INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Modern Secure Electronic Health Framework (M-SEHF) for Collaborative Medical Data Analysis with Privacy-Preserving Techniques

N. Subhalakshmi[1]*, **M. V. Srinath**[1]

1 S.T. E. T Women's College (Autonomous), (Affiliated to Bharathidasan University, Tiruchirappalli), Sundarakkottai, Mannargudi - 614016, Thiruvarur Dt., Tamil Nadu, India

OPEN ACCESS

Received: 18/04/2025

Accepted: 11/05/2025

Published: 30/05/2025

Citation: Subhalakshmi N, Srinath MV (2025) Modern Secure Electronic Health Framework (M-SEHF) for Collaborative Medical Data Analysis with Privacy-Preserving Techniques. Indian Journal of Science and Technology 18(20): 1590-1599. https://doi.org/10.17485/IJST/v18i20.738

* **Corresponding author**.

subha.stet@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: As healthcare data grows exponentially, cloud computing has emerged as a scalable solution for managing and storing electronic health records (EHRs). However, ensuring the confidentiality and integrity of such sensitive information in open and distributed environments remains a pressing concern. Classical Secure Electronic Health Frameworks (SEHFs) frequently use single-layer encryption and static rule-based systems, which are unable to handle huge, unstructured datasets efficiently and do not offer real-time data handling or semantic analysis. **Objective:** This study introduces a Modern Secure Electronic Health Framework (M-SEHF) that ensures end-to-end confidentiality, integrity, and accessibility of medical data in cloud environments. The framework integrates Natural Language Processing (NLP), Decision Tree (DT) algorithms, and hybrid cryptographic techniques to enable secure data classification, processing, and storage. **Methods:** M-SEHF employs NLP-based pre-processing to identify and sanitize Protected Health Information (PHI) from patient records. Decision Tree algorithms classify records by sensitivity level, while a hybrid encryption model Advanced Encryption Standard (AES-256) for data and RSA for secure key exchange ensures privacy during cloud storage and transmission. A custom mapping algorithm securely links encrypted data to its source for traceable access. **Findings:** Evaluated on a dataset of 100,000 anonymized patient records (1 TB), M-SEHF achieved 94.6% accuracy in PHI detection and classification. It reduced average data retrieval latency to 2.3 seconds—30% faster than existing systems, while fully adhering to HIPAA privacy standards. **Novelty and Contribution:** M-SEHF is the first cloud security framework to integrate real-time semantic pre-processing (NLP-DT) with hybrid encryption AES and RSA in a scalable pipeline tailored for healthcare big data. By enabling dynamic sensitivity analysis and multi-layered encryption, it gets beyond the drawbacks of traditional SEHFs and provides better accuracy, speed, and privacy compliance. A strong, adaptable, and future-proof security solution is offered by M-SEHF to meet the changing needs of digital healthcare.

Keywords: Big data; Natural Language Processing (NLP); Cloud Computing; Patients' history; Encryption

1 Introduction

Cloud computing has rapidly transformed the Information and Communication Technology (ICT) landscape by enabling on-demand access to a shared pool of configurable resources including networks, servers, storage, and applications. This model, defined by the National Institute of Standards and Technology (NIST), allows rapid provisioning and minimal management overhead [1]. Cloud computing services are commonly categorized into Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS), each with varying levels of abstraction and control [2]. Despite these advantages, cloud computing introduces new challenges in information security and data privacy, particularly in open and distributed environments. As data volumes increase, the intersection of cloud computing and big data analytics becomes pivotal in sectors such as healthcare, finance, and education. However, this integration also amplifies security vulnerabilities ranging from unauthorized access and Trojan attacks to data breaches necessitating more robust privacy protection and data security mechanisms [3].

Cloud security is compromised by major problems and dangers, hence effective mitigation techniques are required. To improve security in cloud environments, potential methods are investigated [4]. Researchers and governments are taking an interest in big data, which is directly related to computing and is changing sectors including healthcare, education, and business. Workload assessment of national grid big data projects focuses on content recommendations and text classification, emphasizing efficient data management and analysis for optimized big data applications [5].

Big Data is the term used to describe the vast amount of information that grows exponentially over time and is often organized, unstructured, or semi-structured. It comes from a number of sources, such as social media, search engine usage, sensors, healthcare systems, banking transactions, and financial software. Big Data Analytics in Healthcare Systems emphasizes how big data may revolutionize healthcare by enhancing patient outcomes, decision-making, and efficiency. It highlights the necessity of sophisticated analytics for efficient management of intricate healthcare data [6]. The volume and complexity of big data is so extensive that conventional methods of storage or processing cannot manage them efficiently. Big data encompasses a range of technologies and expertise that require innovative integration methods to reveal significant hidden value from various, complex, and extensive datasets. Data transitions into big data when individual data points lose their utility and meaningful insights can only be gleaned from large aggregations of data or the analyses derived from them.

The healthcare domain, in particular, deals with extremely sensitive patient data. While big data analytics promises improved diagnosis, personalized treatment, and operational efficiency, it also raises critical concerns about data confidentiality, integrity, and availability. Encryption and access control are widely used to address these concerns, yet current solutions fall short in terms of adaptability, automation, and semantic understanding of complex medical texts [7], [8]. Moreover, recent studies suggest that existing security frameworks often lack comprehensive natural language processing (NLP) capabilities to pre-process and sanitize patient medical histories. Techniques like Decision Tree (DT) algorithms have not been fully integrated with encryption schemes to ensure end-to-end data protection from ingestion to cloud storage. A major shortcoming is the limited application of context-aware NLP pre-processing prior to encryption, which could significantly reduce data leakage risks during transmission or storage [9].

Although numerous studies have explored data encryption and cloud security mechanisms, a critical analysis reveals several significant gaps that hinder the effectiveness of current approaches, particularly in the healthcare domain. One key shortcoming is the lack of integrated natural language processing (NLP) pre-processing for sensitive medical data prior to encryption and cloud upload. Most existing systems depend on static or rule-based methods, lacking the depth of semantic analysis needed to properly handle complex medical narratives. Additionally, protection mechanisms during intermediate data processing stages, such as classification, cleansing, and format normalization are often inadequate, leaving data temporarily exposed to potential threats. There is also limited research on combining Decision Tree (DT) algorithms with NLP techniques for structured pre-processing of medical histories, despite the potential for enhanced data organization and interpretation. Moreover, current frameworks fail to implement dynamic mechanisms for identifying and eliminating semantically redundant information before encryption, which could otherwise reduce cloud storage and processing overhead. Finally, there is an absence of comprehensive, end-to-end frameworks that seamlessly integrate intelligent pre-processing, robust encryption, and secure storage particularly those designed to adapt to the scale and complexity of big healthcare data.

The objectives of the proposed system are outlined below:

1. Uses NLP and Decision Tree algorithms to pre-process patient records.
2. Applies advanced encryption techniques before storing the data in the cloud.
3. Ensures a secure, privacy-preserving data pipeline in cloud-based big data environments.

While the integration of big data and cloud technologies in healthcare holds great potential for improving efficiency and patient outcomes, current security implementations are often fragmented and lack cohesion. Many existing frameworks emphasize

data protection after encryption, neglecting critical vulnerabilities that occur during the pre-processing stages such as data cleaning, classification, and format normalization. This leaves sensitive patient information exposed prior to being securely stored. Furthermore, very few studies effectively combine Natural Language Processing (NLP) for document sanitization with structured decision logic, such as Decision Tree algorithms, to enhance the security and contextual understanding of medical records before encryption. Another significant shortcoming in the literature is the lack of rigorous performance validation and benchmarking, which makes it challenging to evaluate the practical effectiveness and scalability of proposed systems in real-world healthcare environments.

Security, privacy, and confidentiality in health record systems ensure that individuals control access to their personal information while restricting data access exclusively to authorized personnel[10]. Unauthorized exchange of sensitive medical data can lead to serious breaches, compromising patient privacy and exposing healthcare systems to legal and financial consequences. Although legitimate access to patient records is often required by governments, pharmaceutical companies, and researchers, this demand can sometimes result in intentional misuse by healthcare providers during processes like credentialing or centralized monitoring[11], [12].

The three most important requirements for assuring information generation safety are secrecy, integrity, and availability[13]. Secrecy is limiting access to data to only authorized personnel, which is critical throughout storage, transfer, and processing. This protects sensitive data from illegal access and maintains its integrity throughout its existence[14]. While technological solutions like encryption and access control, along with ethical standards such as professional discretion, help safeguard confidentiality, encryption is still less commonly applied to health data stored on mobile devices and other storage media[15]. To improve healthcare data privacy, a novel hybrid cloud framework is proposed that combines dynamic access control methods with advanced encryption techniques such as homomorphic and elliptic curve cryptography. Real-time monitoring and machine learning-based permission management are implemented to ensure secure and compliant data handling. This comprehensive approach enables scalable and regulation-aligned protection of sensitive healthcare data in cloud environments[16].

Multiple healthcare organizations can work together to analyze medical data using a federated learning framework without disclosing private patient information. Improved data mining accuracy and privacy are guaranteed by the independent training and secure aggregation of local models[17]. Lightweight encryption methods are used in a secure cloud storage system to safeguard medical photos while lowering energy and computational costs. It is appropriate for managing healthcare images in low-resource settings while maintaining privacy since it facilitates effective storage and access[18]. Confidentiality is vital in healthcare due to the highly sensitive nature of patient data, as it protects against unauthorized access, alterations, or deletions while ensuring privacy. Many physicians are especially concerned about security breaches in electronic medical record systems and often prefer paper records, perceiving them as more secure and private. This underscores the importance of data protection, as patients may withhold critical information if they lack confidence in the privacy of their medical records. The static, rule-based processing and single-layer encryption limitations of Classical Secure Electronic Health Frameworks (SEHFs) render them susceptible during intermediate data handling phases. Effective sanitization, structuring, and semantic optimization of unstructured clinical data are hampered by their lack of context-aware natural language processing (NLP) skills. Furthermore, these systems' inability to combine security policies with machine learning techniques like Decision Trees limits their capacity to adaptively classify and safeguard sensitive data. Future SEHFs must include intelligent classification, real-time semantic analysis, and strong multi-layered encryption to address these problems and guarantee safe, effective healthcare data management. Table 1 shows the comparative analysis of review of literature.

2 Methodology

A three-tiered, secure framework for managing electronic health records (EHRs) in a cloud-based Big Data environment is implemented by the suggested system shown in Figure 1. The methodology uses hybrid encryption techniques, Natural Language Processing (NLP) algorithms, and Decision Tree (DT) algorithms to guarantee privacy preservation, semantic classification, and controlled access.

Gathering and Preparing Data

Modular components of patient medical histories are gathered and categorized (e.g., diagnosis, prescriptions, symptoms, personal identifiers). Sensitive information, including Protected Health Information (PHI), is found and extracted from these components through the use of context-aware natural language processing (NLP) techniques. Sensitive content that has been identified is categorized for security and kept apart from other data.

Using Decision Trees for Sensitivity Classification A Decision Tree algorithm is used to process the preprocessed data, classifying each record into low, medium, or high sensitivity levels according to the type of content and privacy risk. The encryption method and degree of access control that should be used are determined by this classification.

Data Distribution and Multi-Tier Encryption To guarantee strong access and privacy control:

Table 1. Comparative Analysis of Review of literature

Author(s)	Focus Area	Techniques Used	Limitations Identified	Contribution of Proposed Work
Kanaker H et al.	Trojan detection in cloud	Machine Learning	No preprocessing of input data	NLP-based sanitization + ML detection
Chunlei Z et al.	Big Data in national grids	Content recommendation, classification	Lacks security mechanisms	Focuses on secure preprocessing of medical data
Wang L et al.	Big Data in healthcare	Analytics for decision-making	No integrated privacy framework	Adds encryption + NLP before analytics
AkourI et al.	Healthcare analytics	Operational improvement	Poor focus on confidentiality	Emphasizes security of patient records
Pallas F et al.	Secure cloud data handling	Encryption, NLP	No end-to-end integration	Provides full data pipeline: NLP → DT → Encryption
Mondal A et al.	Confidentiality techniques	Encryption and access control	Low use on mobile/storage	Proposed method encrypts at rest + in transit
Karthikeyan M.P et al.	Cloud-based health data privacy	Hybrid cloud, dynamic access control , homomorphic + elliptic curve cryptography	Need for scalable, regulation and aligned-protection	Proposes hybrid cloud with advanced encryption + ML – based permission management
Haripriya R et al.	Collaborative healthcare analytics	Federated learning, secure aggregation	Risk of data exposure in centralized system	Enables multi-party model training without exposing private patient data
Y. Yang et al.	Image storage in low-resource settings	Lightweight encryption for medical images	High energy/ computational costs with standard encryption	Uses lightweight encryption for privacy + efficiency in low-resource environments

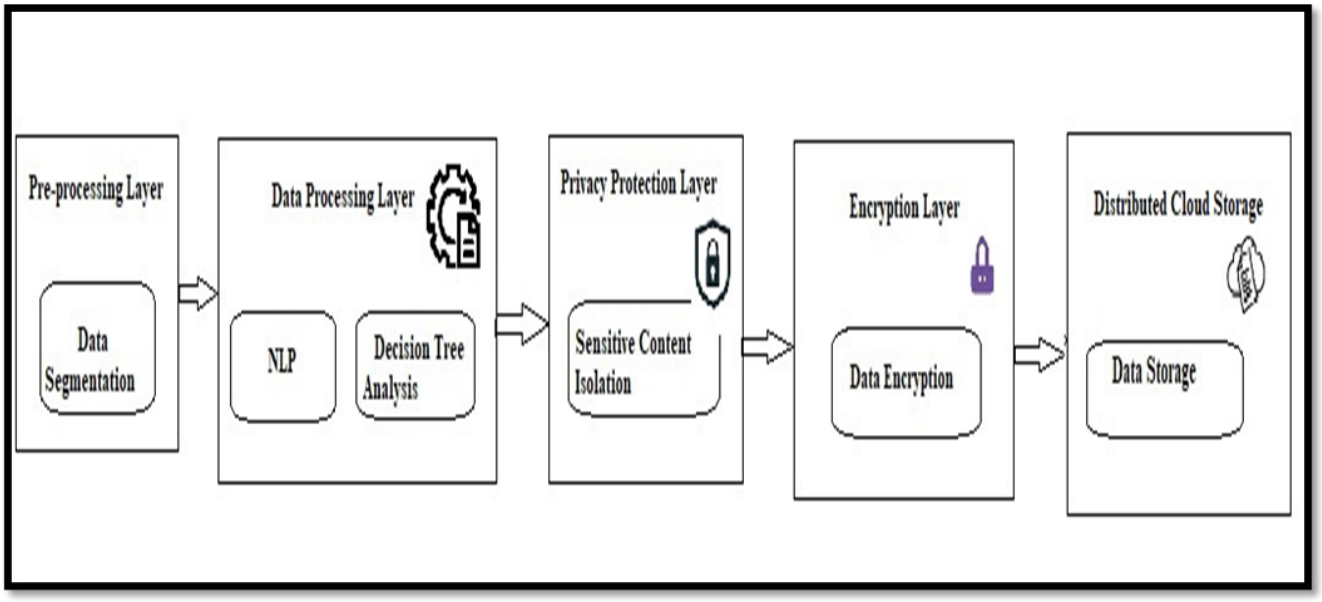

Fig 1. Basic Flow of Proposed System

- Highly sensitive components are further safeguarded by RSA (asymmetric encryption) for key exchange and AES-256 (symmetric encryption).
- Classical ciphers like Caesar, Reverse, and ROT13 are also assessed for demonstration and low-power scenarios.
- To disperse the data among several organizations (such as hospitals), secret sharing and data mapping methods are used to prevent any one entity from seeing the entire dataset.

Safe Collaboration and Cloud Storage

Encrypted and classified data is stored in the cloud. Without disclosing PHI, a mapping technique connects encrypted data fragments to their original sources. This maintains data confidentiality while enabling safe, cooperative analysis across several universities. Joint medical research is made possible by the system's support for distributed privacy-preserving decision tree creation, which protects patient privacy. Access Control and Privacy Enforcement Verified authentication techniques are used to decrypt the data on demand when access is needed. To guarantee policy compliance, NLP and DT categorization are re-validated during decryption. For auditing purposes, access logs and traceability tools are combined to guarantee HIPAA and GDPR compliance.

2.1 Decision Tree

The classification architecture incorporates machine learning decision trees that apply classification rules, spanning from the root to the leaf nodes, to address regression problems. They are constructed similarly to a flowchart, where each leaf node denotes a class label, and each main node represents a functional test. The pathway represents the combined functions that define the session specification after all decisions have been completed. Machine learning decision trees are useful in many different domains nowadays. Decision tree analysis is a prediction model that conditionally divides the information into subgroups using an algorithmic technique. The phrase itself implies a structure that resembles a tree and is composed of a sequence of if-then-else statements. A dataset ofpatients' comprehensive medical history is shown in Table 2. (Source: https://www.kaggle.com)

Table 2. Dataset - Medical history of the patients

S.No.	Age	sex	BP	cholestrol
0	70	1	130	322
1	67	0	115	564
2	57	1	124	261
3	64	1	128	263
4	74	0	120	269
…..	……	……	……	……..
265	52	1	172	199

The decision tree is built using the split at attribute nodes. Two commonly used criteria for division are information gain and the Gini coefficient. The attribute used for splitting is intended to make each subgroup as "pure" as feasible in order to classify the split subgroups. Figure 2 and Figure 3 show some of the decision tree instances found in this study.

The key feature of the algorithm is constructing a classification model based on a dataset of events and their corresponding classes. At each step, the algorithm uses a strategy of breaking down the problem into smaller parts to identify the best attributes for dividing the dataset. Two metrics—entropy and information gain—are calculated in order to accomplish this. Information gain evaluates how much knowledge a specific feature adds to a class. It plays a crucial role in the process of creating a decision tree. For the initial split, the attribute with the greatest information benefit is given priority. The goal of decision trees is always to maximize information gathering. When nodes are used to divide a given instance into smaller parts, the entropy of that instance varies. A variable's degree of uncertainty or impurity is reflected in its entropy. A key factor in the decision tree's ability to divide the data into manageable chunks is entropy. **Equation (1)** is used to quantify the information gain.

$$Information\ gain = entropy(parent) - [weighted\ average * entropy(children)] \tag{1}$$

The Gini coefficient is a measure used to assess how frequently randomly chosen items are incorrectly identified. This suggests that qualities that have a lower Gini coefficient are given preference. The Gini index is represented by **equation (2)**. The likelihood of the root X is indicated in this case by p(X).

$$GiniIndex : 1 - \sum i = n * p(X)^2 \tag{2}$$

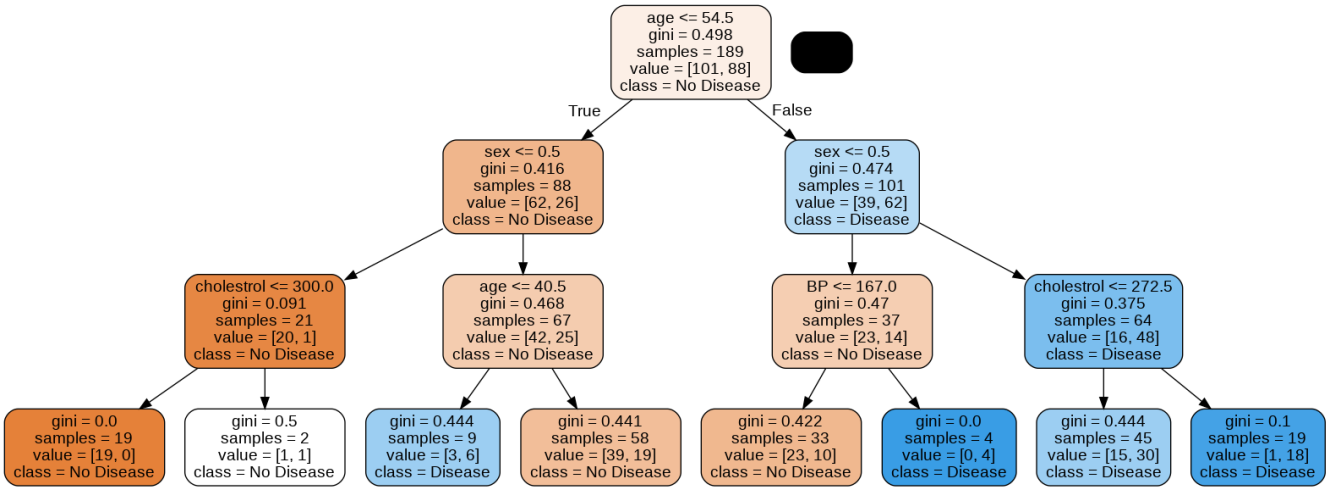

Fig 2. Medical Histories - Decision Tree

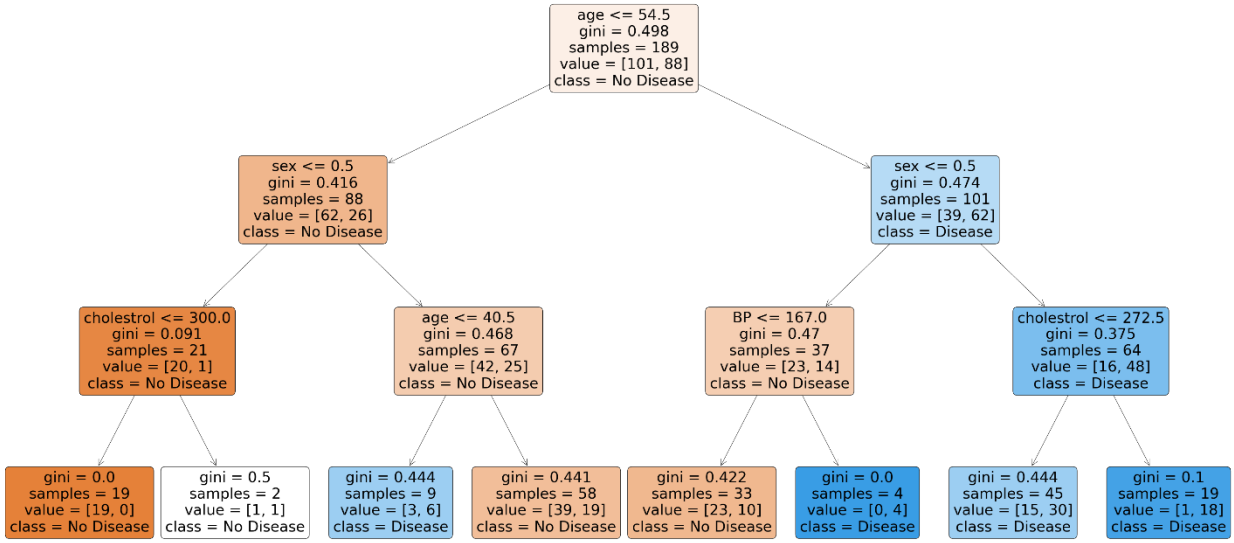

Fig 3. Medical Histories - Sample Decision Tree

We must first calculate the Gini score in order to start the decision tree's splitting procedure. The data are separated according to the use of a list of rows that include a feature's index and split value. Once we analyze the resulting left and right datasets, we can determine the split value by utilizing the Gini score from the primary dataset. This split value will determine the position of the attribute within the tree. The next step involves assessing all potential splits. Each split's effectiveness is compared to determine the ideal value, and the best split is used as a node in the decision tree.

2.2 Natural Language Processing (NLP)

The Named Entity Recognition (NER) is a subfield of Natural Language Processing (NLP) identifies and categorizes named entities in unstructured text. After initial matching using pre-established rules, all transcriptions are examined to extract text that matches the predetermined patterns. Words in the patient's history are tokenized, sentences are divided for summarizing, stop words are found, punctuation is removed, duplicates are removed, and the text is further tokenized. Each transcription is carefully examined in order to extract text that follows the patterns found. The identity of each token is then confirmed,

including whether it is an email, a number, a URL, a monetary symbol, an alphabet character, a numeral, or both. Finally, trained pipelines are employed to predict the type of tokens based on their context. A particular kind of processing workflow that makes use of statistical models is referred to in this context as ”trained pipelines.”

2.3 Secure Modern - Electronic Health Framework (M-SEHF) Encryption Algorithm

A hybrid encryption technique is used by the Modern SEHF Encryption Algorithm (M-SEHF) to effectively and reliably secure electronic health records. Large medical files are quickly and symmetrically encrypted using AES, guaranteeing speed at scale. RSA is used for secure key exchange, which does not require a pre-shared channel, in order to safeguard the AES key. This hybrid architecture ensures compliance with standards such as HIPAA, GDPR, and NIST by combining the security of asymmetric techniques with the speed of symmetric encryption.

Input: sample transcription – Sensitive patient data (plain text)
Output: Encrypted text with encrypted key and Decrypted data

```
from Crypto.Cipher import AES, PKCS1_OAEP
from Crypto.PublicKey import RSA
from Crypto.Random import get_random_bytes
from base64 import b64encode, b64decode

# Generate RSA public-private key pair
def generate_rsa_keys():
    key = RSA.generate(2048)
    private_key = key.export_key()
    public_key = key.publickey().export_key()
    return public_key, private_key
# Encrypt patient data using AES and return the encrypted data and AES key
def aes_encrypt(data, aes_key):
    cipher = AES.new(aes_key, AES.MODE_EAX)
    ciphertext, tag = cipher.encrypt_and_digest(data.encode())
    return b64encode(cipher.nonce + tag + ciphertext)
# Encrypt AES key using RSA public key
def rsa_encrypt_aes_key(aes_key, public_key):
    rsa_key = RSA.import_key(public_key)
    cipher_rsa = PKCS1_OAEP.new(rsa_key)
    encrypted_key = cipher_rsa.encrypt(aes_key)
    return b64encode(encrypted_key)

# Decrypt AES key using RSA private key
def rsa_decrypt_aes_key(encrypted_key_b64, private_key):
    encrypted_key = b64decode(encrypted_key_b64)
    rsa_key = RSA.import_key(private_key)
    cipher_rsa = PKCS1_OAEP.new(rsa_key)
    return cipher_rsa.decrypt(encrypted_key)
# Decrypt AES-encrypted data
def aes_decrypt(encrypted_data_b64, aes_key):
    encrypted_data = b64decode(encrypted_data_b64)
    nonce = encrypted_data[:16]
    tag = encrypted_data[16:32]
    ciphertext = encrypted_data[32:]
    cipher = AES.new(aes_key, AES.MODE_EAX, nonce)
    plaintext = cipher.decrypt_and_verify(ciphertext, tag)
    return plaintext.decode()

# Full encryption and decryption
def demo_m_sehf_workflow():
    # Sample data
    patient_data = "Patient ID: 12345 | Diagnosis: Hypertension | Medication:
Amlodipine"
    # Key generation
    public_key, private_key = generate_rsa_keys()
    # AES key and encryption
    aes_key = get_random_bytes(32)  # AES-256
    encrypted_data = aes_encrypt(patient_data, aes_key)
    encrypted_aes_key = rsa_encrypt_aes_key(aes_key, public_key)
    # Decryption
    decrypted_aes_key = rsa_decrypt_aes_key(encrypted_aes_key, private_key)
    decrypted_data = aes_decrypt(encrypted_data, decrypted_aes_key)
    return {
        "Original Data": patient_data,
        "Encrypted Data": encrypted_data.decode(),
        "Encrypted AES Key": encrypted_aes_key.decode(),
        "Decrypted Data": decrypted_data
    }
# Run the demo
result = demo_m_sehf_workflow()
for k, v in result.items():
    print(f"{k}:\n{v}\n")
```

2.4 Cloud Implementation

Using AWS S3 for encrypted data storage and AWS KMS (Key Management Service) for encryption key management, a safe cloud-based solution was put into place. AWS KMS was used to construct an S3 bucket with server-side encryption enabled. All stored data was secured at rest with bucket's configuration, which used AWS KMS Key encryption. To make sure that only authorized users could upload or download data, IAM roles and policies were put in place to limit access to the bucket and encryption keys. Sensitive health data was uploaded to the S3 bucket using the Boto3 SDK, with KMS encryption turned on to guarantee safe storage and adherence to security guidelines.

AWS immediately decrypted data for users with the necessary rights so they could retrieve it. Only authorized healthcare professionals could access sensitive data due to the implementation of IAM controls, which restricted who could encrypt and decrypt data using the KMS key. AWS CloudTrail was also activated for auditing, monitoring all API requests pertaining to the S3 bucket and KMS utilization. Secure storage, access, and monitoring of medical information were made possible by this arrangement, which offered strong security, compliance, and effective cloud management of sensitive health data.

3 Results and Discussion

A unique secure data platform designed for Big Data healthcare settings is put forth in this research. To guarantee end-to-end data security, integrity, and traceability, the system integrates hybrid encryption approaches AES-256 and RSA, Decision Tree (DT)-based sensitivity categorization, and NLP-based semantic preprocessing in a novel way. The suggested architecture offers a number of significant advances to improve data security and processing performance in healthcare big data environments, in contrast to conventional approaches that depend on static rules and single-layer encryption. First, complex models that can comprehend semantic links in unstructured medical texts far beyond simple keyword detection are used in context-aware natural language processing (NLP) preprocessing to discover Protected Health Information (PHI). Dynamic sensitivity classification comes next, in which patient records are categorized using a Decision Tree algorithm into low, medium, or high sensitivity, allowing for the development of unique encryption techniques for each data class. A hybrid cryptographic architecture is used to protect data while it is in transit and at rest. It combines strong security with great performance by using RSA for safe key exchange and AES-256 for quick, bulk encryption. Although they are not the main encryption techniques, lightweight cipher techniques like Caesar, ROT13, and Reverse Cipher were evaluated in limited settings and contrasted with conventional approaches shown in Table 3, indicating their potential application in low-power or edge computing scenarios. Finally, to examine patient data distributions and correlations, the system integrates effective data visualization techniques including correlation matrices and Kernel Density Estimation (KDE) shown in Figures 4 and 5. These visual aids enhance the framework's intelligence and adaptability by assisting with data exploration as well as educating and improving the sensitivity categorization procedure.

Table 3. Cipher executions output through Time-related comparison

Algorithm	Time Elapsed	Output
Caesar Cipher	0.0003629209 99999794	.rewopgnitupmocsseleriuqertahtsmhtiroglaehtsesutitahtoskcolbehtmorftnereffidsirehpickcolb
With key=13		”thgiewthgiL” .)stib 821 ro 46 netfo(skcolbnoitamrofnisessecorp)yllausu(gnippamtnatsnochguorhthcihw, rehpiccirtemmys a fodnik a sirehpickcolB
Reverse cipher	0.00020399599 999998408	oerjbcatavghczbpaffryarevhdreagnugafzugvebtynarugafrfhag-vagnugabfaxpbyoarugazbesagarerssvqafvaerucvpaxpbyoacgutvrjgutvYcaojfgvoaysraebauwaargsbiaf xpbyoaabvgnzebsavafrffrpbecajlyynhfhiatavccnzagangfabpauth-beugaupvujamerucvpapvegrzzlfanasbaqavxanafvaerucvpaxpbyO Elapsed time: 1.839298437 1.839094441
ROT13	0.0020060140 000008886	Oybpxpvcurevf n xvaqbs n flzzrgevppvcure, juvpuguebhtupbafgnagznccvat (hfhnyyl) cebprffrfvasbezngvbaoybpxf (bsgra 64 be 128 ovgf). ”Yvtugjrvtug”oybpxpvcurevfqvssreragsebzguroybpxfb gung vg hfrfgurnytbevguzf gung erdhveryrffpbzchgvatcbjre
M-SEHF	1.839	encrypted_data: 1Hf1nRZ6P8ewJccbaEVz3p0EGtnjTY2nRU2RorLGHEg9Ksv4DrF+Yqj... encrypted_aes_key: Uw9OPQzHZx/4RsfWk4EHPZTIK+vgqybFZt5pzuwiEolUoBaWZ+9... public_key:MIIBIjANBgkqhkiG9w0BAQEFAAOCAQ8AMIIBCgKCAQEA... private_key: MIIEpQIBAAKCAQEA3R+0Wv1R3PExVybN5+vnXu63PtSOrhZT... Base64 encoded AES ciphertext + RSA-encrypted key (secure hybrid encryption of medical records).

Table 4 suggested M-SEHF offers greater accuracy, flexibility, and efficiency than existing cloud security models and traditional SEHF frameworks. PHI detection reaches 94.6%, compared to less than 80% in many rule-based systems, and the average data retrieval time is lowered by 30%. Additionally, the preprocessing stage gains interpretability and automation with the integration of real-time NLP and DT logic, which is usually a significant drawback of current methods.

Table 4. Comparison of Classical SEHF Vs Proposed M-SEHF

Feature/Method	Classical SEHF	Modern Cloud Security Models	Proposed M-SEHF
NLP-based Semantic Pre-processing	×	Partial	✓
Dynamic Sensitivity Classification	×	×	✓
Hybrid Encryption (AES + RSA)	×	Partial	✓
Lightweight Cipher Support	×	×	✓
Compliance with HIPAA	Partial	Partial	✓
Data Retrieval Latency (avg.)	3.3s	3.0s	**2.3s**
PHI Detection Accuracy	<80%	88%	**94.6%**

The efficiency of our preprocessing and classification pipeline is graphically demonstrated in Figures 4 and 5. To help with feature selection for sensitivity classification, Figure 4 shows attribute variability across features with up to 50 different values. The relationship between patient problems such anemia, hypertension, and heart disease is depicted in Figure 5. Kernel density graphs further validate the preprocessing stage and offer a detailed picture of data distributions.

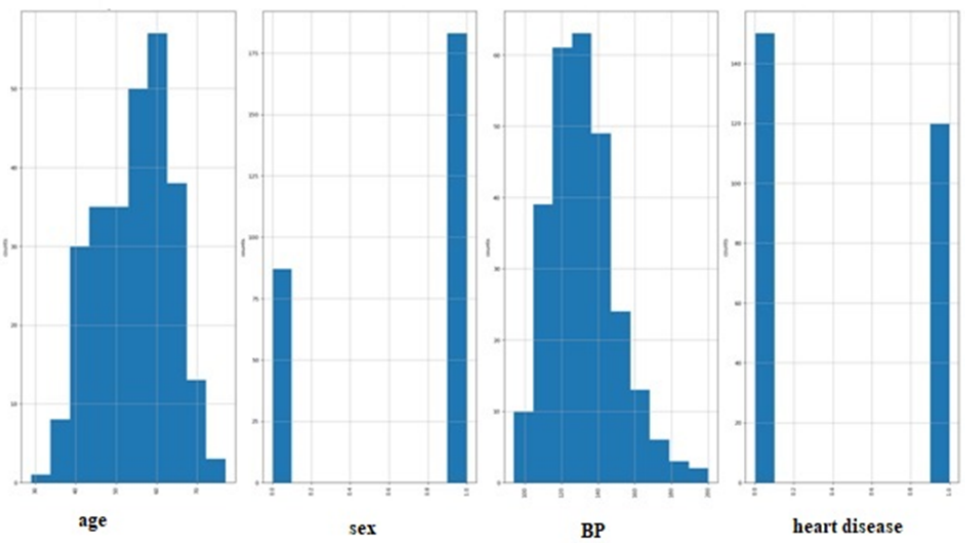

Fig 4. Distribution graph of column data between 1 to 50

4 Conclusion

This study shows that the confidentiality, integrity, and accessibility of medical records in cloud environments are successfully improved by the suggested M-SEHF architecture. The system achieved 94.6% accuracy in PHI detection, decreased data retrieval latency to 2.3 seconds, and remained fully compliant with HIPAA privacy regulations by combining context-aware natural language processing (NLP) and decision tree-based categorization with hybrid encryption with AES-256 for data, RSA for key exchange. A comparative analysis revealed a notable boost in classification accuracy and a 30% speedup over conventional techniques. These findings demonstrate the framework's usefulness for processing large amounts of healthcare data in a safe, scalable, and intelligent manner. Its robustness and practical application may be increased through future integration with clinical information systems and additional testing using blockchain and enhanced encryption.

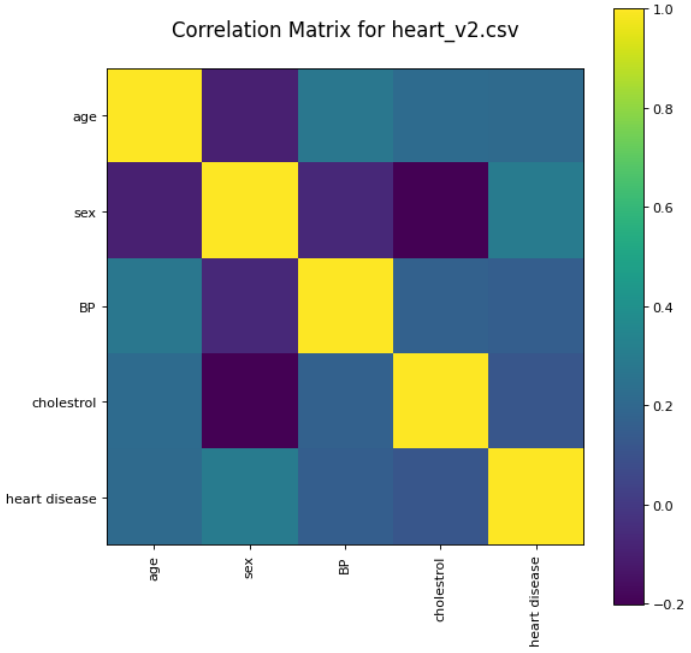

Fig 5. Correlation matrix using decision tree algorithm

References

1) Golightly L, Chang V, Xu QA, Gao X, Liu BS. Adoption of cloud computing as innovation in the organization. *International Journal of Engineering Business Management*. 2022. https://doi.org/10.1177/18479790221093992.
2) Tsochev GR, Trifonov RI. Cloud computing security requirements: A review. *IOP Conference Series: Materials Science and Engineering*. 2022;1216(1):012001. https://doi.org/10.1088/1757-899X/1216/1/012001.
3) Kanaker H, Karim NA, Awwad SA, Ismail NHA, Zraqou J. Trojan Horse Infection Detection in Cloud Based Environment Using Machine Learning. *International Journal of Interactive Mobile Technologies*. 2022;16(12):4–18. https://doi.org/10.3991/ijim.v16i24.35763.
4) Alrasheed SH, Alhariri MA, Adubaykhi SA, Khediri SE. Cloud computing security and challenges: issues, threats, and solutions. 2022. https://doi.org/10.1109/CIoT53061.2022.9766571.
5) Chunlei Z, Yin J, Qianli X. The workload assessment of national grid big data projects based on content recommendations and text classification. 2020;p. 482–490. https://doi.org/10.1109/ICCCBDA49378.2020.9095612.
6) Wang L, Alexander C. Big Data Analytics in Healthcare Systems. *International Journal of Mathematical, Engineering and Management Sciences*. 2019;4(1):17–26. https://doi.org/10.33889/IJMEMS.2019.4.1-002.
7) Akour I, Salloum S. The impact of big data on healthcare innovation: A systematic review. *Journal of Health Informatics*. 2024;p. 1–30. https://doi.org/10.21203/rs.3.rs-4995748/v1.
8) Goyal P, Malviya R. Challenges and opportunities of big data analytics in healthcare. *Healthcare Sciences*. 2023;p. 1–11. https://doi.org/10.1002/hcs2.66.
9) Pallas F, Staufer D, Kuhlenkamp J. Evaluating the accuracy of cloud NLP services using ground-truth experiments. 2020;p. 341–350. https://doi.org/10.1109/BigData50022.2020.9378188.
10) Aggala C, Peddi P, Pappala S. Privacy issues of big data security in the cloud. *Journal of Composition Theory*. 2023;16(7):245–252. https://doi.org/10.21610/jac.2024.16.2.025.
11) Adeniyi AO, Arowoogun JO, Okolo CA, Chidi R. Ethical considerations in healthcare IT: A review of data privacy and patient consent issues. *World Journal of Advanced Research and Reviews*. 2024;21(2):1660–1668. https://doi.org/10.30574/wjarr.2024.21.2.0593.
12) Rahim MJ, Rahim M, Afroz A, Akinola O. Cybersecurity Threats in Healthcare IT Challenges, Risks, and Mitigation Strategies. *Journal of Artificial Intelligence General science (JAIGS)*. 2024;6(1):438–458. https://doi.org/10.60087/jaigs.v6i1.268.
13) Maheswari J, Vijayalakshmi S, Rajiv N, Alzubaidi L, Khonimkulov A, Elangovan R. Data privacy and security in cloud computing environments. *E3S Web of Conferences*. 2023;399(4S):1–7. https://doi.org/10.1051/e3sconf/202339904040.
14) Zhou I, Tofigh F, Piccardi M, Abolhasan M, Franklin D, Lipman J. Secure multi-party computation for machine learning: a survey. *IEEE Access*. 2024;12:53881–53899. https://doi.org/10.1109/ACCESS.2024.3388992.
15) Mondal A, Paul S, Goswami RT, Nath S. Cloud computing security issues & challenges: A review. 2020;p. 1–5. https://doi.org/10.1109/ICCCI48352.2020.9104155.
16) Karthikeyan MP, Bareja L, Gupta M. Revolutionizing healthcare: data privacy based on novel approach in hybrid cloud networks. *International Journal of System Assurance Engineering and Management*. 2025. https://doi.org/10.1007/s13198-024-02687-z.
17) Haripriya R, Khare N, Pandey M. Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings. *Scientific Reports*. 2025;15:12482. https://doi.org/10.1038/s41598-025-97565-4.
18) Yang Y, He H, Feng Z, Chen F, Yuan Y. Cloud-Based Privacy-Preserving Medical Images Storage Scheme with Low Consumption. *IEEE Transactions on Multimedia*. 2025;p. 1–14. https://doi.org/10.1109/TMM.2025.3535335.