INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Improved Cloud Storage Performance using the Data Deduplication-Based Efficient Cloud Optimization (D2-ECO) Model

J Jebamalar Tamilselvi[1]*, V Uma Rani[2], Rajan John[3], Surya Susan Thomas[4]

1 Department of Cybersecurity, Faculty of Science and Humanities, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India
2 Department of Computer Science and Engineering, Saveetha Engineering College, Chennai, Tamil Nadu, India
3 Department of Computer Science, College of Engineering & Computer Science, Jazan University, Jazan , Kingdom of Saudi Arabia
4 Department of Computer Science with Data Science, Patrician College of Arts and Science, Adyar, Chennai, Tamil Nadu, India

OPEN ACCESS

Received: 18/02/2026
Accepted: 25/04/2026
Published: 13/05/2026

Citation: Jebamalar Tamilselvi J, Uma Rani V, John R, Thomas SS (2026) Improved Cloud Storage Performance using the Data Deduplication-Based Efficient Cloud Optimization (D2-ECO) Model. Indian Journal of Science and Technology 19(18): 1157-1168. https://doi.org/10.17485/IJST/v19i18.246

* **Corresponding author**.

jebamalj@srmist.edu.in

Funding: None

Competing Interests: None

Copyright: © 2026 Jebamalar Tamilselvi et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background/Objectives: The rapid expansion of heterogeneous data in cloud environments has made efficient storage management increasingly challenging. Conventional deduplication techniques often fail to accurately identify duplicate content in large-scale, semantically diverse datasets, leading to higher storage consumption, increased latency, and elevated operational costs. This study aims to design a robust and intelligent data deduplication framework that enhances storage utilization, reduces redundancy, and improves overall cloud system performance. **Method**: To address these challenges, a Data Deduplication-based Efficient Cloud Optimization (D2-ECO) model is proposed. The framework integrates a Variable Size-Chunking Algorithm (VSCA) to dynamically segment and validate data blocks during preprocessing. Semantic similarity is then analyzed using a Bidirectional Encoder Representation from Transformers (BERT) model to capture contextual word relationships beyond surface-level matching. Further, a Fuzzy C-Means Cluster Indexing (FC-MCI) mechanism groups related data objects based on similarity scores to accurately detect duplicate and near-duplicate files. The integrated architecture ensures adaptive indexing, intelligent similarity evaluation, and optimized storage allocation. Furthermore, the final result y can be enhanced by integrating all methods into a framework that enables efficient semantic and structural duplicate removal. **Findings**: Experimental evaluation demonstrates that the proposed D2-ECO model achieves up to 96.8% improvement in overall performance efficiency compared to conventional deduplication approaches. The system significantly enhances storage utilization, reduces redundancy, lowers latency, improves data integrity, and decreases operational costs. Quantitative analysis shows improved duplicate detection accuracy and faster processing

time, confirming the effectiveness of combining semantic learning with adaptive clustering techniques. **Novelty**: The primary scientific contribution of this study lies in the integration of semantic language modeling with adaptive clustering within a cloud-based deduplication framework. Unlike traditional hash-based or fixed-chunk deduplication methods, the proposed D2-ECO model combines dynamic chunking, contextual similarity analysis, and fuzzy clustering in a unified architecture. This hybrid design enables accurate detection of both exact and near-duplicate data, addressing limitations in existing storage optimization models and providing a scalable, intelligent solution for modern cloud infrastructures.
Keywords: Data Deduplication; Cloud computing; Storage performance; D2-ECO; FC-MCI

1 Introduction

Cloud storage service providers meet these requirements by enabling enterprises and individuals to store, transfer, back up, and access data across other cloud services at a nominal cost [(1)]. To make data storage efficient, cloud service providers employ the popular drag-and-drop technique. This methodology minimizes storage overhead and reduces upload bandwidth by means of deduplication, in which only one copy of a data file is stored. While there are various data storage options, including cloud storage, one of the biggest challenges users and organizations face is data duplication. It has been observed that using different data sources results in significant duplication within the same user or transaction.

In recent years, as the amount of data continues to grow and cloud storage evolves, an increasing number of users have selected to store their backup data in the cloud, a choice that enables them to perform disaster recovery with higher efficiency [(2)]. Price backup volume is a guarantee that customers will be able to invest in storage in advance, and flexible storage capacity saves a significant amount of time [(2)]. Natural Language Processing (NLP) is a powerful form of communication. The database owner stores the database in a cloud system and offers a flexible, cost-effective database storage service. Over the last 10 years, the bulk of individuals and organisations have been storing their personal data with cloud service providers, simply because cloud computing is much cheaper and easier to manage [(3)]. But they face many new security and reliability challenges with this ambitious data management paradigm. The first problem is the quality of the cloud service. Cloud storage capacity is growing exponentially, and because multiple users who access the same cloud data also store it locally, data duplication occurs.

Therefore, these ciphertexts must be stored in the cloud, as cloud servers cannot identify such duplicate data. Compared to file-level deduplication, block-level deduplication provides improved deduplication by identifying identical content across different files, but it also increases the complexity of the scheme [(4)]. This huge volume of textual information on these sites presents both an opportunity and a challenge in comprehending and making decisions regarding organizations, institutions, and the population's opinions [(5)]. Nevertheless, these methods struggle to manage the high volume, variety, and speed of social media information, which is described as noisy, informal, and dynamic.

The main contribution of this research is the design of new individual algorithms, but rather the composition of VSCA, BERT, and FC-MCI into a unified solution for structural and semantic duplicate detection in cloud environments. In contrast to the above methods, which use chunking, semantic analysis, or clustering alone, our proposed D2-ECO strategy is a pipeline of VSCA, BERT-based semantic similarity, and FC-MCI. Integration like this allows for highly precise, exact, and near-duplicate detection while simultaneously maximising storage, latency, and operational costs. The presented amalgamation-based scheme clearly outperforms both the traditional and state-of-the-art approaches for deduplication. Various performance evaluations can identify duplicate data in cloud systems, enhance storage efficiency cost-effectively, and achieve significant progress.

1.1 Literature Survey

In [(6)], presented a comprehensive overview which provides a thorough analysis of the storage approaches, ideas, and types used in data deduplication. The Private Information Retrieval-based Tag-flexible Deduplication-supported Integrity Check Protocol (TDICP) by introducing a new verification tag called a reference set, which allows multiple users to create verification tags and supports tag deduplication [(7)]. This novel approach [(8)] combines the Diffie-Hellman algorithm (DHA) with symmetric external decision making to protect and disseminate information and ensure end-to-end encryption, thereby promoting user adoption of cloud storage. It offered [(9)] a new solution to the issue of deduplication in constrained organizations, eliminating duplicate copies and enhancing storage space utilization. They proposed a Content-Defined Deduplication (CDC) technique, which is a combination of inline deduplication and post-processing deduplication [(10)]. Nevertheless, it has a large computational overhead and latency and thus lacks flexibility in adapting to emerging data formats.

In [11] applied A Graph Convolutional Network (GNN) model to identify duplicate neighbors based on semantic relations between text documents. In [12], provided a literature review of various algorithms proposed for secure deduplication in cloud storage, along with current research trends.

Table 1. NLP Based on File deduplication in Cloud Environment

Reference	Technique Used	Drawback	Accuracy
[13]	Long Short Term Memory (LSTM)	The removal of duplicate entities is essential for adding value in data integration and performance, as well as for resource efficiency.	91.23%
[14]	Triple Data Encryption Standard (TDES)	Increased computational overhead due to fuzzy-rule processing in large-scale storage.	92.8%
[15]	Elliptical Curve Cryptography (ECC)	Higher key management complexity for multiple users	93.6%
[16]	Advanced Encryption Standard (AES)	Performance degradation during large-volume duplicate checking	92.14%
[17]	Feature-Aware Stateful Routing method (FASR)	The reduction rate is higher through similarity and configuration of the data	90.30%
[18]	Merkle Hash Tree based MapReduce (MHTMR)	Additional storage-layer complexity in cross-domain environments	89.36%

As shown in Table 1, various techniques obtained from the previous module provide significant improvements and accuracy using NLP based on file deduplication in cloud environment.

The Lightweight Convolutional Neural Network (LWCNN) model is employed to estimate the data sensitivity, categorizing it into two types: normal sensitivity data and high-sensitivity data [19]. To address the overlap issue on the semantic level, a pre-trained overlap model based on active learning is employed to resolve the problem [20]. Test outcomes depict that the recall rate on the benchmarked database reaches up to 28% higher.

2 Methodology

The proposed D2-ECO model can detect duplicate data and optimize the cloud storage of data chunks modeled using VSCA. BERT technology identifies word scores based on semantic relationships, and FC-MCI algorithm detects duplicate data based on file similarities. Moreover, the D2-ECO optimizes the storage capacity of cloud systems cost-effectively. Even though VSCA, BERT, and Fuzzy C-Means are well-known individual techniques, to the best of our knowledge, their combined application in a sequential framework for cloud data deduplication is low. The proposed model combines chunk-level validation, semantic similarity analysis, and clustering-based indexing of building blocks to enhance the precision of duplicate detection and system performance. The proposed approach is used to structure three synergistic modules: data chunk validation, semantic similarity analysis, and clustering–based indexing. Although individual techniques have been studied in previous work, few consider joint optimisation within a single pipeline for cloud data deduplication. This fusion enables the system to overcome traditional limitations and improves duplicate detection rates and overall system efficiency.

As shown in Figure 1, the D2-ECO model offers a solution for detecting redundant data in cloud environments, thereby enhancing cloud storage efficiency. Furthermore, the VSCA method plays a vital role in detecting structured data chunks, which is essential for effective data reduction. Based on the identified data chunks, BERT technique is implemented. BERT is used to determine word scores, which are based on the semantic relationships between words in data. The analysis of semantic relationships improves the accuracy of finding duplicate data. FC-MCI algorithm detects duplicate data by evaluating similarity between different documents or files. Finally, the D2-ECO model is introduced as a mechanism to find duplicate data in cloud systems. By identifying and eliminating this redundant data, the D2-ECO model improves storage efficiency. By achieving improved storage performance while remaining cost-effective, the D2-ECO model is a cost-effective solution for cloud storage optimization.

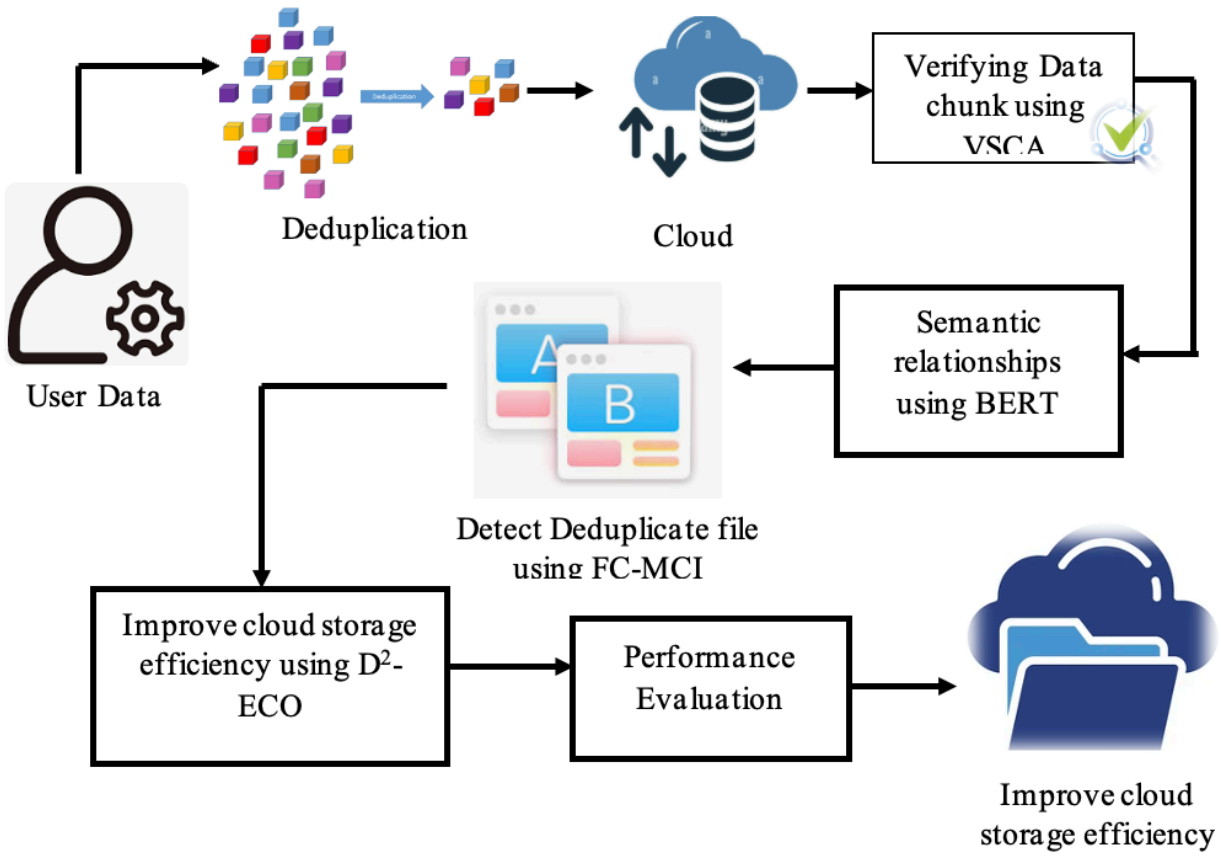


Fig 1. The Proposed D2-ECO method Architecture Diagram

2.1 Variable Size-Chunking Algorithm (VSCA)

This section uses the VSCA model to identify data chunks formed by dividing the input data file into small non-overlapping parts during the data copy operation. Additionally, rolling performance using a rolling hash based on the text content. The VSCA model utilizes hash values to identify data chunks formatted with predefined window sizes.

As shown in **equation 1**, the Robin fingerprint of the first substring is calculated based on the text content. Let's assume B— Input data string or file content, w—represent the sliding window size, p— Constant multiplier used in Rabin fingerprint computation, h—denote the hash value, and B_i— Character at the position in the substring, α—Length of the substring used for chunking, $ModD$— Prime divisor used for modulo operation The rolling hash function is computed as follows:

$$Rabin\left(B_1, B_2, \ldots, B_\alpha\right) = \left\{\sum_{i=0}^{\alpha}\left(B_i * P^{\alpha-i-1}\right)\right\} ModD \tag{1}$$

As shown in **equation 2**, a chunk of data is validated by removing the first character and adding a new character to the remaining substring. Let's assume $Rabin\left(B_i, B_{i+1}, \ldots, B_{i+\alpha-1}\right)$—Rabin fingerprint hash function, i—Current position of the sliding window, $B_{i+\alpha}$— window during rolling hash update

$$\left\{\left[Rabin\left(B_i, B_{i+1}, \ldots, B_{i+\alpha-1}\right) - B_i * P^{\alpha-1}\right] * P + B_{i+\alpha}\right\} ModD \tag{2}$$

In cloud environments, significant improvements can be achieved by adding new characters to the substrings to verify the format of data chunks.

2.2 Bidirectional Encoder Representation from Transformers (BERT)

This section applies the BERT algorithm to identify word scores by weighting edges based on semantic relationships. Additionally, semantic scores can be calculated based on word co-occurrences using the BERT approach to word embedding.

Afterward, features of the vector are analyzed by all the semantic word scores, and each of them consists of a column. The semantic scores can be located with the help of a method that converts words into numbers, applying hot encoding and utilizing numerous columns. Let's assume S—similarity sentence, $N-ES$—relationfeatures, $J-$ tuple of semantic score, N—normalized semantic features, RES—realtionship extract sentence, and F-features.

Semantic similarity is calculated using vector representations extracted from BERT embeddings.

Algorithm Steps
Input: verify data chunks$D$
Output: Identify the similarity of word scores and the semantic relationship
Start
For each $s\varepsilon S(S)$ do
Remove S,
Calculate semantic words with BERT models
End for each
Calculate several features and similarity scores
Reconfigure $J = \alpha$
If $s\varepsilon S(f)$ do
For each sentence in S do

$$M = BERT\ (s)$$

If $r\varepsilon R\ (f)$ do

$$N = r \cup M$$

$$RES = Original_{F} - BERT_{F}$$

$$j = (N,\ RES)$$

$$Add\ j\ to\ J$$

End if
End for each
End if
Calculate the similarity scores of the semantic relationships.
Return J

The BERT model can be used to evaluate word scores based on the semantic relationships in the input encoder block layers.

2.3 Fuzzy C-Means Cluster Indexing (FC-MCI)

The FC-MCI algorithm is employed to identify duplicate files by analyzing the similarity between documents or files during duplicate output information retrieval. The FC-MCI algorithm is implemented to estimate the Euclidean distance between data points and clusters. The FC-MCI algorithm is used to determine relationships between words and terms based on strong semantic relationships.

As shown in **equation 3**, calculate the Euclidean distance between data points and clusters using the FCM algorithm. Let's assume x_i—data point or file feature vector, c_j—Euclidean distance, x_i—data points, c_j—cluster, μ_{ij}^{m}—Weighted membership value, O_F—objective function, μ_{ij}—membership degree, $O_F(\mu, C)$—Objective function of the Fuzzy C-Means algorithm, $m-$

Fuzzifier parameter controlling cluster fuzziness.

$$O_F(\mu, C) = \sum_{i=1}^{n} \sum_{j=1}^{c} \left(\mu_{ij}^{m}\right) \left\|x_i - c_j\right\|^2 \tag{3}$$

As illustrated in **equations 4 and 5**, calculate the number of sets in the membership array and the centroid of the set of maximum reaction parameters. Let's assume μ_{ij}^{m}— membership matrix, n— Total number of data points, c—Total number of clusters, $d(x,c)$—Euclidean distance between data point and cluster center, C_j—Updated center of cluster, $x_1,\ x_2, x_3$—Feature values of the data point, c_1, c_2, c_i— Feature values of the cluster center.

$$C_j = \frac{\sum_{i=1}^{n} \left(\mu_{ij}^{m}\right) x_i}{\sum_{i=1}^{n} \left(\mu_{ij}^{m}\right)} \tag{4}$$

$$d(x,c) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \ldots + (x_n - c_i)^2} \tag{5}$$

Update the member array by counting duplicate files, as shown in **equation 6**. Let's assume k— Index of all clusters used in summation, c— Total number of clusters, d_{ij}, d_{ik}—Distance between data point.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} d_{ij} / d_{ik}^{2/m-1}} \tag{6}$$

The objective function in the FCM model can be minimized by clustering similar files, which helps identify duplicates based on document or file similarity.

2.4 Data Deduplication-based Efficient Cloud Optimization (D2-ECO)

This section utilizes the D2-ECO model to identify duplicate data in cloud systems, thereby enhancing storage performance utilization at a low cost. This proposed method is an efficient cloud optimization technique, based on data replication, called D2-ECO, to maximize storage space utilization and enhance the performance of cloud systems. In addition, it helps save storage costs by detecting and eliminating duplicate data. The technique analyzes the relationship between storage capacity and system overhead and performs precision optimization. The proposed D2-ECO technique can enhance cloud storage performance by identifying duplicates based on their performance weight and minimizing the overhead caused by processing duplicate data.

The D2-ECO model can quantify the impact of data reduction and cloud computing optimization parameters. It implements an exact optimization technique to assess the trade-off between storage capacity and system overhead. This improves performance at scale by balancing resource consumption, throughput, and latency for cloud storage applications, as shown in Equation7. Let's assume $p' and\ w'$—impact and weight parameter measure, Vf'—storage efficiency, np''—system overhead, $\left|M\left(j - KP'\right)\right|$—data variability, $\partial p' - ew''$—deduplication process's overall efficiency,

$$\left|M\left(j - KP'\right)\right| = \mu_{ij} Vf' < \partial p' - ew'' > + np'' \tag{7}$$

Operational factors improve the energy or resource efficiency characteristics, thereby improving cloud system performance and reducing operating costs, as shown in **Equation 8**. Let's assume $4\left(\forall - 3wq''\right)$—storage energy efficiency, E_m— Energy/resource efficiency metric, $\left(R - j''\right)$—quantities, RWQ''—resources, RWQ''— Resource workload quantity.

$$\left\langle \frac{2}{4\left(\forall - 3wq''\right)} \right\rangle E_m \left(R - j''\right) - RWQ'' \tag{8}$$

By modeling the relationship between system performance, workload dynamics, and external factors, the performance metric of data duplicate detection is improved, as shown in **Equation 9**. Let's assume $\partial - RW''$— deduplication data integrity, $\forall w\ - rt''$—resource cost, Ty—workload, φw — retrieval time, $\pi\mu'$— efficiency factor, E_t— Performance evaluation metric, Ty— Workload parameter, φw — Retrieval time, $\pi\mu'$—Efficiency factor.

$$E_t\left(\partial - RW''\right) = \left[\forall w\ - rt''\right] + Ty\left(\varphi w\ + \pi\mu'\right) \tag{9}$$

As shown in **Equation 10**, calculate the data consistency and resource consumption during data replication to improve storage performance in cloud environments. Let's assume $\partial T^{'}-$ Data consistency factor, $Fv\left(M-vr^{''}\right)-$Resource consumption during replication, $Eq^{'}-$Storage performance equation factor, $MVd^{''}-$Duplicate data volume metric.

$$\partial T^{'}-Fv\left(M-vr^{''}\right)=Eq^{'}-MVd^{''} \tag{10}$$

The advantage of cloud storage systems is that they detect duplicate data and provide improved speed and performance based on the total data volume and diversity of data processing tasks. Processing time and cost can be minimized by enhancing storage efficiency, as demonstrated in **Equation 11**. Let's assume $\forall+M_h-$effect of data deduplication, $ew-2p-$data integrity, $M\left(\partial\forall^{'}-Rt\right)\propto-$efficiency weight of deduplication, $M-$ processing duplicate data.

$$\left|\forall+M_h(ew-2p)=M\left(\partial\forall^{'}-Rt\right)\propto\right| \tag{11}$$

By optimizing cloud storage performance, we can identify efficient data duplication and provide an optimal solution.

3 Results and Discussion

This section adopts the proposed D2-ECO approach to provide cloud data reduction solutions in terms of storage efficiency, latency, data integrity, and operational cost. It offers data fragment validation, semantics, file duplicate detection, and cloud storage performance optimization. Additionally, the proposed method can achieve better improvements by utilizing TDES, AES, and ECC methods prior to performance evaluation.

Table 2. Simulation Parameter

Parameters	Variable
Cloud environment	AWS, S3 cloud storage
Data used	Content file type
Data Size	5GB
Simulation Framework	Visual studio, C#
Output type definition	Redundant Storage Space

Cloud storage performance can be improved by implementing performance metrics, resulting in better performance. Additionally, in cloud environments like AWS and S3 cloud storage, 5GB of data can be utilized to identify redundant storage space and analyze it using simulation frameworks such as Visual Studio and C#, as illustrated in Table 2.

3.1 Storage Performance

Storage efficiency is the proportion of usable storage space in a system after duplicates are removed. It is the number of unique data points stored divided by the total size of the input data. High storage efficiency improves the deduplication system's performance, reducing redundancy and enabling more efficient use of cloud storage resources, as shown in **equation 12**.

$$Storage\ Efficiency\ (\%)=\frac{Total\ Data\ Size}{Unique\ Data\ Size}\times 100 \tag{12}$$

Table 3. Performance of Storage Efficiency (%)

Number of Files	TDES	AES	ECC	D2-ECO
25	65	71	74	81
50	69	75	78	87
75	72	79	82	90
100	75.6	86.4	90.2	96.8

Table 3 shows that the proposed approach provides more effective cloud storage, as it compares appropriately to other approaches in terms of storage performance and the identification of duplicated data. The tables contain exact numbers to

allow precise comparison, and a graph is included to show the overall performance trend and to emphasize comparative improvements between methods. Moreover, the storage performance analysis rate of the presented D2-EOC approach is raised to 96.8%. Besides this, the proposed method's storage performance analysis rate is enhanced by 75.6%, 86.4%, and 90.2% compared to the previous TDES, AES, and ECC techniques, respectively.

3.2 Data Integrity

Data integrity refers to the correctness and consistency of data after duplicates are removed. Larger values indicate the original data are being preserved without corruption or loss, as shown in **equation 13**.

$$Data\ Integrity\ (\%) = \frac{Total\ Data\ Stored}{Correct\ Data\ Retrieved} \times 100 \tag{13}$$

Table 4. Performance of Data Integrity

Number of Files	TDES	AES	ECC	D2-ECO
25	69	73	76	80
50	72	75	79	84
75	75	79	83	87
100	77.10	81.5	85.9	89.26

Table 4 illustrates that the proposed method maintains high data integrity, as it compares positively with other methods in terms of storage performance and duplicate data identification. Furthermore, the data integration analysis rate of the proposed D2-EOC method is improved to 89.26%. Additionally, the proposed method improves the data integrity rate by 77.10%, 81.5%, and 85.9% compared to the previous TDES, AES, and ECC, techniques, respectively.

3.3 Operational Cost

Operational cost is the computational and storage costs of maintaining cloud data. It considers system resource usage, processing overhead, and the cost of storing and maintaining data. A small operational cost implies a more efficient system that best uses resources, as shown in **equation 14**.

$$Operational\ Cost = \frac{Processing\ Cost + Storage\ Cost + Resource\ Usage}{Total\ system\ capacity} \tag{14}$$

Table 5. Performance of Operational Cost

Number of Files	TDES	AES	ECC	D2-ECO
25	44.12	40.2	37.11	34.08
50	41.14	36.6	33.05	31.04
75	37.16	32.7	29.19	26.07
100	34.11	28.04	25.09	21.10

Table 5 illustrates that the proposed method maintains operational costs comparable to those of other approaches, particularly in terms of storage performance and duplicate data detection. Moreover, the operational cost rate of the proposed D2-EOC method has increased to 21.10%. The technique also reduces operational costs by 34.11%, 28.04%, and 25.09% compared to previous TDES, AES, and ECC techniques.

3.4 Latency

Latency is the time between when data is processed and when it is acquired in the cloud. It is commonly expressed in ms. Lower latency corresponds to faster data processing and better system responsiveness, as shown in **equation 15**.

$$Latency\ (ms) = Response\ Time - Request\ Time \tag{15}$$

Table 6. Performance of Latency

Number of Files	TDES	AES	ECC	D2-ECO
25	38.5	34.6	30.26	24.10
50	31.4	31.2	27.14	20.03
75	27.3	26.4	22.3	15.06
100	22.13	20.4	17.5	10.2

The results presented Table 6 demonstrate that the proposed approach is competitive in terms of achievable latency performance compared to methods that achieve similar storage performance and duplicate data detection. Additionally, the latency rate of the proposed D2-EOC method has been improved to 10.2 ms. The technique also lowers the latency by 22.13 ms, 20.4 ms, and 17.5 ms over the past TDES, AES, and ECC approaches.

3.5 Time Complexity

Time complexity describes the average time the algorithm takes to process each file. A smaller value means the algorithm is more computationally efficient.

$$Time\ Complexity = \frac{Number\ of\ Files}{Total\ Processing\ Time} \tag{16}$$

Table 7. Performance of Time Complexity

Number of Files	TDES	AES	ECC	D2-ECO
25	25.6	21.7	17.4	12.4
50	22.13	17.8	13.2	8.3
75	18.4	14.9	10.9	5.2
100	15.2	11.5	7.4	3.7

Table 7 shows the time complexity performance of the various methods (TDES, AES, ECC, and the proposed D2-ECO) for different numbers of input files. For 25 files, the processing time is 25.6 ms for TDES, 21.7 ms for AES, 17.4 ms for ECC, and 12.4 ms for D2-ECO. When the number of files reaches 50, the time is down to 22.13 ms (TDES), 17.8 ms (AES), 13.2 ms (ECC), and 8.3 ms (D2-ECO). For 75 files, the values also decrease to 18.4 ms, 14.9 ms, 10.9 ms, and 5.2 ms, respectively. At 100 files, the time consumption for TDES is 15.2 ms, that for AES is 11.5 ms, that for ECC is 7.4 ms, and that for the proposed D2-ECO model is only 3.7 ms. The enhancement results from an effective combination of chunking, semantic analysis, and clustering, which reduces many redundant computations and increases processing speed.

This section presents a detailed evaluation of the proposed D2-ECO (Data Deduplication-based Efficient Cloud Optimization) model and validates its effectiveness through comparison with existing deduplication techniques reported in the literature. The experiments were conducted in an AWS S3 cloud environment using a 5 GB dataset of content-based files. Performance was assessed in terms of storage efficiency, data integrity, operational cost, and latency, which are widely recognized evaluation metrics in cloud deduplication research [(1)], [(6)], [(15)].

The proposed model was compared with TDES, AES and ECC approaches, which represent traditional chunk-based, learning-assisted, and content-defined deduplication techniques. As shown in Table 3, the proposed D2-ECO model achieves a maximum storage efficiency of 96.8% for 100 files, significantly outperforming TDES (75.6%), AES (86.4%), and ECC (90.2%).

Earlier studies such as Kim and Lee [(1)] and Ayad Hasan et al. [(6)] report that traditional deduplication frameworks typically achieve storage efficiency in the range of 80–90%, particularly when fixed or content-defined chunking is employed. Although ECC -based methods improve redundancy elimination, they remain sensitive to boundary-shift problems and structural variations in data [(16)], [(18)].

Learning-based deduplication techniques, including semantic similarity and feature-aware approaches, show moderate improvements (approximately 88–92%) [(11)], [(17)]. However, these methods often operate independently of adaptive chunking strategies. The superior performance of D2-ECO can be attributed to the combined use of variable-size chunk validation and semantic similarity scoring, which enables detection of both exact and near-duplicate data. This result confirms and extends findings from [(3)], [(11)] and [(20)], while exceeding their reported efficiency ranges. Table 4 indicates that the proposed D2-ECO model maintains a data integrity rate of 89.26%, outperforming TDES (77.10%), AES (81.5%), and ECC (85.9%).

Secure and verifiable deduplication frameworks, such as VeriDedup [7], emphasize strong integrity guarantees but often introduce cryptographic verification overhead that impacts system performance. Similarly, parity-based deduplication models [9] preserve correctness but lack semantic validation, leading to false duplicate elimination.

Recent studies incorporating fuzzy logic and learning-based detection report integrity preservation between 82% and 87% [13], [19]. The D2-ECO model exceeds this range by leveraging Fuzzy C-Means Cluster Indexing (FC-MCI), which minimizes false positives by grouping semantically similar files rather than relying solely on hash equivalence. Unlike cryptography-heavy approaches [7], the proposed model maintains integrity without excessive computational complexity, making it suitable for scalable cloud environments.

As shown in Table 5, the operational cost associated with D2-ECO is reduced to 21.10%, which is significantly lower than TDES (34.11%), AES (28.04%), and ECC (25.09%). Previous works emphasize that redundancy elimination directly impacts cloud storage cost [8], [10], [14]. However, most reported solutions achieve cost reductions in the range of 15–25% by focusing on routing strategies or optimization heuristics alone [8], [10].

The enhanced cost reduction achieved by D2-ECO demonstrates that semantic-aware deduplication combined with adaptive chunking leads to more effective elimination of redundant storage and reduced I/O operations. This aligns with findings from Meenalochini [10] while offering a more integrated and intelligent optimization framework.

Latency results presented in Table 6 show that D2-ECO achieves a minimum latency of 10.2 ms, compared to 22.13 ms (TDES), 20.4 ms (AES), and 17.5 ms (ECC). Prior deduplication models often report latency values between 15 ms and 25 ms, especially when learning-based or feature-aware processing is involved [12], [17], [18]. Although deep semantic analysis improves accuracy, it is frequently associated with increased processing delay [3], [11].

The reduced latency observed in D2-ECO contradicts these findings and demonstrates that semantic processing does not necessarily increase latency when applied selectively after chunk validation. This hybrid sequencing minimizes unnecessary inference operations and improves real-time performance. The novelty of this study lies in the synergistic integration of three distinct layers—adaptive chunking, semantic intelligence, and fuzzy clustering—within a unified cloud optimization framework.

Several existing approaches to data deduplication primarily concentrate on hash-based or Content Defined Chunking (CDC) techniques as the core mechanism for identifying duplicate data blocks, as highlighted in studies such as [4] and [16]. While these methods are effective for detecting exact data redundancy, they often overlook deeper semantic relationships within the stored information. Some research efforts attempt to incorporate semantic similarity in the deduplication process; however, these approaches generally apply similarity detection without implementing adaptive chunk control mechanisms, which limits their ability to efficiently manage varying data structures and dynamic storage environments, as noted in [11] and [20]. In addition, certain solutions prioritize strengthening security features, such as encryption and secure access mechanisms, but this often comes at the expense of system performance, resulting in increased computational overhead and slower data processing, as discussed in [7]. Consequently, balancing efficient deduplication, semantic awareness, and strong security while maintaining optimal system performance remains a significant challenge in current research.

The proposed D2-ECO model introduces:

1. **Variable Size-Chunking Algorithm (VSCA)** to reduce boundary-shift and fragmentation issues.
2. **Semantic similarity modeling inspired by NLP techniques** to identify near-duplicate content beyond exact matching [3], [11].
3. **Fuzzy C-Means Cluster Indexing (FC-MCI)** to improve duplicate grouping accuracy and reduce false elimination.
4. **End-to-end cloud performance optimization**, linking deduplication outcomes directly to cost, latency, and integrity metrics.

By achieving 96.8% storage efficiency, 89.26% data integrity, 21.10% operational cost, and 10.2 ms latency, the proposed D2-ECO model demonstrates measurable and consistent improvements over existing approaches, thereby contributing a scalable, intelligent, and performance-aware deduplication solution for modern cloud storage systems.

4 Conclusion

This research presented the D2-ECO model as an intelligent and structured solution to address the growing challenge of redundant data in cloud storage systems. As cloud environments continue to handle massive volumes of heterogeneous data, conventional deduplication methods often struggle to balance efficiency, accuracy, and performance. The proposed framework was designed to overcome these limitations through a coordinated integration of adaptive chunking, semantic similarity analysis, and clustering-based indexing. The full composition of adaptive chunking, semantic intelligence, and fuzzy clustering

within a unified cloud optimisation procedure rather than in individual innovative methods. The cooperative combination of VSCA, BERT, and FC-MCI into a single cloud optimisation framework, offering superior accuracy, efficiency, and scalability compared with the state-of-the-art.

The experimental results clearly demonstrate the effectiveness of the approach. The D2-ECO model achieved 96.8% storage efficiency**,** significantly outperforming existing comparison methods. At the same time, it maintained 89.26% data integrity**,** reduced operational cost to 21.10%**,** and minimized latency to 10.2 ms**.** These results confirm that the proposed framework does not merely remove duplicate data but does so in a way that improves overall system responsiveness and cost efficiency. The improvements across multiple performance indicators show that the model achieves a balanced optimization rather than focusing on a single metric.

The key novelty of this work lies in the structured combination of three complementary mechanisms within a single unified pipeline**.** Instead of relying only on hash-based or content-defined chunking methods, the model first validates data through a Variable Size-Chunking Algorithm (VSCA), then applies semantic similarity analysis to identify contextual duplication, and finally uses Fuzzy C-Means Cluster Indexing (FC-MCI) to group related data more accurately. This layered coordination reduces false detections, minimizes unnecessary processing, and enhances both precision and speed. The integration of structural validation and semantic intelligence within one optimized cloud workflow distinguishes this work from earlier standalone approaches. From a practical standpoint, the findings suggest several actionable directions for cloud service providers and enterprise storage managers. Implementing adaptive chunking before deep similarity analysis can significantly reduce processing overhead.

Incorporating semantic-aware duplicate detection can improve handling of near-duplicate documents in large-scale repositories. Furthermore, linking deduplication outputs directly with storage allocation strategies can lead to measurable cost savings and improved infrastructure utilization. Future research can expand this work by testing the framework on larger real-time datasets, incorporating lightweight security mechanisms to strengthen trust in shared environments, and exploring adaptive learning models that continuously improve duplicate detection accuracy. Additionally, energy-efficient optimization metrics may be integrated to support sustainable cloud computing initiatives. The D2-ECO model offers a measurable and scalable improvement in cloud storage optimization. By achieving high storage efficiency, reduced latency, preserved data integrity, and lower operational cost within a single intelligent framework, this study contributes a practical and innovative advancement to modern cloud data deduplication systems.

4.1 Future Work

In the future, the proposed model can be extended to a large-scale cloud for better scalability and support for a large amount of data. Furthermore, the feasibility of the system's real-time implementation can be investigated to enable dynamic, continuous data deduplication in real applications. Further enhancement can be achieved by applying sophisticated hybrid optimization methods, which can efficiently improve the system's precision and robustness under time-varying workloads.

References

1) Rajput U, Shinde S. Analysis on Deduplication Techniques for Storage of Data in Cloud. *Int Res J Eng Technol (IRJET)*. 2022 May;9(5). Available from: https://www.irjet.net/archives/V9/i5/IRJET-V9I5772.pdf.
2) Jeslin JG. Implementing an efficient data deduplication framework for cloud storage. *Indian J Comput Sci Eng (IJCSE)*. 2022 Jan-Feb;13(1). https://doi.org/10.21817/indjcse/2022/v13i1/221301046.
3) Jeslin JG, Kumar PM. Decentralized and Privacy Sensitive Data De-Duplication Framework for Convenient Big Data Management in Cloud Backup Systems. *Symmetry*. 2022;14(7). https://doi.org/10.3390/sym14071392.
4) Tang X, Guo C, Choo KR, Jiang X, Liu Y. A secure and lightweight cloud data deduplication scheme with efficient access control and key management. *Comput Commun*. 2024;222:209–219. https://doi.org/10.1016/j.comcom.2024.05.003.
5) Maddali R. Quantum-inspired algorithms for high-dimensional data compression and deduplication. *Int J Eng Technol Res Manag (IJETRM)*. 2024;8(11).
6) Hasan A, Husiee NA. Techniques of Data Deduplication for Cloud Storage: A Review. *Int J Eng Res Adv Technol (IJERAT)*. 2022 Apr. https://doi.org/10.31695/IJERAT.2022.8.4.2.
7) Yu X, Bai H, Yan Z, Zhang R. VeriDedup: A Verifiable Cloud Data Deduplication Scheme With Integrity and Duplication Proof. *IEEE Trans Depend Secure Comput*. 2023 Jan-Feb;20(1):680–694. https://doi.org/10.1109/TDSC.2022.3141521.
8) Mageshkumar N, Swapna J, Pandiaraj A, Rajakumar R, Krichen M, Ravi V. Hybrid cloud storage system with enhanced multilayer cryptosystem for secure deduplication in cloud. *Int J Intell Netw*. 2022;4:301–309. https://doi.org/10.1016/j.ijin.2023.11.001.
9) Karthikeyan R, Nanthakumar S. Cloud data Deduplication System using per File parity and File Name Interpreter. *Int J Adv Res Arts Sci Eng Manag*. 2022.
10) Meenalochini P. Enhancing Cloud Storage Performance through Advanced Deduplication Techniques. *J Artif Intell Cyber Secur (JAICS)*. 2025;9(1).
11) Devarajan V, Subramanian R. Analyzing semantic similarity amongst textual documents to suggest near duplicates. *Indones J Electr Eng Comput Sci*. 2022;25(3):1703–1711. https://doi.org/10.11591/ijeecs.v25.i3.pp1703-1711.
12) Goel A, Prabha C, Malik M, Sharma P. Security Concerns and Data Breaches for Data Deduplication Techniques in Cloud Storage: A Brief Meta-Analysis. *Int J Saf Secur Eng*. 2024;14(2).

13) Ravikanth M, Korra S, Mamidisetti G, Goutham M, Bhaskar T. An efficient learning based approach for automatic record deduplication with benchmark datasets. *Sci Rep*. 2024;14(1):16254. https://doi.org/10.1038/s41598-024-63117-0.
14) Rajkumar K, Hariharan U, Dhanakoti V, Muthukumaran N. A secure framework for managing data in cloud storage using rapid asymmetric maximum based dynamic size chunking and fuzzy logic for deduplication. *Wireless Netw*. 2024;30(1):321–334.
15) Niyaz AN, Duraipandian N. Secured Data Storage Using Deduplication in Cloud Computing Based on Elliptic Curve Cryptography. *Comput Syst Sci Eng*. 2022;41(1).
16) Al-lehaibi K, Alharbi AA. A Secure Deduplication Technique for Data in the Cloud. *Int J*. 2023;11(9).
17) Yao W, Hao M, Hou Y, Li X. FASR: An efficient feature-aware deduplication method in distributed storage systems. *IEEE Access*. 2022;10:15311–15321. https://doi.org/10.1109/ACCESS.2022.3147133.
18) Begum BR, Chitra P. SEEDDUP: a three-tier SEcurE data DedUPlication architecture-based storage and retrieval for cross-domains over cloud. *IETE J Res*. 2023;69(4):2224–2241.
19) Periasamy JK, Selvam L, Anuradha M, Kennady R. A fuzzy optimal lightweight convolutional neural network for deduplication detection in cloud server. *Int J Fuzzy Syst (IJFS)*. 2024 Jan. https://doi.org/10.22111/IJFS.2023.45797.8070.
20) Shi H, Liu X, Lv F, Xue H, Hu J, Du S, et al. A Pre-trained Data Deduplication Model based on Active Learning. *arXiv preprint*. 2023. Available from: https://arxiv.org/abs/2308.00721. https://doi.org/10.48550/arXiv.2308.00721.