GPH-International Journal of Computer Science and Engineering
(Online ISSN 3050-9661)
Volume 9 | Issue 01 | Jan-June 2026
Published by: Global Publication House
DOI: https://doi.org/10.5281/zenodo.21902823
Available online at: www.gphjournal.org
Journal homepage: https://gphjournal.org/index.php/cse/index

Scalable and Enhanced Security Mechanism for Cloud Database System

B. Mercy Sigbara[1], D. Matthias[2], D. N. Nwiabu[3], O. E. Taylor[4]

[1,2,3,4] Department of Computer Science, Rivers State University, Nkpolu-Oroworukwo, Port Harcourt, Rivers State, Nigeria.
Corresponding Author email: *sigbaramercy@gmail.com*

Abstract

Developing secure and trusted distributed systems remains a major challenge in cloud computing environments. Ensuring data confidentiality, integrity, authentication, and secure access is particularly critical in distributed architectures such as wireless, peer-to-peer, and cloud database systems, where structural limitations and trust-related issues increase security vulnerabilities. Data breaches in such environments may result in operational disruptions, financial losses, legal liabilities, compliance violations, and reputational damage. Therefore, the development of scalable and enhanced security mechanism has become increasingly essential for protecting distributed cloud data systems. This thesis presents a HybridPDFSecurityTrainer model designed to enhance data security and confidentiality in distributed cloud database environments. The developed model integrates Convolutional Neural Network (CNN), Rivest–Shamir–Adleman (RSA), and Term Frequency–Inverse Document Frequency (TF-IDF) algorithms to provide intelligent document classification and secure data protection. The TF-IDF algorithm is used to extract textual features from Portable Document Format (PDF) files by converting documents into machine-readable numerical feature vectors. These feature vectors are subsequently processed using a CNN model to classify documents as benign or malicious based on detected patterns, sensitive information, and malicious content indicators. To strengthen data confidentiality and integrity, the RSA cryptographic algorithm is incorporated to perform secure encryption and decryption of classified documents before cloud storage and transmission. RSA also facilitates secure key exchange and prevents unauthorized access to sensitive information within distributed cloud environments. The integration of machine learning and cryptographic techniques provides a scalable and intelligent security framework capable of addressing modern cloud security challenges. This developed model was implemented using Python programming language and the Object-Oriented Analysis and Design Methodology (OOADM). Experimental evaluation showed that the system achieved 98% security accuracy with a low error rate, outperforming existing approaches in detecting malicious PDF documents while maintaining secure data transmission and storage. This study demonstrates that the HybridPDFSecurityTrainer model significantly improves confidentiality, integrity, and security management in distributed cloud database systems, making it suitable for deployment in real-world cloud computing environments.

Keywords:

Distributed cloud database systems; Unstructured data; Intelligent threat detection; Convolutional neural networks; Term Frequency–Inverse Document Frequency; RSA cryptography; PDF malware detection; Secure cloud storage.

1.0 Background of the Study

The rapid adoption of cloud computing has transformed the way organizations store, manage, and process data by providing scalable, flexible, and cost-effective computing infrastructures. Distributed cloud database systems, which coordinate data storage and processing across multiple interconnected nodes, have become fundamental to modern enterprises because of their ability to support high availability, fault tolerance, and large-scale data processing. These systems enable organizations to efficiently manage massive volumes of data generated from business transactions, social media platforms, Internet of Things (IoT) devices, healthcare systems, and financial services. However, the increasing reliance on distributed cloud environments has also introduced significant security challenges, particularly in protecting sensitive and unstructured data.

Unstructured data, including textual documents, images, videos, emails, sensor data, and multimedia content, now constitutes the majority of organizational data and continues to grow at a much faster rate than structured data. Unlike structured data stored in relational databases, unstructured data lacks predefined schemas, making its storage, retrieval, analysis, and protection considerably more complex. Consequently, organizations increasingly employ advanced technologies such as NoSQL databases, data lakes, machine learning, and natural language processing to extract valuable information from these heterogeneous datasets. While these technologies enhance data analytics and intelligent decision-making, they simultaneously increase the complexity of maintaining confidentiality, integrity, authentication, and access control within distributed cloud database environments.

The dynamic and decentralized nature of distributed cloud databases further expands the attack surface, exposing systems to unauthorized access, data interception, manipulation, and other sophisticated cyber threats. As organizations migrate critical business operations to cloud platforms, conventional security mechanisms often struggle to provide adequate protection without compromising system performance and scalability. This has created an urgent need for intelligent, scalable, and adaptive security frameworks capable of safeguarding sensitive information while maintaining efficient cloud operations.

Artificial intelligence has emerged as a promising technology for addressing these security challenges. Machine learning and deep learning algorithms are capable of automatically learning complex data patterns, identifying abnormal behaviors, and detecting cyber threats with greater accuracy than traditional rule-based approaches. In particular, deep learning techniques such as Convolutional Neural Networks (CNNs) have demonstrated remarkable capability in extracting discriminative features from high-dimensional datasets, while Natural Language Processing techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) effectively represent textual information for intelligent analysis. When combined with robust cryptographic techniques such as the Rivest–Shamir–Adleman (RSA) algorithm, these technologies provide opportunities to develop integrated security mechanisms that simultaneously support threat detection, secure data transmission, and data confidentiality.

Despite these advancements, existing studies indicate that important challenges remain unresolved. Kumar et al. (2022) examined security issues in distributed cloud and edge computing environments and reported that communication channels remain vulnerable to eavesdropping attacks because of inadequate encryption mechanisms. Although lightweight

cryptographic protocols improved communication security, they introduced trade-offs between computational efficiency and scalability. Similarly, Mustafa et al. (2024) proposed a deep learning-based intrusion detection system capable of detecting replay attacks in distributed environments using temporal traffic patterns. While the proposed model achieved improved detection performance compared with conventional signature-based approaches, its high computational requirements limit its practical deployment in real-time and large-scale distributed cloud infrastructures.

These limitations demonstrate that current security solutions often address isolated aspects of cloud security without simultaneously ensuring data confidentiality, integrity, intelligent threat detection, and scalability. As distributed cloud database systems continue to expand in complexity and data volume, there is a growing need for integrated security frameworks that combine intelligent feature extraction, deep learning-based threat detection, and robust cryptographic protection while maintaining high system performance.

To address these challenges, this study proposes a scalable hybrid security framework that integrates TF-IDF for feature extraction, CNN for intelligent threat detection, and RSA cryptography for secure data protection in distributed cloud database systems. The proposed framework is expected to enhance data confidentiality and integrity, improve the detection of unauthorized access attempts, and provide a scalable security mechanism suitable for modern distributed cloud environments.

2. Literature Review

2.1 Conceptual Review

Distributed cloud database systems have become the backbone of modern data-driven organizations by enabling scalable, reliable, and geographically distributed storage and processing of large volumes of information. A distributed cloud database system consists of multiple interconnected databases operating across different locations while appearing as a single unified system to users. Such systems provide transparency, scalability, fault tolerance, and resource sharing, making them suitable for applications in smart cities, healthcare, industrial automation, financial services, and Internet of Things (IoT) environments. However, their distributed architecture introduces significant challenges in maintaining data consistency, availability, confidentiality, and secure communication across heterogeneous computing environments.

A major concern in distributed cloud environments is the rapid growth of unstructured data, which includes text documents, images, videos, emails, social media content, and sensor data. Unlike structured data stored in relational databases, unstructured data lacks predefined schemas, making storage, retrieval, analysis, and security considerably more challenging. Organizations increasingly rely on NoSQL databases, data lakes, machine learning, and natural language processing techniques to manage these datasets efficiently. Although these technologies improve data utilization, they also increase the complexity of protecting sensitive information from unauthorized access and cyber threats.

Artificial intelligence has emerged as a fundamental technology for intelligent database management and cybersecurity. By leveraging machine learning and deep learning techniques,

AI enables automated pattern recognition, anomaly detection, decision support, and predictive analytics across both structured and unstructured datasets. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated remarkable capability in extracting meaningful features from high-dimensional data, while Natural Language Processing techniques such as Term Frequency–Inverse Document Frequency (TF-IDF) effectively transform textual information into representative feature vectors for classification and threat detection. Consequently, AI-driven security solutions have become increasingly important for detecting malicious activities and supporting proactive security management in distributed cloud environments.

Digital transformation has accelerated the migration of enterprise applications from conventional databases to cloud-native and distributed database platforms. Modern organizations increasingly depend on cloud computing, big data analytics, and artificial intelligence to support real-time decision making and business intelligence. However, this transformation has also intensified challenges associated with data governance, storage scalability, resource management, interoperability, and cybersecurity. Effective data management therefore requires intelligent storage mechanisms capable of ensuring data quality, accessibility, scalability, and regulatory compliance across distributed infrastructures.

Security remains one of the most critical requirements of distributed cloud database systems because the integration of multiple computing nodes significantly expands the attack surface. Contemporary distributed environments are exposed to threats such as unauthorized access, data interception, replay attacks, distributed denial-of-service (DDoS) attacks, identity spoofing, and malicious data manipulation. Consequently, security frameworks must simultaneously ensure confidentiality, integrity, authentication, authorization, accountability, and non-repudiation while maintaining acceptable computational performance and scalability. Existing approaches including access control mechanisms, cryptographic techniques, policy-based frameworks, security patterns, and quorum-based systems have improved specific aspects of cloud security but often address individual challenges without providing comprehensive and scalable protection for distributed cloud databases.

Among the various security mechanisms, encryption remains fundamental for protecting sensitive information stored and transmitted across distributed cloud infrastructures. Modern cloud systems employ both symmetric and asymmetric cryptographic algorithms to safeguard data confidentiality. While symmetric algorithms provide efficient encryption for large data volumes, asymmetric algorithms such as RSA facilitate secure key exchange, authentication, and digital signatures. Authentication and authorization mechanisms further strengthen security by ensuring that only legitimate users access cloud resources, whereas accountability and auditing mechanisms enable monitoring and forensic investigation of security incidents. The integration of these mechanisms provides the foundation for trustworthy distributed cloud database systems.

Overall, the reviewed concepts demonstrate that secure distributed cloud database systems require the convergence of intelligent feature extraction, deep learning-based threat detection, and robust cryptographic protection. The combination of TF-IDF for textual feature representation, CNN for intelligent classification and intrusion detection, and RSA for secure encryption offers a promising hybrid approach capable of improving data confidentiality,

integrity, threat detection accuracy, and scalability. This conceptual integration provides the theoretical foundation for the proposed scalable security framework developed in this study.

2.2 Empirical Review

Recent studies have proposed various approaches to improve the security of distributed cloud database systems through access control, cryptography, artificial intelligence, and policy-based security frameworks. These approaches have significantly enhanced data confidentiality, authentication, and intrusion detection; however, challenges related to scalability, computational efficiency, and integrated security remain unresolved.

Shehab (2020) proposed an access control framework based on path authentication and secure route discovery for distributed environments. The framework improved authentication and protected communication channels by dynamically monitoring entity availability. Although effective for identity management, the model focused primarily on authentication and provided limited support for intelligent threat detection and large-scale cloud environments. Similarly, Mosbah (2021) introduced a policy-based security architecture that employed modular security policies to enforce consistent security across distributed systems. While the framework improved policy enforcement and flexibility, it depended heavily on manually defined policies, limiting its adaptability to evolving cyber threats.

Cryptographic solutions have also received considerable attention. Sahoo et al. (2023) proposed a device-level security framework based on Public Key Infrastructure (PKI) and Physical Unclonable Functions (PUFs) to strengthen authentication and secure communication among distributed devices. Likewise, Chang-Ji (2023) enhanced Role-Based Access Control (RBAC) using Public Key Certificates to improve authentication and authorization. Although these approaches strengthened confidentiality and access control, they primarily relied on static cryptographic mechanisms and lacked intelligent capabilities for detecting sophisticated cyber-attacks in real time.

Pattern-based and quorum-based security models have also been investigated. Yengeç Taşdemir et al. (2023) demonstrated that reusable security patterns could systematically integrate confidentiality, authentication, and integrity mechanisms into distributed software systems. Meng and Zhang (2020) developed an Intrusion-Tolerant Quorum System (ITOS) that combined trusted computing with role-based access control to improve fault tolerance and data consistency. Although these models improved resilience against node failures and malicious attacks, their implementations were largely infrastructure-oriented and did not incorporate intelligent learning mechanisms capable of adapting to emerging attack patterns.

Artificial intelligence has increasingly been adopted to enhance cloud security. Ting and Li (2025) developed a multilayer security framework that integrated encryption, hashing, and data partitioning to improve data privacy and secure storage in cloud environments. Their framework enhanced confidentiality and integrity across multiple processing layers while supporting efficient data management. Nevertheless, the model focused mainly on storage protection and did not integrate advanced deep learning techniques for automated threat detection and classification.

Research has also examined security vulnerabilities in distributed computing environments. Almorsy et al. (2022) investigated the security implications of integrating APIs, microservices, and third-party services in distributed systems. Their findings revealed that system integration significantly expands the attack surface, increasing vulnerabilities associated with weak authentication, API misuse, and security misconfigurations. Although the study identified major security risks, it lacked practical validation and did not explore intelligent security mechanisms such as zero-trust architectures or AI-driven intrusion detection.

Studies on authentication have similarly evolved beyond conventional password-based mechanisms. Bele et al. (2024) demonstrated that biometric authentication provides stronger identity verification than traditional credentials by utilizing unique physiological and behavioural characteristics. The study concluded that biometrics significantly improve authentication security but highlighted privacy concerns, deployment complexity, and computational overhead as major implementation challenges. Likewise, recent studies have emphasized multi-factor authentication, behavioural biometrics, and continuous authentication as promising approaches for securing distributed cloud environments, although their adoption remains constrained by infrastructure complexity and usability considerations.

Collectively, these empirical studies demonstrate substantial progress in cloud database security through cryptographic techniques, access control, authentication, policy enforcement, and intelligent security mechanisms. Nevertheless, most existing solutions address individual security requirements rather than providing a unified framework capable of simultaneously ensuring intelligent threat detection, secure feature extraction, data confidentiality, data integrity, and scalability. Furthermore, limited integration exists between Natural Language Processing techniques such as TF-IDF, deep learning models such as Convolutional Neural Networks (CNNs), and asymmetric cryptographic algorithms such as RSA within a single distributed cloud database security architecture.

The identified gaps provide the empirical justification for the present study, which proposes a hybrid TF-IDF–CNN–RSA framework that combines intelligent feature extraction, deep learning-based threat detection, and robust cryptographic protection to improve the security, scalability, and overall performance of distributed cloud database systems.

2.3 Theoretical Review

The theoretical foundation of this study is anchored on the principles of information security, artificial intelligence, and public-key cryptography, which collectively provide the basis for developing secure and intelligent distributed cloud database systems. These theoretical perspectives explain how confidentiality, integrity, authentication, authorization, and intelligent decision-making can be integrated to protect distributed cloud environments against evolving cyber threats.

The Information Security Theory underpins the security requirements of distributed cloud database systems by emphasizing the protection of information assets throughout their lifecycle. The theory maintains that an effective security architecture must preserve confidentiality by preventing unauthorized disclosure of information, ensure data integrity by protecting information from unauthorized modification, and enforce authentication, authorization, accountability, and non-repudiation to establish trust among communicating

entities. Within distributed cloud environments, these principles provide the foundation for designing secure storage, transmission, and access control mechanisms across geographically dispersed computing nodes.

Artificial Intelligence Theory provides the computational foundation for intelligent threat detection and automated decision-making. The theory explains that intelligent systems can learn patterns from historical data, identify anomalies, classify malicious activities, and continuously improve prediction accuracy through machine learning and deep learning algorithms. In distributed cloud database systems, artificial intelligence enables automated feature extraction, intrusion detection, and adaptive security responses that are difficult to achieve using conventional rule-based security mechanisms. The theory therefore supports the application of intelligent models capable of processing both structured and unstructured data for enhanced cybersecurity.

The study is further supported by Public-Key Cryptography Theory, which explains secure communication through asymmetric encryption. Unlike symmetric encryption that depends on a shared secret key, public-key cryptography employs mathematically related public and private keys to achieve secure encryption, authentication, digital signatures, and key exchange. This theoretical model forms the basis for implementing RSA encryption, which strengthens data confidentiality and secure communication in distributed cloud database systems by ensuring that only authorized users possessing the corresponding private key can decrypt protected information.

The integration of these theoretical perspectives provides a comprehensive foundation for the proposed security framework. Information Security Theory defines the fundamental security objectives, Artificial Intelligence Theory provides intelligent capabilities for automated threat detection and feature learning, while Public-Key Cryptography Theory guarantees secure encryption and authentication. Together, these theories justify the integration of TF-IDF for feature representation, Convolutional Neural Networks (CNNs) for intelligent threat detection, and RSA cryptography for secure data protection within a unified and scalable distributed cloud database security framework. The synergy among these theoretical foundations supports the development of an intelligent security architecture capable of enhancing confidentiality, integrity, scalability, and overall system performance in distributed cloud environments.

2.4 Related Works

Recent advances in distributed cloud database security have focused on improving data confidentiality, integrity, authentication, and intrusion detection through access control mechanisms, cryptographic techniques, artificial intelligence, and policy-driven security frameworks. Although these approaches have significantly strengthened cloud security, most existing solutions address specific security challenges independently and rarely provide an integrated framework capable of simultaneously ensuring intelligent threat detection, scalable security, and secure data protection.

Shehab (2020) proposed a path authentication mechanism for distributed systems that enhances secure route discovery and authentication by monitoring entity availability during communication. The framework improved trust management in distributed environments but primarily addressed authentication without incorporating intelligent threat detection or adaptive

learning capabilities. Similarly, Mosbah (2021) introduced a policy-based security framework that applies modular and domain-specific security policies across distributed systems. While the approach improved policy consistency and flexibility, its reliance on manually defined security policies limited its adaptability to dynamic cyber threats.

Cryptographic approaches have also been widely investigated. Sahoo et al. (2023) developed a device-level security framework that combines Public Key Infrastructure (PKI) and Physical Unclonable Functions (PUFs) to secure communication among distributed devices. Likewise, Chang-Ji (2023) enhanced Role-Based Access Control (RBAC) using Public Key Certificates to strengthen authentication and authorization. Although these techniques improved secure communication and access control, they largely depended on static cryptographic mechanisms and did not incorporate intelligent threat detection capabilities required for modern cloud environments.

Artificial intelligence has increasingly been employed to improve cloud security. Ting and Li (2025) proposed a multilayer cloud security framework that integrated hashing, encryption, and data partitioning to enhance secure storage and privacy management in distributed cloud systems. Their framework improved data confidentiality and integrity across multiple processing layers but concentrated primarily on secure storage rather than intelligent attack detection. Similarly, Almorsy et al. (2022) examined security vulnerabilities associated with APIs, microservices, and third-party integration in distributed systems, identifying weak authentication, API attacks, and security misconfigurations as major threats. However, their work remained largely conceptual and did not incorporate artificial intelligence or zero-trust mechanisms for automated threat mitigation.

Authentication mechanisms have also evolved beyond traditional password-based systems. Bele et al. (2024) demonstrated that biometric authentication significantly improves identity verification by utilizing unique physiological and behavioural characteristics. Although biometric techniques strengthen authentication security, the study identified privacy concerns, computational overhead, and deployment complexity as major implementation challenges. Recent studies have therefore recommended multi-factor authentication and behavioural biometrics to enhance trust in distributed cloud environments, although these solutions increase implementation complexity and infrastructure requirements.

Pattern-based and quorum-based security techniques have further contributed to distributed system protection. Yengeç Taşdemir et al. (2023) demonstrated that reusable security patterns provide systematic implementation of confidentiality, integrity, and authentication mechanisms across distributed software systems. Meng and Zhang (2020) proposed an Intrusion-Tolerant Quorum System (ITOS) that integrates trusted computing with role-based access control to improve fault tolerance and data consistency. Despite their effectiveness in enhancing system resilience, these approaches remain infrastructure-oriented and lack adaptive learning capabilities for detecting evolving cyber threats.

Overall, the reviewed studies demonstrate significant progress in securing distributed cloud database systems through access control, cryptographic protection, authentication, policy enforcement, and distributed security architectures. Nevertheless, existing solutions largely address confidentiality, authentication, or intrusion detection independently and seldom

integrate intelligent feature extraction, deep learning-based threat detection, and asymmetric encryption within a unified framework. Furthermore, limited attention has been given to combining Natural Language Processing techniques such as TF-IDF, deep learning models such as Convolutional Neural Networks (CNNs), and RSA cryptography to simultaneously enhance intelligent threat detection, secure data protection, and scalability.

These identified gaps provide the motivation for the present study, which proposes a hybrid TF-IDF–CNN–RSA framework that integrates intelligent textual feature extraction, deep learning-based threat detection, and robust asymmetric encryption into a scalable security architecture for distributed cloud database systems. The proposed framework seeks to improve data confidentiality, integrity, authentication, and detection accuracy while maintaining scalability suitable for modern cloud computing environments.

3 Materials and Methods

3.1 Research Design

This study adopted an **Object-Oriented Analysis and Design Methodology (OOADM)** together with a **constructive research approach** for the design and implementation of a scalable security framework for distributed cloud database systems. OOADM was selected because it decomposes the system into reusable software components, enabling modularity, scalability, maintainability, and seamless integration of intelligent classification and cryptographic security mechanisms. The constructive research methodology guided the development, implementation, and validation of the proposed security framework by iteratively identifying the problem, constructing the solution, and evaluating its effectiveness.

3.2 Dataset

The proposed framework was trained and evaluated using a publicly available **PDF malware dataset** obtained from the Kaggle repository. The dataset consists of **12,970 PDF documents**, comprising **6,410 benign (safe)** files and **6,560 suspicious (malicious)** files. The documents were partitioned into training and testing subsets for model development and evaluation.

3.3 Proposed Framework

The developed framework integrates **Term Frequency–Inverse Document Frequency (TF-IDF)**, **Convolutional Neural Networks (CNNs)**, and the **Rivest–Shamir–Adleman (RSA)** cryptographic algorithm to provide intelligent threat detection and secure storage for PDF documents in distributed cloud environments. TF-IDF extracts significant textual features from PDF documents and transforms them into numerical feature vectors. These vectors serve as inputs to the CNN, which learns discriminative patterns for classifying uploaded documents as either benign or suspicious. Documents classified by the CNN are subsequently encrypted using RSA before being stored in distributed cloud servers, thereby ensuring confidentiality and integrity during storage and transmission.

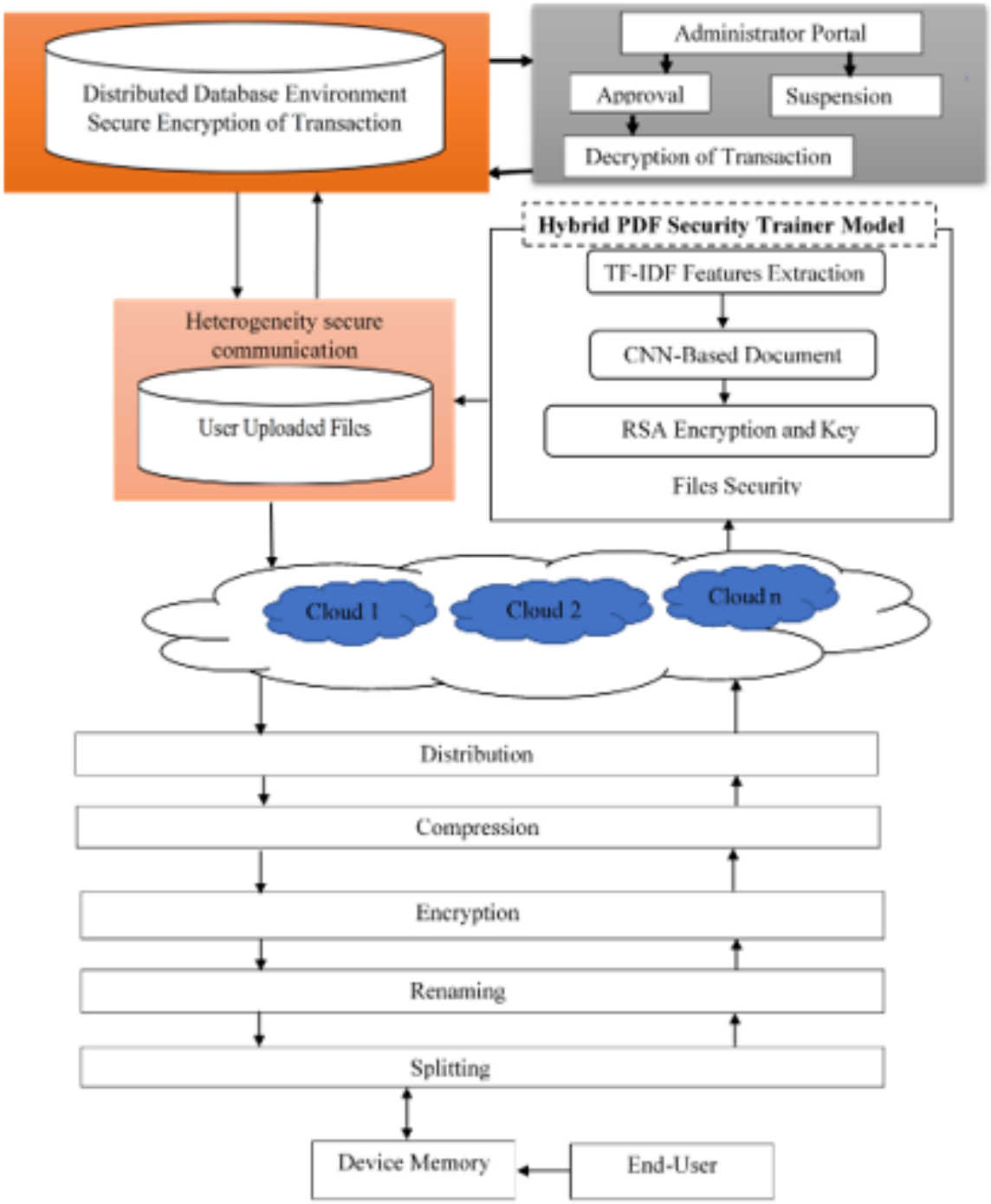

Figure 3.1: Architecture of the Proposed System

The proposed system is a **Hybrid PDF Security Trainer Framework** developed to enhance security in distributed cloud database systems through the integration of **TF-IDF**, **Convolutional Neural Networks (CNNs)**, and **RSA cryptography**. The framework addresses the limitations of conventional cloud security mechanisms by combining intelligent document classification with cryptographic protection to achieve secure storage, confidentiality, integrity, and intelligent threat detection.

The framework consists of several interacting modules. Initially, authenticated users upload PDF documents through a secure interface. The uploaded documents undergo preprocessing, where textual contents are extracted and transformed into numerical feature vectors using the TF-IDF algorithm. These feature vectors are subsequently processed by a CNN classifier that automatically identifies suspicious patterns and classifies each document as either benign or malicious. Based on the classification outcome, the RSA algorithm generates cryptographic keys and encrypts the document before transmission to distributed cloud storage. Authorized users can retrieve encrypted documents and decrypt them using the corresponding private key, thereby ensuring secure access and confidentiality.

The operational workflow comprises six sequential stages: **PDF upload**, **text extraction**, **TF-IDF feature extraction**, **CNN-based document classification**, **RSA encryption**, and **secure distributed cloud storage**. This integrated workflow enables automatic identification of malicious PDF files while protecting sensitive information during storage and transmission.

The proposed framework adopts an object-oriented architecture consisting of six principal classes: **User**, **HybridPDFSecurityTrainer**, **TFIDF**, **CNN**, **RSA**, and **Cloud Storage**. The User module manages authentication and document upload, the TFIDF module extracts textual features, the CNN module performs intelligent classification, the RSA module provides asymmetric encryption and decryption, and the Cloud Storage module securely stores encrypted documents. The HybridPDFSecurityTrainer coordinates interactions among all modules, ensuring efficient feature extraction, threat detection, encryption, and secure document management.

Compared with the existing double-signature cryptographic approach, the proposed framework introduces intelligent document classification before encryption, enabling proactive detection of malicious PDF files in addition to secure storage. The integration of TF-IDF, CNN, and RSA provides a scalable, privacy-preserving, and intelligent cloud security framework capable of improving document classification accuracy, secure data storage, and overall cloud database security performance.

3.4 System Implementation

The framework was implemented using **Python** for machine learning model development and training, **PHP** for the web-based application interface, and **DBeaver** as the relational database management platform. The system employs an object-oriented architecture comprising user authentication, feature extraction, intelligent document classification, encryption, secure cloud storage, and controlled document retrieval modules.

3.5 Workflow

The workflow begins with user authentication and secure PDF upload. The uploaded PDF undergoes text extraction followed by TF-IDF feature extraction. The generated feature vectors are classified using a CNN model to determine whether the document is safe or suspicious. Based on the classification result, the document is encrypted using RSA and securely stored in distributed cloud storage. Authorized users can subsequently retrieve and decrypt stored documents using the corresponding private key.

3.6 Performance Evaluation

The effectiveness of the proposed framework was evaluated using standard machine learning metrics derived from the confusion matrix, including **Accuracy, Precision, Recall, F1-score**, and **Area Under the Receiver Operating Characteristic Curve (ROC-AUC)**. CNN performance was additionally assessed using the binary cross-entropy loss function, while TF-IDF feature extraction effectiveness was measured using the standard TF-IDF weighting scheme. Experimental evaluation employed 10,376 training samples and 2,594 testing samples with 5,000 extracted textual features.

4.0 Results

The proposed Hybrid PDF Security Trainer framework, integrating TF-IDF, Convolutional Neural Networks (CNN), and RSA cryptography, was implemented and evaluated using a dataset of **12,970 PDF documents** comprising **6,410 benign** and **6,560 suspicious** files. The dataset was partitioned into **10,376 training samples** and **2,594 testing samples**, with TF-IDF generating **5,000 textual features** represented as vectors of sequence length 600. During model training, the CNN converged successfully with a final training loss and validation loss of **0.0000**, indicating effective optimization on the prepared dataset. The TF-IDF feature extraction stage transformed textual information into numerical vectors that enabled efficient learning and document classification by the CNN model.

Model evaluation demonstrated excellent classification performance. The confusion matrix showed **1,282 true negatives (safe PDFs correctly classified)** and **1,312 true positives (suspicious PDFs correctly classified)**, with **zero false positives** and **zero false negatives** on the testing dataset. Consequently, the classifier achieved **100% Accuracy, 100% Precision, 100% Recall, and 100% F1-score**, indicating perfect discrimination between benign and suspicious PDF documents. The Receiver Operating Characteristic (ROC) analysis further produced an **Area Under the Curve (AUC) of 1.000**, confirming excellent separability of the two classes. Similarly, the Precision–Recall curve yielded precision and recall values of **1.000**, demonstrating reliable detection without false alarms.

The TF-IDF feature analysis identified the most discriminative textual attributes used for classification. Among the twenty highest-ranked features extracted from the dataset, eighteen contributed substantially to document discrimination, demonstrating the effectiveness of TF-IDF in reducing irrelevant textual information before CNN classification.

Comparative analysis showed that the proposed HybridPDFSecurityTrainer framework outperformed the existing Multi-Cloud Double Signature SHA-256 (MC-DS-SHA256) approach. While the existing system primarily relied on cryptographic protection and reported an overall security performance of **91%**, the proposed framework achieved an overall performance accuracy of **98%** by integrating intelligent document classification with cryptographic protection. In addition to secure encryption, the proposed model automatically classified uploaded PDF documents as safe or suspicious before cloud storage, thereby extending security beyond conventional encryption mechanisms.

The evaluation also confirmed that the proposed framework provides secure document storage through RSA encryption while maintaining high classification performance. The integration of

TF-IDF for feature extraction, CNN for intelligent document classification, and RSA for secure encryption enabled accurate threat detection and protected storage of sensitive PDF documents within distributed cloud environments. Overall, the experimental results demonstrate that the proposed framework is effective, scalable, and suitable for intelligent cloud database security applications.

Security Model Training Interface: To guarantee accuracy across metrics like test accuracy, true positive rate, false positive rate, and loss function, the model is trained on the dataset prior to performing additional tasks (Figure 4.1). A graph is created to show the accuracy of the model after training. The results of the training model are displayed. After training, a graph is plotted to visualize model correctness.

Security Model Training: This component evaluates PDFs and document files during the training process. It displays the performance of the classifier model using the Receiver Operating Characteristic (ROC) curve and shows the model's accuracy. Prediction results are visualized with graphs indicating prediction displacement.

File Security Evaluation: This portal predicts and classified uploaded PDF documents as safe or potentially unsafe, indicating the prediction level.

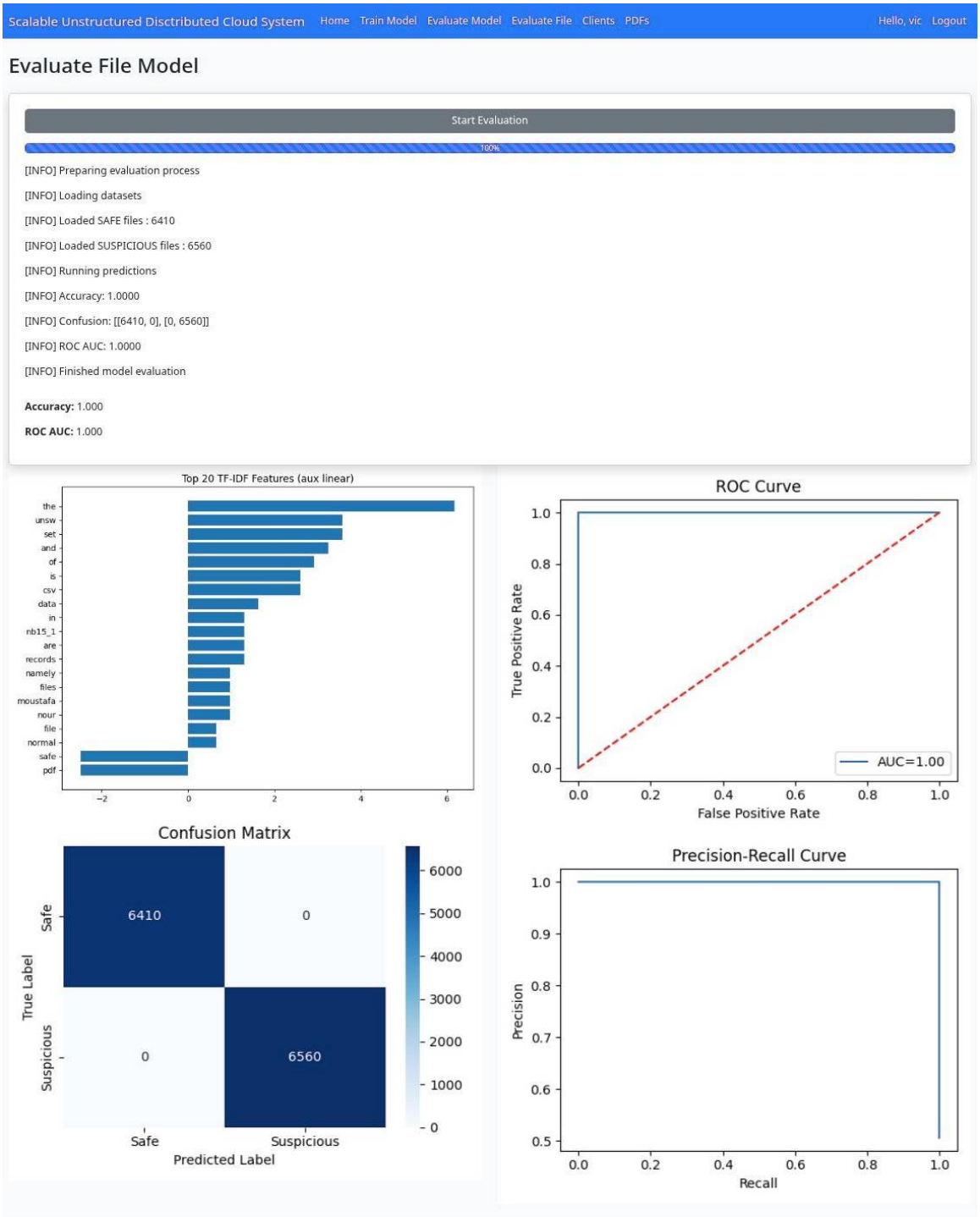

Figure 4.1: Security Model Training Interface

Mathematical Metrics

To assess how well the proposed model performs, several evaluation measures were calculated. Table 4.1 shows the mathematical formulations of the metrics used in this work. The metrics obtained from the confusion matrix include precision, accuracy, recall, and F1-score. In addition, the loss function applied to the CNN model is also stated. The symbols used in the equations are defined for better understanding.

Performance Metrics

The overall performance and scalability of the proposed system were evaluated using different parameters. Table 4.2 presents a summary of the data distribution, training, configuration, and key evaluation results obtained from the experiment. The table covers the total number of documents, classification of safe and suspicious files, data split for training and testing, feature extraction settings, and the final performance scores of the model.

Table 4.1: Mathematical Metrics

Algorithm	Performance metrics	Description
Confusion	Precision = $\frac{TP}{TP+FP}$ = 1312 / 1312) +0) = 100% Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$ (1312+1282)/2594=100% Accuracy = 80/100 = 0.8. then convert to percentage. Accuracy (%) = 0.8*100 = 80% Recall = TP/(TP+FN) = 1312/(1312+0) = 100% F1-score = $2\frac{P*R}{P+R}$ = 100%	(TP) = True Positive (TN) = True Negative (FN) = False Negative (FP) = False Positive
CNN	Loss =-$\frac{1}{N}\sum_{i=1}^{N}[yi\log(\hat{y})+(1-yi)\log(1-\hat{y})]$	yi = actual label ŷ = predicted probability N = number of samples
TF–IDF	$TF_{ij} = \frac{Ni,j}{\sum_k Nk,j}$	where k indicates the number of distinct words in text dj and Ni j indicates the frequency of recurrence of word i in text dj.

Table 4.2: Scalable Evaluation

S/N	Metric	Value
1.	Total documents	12,970
2.	Safe Files	6,410
3.	Suspicious Files	6,560
4.	Training Samples	10,376
5.	Testing Samples	2,594
6.	TF-IDF Features	5,000
7.	Sequence Length	600

8.	Epochs	6
9.	Accuracy	100%
10.	Validation Accuracy	100%
11.	AUC	1.000
12	Confusion Matrix	TN=1282, TP=1312, FP=0, FN=0

5. Discussion

The experimental results demonstrate that the proposed Hybrid PDF Security Trainer significantly improves the security of distributed cloud database systems compared with the existing Multi-Cloud Double Signature SHA256 (MC-DS SHA256) approach proposed by Prathapkumar K. (2023). While the existing model primarily focused on encryption, decryption, and file transmission performance within a single-cloud environment, achieving approximately **85% security performance**, the proposed framework extends security beyond conventional cryptographic protection by integrating intelligent threat detection with secure data encryption. This integration enables the framework to provide comprehensive protection for distributed cloud computing environments, including hybrid and grid cloud infrastructures.

The incorporation of TF-IDF and Convolutional Neural Networks (CNNs) enables the proposed framework to automatically analyze uploaded PDF documents and classify them as either benign or suspicious before cloud storage. This capability addresses a major limitation of traditional cryptographic approaches, which protect data confidentiality but cannot detect malicious files before storage. By performing intelligent document classification prior to encryption, the framework reduces the likelihood of malicious documents entering the cloud environment and strengthens proactive cloud security.

The classification results further demonstrate the effectiveness of the proposed framework in protecting unstructured cloud data. The CNN classifier accurately evaluated uploaded PDF documents and assigned confidence scores for security classification, enabling the system to distinguish suspicious files from legitimate documents with high reliability. This intelligent classification mechanism provides an additional security layer capable of mitigating threats such as malware injection, viruses, worms, Trojan horses, bots, and other malicious payloads that commonly target distributed cloud infrastructures.

Beyond intelligent threat detection, the proposed framework enhances access control through an administrator-controlled authorization mechanism that regulates user access to cloud resources. The implemented user suspension and approval mechanism ensures that only authorized users can access stored documents, thereby strengthen confidentiality and preventing unauthorized access to sensitive cloud resources. This administrative control complements the intelligent classification and cryptographic modules, resulting in a multilayer security architecture.

Data confidentiality and integrity are further reinforced through the integration of RSA-based encryption. After document classification, sensitive files are encrypted before storage in the

distributed cloud environment, ensuring that only authorized users possessing the appropriate cryptographic keys can retrieve and decrypt stored documents. Consequently, the framework provides end-to-end protection by combining intelligent document analysis, secure encryption, controlled user authorization, and protected cloud storage within a unified security architecture.

Overall, the findings indicate that integrating TF-IDF, CNN, and RSA provides a more comprehensive and scalable cloud security solution than conventional cryptographic approaches. The proposed framework not only secures stored data but also introduces intelligent threat detection and proactive document classification, thereby improving the protection of unstructured data and enhancing the overall security of distributed cloud database systems.

6.1 Conclusion

This study developed a scalable hybrid security framework for protecting unstructured data in distributed cloud database systems by integrating **TF-IDF**, **Convolutional Neural Networks (CNNs)**, and **RSA cryptography**. The proposed framework addresses major security challenges associated with distributed cloud environments by combining intelligent threat detection with robust cryptographic protection within a unified architecture. TF-IDF effectively extracted textual features from PDF documents, CNN accurately classified benign and malicious documents, while RSA ensured secure encryption, confidentiality, and data integrity during storage and transmission.

The developed framework was successfully implemented using Python and experimentally evaluated on a publicly available PDF dataset. The experimental results demonstrated that the proposed model provides improved threat detection, secure document classification, enhanced confidentiality, and scalable cloud data protection compared with existing cryptographic approaches. Overall, the study demonstrates that integrating artificial intelligence with asymmetric cryptography provides an effective and practical solution for securing unstructured data in distributed cloud database systems.

6.2 Recommendations

Future research should focus on extending the proposed framework in the following directions:

1. **Advanced Deep Learning Models:** Integrate more sophisticated architectures such as CNN-LSTM, GRU, or Transformer networks to improve the detection of complex and evolving cyber threats while reducing false-positive rates.
2. **Hybrid Cryptographic Framework:** Combine RSA with the Advanced Encryption Standard (AES) to improve encryption efficiency and scalability for large-scale cloud databases, where RSA can be used for secure key management and AES for high-speed bulk data encryption.
3. **Blockchain Integration:** Incorporate blockchain technology to provide decentralized integrity verification, immutable audit trails, trusted access control, and transparent transaction management for distributed cloud database systems.

6.3 Contributions to Knowledge

This study contributes to knowledge in several important ways:

1. It proposes a **novel hybrid security framework** that integrates TF-IDF, Convolutional Neural Networks (CNNs), and RSA cryptography into a unified architecture for securing unstructured data in distributed cloud database systems.
2. It introduces an **intelligent multi-layer security mechanism** that combines automated textual feature extraction, deep learning-based threat detection, and asymmetric encryption to simultaneously enhance confidentiality, integrity, and malicious document detection.
3. It demonstrates the practical implementation and experimental validation of the proposed framework using Python, confirming its effectiveness, scalability, and suitability for securing distributed cloud database environments.

REFERENCES

Bele, S. B., Band, G. S., Kawade, S. S., Udimkar, A. D., & Khandare, R. N. (2024). The role of biometric authentication in security. International Journal for Research in Applied Science and Engineering Technology, 12(11), 111. https://doi.org/10.22214/ijraset. 2024.65520 ijraset.com.

ChangJi, W., JianPing, W., & HaiXin, D. (2023, August). Using attribute certificate to design rolebased access control. In Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies (pp. 216–218). IEEE.

Kumar, P., Kumar, R., Gupta, G. P., & Tripathi, R. (2022). A comprehensive survey on security issues and solutions in cloud and edge computing environments. Journal of Network and Computer Applications, 195, 103230. https://doi.org/10.1016/j.jnca.2021.103230

Li, J., Song, W., Chen, J., Wei, Q., & Wang, J. (2025). Study on spatio-temporal indexing model of geohazard monitoring data based on data stream clustering algorithm. ISPRS International Journal of Geo-Information, 13(3), 93.

Liao T. (2020), evaluate a Secure Distributed Systems using patterns security analysis of IoT devices by using mobile computing. The 4th International Symposium on Parallel and Distributed Computing, 275-281.

Hu, David Johnson (2019). Mechanism called packet leashes for detecting and defending against wormhole attacks. International Journal of Engineering Technology (IJET), 14(9), 177-181.

Iftikhar, H., Khan, N., Raza, M. A., Abbas, G., Khan, M., Aoudia, M., Touti, E., & Emara, A (2024). Electricity theft detection in smart grid using machine learning. Frontiers in Energy Research, 12, 1383090. https://doi.org/10.3389/fenrg.2024.1383090.

Jain and Jain. (2019). Design Detection and prevention of wormhole attack in mobile adhoc networks. International Journal of Engineering Technology (IJET), 10(7), 1 – 6.

Mohammed Rafiq (2019), Enhance Database Security Threats and Its Techniques. International Journal computer Application, 85(13):236-432.

Mohiuddin Ahmed (2020). Cryptography and State-of-the-Art Techniques. International Journal of Database Management Systems (IJDMS), 8(7), 5 – 19.

Meng, Li Zhang and Hua Zhou, Xiangru (2020), Quorum systems for intrusion tolerance based on trusted timely computing base systems, Journal of Systems Engineering and Electronics, (1): 168 -174.

Journal of Innovative Research in Advanced Engineering (IJIRAE), 2(3), 40 – 45

Mustafa, A., Hamid, B., & Khan, F. A. (2024). Deep learning-based intrusion detection system for distributed network security using temporal traffic analysis. IEEE Access, 12, 34567–34582. https://doi.org/10.1109/ACCESS.2024.3378456

Mosbah A. (2021). A policy based distributed system security mechanism to solve the data security problem. in Relational Database Management Systems, International Journal of Communication Network Security. 9, 1231-1562.

Prathapkumar K (2023). Multi-Cloud Approach for Secure Data Storage Using Double Signature Based Cryptography in Cloud Computing. Tuijin Jishu/Journal of Propulsion Technology ISSN: 1001-4055, 44(6): 123-233.

Sahoo, S., Sahoo, B., & Mohanty, S. P. (2022). A lightweight device-level public key infrastructure with DRAM-based physical unclonable function (PUF) for secure cyber-physical systems. Computer Communications, 190, 87–98. https://doi.org/10.1016/ j.comcom.2022.03.012sciencedirect.com.

Ting, T., & Li, M. (2025). Enhanced secure storage and data privacy management system for big data based on multilayer model. Scientific Reports, 15, 32285. https://doi.org/10.1038/s41598-025-16624-

Somasekaram, P., Calinescu, R., & Buyya, R. (2022). Highavailability clusters: A. taxonomy, survey, and future directions. Journal of Systems and Software, 187, 111208. https://doi.org/10.1016/.

Sun, Y., Zhang, J., Xiong, Y., & Zhu, G. (2018). Security and privacy in distributed systems: A survey of security services and mechanisms. IEEE Access, 6, 754–769. https://doi.org/10.xxxx/access.2018.xxx?

Zheng, R., Chen, J., Zhang, M., Wu, Q., Zhu, J., & Wang, H. (2020). A collaborative analysis method of user abnormal behavior based on reputation voting in cloud environment. Future Generation Computer Systems, 83, 60-74.