Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

IoT-based Unstructured Data Deduplication Framework using Attention-based DenseRNN with Encryption

[1]Manjunath Singh H, [2]Dr Tanuja R

[1]Research scholar,

UVCE, Bangalore..

mansh24.singh@gmail.com.

[2]Associate professor,

UVCE, Bangalore..

tanujar.uvce@gmail.com

ARTICLE INFO

Received: 03 Nov 2025

Revised: 10 Dec 2025

Accepted: 20 Dec 2025

ABSTRACT

Due to the rapid expansion of digital data, centralized resources are offered by cloud computing systems to manage the data. Among its various services, data storage is one of the most widely adopted services all over the world. Data deduplication is a fundamental technology in cloud storage systems that help to save space by identifying and eliminating redundant copies of data. To maintain confidentiality, users encrypt their data before uploading it, but most of the traditional deduplication methods are ineffective when data is encrypted. As a result, optimizing deduplication with data encryption has been investigated in many research works to enhance storage efficiency and minimize network bandwidth usage. A major concern in this area is improving the security of deduplication mechanisms to defend against issues like data tampering, file ownership impersonation and fake duplicate generation. Addressing these limitations is essential for advancing robust and secure data deduplication methods in cloud environments. So, an efficient unstructured data deduplication mechanism is designed in the research work with novel techniques. Initially, unstructured data required for the validation is sourced from benchmark resources. Next, the unstructured data is converted into structured data with the help of entity. Then, data deduplication procedures are executed in the structured data using an Attention-based Dense Recurrent Neural Network (ADRNN). Further, the deduplicated data is encrypted using Elliptical Curve Cryptography (ECC) and stored in the cloud platform. Later, various experiments are executed to verify the overall efficiency of the developed framework over existing techniques.

Keywords: Unstructured Data Deduplication; Cloud Storage; Internet of Things; Attention-based Dense Recurrent Neural Network; Elliptical Curve Cryptography

Introduction

As cloud data volumes continue to increase, the system performance and security issues are increasing significantly. The rapid increase in digital data stored in cloud environments is often impacted by the presence of redundant data that increases the burden on cloud storage systems [6].

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The lack of structured format in data is referred as unstructured data and it is difficult to process and analyze [7]. Unstructured data is a collection of information that is not organized in a standard format like tables. The widespread use of sensors and various data-generating devices generates a huge volume of data that leads to duplication and inefficiencies in resource utilization, processing and storage [8]. A wide range of industries are now utilizing Internet of Things (IoT) devices to create, gather, and process data. This method supports efficient and rapid fine-grained data acquisition through various processing technologies. The data generated and captured by IoT systems is not only immense but also constantly changing. Cloud computing is generally used to store this extensive data for further analysis or processing [9]. However, the data from IoT devices often includes confidential information, such as product specifications. As a result, ensuring the privacy of this data before transferring it to the cloud is essential [10].

In order to improve storage efficiency, the data deduplication mechanism is utilized in large-scale storage environments. This technique reduces redundancy by eliminating repeated content and retaining only original documents [11]. Block-level deduplication and file-level deduplication are the two main methods used in deduplication. Block-level deduplication splits the data into smaller segments either fixed-size or variable-length and then removes repeated blocks while keeping the unique contents [12]. File-level deduplication identifies and removes duplicate files across the system by preserving only distinct files. But, block-level deduplication requires high computational demands due to the need for hashing and comparison, and it also leads to internal fragmentation, which affects the system performance [13]. Cloud-based deduplication methods have challenges in maintaining a balance between efficient storage, real-time data processing, and system scalability [14]. Several methods are employed for deduplication processes, including oblivious pseudo-random functions, homomorphic encryption, proof of ownership and Multiple Line Encryption (MLE). Yet, the current deduplication methods are not capable of maintaining strong data security when handling encrypted information [15]. Therefore, a data deduplication model specially for handling unstructured data is introduced to maintain storage issues in cloud platforms.

Contributions of the proposed unstructured data deduplication model are given below:

- To develop an advanced data deduplication framework for handling unstructured data by converting it into structured form through entity extraction. Due to the structured format, the proposed model is capable of enabling the effective processing of various data types such as logs and reports. Deduplicating unstructured data reduces storage overhead and eliminates redundancy in cloud storage applications. This approach does not eliminate the valuable details during this process and ensures that only unique data are stored securely.
- To implement an AD-RNN model for performing deduplication on structured data by capturing complex relationships between different entities. An attention mechanism is incorporated in this network for prioritizing unique and relevant data, while the dense recurrent layers ensure an efficient learning process. This leads to more precise detection of both original and similar data, even when their structure varies.
- To ensure secure storage of unique data, the ECC-based encryption is performed before uploading the data to the cloud. This encryption process is useful in maintaining the confidentiality of the user's original data. ECC does not increase the computational load and offers robust encryption without the need for large key sizes. The utilization of ECC-assisted encryption protects the sensitive information in cloud storage systems.

The executed unstructured data deduplication model is organized as follows: In Sub-part II, the existing data deduplication models are reviewed and some of the main drawbacks are listed. In Sub-part III, the architectural preview of the proposed model and the unstructured data transformation

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

process is given. In Sub-part IV, the data deduplication using the proposed model and the encryption process are described. The result and conclusion are given in Sub-part V and Sub-part VI, respectively.

Literature survey

A. *Related Works*

In 2024, Altowaijri [1] has suggested a grid-based hashing model for removing duplicate data in cloud storage. In order to perform an efficient deduplication process, the author used region-splitting and data aggregation approaches. This hash generation-based method helped to solve issues like energy consumption and computational overhead.

In 2025, Lapmoon and S. Fugkeaw [2] have introduced a secure and verifiable data deduplication model. Moreover, the unwanted data were automatically removed by a temporal data algorithm proposed in the work. The integrity of data in fog-assisted cloud storage was ensured by the developed technique.

In 2023, Mageshkumar *et al.* [3] have performed data deduplication with an authentication process for high data security. To encrypt the data, the Diffie-Hellman algorithm was utilized in this work. Block-level deduplication process has been accomplished in this work and this model was capable of handling external and internal attacks in the cloud.

In 2024, Akbar *et al.* [4] have recommended a Dynamic Prime Coding (DPC) model for performing deduplication. An index method was considered in this work for minimizing the execution time and increasing the throughput. Both privacy preservation and safe data exchange was accomplished by the model.

In 2024, Hamandawana *et al.* [5] have proposed a deduplication model that was capable of enhancing storage efficiency while reducing latency. The system's throughput is also enhanced by using an object replica method during deduplication. This model rectifies system failure with fast recovery time.

B. *Problem statement*

As internet technology advances, the amount of data increases and consumes more and more storage space. Unstructured data are very high dimensional and it is difficult to divide into smaller bits. So, the unstructured duplicate files require a lot of physical space and cost for better maintenance. Many research works have used new methods for performing data deduplication and some of the issues in those models are provided below:

- Most existing data deduplication approaches do not provide accurate results as they struggle to handle unstructured data formats. These traditional models often lack the capability to interpret or transform unstructured data that results in partial deduplication. Moreover, without effective unstructured data transformation mechanisms, removing irrelevant or redundant data entries is difficult and it increases the storage costs.
- Many current models do not incorporate advanced deep-learning mechanisms that capture temporal and contextual relationships in structured data. As a result, they fail to detect complex duplicate entries in the cloud that affects the reliability of storage systems.
- Security is one of the most significant concerns. Most of the existing deduplication systems do not incorporate advanced encryption techniques that affect the integrity of sensitive information stored in cloud environments.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

The merits and demerits of the conventional data deduplication models are given in Table I.

TABLE I. MERITS AND DEMERITS OF THE CONVENTIONAL DATA DEDUPLICATION MODELS

Author [citation]	Methodology	Features	Challenges
Altowaijri [1]	GH-EDA	It is capable of finding duplicate copies in noisy data. Average latency and computational overhead issues are rectified.	It is not capable of mitigating threats.
Lapmoon and S. Fugkeaw [2]	VERDUP	It minimizes the storage cost by automatically removing outdated data.	The adaptability and scalability of the suggested model is low. It is not suitable for handling the privacy of sensitive data.
Mageshkumar *et al.* [3]	Diffie-Hellman algorithm	It is capable of mitigating brute-force attacks. It rectifies operational overhead issues.	User privacy is not protected by this model.
Akbar *et al.* [4]	DPC	It reduces the storage cost and protects the data from channel attacks.	It does not have the ability to overcome security breaches. It is not capable of preventing unauthorized access.
Hamandawana *et al.* [5]	Speed-Dedup	It is most efficient in fault tolerance and the time required for failure recovery is high.	Efficiency in large-scale environments is low.

C. *Architectural Preview of Proposed Model*

Conventional data deduplication algorithms have issues in handling unstructured data. Traditional methods are not applicable in identifying redundancies in logs as well as documents as they are only capable of managing structured data. So, storage is not used efficiently, and duplicate content cannot be removed properly. Furthermore, a lot of current systems use shallow learning approaches, which are unable to recognize data entries that have different syntax. Such models are unable to attain high deduplication accuracy due to the absence of contextual understanding. Another serious issue in the existing model is security and most of the traditional frameworks do not include an encryption mechanism that increases the vulnerabilities of data. These problems result in lower scalability and

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

high computational costs when implemented in cloud-based infrastructures. Therefore, an advanced unstructured data deduplication model is developed and its structural view is depicted in Fig. 1.

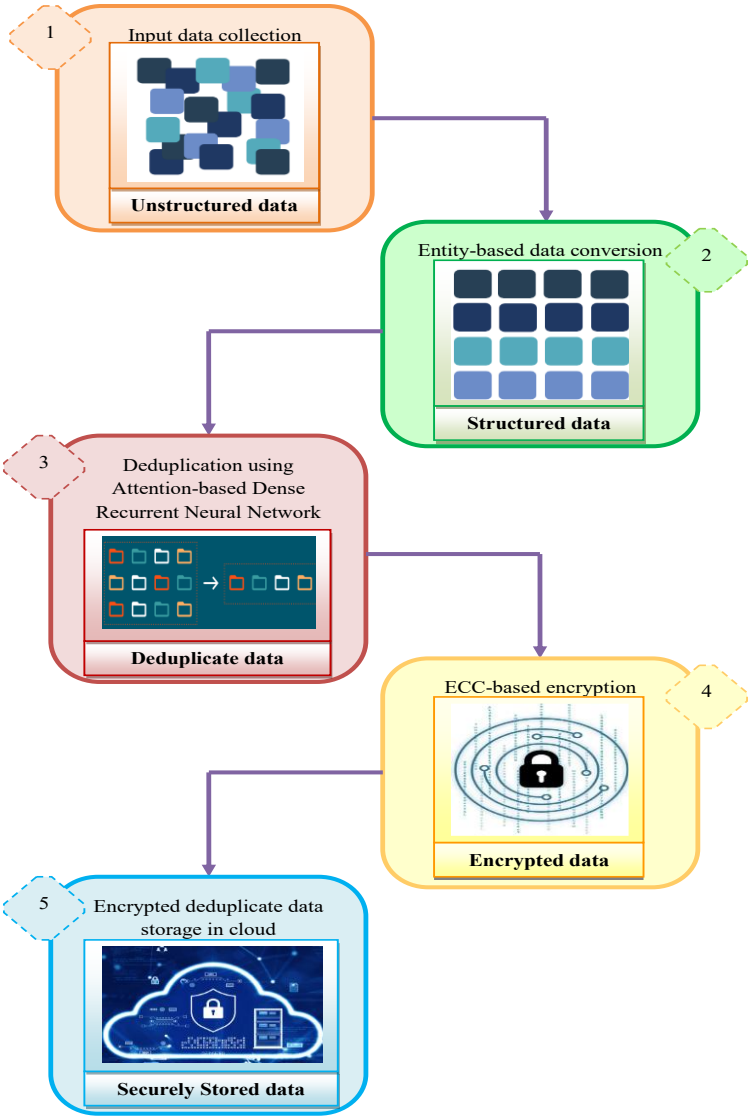


Fig. 1. Structural View of Proposed Unstructured Data Deduplication Model

This research presents an effective deduplication framework for processing unstructured data to improve storage efficiency and data security. The need for storage requirements are highly reduced by eliminating duplicate content from large volumes of unstructured data such as logs, and documents. By removing redundant and irrelevant entries, the deduplication process ensures only unique data is stored in the cloud applications. At first, unstructured data from benchmark datasets are collected. Unstructured data are converted into structured data using an entity-based conversion technique. In order to accurately identify and eliminate redundant information, this transformation is essential. An ADRNN model is used to perform the deduplication process, after the data has been organized in a structured format. The dense recurrent layers are useful for learning the sequential dependencies and

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

recognizing the duplicate data even it is a complex entity, while the attention mechanism concentrates on important data attributes for eliminating the irrelevant or duplicate data. The precision of differentiating the duplicate copies and the original data is high due to the dense architecture of the RNN mechanism. The original data are encrypted securely before storing on cloud storage. Here, the ECC technique is used for performing encryption as it is capable of providing robust security with small key sizes, so it is beneficial for storing data in cloud-based applications. The risk of unauthorized access is highly reduced while maintaining fast encryption as well as decryption. The encrypted data is stored on a cloud platform, which keeps it safe from unwanted access and reduces storage costs by eliminating duplicates. At last, experimental evaluation is carried out to evaluate the effectiveness of the suggested framework in terms of memory size and time consumption.

D. IoT-based Unstructured Data Collection

The IoT-based unstructured data is gathered from a benchmark site and it is described below:

Unstructured Data Analysis Project: This dataset is accessed using https://www.kaggle.com/code/tpvermei/unstructured-data-analysis-project on 2025-04-18. This database presents the details of input, output and logs. Raw text and its labels are given in this dataset. In consists of 5169 unique value for analyzing unstructured data. The collected data are indicated as U_c, here the variable c denotes the total unstructured data.

E. Entity-based Unstructured to Structured Data Conversion

The unstructured data are high-dimensional and it has various contents without a set of format. Entities are used in this work for data conversion. In order to create a structured representation, this entity-based approach is used. Every unstructured data point is examined to find relevant elements, which are arranged properly. Through this transformation, the data with unstructured formatting are transformed into structured data for feasible processing.

Consider a dataset U, which has unstructured data formats. Then, a combined entity is mathematically defined in Eq. (1).

$$S(U)=\{(s_a,s_b)\}|(s_a,s_b)\in U\times U, N(n_a,n_b) \quad (1)$$

Here, the real entity is denoted as n_a and the entity profile is mentioned as s_a. Filtering as well as matching process is done to reduce the computational complexity. Matching $V(U)$ is done based on below Eq. (2) for final conversion.

$$V(U)=V_U\subseteq U_f|MAX|V_U\cap S(U) \quad (2)$$

The entity-based approach searches particular patterns or indicators that denote the contents in unstructured text data. Based on that, it efficiently converts unstructured data into structured information by classifying and extracting pertinent entities from the text. It identifies the important attributes of raw data and aligns them into an attribute-value pair and provides a structured data format. The resultant structured data is termed as Q_y.

F. Description of ADRNN

The ADRNN [16] is used for the data deduplication process. The ADRNN is developed by including an attention layer in the DRNN mechanism to recognize the same copies in cloud storage. In DRNN,

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

the information flow is more efficient as the layers are connected in a feed-forward manner. The dense connections are useful in capturing more rich information. As the network is densely connected, the output from the previous layer is used by the next layer as input like given in Eq. (3).

$$F_n = G_n\left[\left(F_{0.....}, F_{n-1}\right)\right] \quad (3)$$

Here, the nonlinear transformation module is defined as G_n. The variable F_n defines the n^{th} feature. Concatenation of the output is given as $\left(F_{0.....}, F_{n-1}\right)$. The function of DRNN is defined in the following Eq. (4)-Eq. (6).

$$w_d = a\left(w_{d-1}, Q_y\right) \quad (4)$$

$$w_d = a\left[\left(v_{wq}.Q_y + v_{ww}.Q_{y-1}\right) + x_w\right] \quad (5)$$

$$\bar{t}_d = a\left[\left(v_{wq}.Q_y + v_{ww}.Q_{y-1}\right) + x_w\right] \quad (6)$$

In the above expression, the weight connections and the input is denoted as v_{wq} and Q_y, concurrently. The output at time d is mentioned as w_d and the predicted outcome is stated as $\bar{t}_d$. The weight between output and hidden layers are signified as v_{ww}. The dense connections send the information to the hidden layers and the important data are analyzed by the attention module. The attention function is mathematically modeled in Eq. (7) and Eq. (8).

$$w_d = DF\left(v^m A_d\right) \quad (7)$$

$$A_d = \tanh\left(v_g A_d\right) \quad (8)$$

Here, the weight matrices of the attention module is given as v_g and v^m, respectively. The term DF denotes the softmax function.

G. *Data Deduplication in Cloud using ADRNN*

The structured data Q_y is processed by the developed ADRNN for finding the duplicates. The DRNN architecture uses the sequential dependencies between records to find duplicate copies. The attention mechanism concentrates on the most important aspects of every record for improving the system's capacity to recognize the identical entries. In order to preserve only unique records, the model filters out redundant copies by categorizing as the duplicate records.

- At first, the unstructured data are divided into sections or blocks. Depending on the amount and type of data, a block contains a document, a paragraph, or a certain number of records.
- For deduplication, each block is handled as a separate unit, which lowers the computational burden and increases the deduplication accuracy.
- In the developed ADRNN model, an attention layer is added that highlights the significant elements in each block and it modifies the weights to reduce the number of blocks that are incorrectly identified as unique or duplicated.
- The model forecasts the probability that each block is a duplicate of other blocks. When evaluating similarity, the attention mechanism concentrates on important characteristics or entities that indicate duplication.
- Finally, the duplicate data are eliminated and the original data are sent for encryption.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Schematic representation of data deduplication in cloud using ADRNN is visualized in Fig. 2.

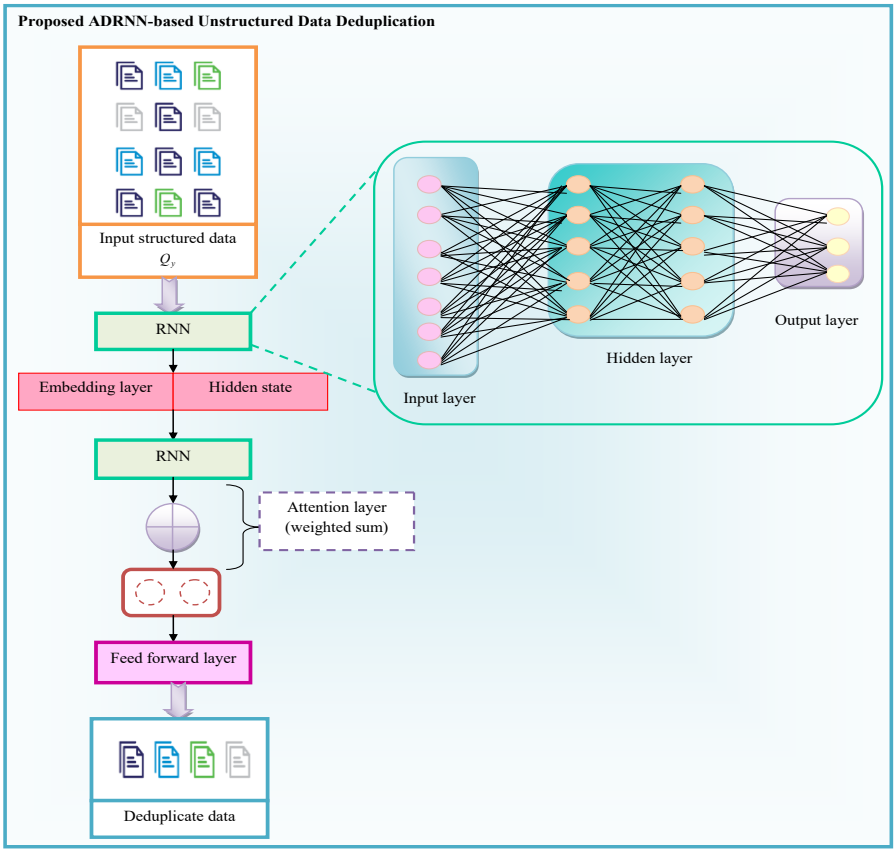


Fig. 2. Schematic representation of data deduplication in cloud using ADRNN

H. ECC-based Data Storage

Cloud storage is a service that allows data to be remotely managed and maintained. Users can access the service over internet connections. It enables users to store their data online so they may access it from anywhere in the world through the internet. Cloud storage reduces the user's hardware and software requirements. Yet, cloud storage faces several security issues. So, saving the data in cloud storage with encryption helps to protect the data from security risks.

ECC [17] is a modern and efficient public-key encryption technique. It works based on the mathematical properties of elliptic curves and it is defined in the below Eq. (9).

$$z^2 = w^3 + pw + q \tag{9}$$

Here, the terms p and q are constants. ECC offers high level of security while using smaller keys. The strength of ECC lies in the complexity of the Elliptic Curve Discrete Logarithm Problem (ECDLP). In this problem, for a given point m, the integer n is determined by the scalar values v are multiplication $m*n$. Various operations in ECC-based key generation are described in the following:

Point Addition: Merging two points to obtain a new point in the elliptic curve is the process of point addition. Consider two points $A(i_1, j_1)$ and $B(i_2, j_2)$, then the third point $N(i_3, j_3)$ is obtained by computing $A + B$.

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Point Multiplication: In this process, a new point is determined by multiplying an integer n with point A. So, a resultant point nA is obtained. In ECC, the point addition operation is repeated iteratively for performing scalar multiplication like in Eq. (10).

$$4A = A + A + A + A \tag{10}$$

Point Subtraction: The inverse function of the point addition operation is termed as point subtraction. If the point addition of two points is $B = A + N$, then the new point N is obtained by subtracting A from B. This process is arithmetically denoted in Eq. (11).

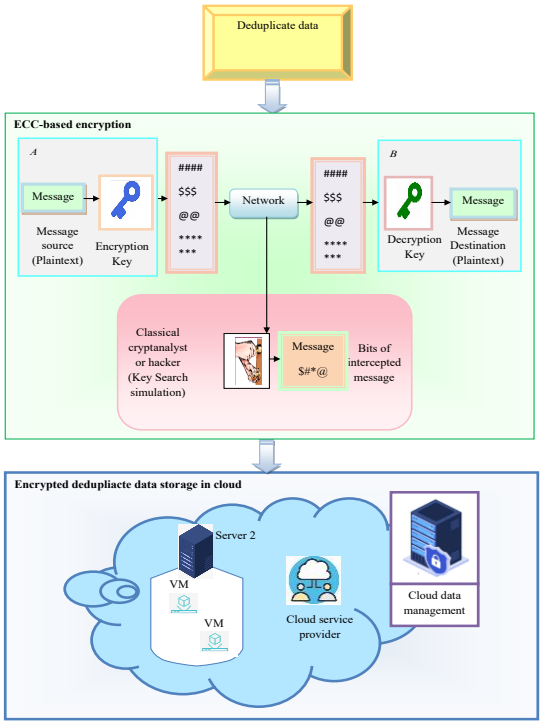
$$B - A = B + (-A) \tag{11}$$

Here, the negation of point A is given as $-A$.

Key Generation: The key generation begins with selecting a private key in a specific range. The corresponding public key is then calculated by multiplying this private key with a designated base point on the elliptic curve. This operation is known as scalar multiplication and it forms the basis for the public key as represented in Eq. (12).

$$LK = SK \times TW \tag{12}$$

In the above expression, the variables TW, SK and LK denotes the base point, private key and public key. An integer is multiplied with the base point to generate other points. The scalar value is the private key and it is not sharable, while the public key is point in the elliptic curve and it can be shared with anyone. After encryption, all the original files are uploaded in the cloud securely. Visual representation of ECC-based encryption for cloud data storage is shown in Fig. 3.



Visual Representation of ECC-based Encryption for Cloud Data Storage

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

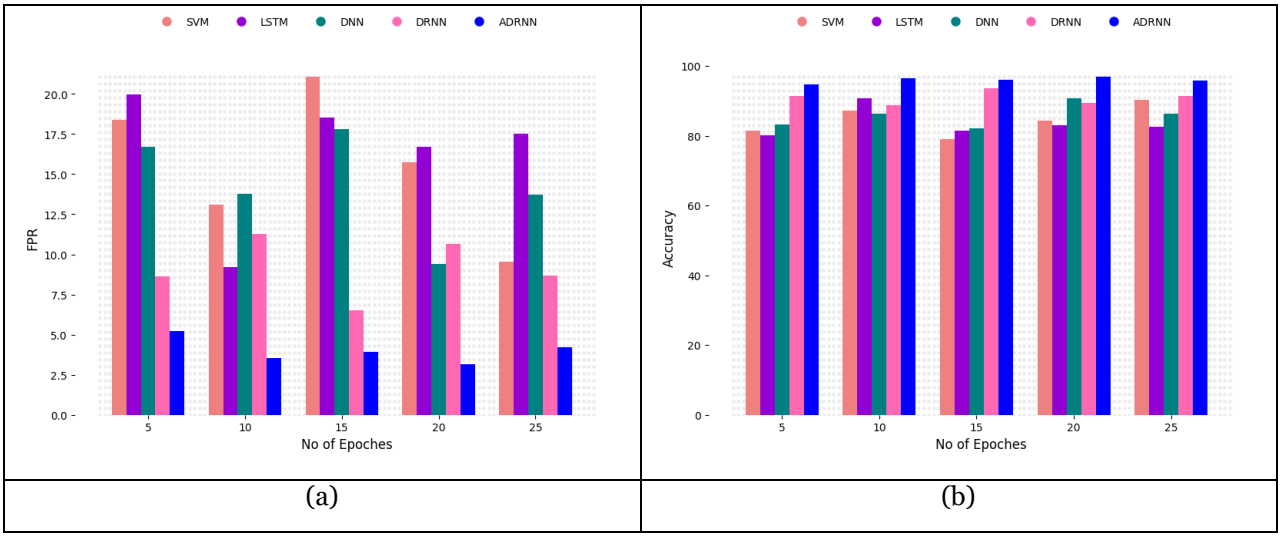
RESULTANT OBSERVATIONS AND DISCUSSIONS

I. Simulation Setup

The proposed unstructured data deduplication model was implemented in Python. The deduplication ability of the developed approach was estimated among encryption models like Fully Homomorphic Encryption (FHE) [19], Rivest-Shamir-Adleman (RSA) [18], Advanced Encryption Standard (AES) [20] and Data Encryption Standard (DES) [18]. Classifiers like Support Vector Machine (SVM) [22], Long Short-Term Memory (LSTM) [23] and Deep Neural Networks (DNN) [21] were used in this study for performance comparison.

J. Data Deduplication Performance Analysis

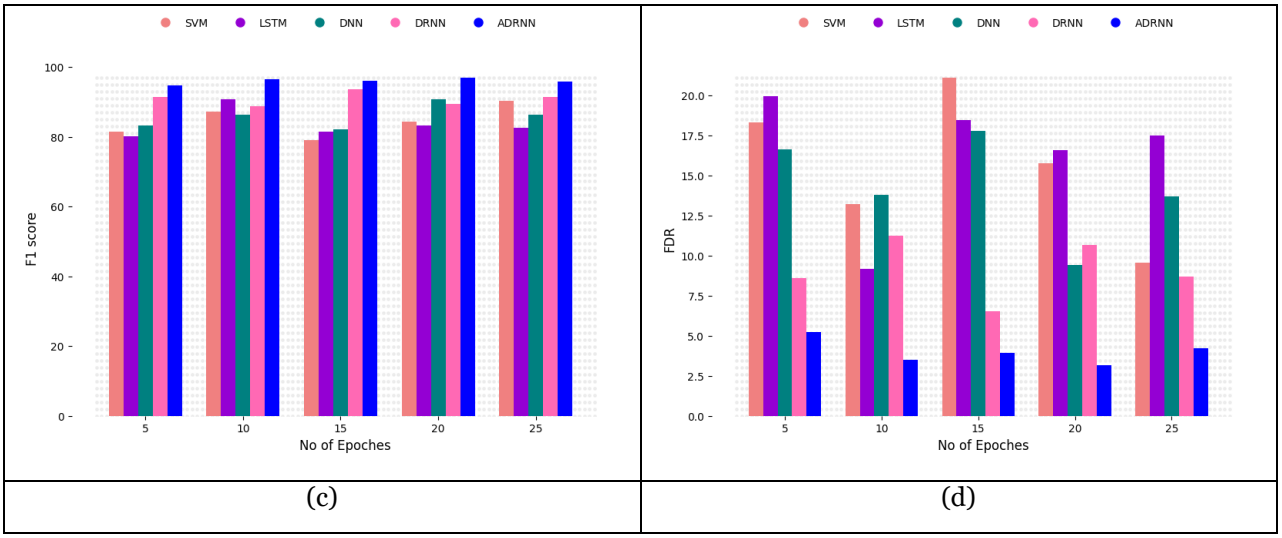
The data deduplication efficiency of the proposed model is validated among conventional deep learning techniques and the results are showcased in Fig. 4. The accuracy of deduplication process of the recommended ADRNN is higher than 90% in all epoch range that ensures its ability in eliminating redundant data. Existing approach like SVM is not applicable for handling large and complex data that result in poor performance. The DNN models might require additional feature extraction processes for finding the duplicate entities in structured data that increase the computational time and computational overhead of cloud storage systems. The integration of DRNN with the attention mechanism reduced the False Positive Rate (FPR) as it is capable of handling complex data files with high accuracy. The F1-score of the suggested ADRNN-based deduplication is 14.82%, 19.83%, 6.81% and 9.81% higher than SVM, LSTM, DNN and DRNN. The DRNN's accuracy is enhanced by incorporating the attention layer for analyzing the important as well as unique data.



(a) (b)

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

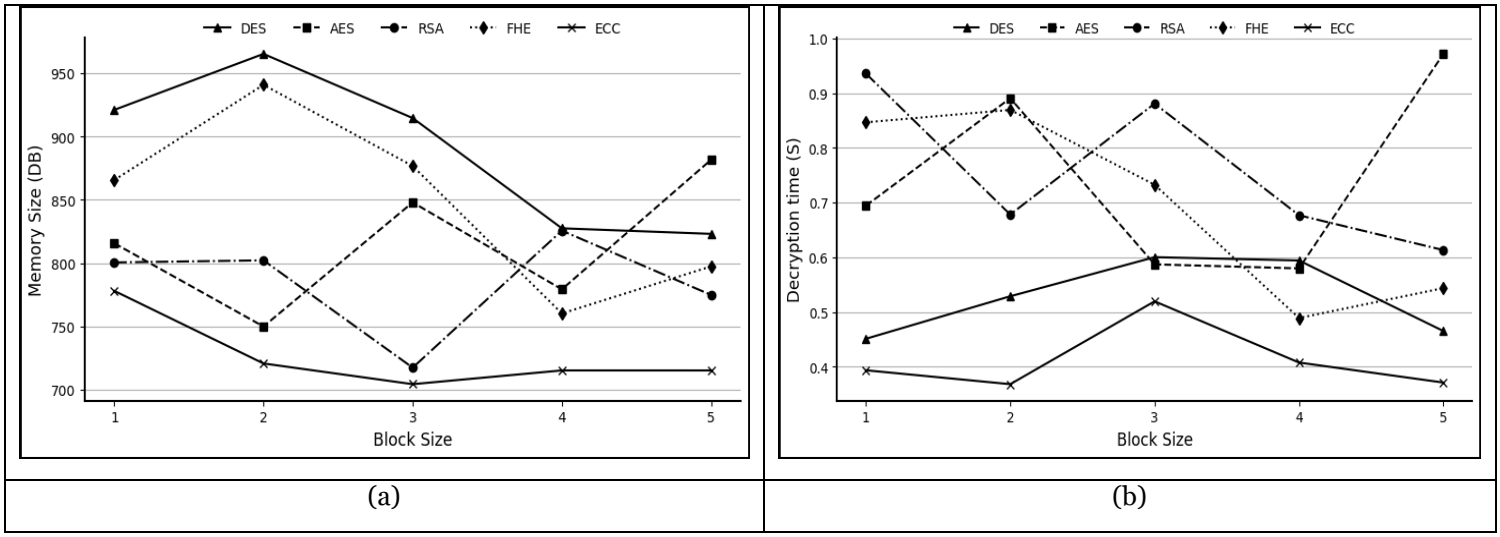


(c) (d)

Fig. 3. Data Deduplication Performance Analysis Among Other Classifiers with Respect to a) FPR, b) Accuracy, c) F1-score and d) FDR

K. Efficiency Validation of Encryption Task

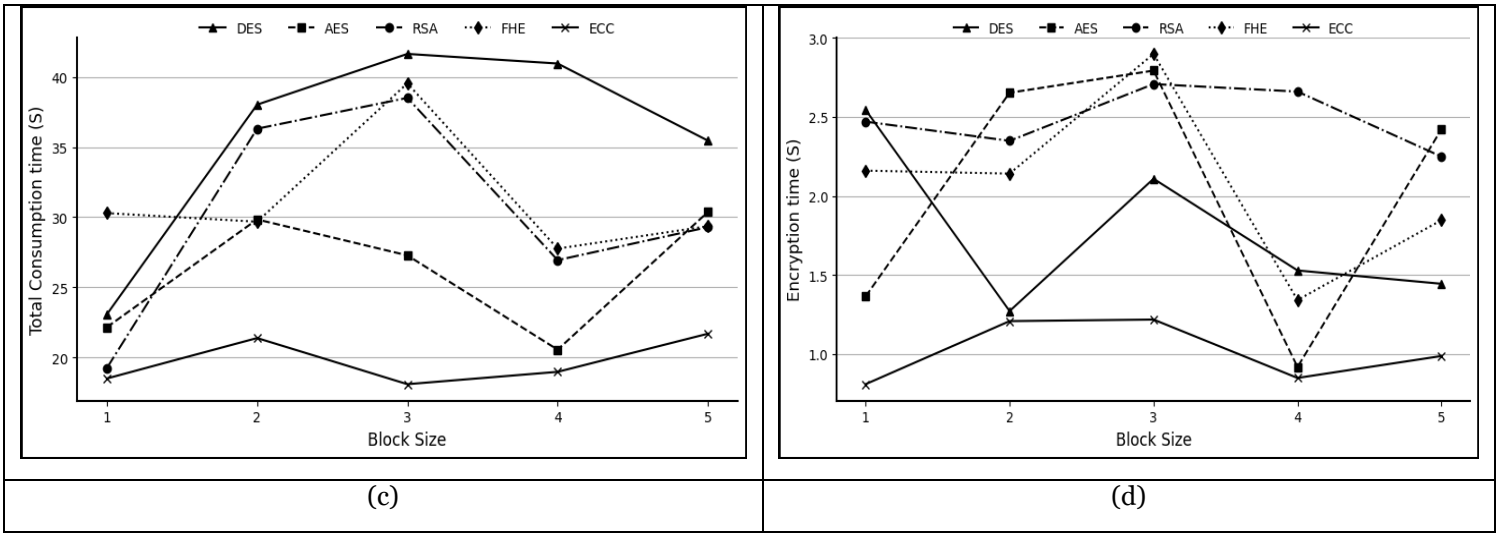
The efficiency validation of the ECC-based encryption task is analyzed and the outcomes are provided in Fig. 5. As per the graph, the time needed for the encryption technique is much less than conventional cryptography models like DES, AES, RSA and FHE. The computational overhead is highly minimized by the ECC-assisted encryption as it uses small key sizes. The encryption time of AES and RSA are higher as they utilize large key sizes that lead to an increase in computational complexity. Moreover, the memory size of the FHE is also higher than ECC algorithm in all block sizes and the FHE approach is not feasible in the encryption of large volumes of data. Therefore, the ECC approach offers high security while reducing the time and memory size.



(a) (b)

Journal of Information Systems Engineering and Management
2025, 10(63s)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article



(c) (d)

Fig. 4. Efficiency Validation of Encryption Performance in Terms of a) Memory size, b) Decryption time, c) Total consumption time and d) Encryption time

L. ROC Evaluation

The ROC evaluation to prove the efficiency of ADRNN in differentiating original and duplicate data is carried out and the graph is shown in Fig. 6. ADRNN achieves the highest true positive rate in finding the redundant data when compared to existing techniques like SVM, LSTM, DNN and DRNN. ROC graphs confirm the efficacy of using ADRNN for removing duplicates from unstructured data. Cloud data storage is managed efficiently without an increase in computational cost. Moreover, the proposed ADRNN and ECC-based encryption approach helps to handle large scale environments without any complexity issues.

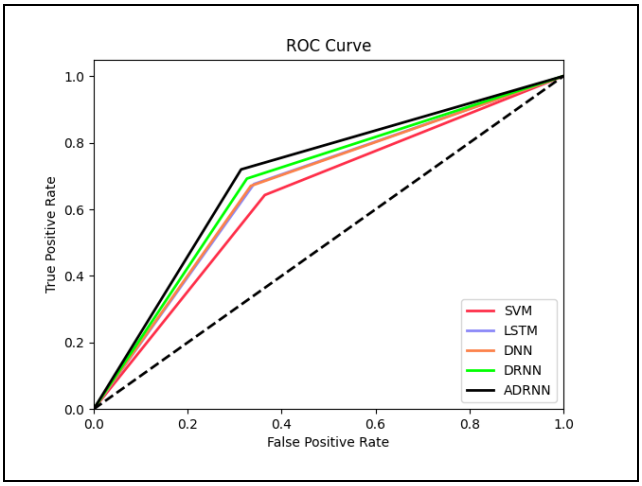


Fig. 5. ROC Evaluation of Proposed Unstructured Data Deduplication Model

Conclusion

In this research, an effective deduplication framework for processing unstructured data was developed. The proposed model was useful in improving the storage efficiency and data security. The collected unstructured data from benchmark datasets were converted into a structured format using entity. In order to accurately identify and eliminate redundant information, this transformation was essential. An ADRNN model was used to perform the deduplication process after the data has been

organized in a structured format. The precision of differentiating the duplicate copies and the original data was high due to the dense architecture of RNN mechanism. Then, an ECC technique was used for performing encryption before storing the unique data in the cloud. At last, experimental evaluation was carried out to evaluate the effectiveness of the suggested framework. The accuracy of the proposed unstructured data deduplication model using ADRNN is 13.68%, 15.78%, 11.57% and 5.26% higher than SVM, LSTM, DNN and DRNN at an epoch range of 5. Therefore, the proposed model was capable of finding duplicate unstructured data with high accuracy. In the future, attack detection models will be executed to protect the data stored in the cloud. Moreover, other inputs like images and texts will be considered to improve the generalizability of the proposed model.

References

[1] S. M. Altowaijri, "Efficient Data Aggregation and Duplicate Removal Using Grid-Based Hashing in Cloud-Assisted Industrial IoT," IEEE Access, vol. 12, pp. 145350-145365, 2024.

[2] J. Lapmoon and S. Fugkeaw, "A Verifiable and Secure Industrial IoT Data Deduplication Scheme With Real-Time Data Integrity Checking in Fog-Assisted Cloud Environments," IEEE Access, vol. 13, pp. 11969-11988, 2025.

[3] Mageshkumar, Nagappan, J. Swapna, A. Pandiaraj, R. Rajakumar, Moez Krichen, and Vinayakumar Ravi, "Hybrid cloud storage system with enhanced multilayer cryptosystem for secure deduplication in cloud," International Journal of Intelligent Networks, vol. 4, pp. 301-309, 2023.

[4] Akbar, Mohd, Irshad Ahmad, Mohsina Mirza, Manavver Ali, and Praveen Barmavatu, "Enhanced authentication for de-duplication of big data on cloud storage system using machine learning approach," Cluster Computing, vol. 27, no. 3, pp. 3683-3702, 2024.

[5] Hamandawana, Prince, Da-Jung Cho, and Tae-Sun Chung, "Speed-Dedup: A New Deduplication Framework for Enhanced Performance and Reduced Overhead in Scale-Out Storage," Electronics, vol. 13, no. 22, pp.4393, 2024.

[6] Cao, Huan, Shengdong Du, Jie Hu, Yan Yang, Shi-Jinn Horng, and Tianrui Li, "Graph deep active learning framework for data deduplication," Big Data Mining and Analytics, vol. 7, no. 3, pp.753-764, 2024.

[7] Hamandawana, Prince, Awais Khan, Jongik Kim, and Tae-Sun Chung, "Accelerating ML/DL applications with hierarchical caching on deduplication storage clusters," IEEE Transactions on Big Data, vol. 8, no. 6, pp.1622-1636, 2021.

[8] Rajkumar, K., U. Hariharan, V. Dhanakoti, and N. Muthukumaran, "A secure framework for managing data in cloud storage using rapid asymmetric maximum based dynamic size chunking and fuzzy logic for deduplication," Wireless Networks, vol. 30, no. 1, pp.321-334, 2024.

[9] Liu, Xinyao, Shengdong Du, Fengmao Lv, Hongtao Xue, Jie Hu, and Tianrui Li, "A Pre-trained Data Deduplication Model based on Active Learning," arXiv preprint arXiv:2308.00721, 2023.

[10] Ahmad, Shahnawaz, Mohd Arif, Javed Ahmad, Mohd Nazim, and Shabana Mehfuz, "Convergent encryption enabled secure data deduplication algorithm for cloud environment," Concurrency and Computation: Practice and Experience, vol. 36, no. 21, pp.e8205, 2024.

[11] Amrita Raj, and G. Uma Devi, "Optimizing Storage Efficiency in Hadoop Distribution File System Through Data Deduplication," SCIENCE AND TECHNOLOGY PUBLICATIONS, 2024.

[12] Venkatesh, Sai Vishwanath, Atra Akandeh, and Madhu Lokanath, "MetaPix: A Data-Centric AI Development Platform for Efficient Management and Utilization of Unstructured Computer Vision Data," arXiv preprint arXiv:2409.12289, 2024.

[13] Tang, Xinyu, Cheng Guo, Kim-Kwang Raymond Choo, Xueru Jiang, and Yining Liu, "A secure and lightweight cloud data deduplication scheme with efficient access control and key management," Computer Communications, vol. 222, pp.209-219, 2024.

[14] Zhao, Mark, Dhruv Choudhary, Devashish Tyagi, Ajay Somani, Max Kaplan, Sung-Han Lin, Sarunya Pumma et al, "RecD: Deduplication for end-to-end deep learning recommendation model training infrastructure," Proceedings of Machine Learning and Systems, vol. 5, pp. 754-767, 2023.

[15] Sylvana Yakhni, Joe Tekli, Elio Mansour, and Richard Chbeir, "Using fuzzy reasoning to improve redundancy elimination for data deduplication in connected environments," Soft Computing, vol. 27, no. 17, pp.12387-12418, 2023.

[16] Kasongo, Sydney Mambwe, "A deep learning technique for intrusion detection system using a Recurrent Neural Networks based framework," Computer Communications, vol. 199, pp.113-125, 2023.

[17] Kumar, Sanjay, and Deepmala Sharma, "A chaotic based image encryption scheme using elliptic curve cryptography and genetic algorithm," Artificial Intelligence Review, vol. 57, no. 4, pp.87, 2024.

[18] Yadav, Rajan Kumar, Munish Saran, Pranjal Maurya, Sangeeta Devi, and Upendra Nath Tripathi, "Hybrid DES-RSA Model for the Security of Data over Cloud Storage," Computer Integrated Manufacturing Systems, vol. 29, no. 7, pp. 177-190, 2023.

[19] de Castro, Leo, Antigoni Polychroniadou, and Daniel Escudero, "Privacy-Preserving Large Language Model Inference via GPU-Accelerated Fully Homomorphic Encryption," Neurips Safe Generative AI Workshop, 2024.

[20] Li, ZhenQiang, BinBin Cai, HongWei Sun, HaiLing Liu, LinChun Wan, SuJuan Qin, QiaoYan Wen, and Fei Gao, "Novel quantum circuit implementation of advanced encryption standard with low costs," Science China Physics, Mechanics & Astronomy, vol. 65, no. 9, pp.290311, 2022.

[21] Ghosh, Soumendu Kumar, Arnab Raha, Vijay Raghunathan, and Anand Raghunathan, "Partnner: Platform-agnostic adaptive edge-cloud dnn partitioning for minimizing end-to-end latency," ACM Transactions on Embedded Computing Systems, vol. 23, no. 1, pp.1-38, 2024.

[22] Mansour, Mohamed, Jan Martens, and Jörg Blankenbach, "Hierarchical SVM for semantic segmentation of 3D point clouds for infrastructure scenes," Infrastructures, vol. 9, no. 5, pp.83, 2024.

[23] Bi, Jing, Haisen Ma, Haitao Yuan, and Jia Zhang, "Accurate prediction of workloads and resources with multi-head attention and hybrid LSTM for cloud data centers," IEEE Transactions on Sustainable Computing, vol. 8, no. 3, pp.375-384, 2023.