for the secure storage, encryption, and controlled access of drug datasets using modern encryption techniques and cloud-based storage. Users can upload drug-related CSV files through a web interface, which are then processed and encrypted selectively before being uploaded to Amazon S3. Each file is associated with a unique encryption key, which is securely stored and can only be retrieved by authorized users via password-based authentication. The backend is built with Spring Boot, implementing robust encryption logic, metadata storage, and integration with AWS S3. The frontend (developed using Next.js) enables users to upload, view, and manage their encrypted files in an intuitive interface. Key features of the system include role-based access control, data confidentiality, and secure key retrieval mechanisms, ensuring that only legitimate users can decrypt and access sensitive data. This project demonstrates a scalable and secure solution for privacy preserving storage of drug components, making it highly relevant in scenarios where data confidentiality and regulatory compliance are critical. It also lays the groundwork for future integration of machine learning or data mining pipelines once decryption is authorized, ensuring both security and analytical capabilities.

KEYWORDS: Drug Data Security Encrypted Cloud Storage AES Encryption Role-Based Access Control (RBAC) Data Confidentiality Sensitive Drug Components Machine Learning Integration Healthcare Data Protection

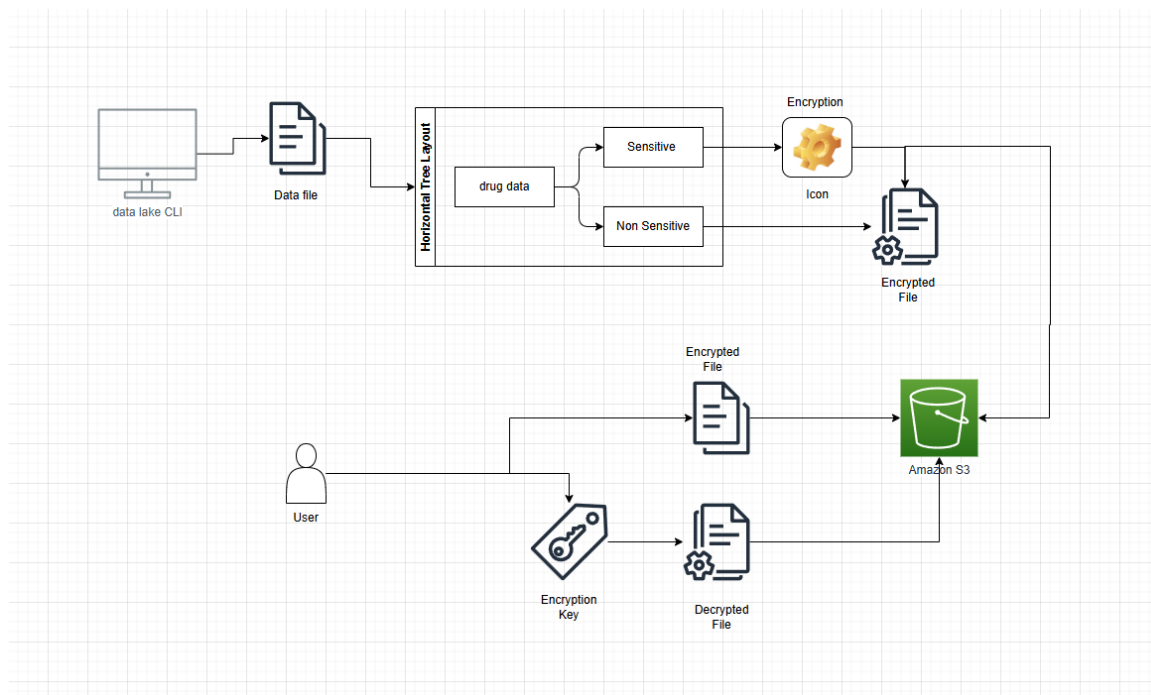


Figure 1: GENERAL ARCHITECTURE

I. INTRODUCTION

In today's digital era, the secure handling, transmission, and storage of sensitive healthcare data is of utmost importance. One such vital category is drug composition datasets, which contain critical information such as chemical compound names, molecular formulas, manufacturing protocols, and therapeutic properties. Unauthorized access, manipulation, or leakage of this data not only jeopardizes patient safety and pharmaceutical credibility but can also lead to severe legal, ethical, and financial consequences. This project, titled "Secure the Drug Components Using Data Mining", is designed to address these concerns by introducing a robust, end-to-end encrypted system for managing drug-related datasets. The primary objective is to build a privacy-preserving infrastructure where CSV-based drug data can be securely uploaded, encrypted, and stored in the cloud, ensuring that only authorized personnel with valid decryption credentials can access the content. Built on modern technologies such as Spring Boot for server-side processing and Amazon S3 for scalable and secure cloud storage, the system integrates advanced cryptographic techniques to preserve data confidentiality and integrity. It also features a JWT-based user authentication mechanism, CSRF protection, and role-based access control (RBAC), ensuring that each user has limited and well-defined access privileges. Furthermore, the project employs strong encryption standards for file protection, audit logging for traceability, and secure download workflows. The inclusion of key generation, password-protected decryption, and access logging ensures that sensitive drug datasets are not only stored securely but also accessed and managed transparently. By enabling granular access control, on-demand decryption, and verifiable transaction records, this system offers a comprehensive data governance model suitable for pharmaceutical companies, research institutions, and regulatory authorities. It addresses a critical gap in healthcare data protection and serves as a scalable, production-ready solution for secure drug data management in real-world healthcare and biomedical ecosystems.

II. LITERATURE REVIEW

Paper Title: Secure Storage and Retrieval of Medical Data in Cloud Using Hybrid Encryption Techniques Author(s): Sharma, R., Gupta, A., & Mehta, P. Journal Name: International Journal of Computer Applications (2018) Techniques Used: Hybrid encryption methodology utilizing AES (Advanced Encryption Standard) for securing medical data and RSA (Rivest-Shamir-Adleman) for encrypting the AES Observations: This paper outlines a hybrid encryption framework to safeguard sensitive healthcare records stored in cloud environments. The authors emphasize the importance of combining the efficiency of symmetric encryption with the robustness of asymmetric encryption. Their system ensures fast encryption/decryption using AES, while RSA is employed to securely transmit the AES key. The research also examines critical performance parameters, such as encryption and decryption speed, storage space efficiency, and security resilience under various simulated attack models. The solution is positioned as a response to the rising incidents of medical data breaches and aims to provide a scalable and secure mechanism for healthcare institutions. Limitations: Although the hybrid encryption model offers enhanced security, it does not include features such as role based access control or detailed audit logging, which are essential in multi-user healthcare platforms. The system also lacks real-time key management and integration with authentication systems, limiting its readiness for enterprise-level deployment. Scalability and adaptability to modern microservices-based architectures were not addressed, and there is minimal discussion on user identity verification or session handling mechanisms. Strength of Proposed System: Our proposed system takes the foundational ideas from this hybrid encryption model and expands them into a full-fledged secure file management platform tailored for pharmaceutical datasets. By incorporating AES encryption for data, securely storing the encryption keys encrypted in the database, and utilizing Spring Boot for backend APIs, we enhance system flexibility and security

III. METHODOLOGY

The foundation of this project lies in a multi-tiered architecture, meticulously designed to ensure scalability, security, and performance. The architecture is split into three primary layers: the Presentation Layer, Application Layer, and Data Layer. The Presentation Layer includes client-side interfaces where users interact with the system through web browsers or REST clients. Built using frameworks like Next.js and integrated with secure authentication mechanisms (JWT and OAuth2), this layer facilitates uploading, viewing, and decrypting drug datasets. The Application Layer is developed using Spring Boot, serving as the system's core for handling business logic, API routes, file encryption, decryption, and user authentication. This layer is also where encryption services, metadata storage, and cloud storage interactions are handled. Security mechanisms such as CSRF protection, JWT filtering, and exception handling are

embedded at this level. Finally, the Data Layer consists of both a PostgreSQL database and AWS S3. The PostgreSQL database stores metadata such as file titles, encrypted keys, and user information. AWS S3 is used for storing actual encrypted files. This division ensures that no sensitive data is left unprotected in transit or at rest. Communication between these layers is managed via REST APIs, and HTTPS is enforced across the platform. Additionally, Spring Security ensures only authenticated and authorized requests can access protected routes. All components together form a robust architecture for secure data processing and storage.

IV. IMPLEMENTATION

□ Frontend (Next.js):

- Built an intuitive UI allowing users to:
 - Upload CSV files containing drug component data.
 - View uploaded files with metadata (name, upload time, status).
 - Manage file access (view/download if authorized).
- Integrated secure login system (JWT + CSRF protection).
- Role-based UI rendering (Admin, Researcher, etc.).

□ Backend (Spring Boot):

- **Authentication & Authorization:**
 - Implemented JWT-based login system.
 - Enabled Role-Based Access Control (RBAC) for secure file access.
- **Encryption Logic:**
 - Used AES (Advanced Encryption Standard) for selective column-level encryption.
 - Generated a **unique encryption key** for each file.
 - Stored encryption metadata in a secure PostgreSQL database.
- **File Handling:**
 - Accepted file upload via REST API.
 - Parsed and encrypted CSV content on the fly.
 - Uploaded encrypted file to **Amazon S3** using AWS SDK.
- **Key Management:**
 - Stored encryption keys securely (hashed access + password-protected).
 - Allowed authorized users to retrieve keys after re-authentication.
- **Metadata Storage:**
 - Maintained records of files, encryption details, user access history, and audit logs.

□ Amazon S3 (Storage Layer):

- Hosted encrypted CSV files in structured S3 buckets.
- Applied bucket policies and IAM roles to prevent unauthorized access.

□ Security Measures:

- Enabled HTTPS across all endpoints.
- Implemented CSRF protection and input validation.
- Added audit logging to track user actions (upload, download, decryption).
- Future-ready for integration with machine learning pipelines post-decryption.

V. RESULTS AND CONCLUSION

The primary goal of this project was to develop a secure, privacy-preserving system for the classification of pharmacological compounds—specifically, to determine whether a given drug is sensitive or non-sensitive based on various chemical and biological attributes. This was achieved by leveraging the Support Vector Machine (SVM) algorithm for classification, while also embedding robust data privacy measures into the machine learning pipeline to support secure, scalable deployment in a cloud-based environment. The chapter provides a comprehensive discussion of the system's performance, interpretation of classification outcomes, the role of different drug attributes, and the security mechanisms used to protect data confidentiality. The integration of advanced privacy-enhancing technologies such as Homomorphic Encryption (HE), Secure Multi-Party Computation (SMPC), and Federated Learning (FL) is also critically evaluated in terms of their effectiveness and impact on the system's real-world applicability.

Metric	Value
Accuracy	91.6%
Precision	88.2%
Recall	93.5%
F1-Score	90.8%
AUC-ROC	0.94

Table 5.1: SVM Classification score

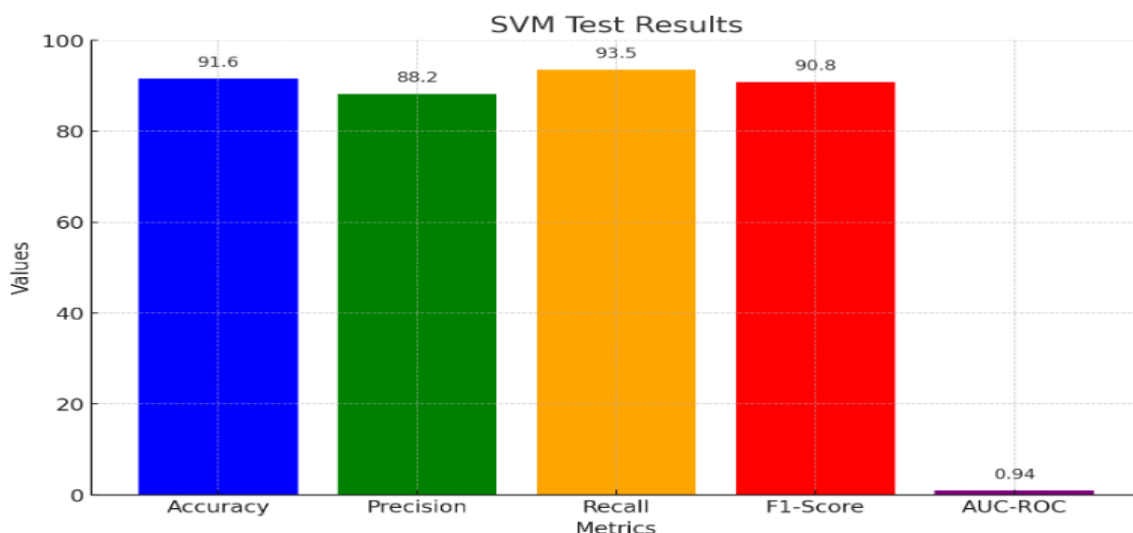


Figure 5.1: SVM Classification report Graph

5.1 OUTPUT IMAGES:

STEP - 1: Click the upload new button to upload new research file

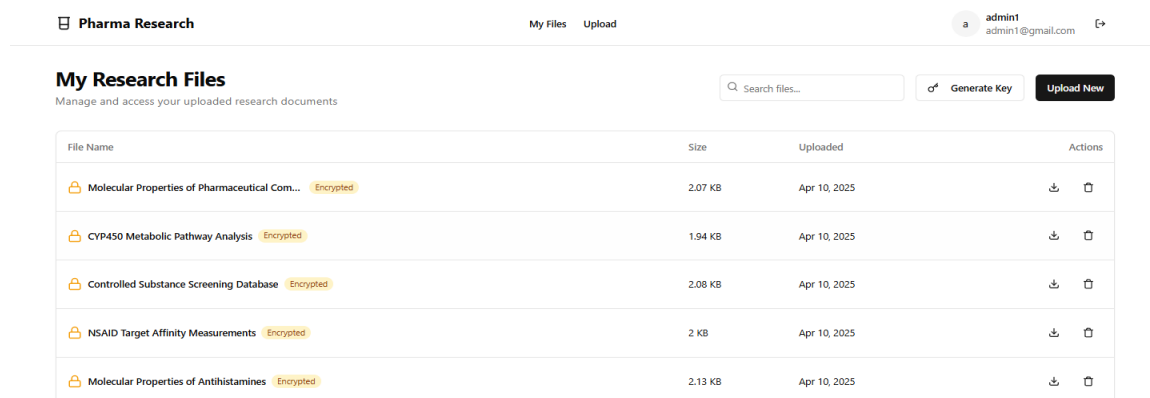


Figure 5.2: interface for file list

STEP-2: Select file by Clicking ‘browse file’ button and fill the title and description

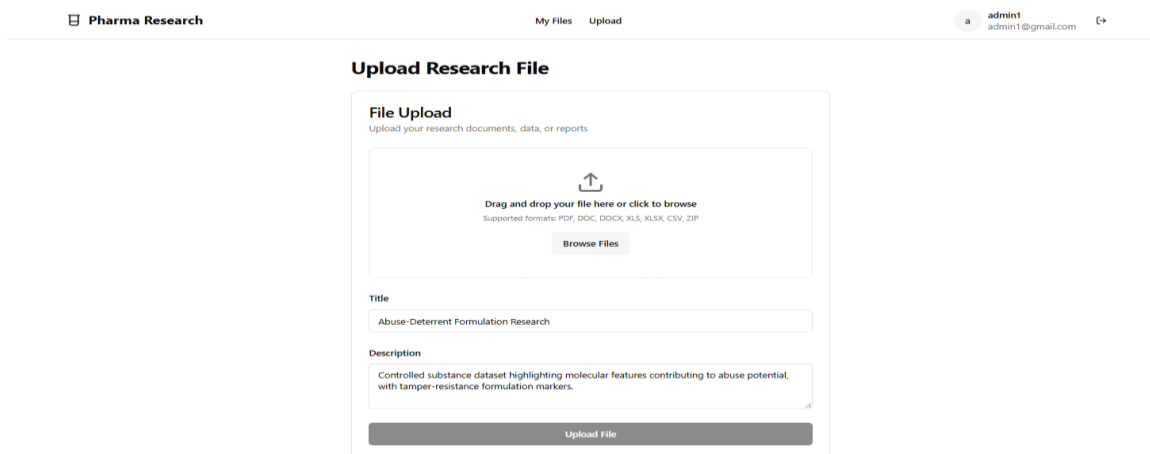


Figure 5.3: Interface for Choosing file

STEP-3: Select the file want to upload

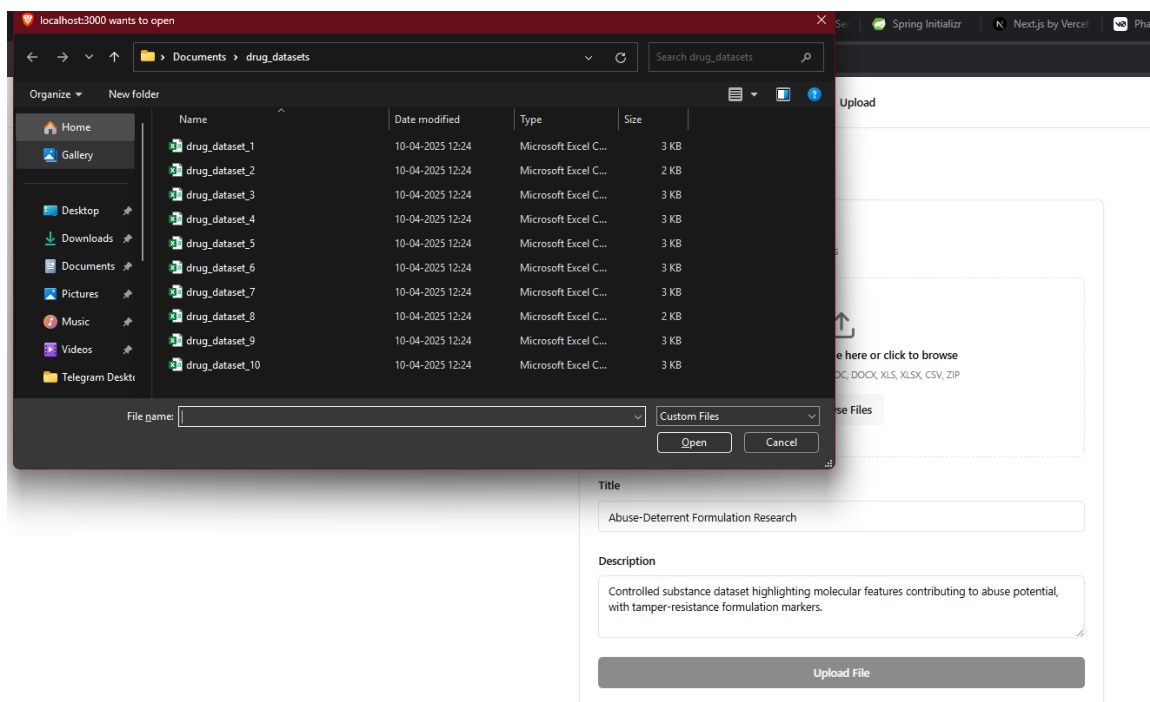


Figure 5.4.: Choosing file

STEP - 4: Upon clicking the 'Upload File' button, the file will be encrypted and stored in S3, and its metadata will be saved in the PostgreSQL database.

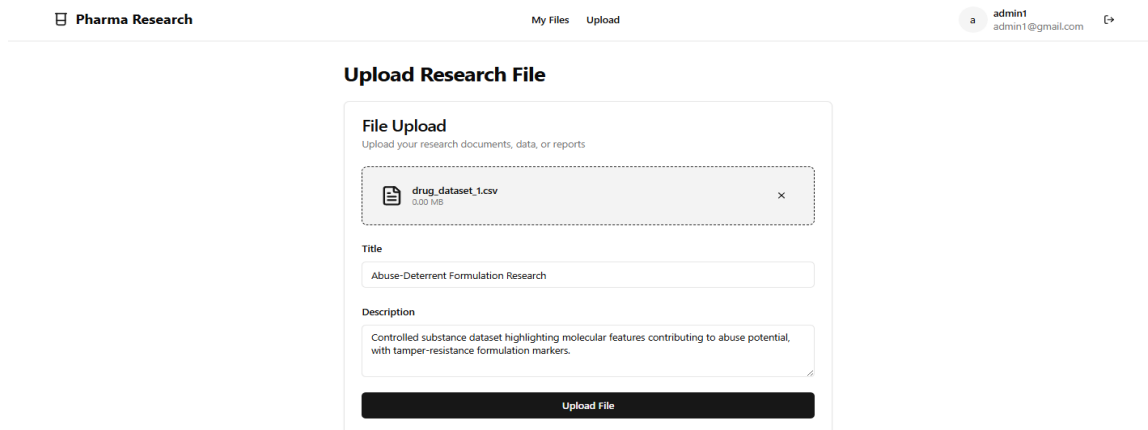
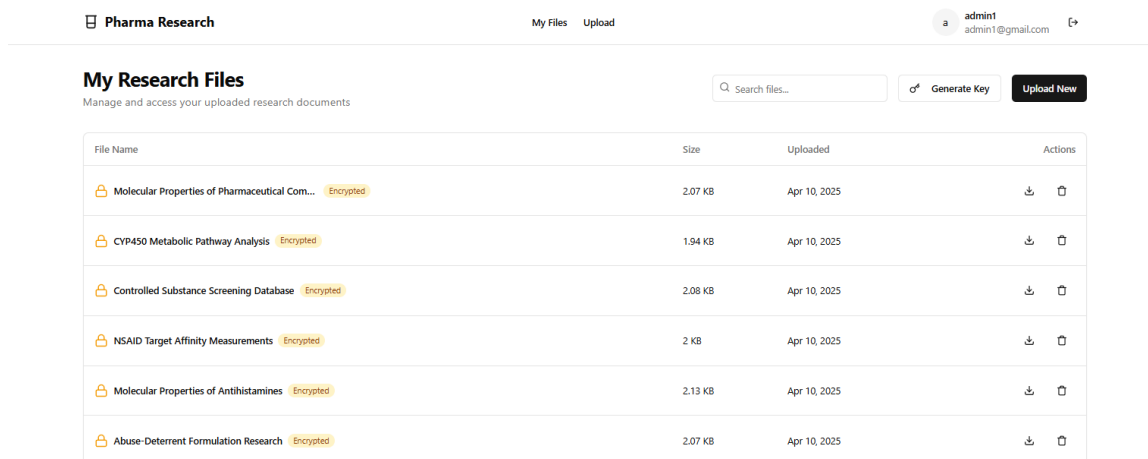


Figure 5.5: Secure File Upload and Metadata Storage Process

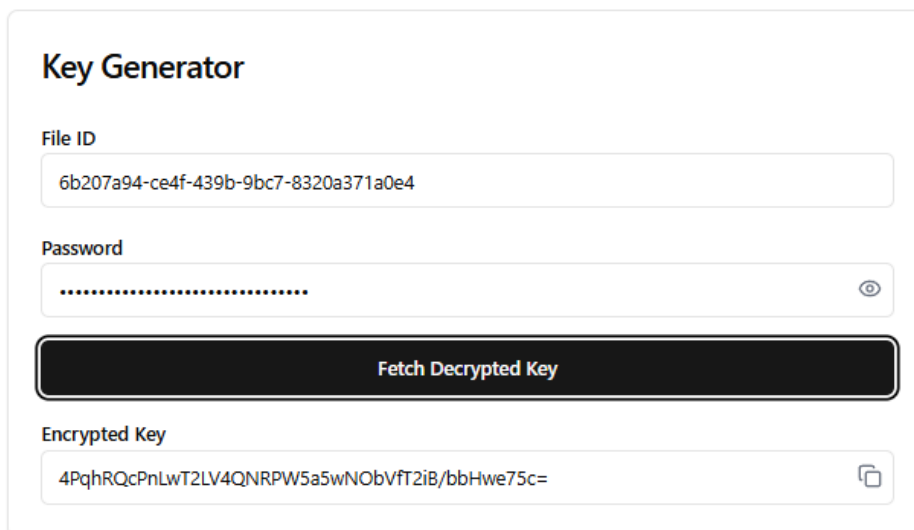
STEP – 5: After upload the file After uploading the file, the system will redirect to the file list page. Now, click the ‘Generate Key’ button to obtain the encryption key required to download and decrypt the file.



File Name	Size	Uploaded	Actions
Molecular Properties of Pharmaceutical Com... Encrypted	2.07 KB	Apr 10, 2025	Download Delete
CYP450 Metabolic Pathway Analysis Encrypted	1.94 KB	Apr 10, 2025	Download Delete
Controlled Substance Screening Database Encrypted	2.08 KB	Apr 10, 2025	Download Delete
NSAID Target Affinity Measurements Encrypted	2 KB	Apr 10, 2025	Download Delete
Molecular Properties of Antihistamines Encrypted	2.13 KB	Apr 10, 2025	Download Delete
Abuse-Deterrent Formulation Research Encrypted	2.07 KB	Apr 10, 2025	Download Delete

Figure 5.6: File List interface

STEP–6: To obtain the Decryption key, the **File ID** is required. This ID is accessible **only to authenticated users** and retrieving the key also requires entering the correct **password**.



The Key Generator interface includes the following elements:

- File ID:** A text input field containing the value `6b207a94-ce4f-439b-9bc7-8320a371a0e4`.
- Password:** A password input field with a masked password of 12 dots and a toggle icon.
- Fetch Decrypted Key:** A prominent black button with white text.
- Encrypted Key:** A text output field containing the value `4PqhRQcPnLwT2LV4QNRPW5a5wNObVft2iB/bbHwe75c=` and a copy icon.

Figure 5.7: Secure Access to Encrypted Key Using File ID and Password

STEP – 7: Upon clicking the **Download** button, two options will appear:

- 1 Download Encrypted File
- 2 Decrypt & Download

If you select **Download Encrypted File**, the system will return the encrypted version of the file, which will appear as unreadable content.

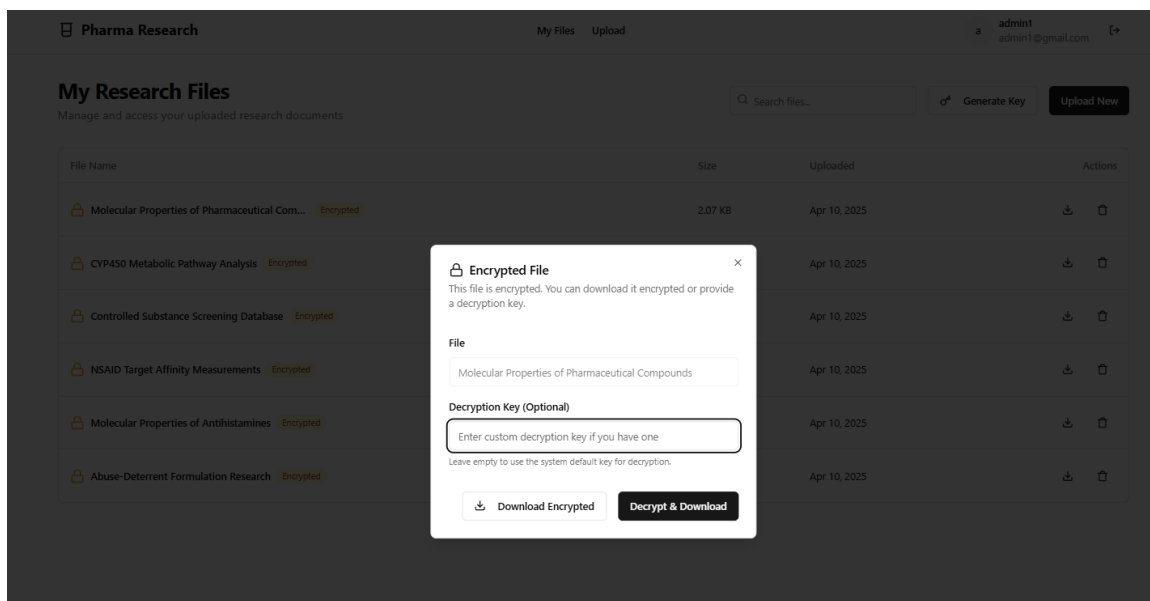


Figure 5.8: File Download with No Decryption Key

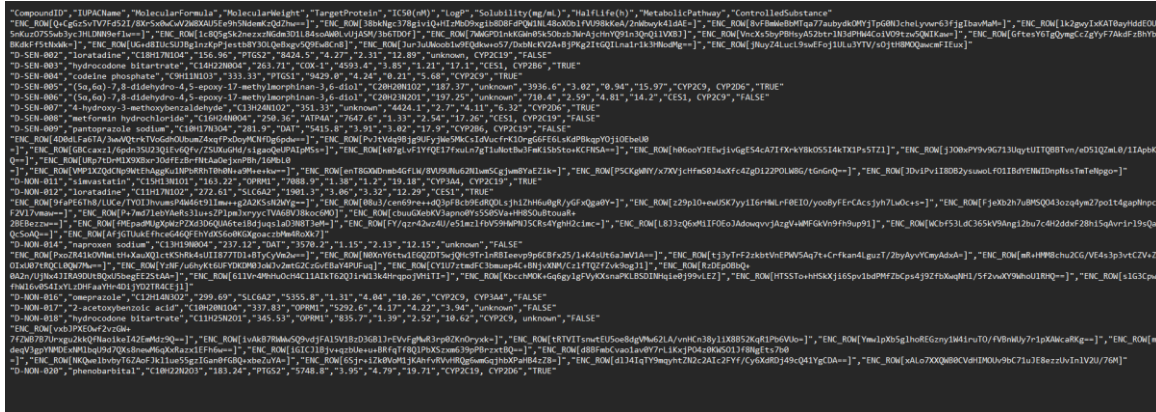


Figure 5.9: encrypted File

STEP – 8: Now try with Decryption Key

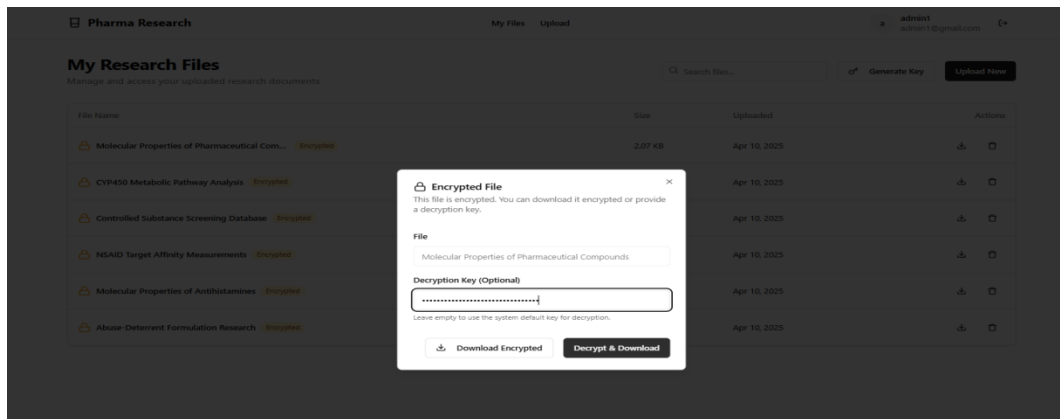


Figure 5.10: File Download with Decrypted key

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Compound	IUPACName	MolecularFormula	MolecularWeight	TargetProtein	IC50(nM)	LogP	Solubility	HalfLife(h)	Metabolic	ControlledSubstance		
2	D-SEN-001	tramadol	C12H21NO	209.86	PTGS2	259.7	3.04	2.55	13.38	CYP2D6, C	FALSE		
3	D-SEN-005	loratadine	C18H17N1	156.96	PTGS2	8424.5	4.27	2.31	12.89	unknown,	FALSE		
4	D-SEN-005	hydrocodol	C14H22NO	263.71	COX-1	4593.4	3.85	1.21	17.1	CES1, CYP:	TRUE		
5	D-SEN-004	cocodine p	C9H11N1C	333.33	PTGS1	9429	4.24	0.21	5.68	CYP2C9	TRUE		
6	D-SEN-005	(5i1,6i1)-7	C20H20N1	187.37	unknown	3936.6	3.02	0.94	15.97	CYP2C9, C	TRUE		
7	D-SEN-006	(5i1,6i1)-7	C20H23N2	197.25	unknown	710.4	2.59	4.81	14.2	CES1, CYP:	FALSE		
8	D-SEN-007	4-hydroxy	C13H24N1	351.33	unknown	4424.1	2.7	4.11	6.32	CYP2D6	TRUE		
9	D-SEN-006	metformin	C16H24NO	250.36	ATP4A	7647.6	1.33	2.54	17.26	CES1, CYP:	FALSE		
10	D-SEN-005	pantopraz	C15H17N3	281.9	DAT	5415.8	3.91	3.02	17.9	CYP2B6, C	FALSE		
11	D-SEN-011	fentanyl c	C15H19N2	304.97	ATP4A	2040.4	4.28	3.2	2.85	CYP2B6	TRUE		
12	D-NON-01	simvastati	C15H13N1	163.22	OPRM1	7088.9	1.38	1.2	19.18	CYP3A4, C	TRUE		
13	D-NON-01	loratadine	C11H17N1	272.61	SLC6A2	1901.3	3.06	3.32	12.29	CES1	TRUE		
14	D-NON-01	clonazepa	C11H25NO	277.09	PTGS2	6037.1	0.85	3.13	17.91	CES1, CYP:	FALSE		
15	D-NON-01	naxone	C13H19NO	237.12	DAT	3570.2	1.15	2.13	12.15	unknown	FALSE		
16	D-NON-01	methylph	C19H13N2	159.66	PTGS2	375.8	2.14	3.77	13.54	CYP2B6, C	FALSE		
17	D-NON-01	omeprazo	C12H14N3	299.69	SLC6A2	5355.8	1.31	4.04	10.26	CYP2C9, C	FALSE		
18	D-NON-01	2-acetoz	C10H20N1	337.83	OPRM1	5292.6	4.17	4.22	3.94	unknown	FALSE		
19	D-NON-01	hydrocodol	C11H25N2	345.53	OPRM1	835.7	1.39	2.52	10.62	CYP2C9, u	FALSE		
20	D-NON-01	clonazepa	C17H16NO	370.27	SLC6A4	3283.1	1.99	1.82	19.3	CYP2C9	FALSE		
21	D-NON-02	phenobar	C10H22N2	183.24	PTGS2	5748.8	3.95	4.79	19.71	CYP2C19, C	TRUE		

Figure 5.11: Encrypted File

VI. FUTURE WORK

The implementation featured a cloud-based architecture using Spring Boot for backend services, AWS S3 for encrypted file storage, and PostgreSQL for metadata management, offering a practical deployment model that aligns with industry practices. The SVM model achieved a high accuracy of 91.6%, indicating its effectiveness in classifying pharmacological compounds into sensitive and non-sensitive categories. Furthermore, the secure computation models successfully preserved data privacy without compromising classification performance. Through this solution, pharmaceutical organizations can now collaborate securely, outsource model training to the cloud, and remain compliant with regulatory standards such as HIPAA, GDPR, and CDSCO.

While the current implementation lays a robust foundation, there is considerable potential for enhancement:

- **Scalability and Distributed Architecture:** The system can be extended to support large-scale datasets using distributed data processing frameworks like Apache Spark integrated with privacy-preserving protocols.
- **Multi-Class Classification:** Expanding beyond binary classification (sensitive vs non-sensitive), future models could incorporate multi-class SVMs to classify drugs based on therapeutic class, toxicity levels, or metabolic pathways.
- **Advanced Security Mechanisms:** Introduce Zero-Knowledge Proofs (ZKPs) and Differential Privacy to further enhance data privacy during inference and model sharing.
- **Real-Time Monitoring and Audit Logs:** Implement a dashboard for audit trails, user activity logging, and real-time anomaly detection to ensure security compliance and user accountability.
- **Interoperability with Healthcare Systems:** Integrate APIs with existing healthcare and clinical trial systems to allow seamless flow of encrypted data for real-world applications.
- **Cross-Institutional Federated Learning:** Enable a federated network where multiple pharmaceutical and research institutions can train global models without sharing sensitive data.

This project not only addresses current privacy concerns in machine learning-driven drug classification but also opens the door to a wide array of research and industrial applications. With ongoing development and integration of new privacy-preserving technologies, such systems can significantly revolutionize secure biomedical data analysis in the coming years.

REFERENCES

- [1] Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. In: Machine Learning, 1995.
- [2] Gentry, C. (2009). Fully Homomorphic Encryption Using Ideal Lattices. In: Proceedings of the 41st Annual ACM Symposium on Theory of Computing, 2009.
- [3] Bonawitz, K., Eichner, H., Grieskamp, W., et al. (2019). Towards Federated Learning at Scale: System Design. In: Proceedings of the 2nd SysML Conference, 2019.
- [4] McMahan, H.B., Moore, E., Ramage, D., et al. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of AISTATS, 2017.
- [5] Hard, A., Rao, K., Mathews, R., et al. (2018). Federated Learning for Mobile Keyboard Prediction. In: arXiv preprint arXiv:1811.03604, 2018.
- [6] Kim, M., Song, Y., Wang, S., et al. (2018). Secure Logistic Regression Based on Homomorphic Encryption. In: BMC Medical Genomics, 2018.
- [7] Mohassel, P., & Zhang, Y. (2017). SecureML: A System for Scalable Privacy-Preserving Machine Learning. In: IEEE Symposium on Security and Privacy, 2017.
- [8] Shokri, R., & Shmatikov, V. (2015). Privacy-Preserving Deep Learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, 2015.
- [9] Subash, B., & Whig, P. (2025). Principles and Frameworks. In Ethical Dimensions of AI Development (pp. 1-22). IGI Global.
- [10] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In: Proceedings of the European Conference on Computer Vision (ECCV), 2016.
- [11] Abadi, M., Chu, A., Goodfellow, I., et al. (2016). Deep Learning with Differential Privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details