Computer Fraud and Security
ISSN (online): 1873-7056

An Unstructured IoT Data Conversion and Secured Data Deduplication System in Cloud Network using Adaptive Deep Learning and Optimal Key-aided Cryptography

Manjunath Singh H,
Research scholar, UVCE, Bangalore..
mansh24.singh@gmail.com.

Dr Tanuja R,
Associate professor, UVCE, Bangalore..
tanujar.uvce@gmail.com

Abstract- Internet of Things (IoT) sensors continuously creates unstructured data, which needs to be transformed into a structured format in order to extract information from it. The designed data is helpful for other kinds of analysis, such as forecasting data that may be analyzed to make predictions about the future. More and more people are adopting cloud computing, due to more convenient, cost-effective, and time-efficient services than traditional architecture. Because cloud storage systems provide consumers with easy and affordable network storage, they are becoming more and more popular. The same data with multiple copies are frequently stored by cloud systems, particularly in backup conditions. In order to ensure that only one unique instance is stored data deduplication finds and eliminates these duplicates, thereby reducing the total memory space. Hence, in this proposed model, a new deep learning approach is proposed to address the issues present in the existing research. Initially, the essential data are gathered from the IoT sensors, where the data is in the unstructured format. At first, the gathered unstructured data is cleaned and given for the entity extraction process with the aid of Adaptive Dilated Conv-Cascaded Long Short-Term Memory (LSTM) with Attention Mechanism (ADCL-AM). Here the parameters present in the model are optimized using the Enhanced Position Updating-based Flamingo Search Algorithm (EPU-FSA) optimization. After extracting the entity, the pattern generation is carried out. Thus, the unstructured data is converted into a structured form, which is further used for the data deduplication process. In this phase, the structured data is subjected to ADCL-AM for data deduplication. Here also, the parameters present in the model are optimized using the same IFSA optimization. Subsequently, the deduplicated data are securely managed by using Hyper-Elliptic Curve Cryptography (HECC), in which the keys are optimally tuned by the developed IFSA algorithm. The competence of the proposed model is analyzed based on an experimental outcome. The result demonstrated that the proposed approach outperformed the standard techniques.

Keywords- *Internet of Things Data Conversion; Secured Data Deduplication; Adaptive Dilated Conv-Cascaded Long Short-Term Memory with Attention Mechanism; Enhanced Arbitrary Value-based Flamingo Search Algorithm;*

I. INTRODUCTION

Cloud computing is used to assess and store data over the internet. It provides end users with enormous amounts of resources in response to their needs [9]. Cloud storage is the most widely used service offered by cloud computing providers [10]. Cloud storage must be used because the amount of data in the globe is growing exponentially. Duplicate data storage is a major contributor to storage waste [11]. The same data with multiple copies are saved to the cloud by several users. Under a deduplication scheme, the users in cloud first determine if the uploaded file is already saved [12]. Then, the duplicated file is removed from the cloud storage. However, cloud users may have certain difficulties in security and privacy [13]. Particularly, the uploaded data are altered, tampered with, or removed by cloud users due to certain reasons. Deduplicated data loss might result in significant losses for all parties involved, including data holders and owners [14]. As a result, it is significant to

confirm the integrity of data stored in the cloud, particularly when it comes to deduplication and duplicate data storage. Unstructured textual data are challenging to manage [15]. Hence to extract information, the unstructured data must be converted into a structured format.

Different retrievability strategies are utilized to solve the problem of integrity checks for cloud data storage. A user adds verification tags and a file in these systems [16]. In order to remove redundant information data deduplication is used, which supports data integrity [17]. This lowers the possibility of errors and inconsistencies because the system only stores unique data [18]. While maintaining data security and integrity, the deduplication method finds and eliminates duplicate records. When data is uploaded to the cloud, a hash of the data is generated and stored [19]. To ensure data integrity during retrieval, the hash is computed and compared to the stored hash [20]. This technique makes sure that any illegal modifications to the data are identifiable. The implementation of a cryptosystem with hashing is the security mechanism to maintain the integrity of data [21]. Deep learning-based data deduplication eliminates redundant data to save storage space and progress data management efficiency [22]. These models are more robust to noise and variations in data. They effectively identify duplicates even when there are minor differences in formatting, encoding, or other attributes [23]. A deduplication process is resource-intensive, requiring significant resources and memory usage, especially when analyzing large datasets [24]. Handling sensitive data during the deduplication process raises security concerns, especially if the data is being transferred to third-party cloud services. Cloud environments often contain unstructured data, which is more challenging to deduplicate compared to structured data [25]. Hence, to overcome these challenges, a secured data deduplication system is developed.

The significant contributions of the designed deduplication model are provided in the below content.

- ✓ The secured data deduplication model is developed for boosting the reliability in cloud storage environments. Several copies of the same data are frequently seen in cloud settings, especially during backup. By eliminating these duplicate copies, a deduplication system helps to decrease the amount of storage space needed, which is essential for cloud services.
- ✓ IoT data conversion is performed in this proposed model through the entity extraction process using the ADCL-AM network for converting data produced by IoT devices into a structured format, which is easy to analyze. Errors in analysis are minimized due to structured data because it is generally more accurate, consistent, and dependable.
- ✓ To enhance the speed of data access and retrieval, an optimization strategy is applied in both the deduplication and encryption process with the aid of EPU-FSA, which successfully meets the demands of modern data management while maintaining high standards of security and efficacy.
- ✓ HECC is crucial in the data deduplication process because it improves data security and privacy, without hampering sensitive data and preserves data integrity with user trust. Organizations safely manage their data in cloud settings while avoiding security threats by utilizing HECC.

The general structure of this proposed data deduplication system is specified in the below section. Different reviews related to this data deduplication scheme with their existing challenges and benefits are provided in section II. The description of unstructured IoT data conversion and secured data deduplication in cloud networks are explained in Section III. The detailed descriptions of the proposed network formation along with pattern creation are specified in Section IV. In Section V, the explanations of deep learning-based optimal key-aided cryptography are provided. Finally, the resultant discussion is provided in Section VI and a quick summary is given in Section VII.

II. LITERATURE SURVEY

A. Related Works

In 2020, Liu *et al*. [1] have developed a new attribute-based keyword search system that protected the confidentiality and integrity of user data while enabling users to search across encrypted cloud data. This strategy provided data deduplication technologies that maximize storage resources and minimize network bandwidth consumption by removing redundant data from cloud storage. To ensure that the data obtained was

correct and unaltered, this model was accomplished through the use of hash functions, and signatures. Thus, the designed scheme was more computationally efficient than existing deduplication schemes.

In 2022, Yu *et al.* [2] have designed a verified cloud data deduplication storage strategy that focused on ensuring the integrity of deduplicated data while enabling effective storage management. The designed model provided sequences of bits as verification tags that were attacked to the encrypted block of uploaded files. It helped to verify the reliability of blocks and minimize the usage of storage needed for integrity checks.

In 2023, Kumar *et al.* [3] have introduced the cloud architecture that successfully performed data encryption to boost the safety of the cloud network. The suggested model successfully achieved high compression factor values for a range of file sizes, demonstrating its capacity to lower storage needs and boost data transfer rates. When compared to conventional models, this suggested security model has a greater level of security and offers a reliable framework for managing ownership and data integrity in cloud environments.

In 2021, Ebinazer *et al.* [4] have developed a deep strategy for authorized deduplication purposes. A designed encryption strategy was used to stop data leaks, and re-encryption was cleverly used to attain approved deduplication. The authorized request was especially handled by the management center, which also created an RT configuration to map the interaction. Additionally, Bloom Filter (BF) was used to execute data updating and improve the effectiveness of confirmation retrieval. A thorough simulation experiment was conducted to show the efficacy and security of the model.

In 2020, Azad *et al.* [5] have examined the IoT's current data management strategies. Three primary groups of data management approaches were used in this proposed model. Furthermore, the comprehensive analysis of the key mechanisms in every category resulted in a suggestion for additional research.

In 2020, Reddy *et al.* [6] have created a model that allowed unstructured data to be transformed into structured data using the Big Data Analytics technique and Hadoop ecosystem technologies like Map Reduce and HBase.

In 2017, Krishnan *et al*. [7] have exhibited Natural Language Processing (NLP) techniques. By analyzing the semantic meaning of the text, the NLP approach identified duplicates more precisely, even if they were not perfect matches. By identifying and removing duplicates, NLP-based deduplication contributes to overall data quality improvement.

In 2017, Anujna and Ushadevi [8] have focused on utilizing the regular expression pattern matching technique to extract pertinent data from the text/word file. Every word or character found in the documents was extracted using a text file with a high amount of free text. Users found the system useful in finding pertinent documents and in organizing all of the unstructured material.

B. Problem statement

Nowadays, a vast amount of data is produced from the IoT sensors in the format of unstructured. In order to boost the analysis, the conversion of unstructured data into structured data is necessary. Further, deduplicating the structured data is another process in the cloud network. The deduplication is a process that helps minimize the storage expenses and enhance the disaster recovery by avoiding the replicated data copies from the device. In the cloud network, this data deduplication process prevents the users from storing the same data more than one time by connecting the data of the client to the conventional data in the cloud. Numerous cloud deduplication mechanisms have been implemented in the past years. However, the existing techniques contain some primary issues that are given below. Additionally, Table I shows the merits and difficulties of the existing data deduplication models using structured data in the cloud.

There are numerous data deduplication techniques have been suggested in the past years. However, very few techniques perform the data conversion process from unstructured data to structured data for the data deduplication process.

Most of the conventional data deduplication techniques didn’t ensure the safety of the data present in the cloud network. Hence, a construction of secured data deduplication is necessary.

Several data deduplication systems simply performed the deduplication with the support of very old mechanisms. This results in inaccurate solutions, Therefore, developing a deep learning-based data deduplication system is highly significant.

The conventional techniques didn't include the optimization process in the deduplication task which improved the computational complexities of the network. Therefore, performing parameter tuning for the deduplication process is important.

Though some data deduplication techniques perform the security enhancement, the conventional techniques couldn't fulfill that requirement efficiently. Hence, an effective cryptography technique is necessary for improving the security of the cloud network.

TABLE I. FEATURES AND CHALLENGES OF CONVENTIONAL DATA DEDUPLICATION SYSTEM USING STRUCTURED DATA IN THE CLOUD

Author [citation]	Methodology	Features	Challenges
Liu *et al.* [1]	Attribute-based encryption	• It offers structural efficiency and provides better access control. • It helps to preserve the integrity of the data.	• Its coordination of keys can be a challenge. • It gives privacy and scalability issues.
Yu *et al.* [2]	Proxy Re-Encryption (PRE)	• It preserves the confidentiality and simplifies the process. • It enhances the bandwidth of the network.	• It consumes more time and storage resources. • It contains a suspicious proxy.
Kumar *et al.* [3]	Diffie-Hellman algorithm	• It exchanges the keys securely and performs secure data transfer.	• It is computationally intensive. • It cannot be employed to encrypt the messages.
Ebinazer *et al.* [4]	BF	• It has high scalability and has better space efficiency. • It recognizes the duplicate events effectively.	• It has high false positive rates. • It is hard to modify the data.
Azad *et al.* [5]	IoT	• It can monitor the data in real time and automate the tasks. • It can generate vast amounts of data.	• It encounters security concerns. • It has low reliability.
Reddy *et al.* [6]	Hadoop ecosystem	• It is cost-effective and fault-tolerant. • It is very flexible and scalable.	• It processes the data very slowly. • It has low security features.
Krishnan *et al.* [7]	NLP	• It can evaluate a large amount of data. • It offers accurate data analysis.	• It gives low data quality and security. • Its computational requirements are high.
Anujna and Ushadevi [8]	Pattern matching technique	• Its accuracy is higher than the manual tasks. • It catches numerous errors and checks the redundancy.	• It is very difficult to execute and very slow. • It doesn't have sufficient samples for the training process.

III. UNSTRUCTURED IOT DATA CONVERSION AND SECURED DATA DEDUPLICATION SYSTEM IN CLOUD NETWORK

A. Proposed Unstructured IoT Data Conversion and Secured Data Deduplication System in Cloud

IoT data conversion and secured data deduplication systems are developed to overcome the existing data deduplication challenges like data diversity, latency issues, and security concerns. IoT data frequently originates from a variety of sensors and devices. By standardizing this data, authorities are able to ensure that various IoT systems and devices successfully connect with one another by transforming data into a standard format. Authorities drastically reduce their storage needs by eliminating duplicate data, which lowers the cost of the infrastructure needed to store data. By ensuring that storage resources are used effectively, secured data deduplication allows valuable resources to be allocated to important tasks. Efficient operations and substantial storage capacity are necessary for the storage of unstructured data. Unstructured data are hard to organize and retrieve from traditional databases because of their potential limitations. To overcome this drawback, an unstructured IoT data conversion-based secured data deduplication scheme is developed. The visualization of the designed model is given in Fig. 1.

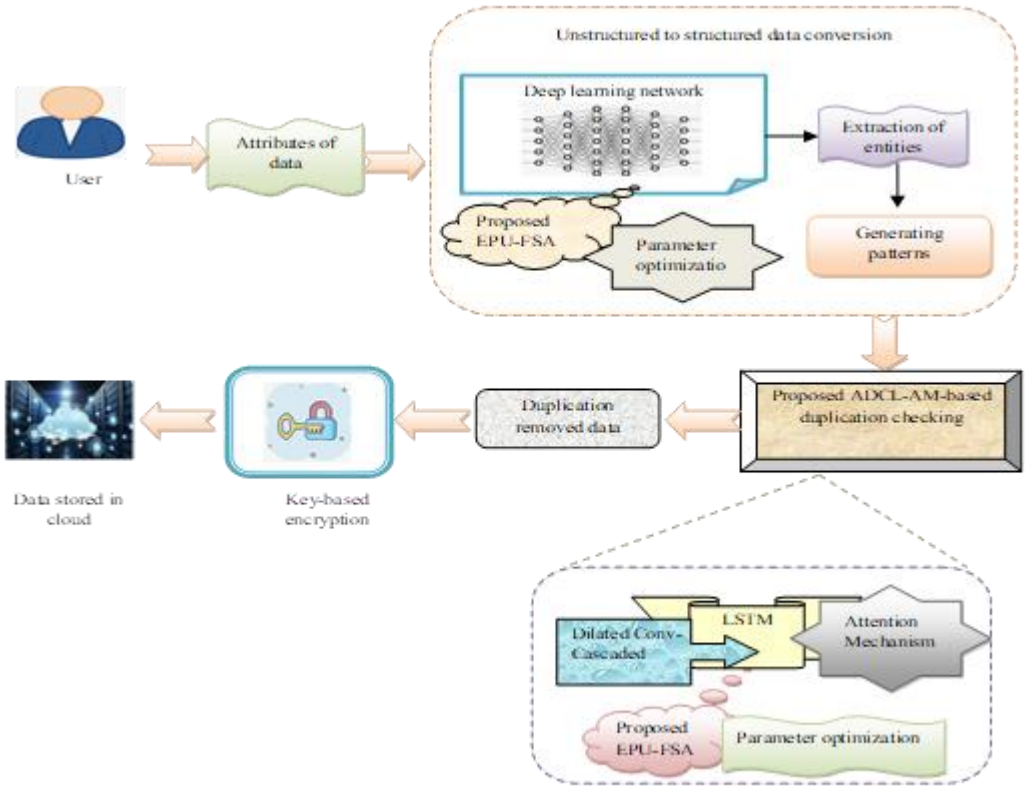


Fig. 1. Diagrammatic Representation of Proposed Data Deduplication System in Cloud

The secured data deduplication model is designed to remove redundant copies of data, which dramatically lowers the amount of storage space in cloud services. IoT-based unstructured data are collected from interconnected devices and sensors for the deduplication process. The key entities from the data are generated using ADCL-AM. The designed model precisely recognizes and categorizes entities in IoT data using an attention mechanism combined with a Dilated Conv-Cascaded LSTM. This method efficiently processes and analyzes the complex data produced by IoT devices with the aid of convolutional networks and LSTMs, which are enhanced by attention mechanisms. Entity extraction is utilized for grouping similar entities to from structured data and the patterns are generated for the training process. During the entity extraction process, the correlation coefficient of the designed system is maximized by optimizing the parameters from LSTM using EPU-FSA. Then, these cloud services reduce the storage expenses by reducing the quantity of redundant data

retained. Finally, generated patterns and entities are fed to ADCL-AM for the data deduplication process. Designed ADCL-AM to eliminate redundant copies of data to maximize storage and enhance data management, which is especially crucial in cloud storage. The precision and efficiency of deduplication processes are enhanced by optimizing the parameters from LSTM using the same EPU-FSA. Optimization helps in minimization of FOR and FDR along with maximization of precision. To ensure data security, cryptography is essential in the data deduplication process. In this suggested model, HECC-based cryptography is performed to ensure that the data remains secure during the deduplication process. Optimal key-based encryption and decryption are performed to protect the sensitive information in cloud services. Different methods and algorithms for data deduplication are contrasted according to their effectiveness, security features, and performance.

B. Deduplication Data Attribute Generation

In the deduplication process, data attributes specify the relevant features of data entries. This procedure is essential for managing databases, and making choices based on that data. The proposed data deduplication model considered attributes like File name, File location, Block name, Hashtag, File Type, Last Modified Date, File size and Data Pattern.

C. Proposed EPU-FSA

In this secure data deduplication framework, data integrity and scalability of the cloud system are boosted by optimizing the variables like hidden neuron count, epoch count, and learning rate from LSTM using EPU-FSA during both entity extraction and deduplication process. Conventional FSA imitates the flamingos' social behavior and movement patterns, especially their flocking and foraging techniques. This algorithm is able to efficiently search the solution space because of its natural inspiration. The choice of input parameters can have a significant impact on how effectively FSA performs. Determining the ideal parameter settings is essential since incorrect values might result in slow convergence. High-dimensional situations demand more computing time and resources due to FSA's scaling issues. To overcome these issues, FSA is enhanced and this enhanced version is named EPU-FSA by updating the final position of the solution based on the current best fitness-based position and random position. Here, the position updating of the flamingo based on the fitness value is done, which is provided in Eq. (2).

$$F1_m^{y+1} = \frac{\left(\left|F1_{sm}^{y} + \omega_1 \times fh_m^{y} + H_2 \times \left|H_1 \times fh_m^{y} + \omega_2 \times F1_{sm}^{y}\right|\right|\right)}{P} \quad (1)$$

Here, the term f_m^y is indicated as m^{th} flamingo at the $(y+1)$ iteration. The random variables of the flamingo are indicated as H_1 and H_2 that follows normal distribution. The terms ω_1 and ω_2 are randomized by $[-1 to 1]$. The term P is the diffusion factor. The current best position is indicated as fh_m^y at m^{th} flamingo at the current iteration C_{it} .

In addition, the second random position of the flamingo is generated and it is denoted by $F2_m^{y+1}$. The developed EPU-FSA is modified by updating the final position based on the current and the randomized position. The mathematical form of the modified EPU-FSA concept is provided in Eq. (2).

$$F = F1_m^{y+1} * 0.45 + F2_m^{y+1} * 0.55 \quad (2)$$

Here, the updated new position is indicated as F . EPU-FSA modifies its search behavior in response to the solutions' fitness by combining random and current placements. Because they dynamically alter their locations, flamingos are able to adapt to environmental changes and come up with the best solutions. FSA can reach optimal solutions more quickly by updating locations based on random and current parameters. While the present position helps to focus the search around potential places, the random component aids the algorithm in escaping from local optima. The pseudocode of the proposed EPU-FSA is provided in Algorithm 1.

Algorithm 1: Proposed EPU-FSA		
Input: Optimization Attributes hidden neuron counts, epoch count Rp_h^{LSTM}, Ej_m^{LSTM} and learning rate Wk_b^{LSTM}		
Initialize the random position.		
Evaluate the optimal spot of the flamingo.		
Empty the global position		
While $C_{it} < J_{\max}$		
	Best optimal spot is identified	
	The new position F is upgraded using Eq. (2).	
	For $y \to 1 to J_{\max}$	
		Assess the fitness of every individual.
		The best spot is obtained.
	End for	
End while		
Display the optimal outcome.		
Output: Optimized attributes $Rp_{h^*}^{opt}$, $Ej_{m^*}^{opt}$, and $Wk_{b^*}^{opt}$		

IV. Adaptive Dilated Conv-Cascaded Long Short-Term Memory with Attention Mechanism for Entity Extraction along with +Pattern Creation

A. Cascaded Long Short-Term Memory

Cascaded LSTM is a neural network that uses LSTM blocks, which perform better in data deduplicated process. By stacking multiple LSTM layers, the proposed model focuses only on relevant information and boosts its ability to identify the duplicated data by reducing the overfitting issues. LSTM [37]: It is the neural network, which is implemented in this proposed data deduplication model to handle the issues related to vanishing gradient in the system. It is a unique structure that includes an input gate, forget gate, memory cell, and output gate. The information flow in a memory cell is controlled by the input gate, and some of the irrelevant data are discarded from the memory cell using the forget gate. Finally, the output of the memory cell is evaluated by the output gate.. The cell states are statistically expressed in Eq. (3).

$$Pi = \xi\left(R_q \cdot \left[S_{p-1}, V_p\right] + M_p\right) \tag{3}$$

Here, the term R_q and M_p are the weighted matrix and bias of the gate functions. The sigmoid function of the system is indicated as ξ. The term S_{p-1} is the hidden state and V_p is the cell current state. The new state $\hat{F}_p$ is formed using the tanh function and the expression is given in Eq. (4).

$$\hat{F}_p = \tanh\left(R_x \cdot \left[S_{p-1}, V_p\right] + M_x\right) \tag{4}$$

The sigmoid layer is statistically expressed in Eq. (5).

$$P_o = \zeta\left(R_j \cdot \left[S_{p-1}, V_p\right] + M_j\right) \tag{5}$$

If the input sequences are fed to the LSTM layer that process is step by step. Then, LSTM updates its hidden state and memory cell at each step based on the current input sequences. Thus, the successful identification and elimination of duplicate data are improved when the LSTM network is used in the data deduplication network. The architecture of cascaded LSTM is given in Fig. 2.

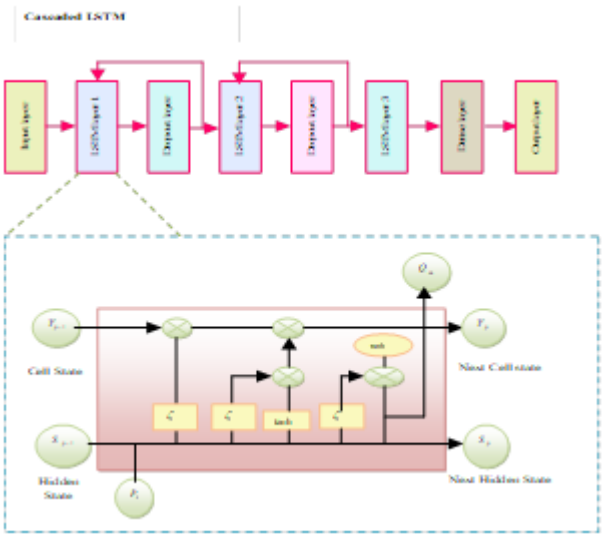


Fig. 2. Basic architecture of Cascaded LSTM

B. Developed ADCL-AM for Entity Extraction

In this proposed secure data deduplication model, an ADCL-AM network is developed for identifying and extracting entities from text. Tokenizing the input text initially results in representations of the text as embeddings, such as word embeddings, which capture word-related data. Dilated convolution layers process the input embeddings and extract features from the text while taking deeper contexts into effect. LSTM layers cascade the output from the convolution layers. When processing sequential input, the cascaded LSTMs preserve information about words, which is essential for recognizing the context of entities. Cascaded LSTM layers output is connected to an attention mechanism. By assessing attention ratings for every word, this method enables the model to deliberate on relevant terms when predicting entities. Entity extraction is used for grouping similar entities to form structured data. From these generated entities, patterns are generated for the training process. The adaptive nature of this model improves the reliability of entity extraction by fine-tuning the variables like hidden neuron count, epoch count, and learning rate from LSTM using EPU-FSA. Optimization helps to maximize the correlation coefficient of the system and the objective function R_{obj} is provided in Eq. (6).

$$R_{obj} = \underset{\left\{Rp_h^{opt}, Ej_m^{opt}, Wk_b^{opt}\right\}}{\arg\min}\left(\frac{1}{C_f}\right) \tag{6}$$

Here, the term Rp_h^{opt} is the hidden neuron count that varies in the range of $[5-225]$, Ej_m^{opt} is the epoch count that varies in the range of $[5-50]$, and the learning rate is indicated as Wk_b^{opt} that varies in the range of $[0.01-0.99]$. The formula used for calculating correlation coefficient C_f is provided in Eq. (7).

$$C_f = \frac{\Sigma\left(P_j - \overline{P}\right)\left(W_j - \overline{W}\right)}{\sqrt{\Sigma\left(P_j - \overline{P}\right)^2 \Sigma\left(W_j - \overline{W}\right)^2}} \tag{7}$$

Here, the term P_j is the values of P variables within the samples, W_j is the values of W variables in a sample, $\overline{W}$ is the mean value of W variable. The diagrammatic visualization of the proposed ADCL-AM-based entity extraction is specified in Fig. 3.

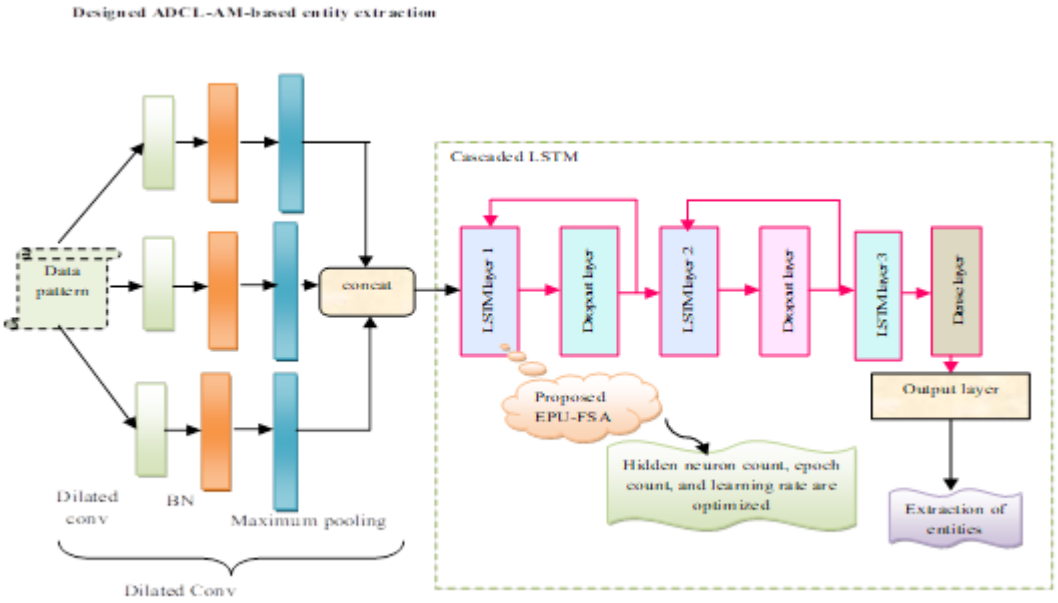


Fig. 3. Designed ADCL-AM-based entity extraction

Secured Data Deduplication System in Cloud Network using Adaptive Deep Learning and Optimal Key-aided Cryptography

C. Secured Data Deduplication Network: ADCL-AM

The designed ADCL-AM network is effectively utilized for the secured data deduplication process. The generated entities represented as T_{vb}^{enty} and data patterns represented as D_{hn}^{ptn} are considered as the input for data deduplication detection and this generated pattern is utilized for the training process. This method efficiently extracts features from the input data by employing dilated convolutions, which aids in the identification of duplicates. Cascading LSTMs are utilized to capture intricate temporal connections in the data by stacking up numerous LSTM layers. There are several levels of complexity that each layer might learn. The memory size of the cascaded LSTM model is enhanced for long-term dependencies, which is important for deduplication activities where duplicates might show after several entries. The attention mechanism in ADCL-AM focuses on specific parts of the input data that are more relevant for making duplicate predictions. This feature highlights the most essential elements while ignoring less important information and increases the model's performance in spotting duplicates.

Deduplication identifying the capacity of the ADCL-AM is improved by optimally tuning the variables from LSTM using EPU-FSA. The objective function helps to maximize the precision along with the minimization of FOR and FDR. The statistical form of the objective function D_{obj} is specified in Eq. (8).

$$D_{obj} = \underset{\left\{Rp_{h^*}^{opt}, Ej_{m^*}^{opt}, Wk_{b^*}^{opt}\right\}}{\arg\min} \left(\left(\frac{1}{S_{pre}} \right) + FOR + FDR \right) \tag{8}$$

Here, the term $Rp_{h^*}^{opt}$ is the hidden neuron count, $Ej_{m^*}^{opt}$ is the epoch count, and the learning rate is indicated as $Wk_{b^*}^{opt}$. The formula for measuring precision rate, False Omission Rate (FOR), and False Discovery Rate (FDR) are provided in the below points.

Precision (S_{pre}): The precision rate of the designed system is assessed using the formula in Eq. (9).

$$S_{pre} = \frac{R_{ve}}{R_{ve} + G_{ve}} \tag{9}$$

FOR of the designed framework is evaluated using Eq. (10).

$$FOR = \frac{G_{ne}}{G_{ne} + R_{ve}} \tag{10}$$

FDR of the designed framework is evaluated using Eq. (11).

$$FDR = \frac{G_{ve}}{R_{ve} + G_{ve}} \tag{11}$$

Here, the true positive and negative values are represented as R_{ve} and R_{ne}. The false positive and negative values are indicated as G_{ve} and G_{ne} respectively.

During training, parameter optimization results in quicker convergence. Training time is decreased by having the model attain an acceptable level of performance faster by determining the ideal learning rate, epoch count, and neuron count. This helps the ADCL-AM to recognize duplicates based on underlying patterns rather than memorizing the training set. Thus, by optimizing parameters, ADCL-AM achieves better performance with fewer resources, which is particularly important in cloud services with limited computational power or memory. The architectural diagram of the ADCL-AM-based data deduplication process is provided in Fig. 4.

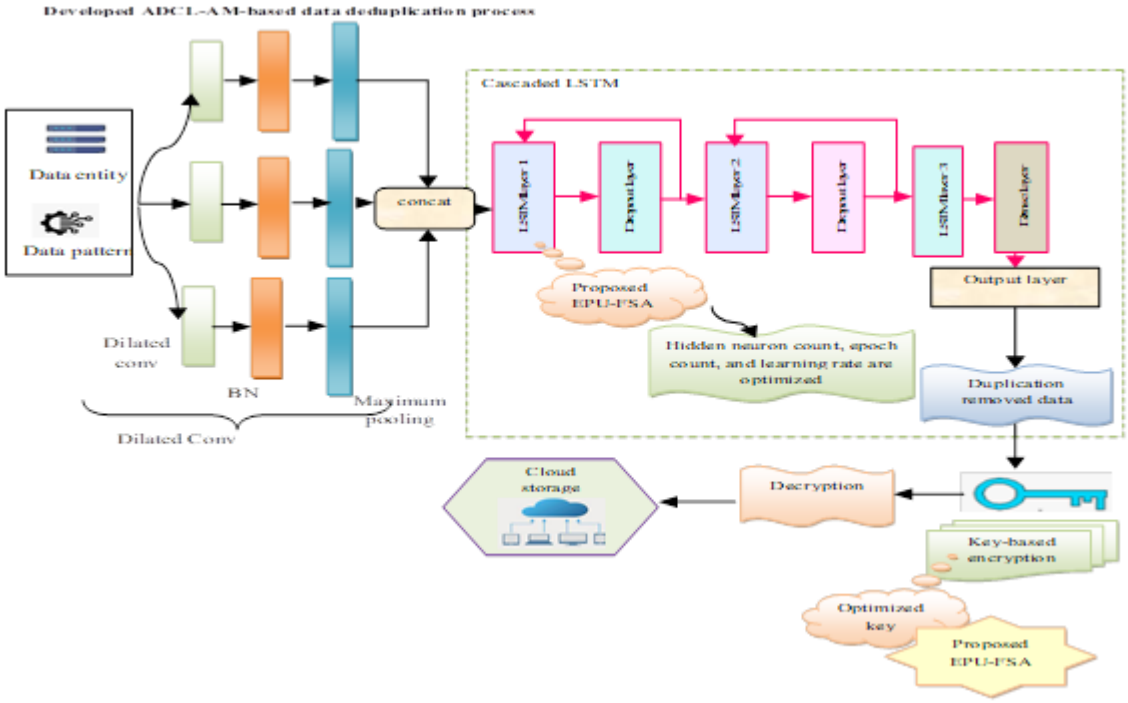


Fig. 4. Architectural diagram of ADCL-AM-based data deduplication process

D. Optimal Key-aware Deduplicated Data Security with HECC Mechanism

To boost the cloud security, the HECC mechanism is developed. This method is appropriate for cloud situations by generating key-based encryption and decryption. ADCL-AM method of removing duplicate data copies from the cloud reduces storage requirements and boosts performance. Traditional deduplication techniques might provide security problems. To overcome these security issues, the proposed ADCL-AM model

uses HECC to boost the confidentiality and integrity of the cloud system using a key-aware technique that permits access to the deduplicated data only to authorized users.

During the deduplication process, the HECC method is used for encrypting the data before it is uploaded to the cloud. Next, duplicates in the encrypted data are examined.

HECC [38]: It is an expansion of the classic ECC. It is capable of providing security levels that are comparable to those of ECC but with lower key sizes, using less computing power and communication overhead. The general form of a hyper elliptic curve can be expressed in Eq. (12).

$$E(b) = F^{2} + K(b)F \tag{12}$$

Here, the term F is the variable representing the output of the curve, $E(b)$ is a polynomial of degree, and $K(b)$ is a monic polynomial of degree. If duplicates are identified, only one copy is kept and the duplicated copies are eliminated from the storage. Using keys, HECC makes sure that access to the deduplicated data is restricted. Data integrity and confidentiality are preserved even in a deduplicated state because only users with the proper private keys which are generated from their identities can decrypt and access the information. Faster key generation and encryption operations are achieved in ADCL-AM by employing HECC, which is especially useful in cloud situations where massive amounts of data are handled. The system becomes more efficient in terms of storage and access times due to this key-based cryptography reduction technique, which also improves the performance of a secured deduplication system. The network representation of the HECC mechanism is provided in Fig. 5.

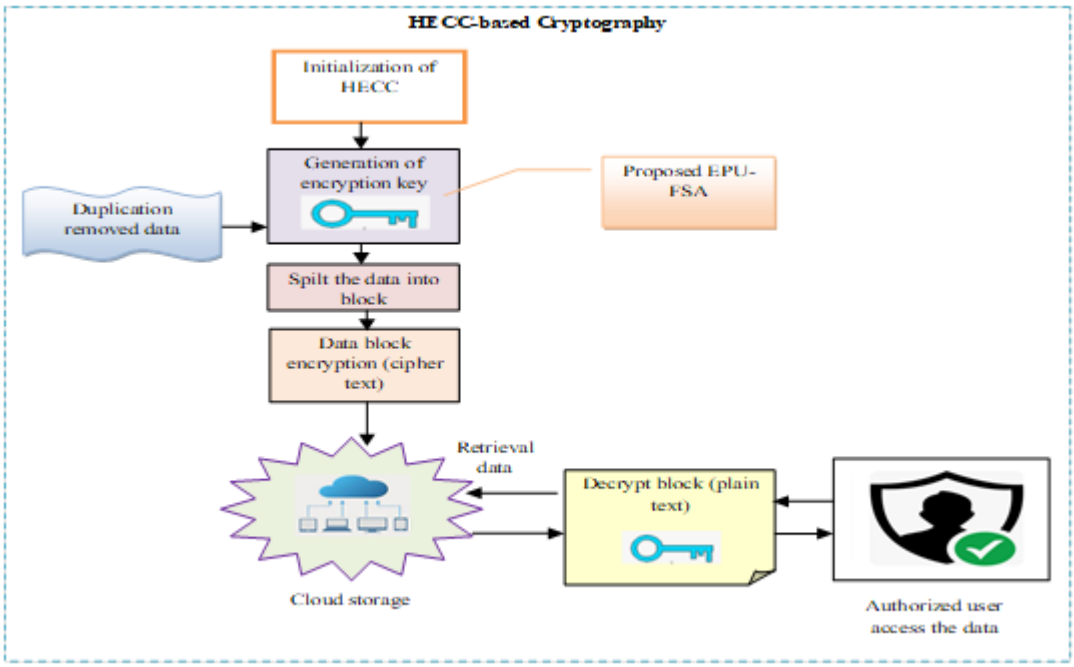


Fig. 5. Diagrammatic representation of HECC mechanism

E. Objective Function of Optimal Key-based Encryption Mechanism

In this proposed deduplication system, the reliability of the cryptographic method is increased by choosing the binary key-based HECC model. In this model, the EPU-FSA approach is selected for creating the key optimally. In the binary key-based HECC model, the encryption mechanism transfers the binary code into unique secure cipher text and during the decryption phase, the authorized user decrypts the cipher text back to plain text using the generated binary key. Thus, the data are securely transmitted over the cloud services. HECC-based encryption and decryption minimize the time and memory size needed for cloud storage. The objective function of the optimal key-based encryption mechanism K_{obj} is provided in Eq. (13).

$$K_{obj} = \underset{\{Key_b^{opt}\}}{\arg\min}\left(T_{me} + M_{sze}\right) \quad (13)$$

Here, the optimal key is indicated as Key_b^{opt}, the time and memory size of the system is specified as T_{me} and M_{sze} respectively.

Time $\left(T_{me}\right)$ is calculated by considering the time taken to generate both encryption and decryption in HECC.

Memory $\left(M_{sze}\right)$ is calculated based on the data entity and its size.

V. RESULTS AND DISCUSSION

A. *Experimental setup*

A secure data deduplication model was implemented on Python software. By implementing this model, cloud systems maintain more data efficiently. This successful deduplication scheme minimizes bandwidth usage by reducing the duplicated files. To boost the scalability of designed model, certain variables like hidden neuron count, epoch count, and learning rate were optimized with maximum iteration as 50, chromosome length as 3, and number of population as 10. The efficiency and security of the proposed data deduplication model were validated by comparing the proposed outcome with existing algorithms and techniques. Conventional algorithms like Black Widow Optimization (BWO) [27], Golden Eagle Optimizer (GEO) [28], Mud Ring Algorithm (MRA) [29], and Flamingo Search Algorithm (FSA) [26] were used for comparison. Traditional methods like BF [4], PRE [2], RNN [35], DBN [36], and LSTM [30] were also used to evaluate the strength of the proposed scheme. Several encryption algorithms like Data Encryption Standard (DES) [31], Advanced Encryption Standard (AES) [32], Rivest-Shamir-Adleman (RSA) [33], and Elliptic Curve Cryptography (ECC) [34] were used to ensure that the system meets the required security standards.

B. *Performance metrics*

Statistical formulas for assessing the performance metrics like Accuracy and FI-score are provided in the below points.

Accuracy (C_{ay}): The accuracy of the proposed model is calculated using Eq. (14).

$$C_{ay} = \frac{R_{ve} + R_{ne}}{R_{ve} + R_{ne} + G_{ve} + G_{ne}} \quad (14)$$

FI-score (F_{scre}): The FI-score of the model is assessed using Eq. (15).

$$F_{scre} = \frac{2 \cdot R_{ve}}{2 \cdot R_{ve} + G_{ve} + G_{ne}} \quad (15)$$

C. *Convergence Analysis*

The convergence analysis of the proposed data deduplication framework is provided in Fig. 6. By analyzing the convergences of data deduplication techniques, the most efficient deduplication method for reducing data duplication is identified through this proposed model. It helps to diminished the memory space and improve the data retrieval times. Convergence analysis helps to identify the strategies that maximize deduplication, leading to lower operating costs related to storage infrastructure. Thus, convergence analysis proved that the deduplication detection accuracy of the proposed EPU-FSA model is better than other deduplication techniques.

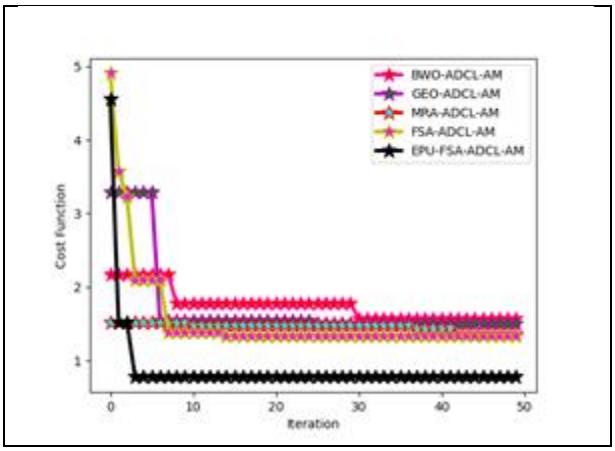
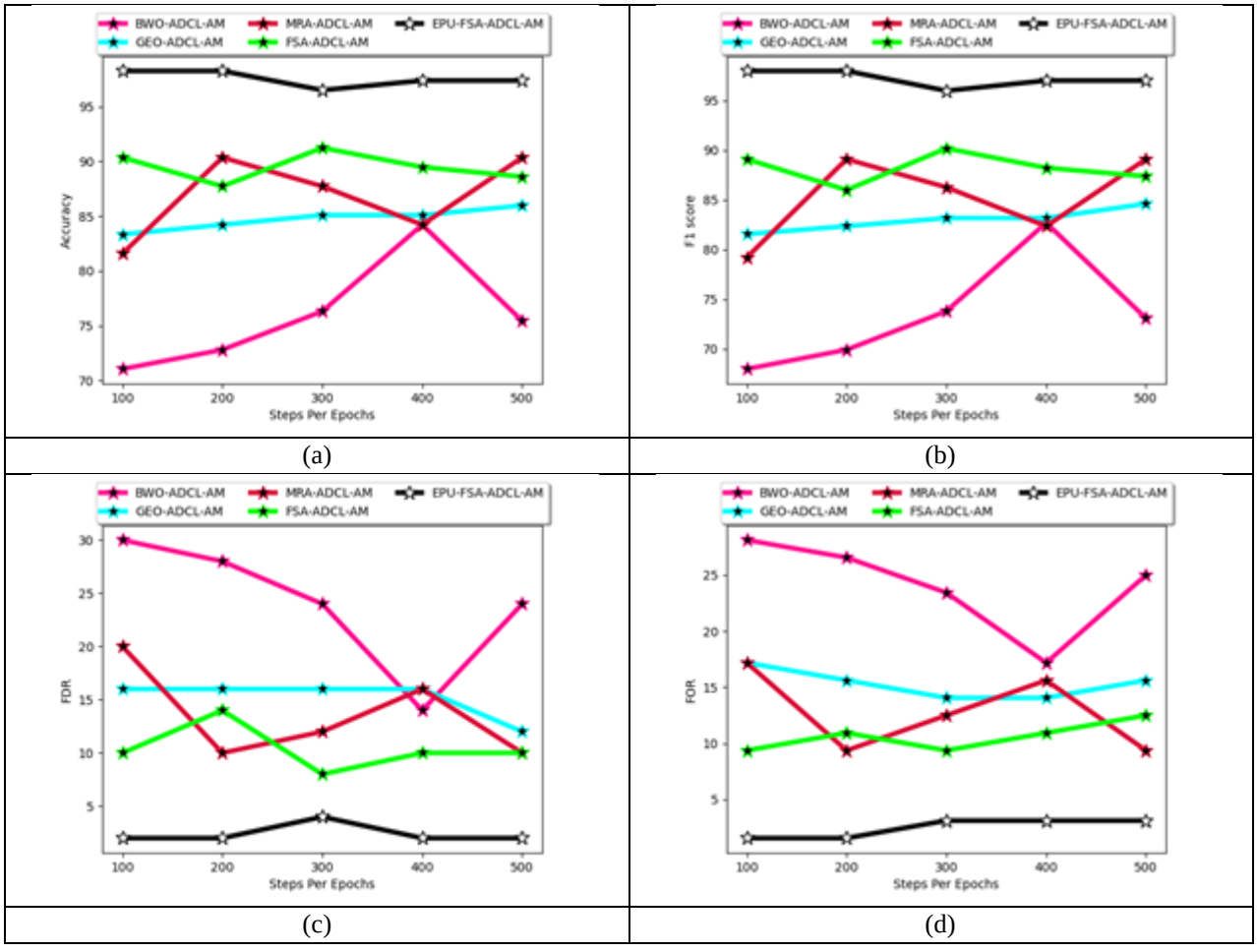
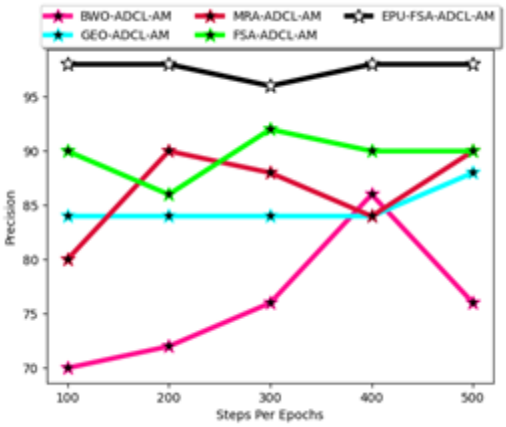


Fig. 6. Convergence analysis of the proposed data deduplication model

D. Performance Analysis of the Proposed data deduplication Model

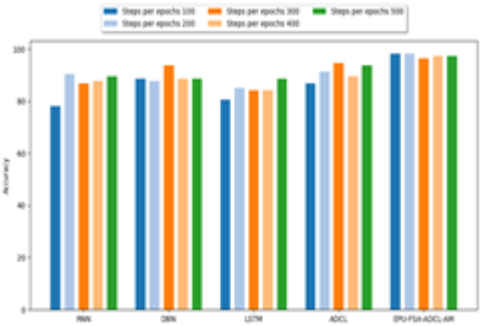
Analyzing various metrics helps to identify issues in the system such as slow processing times and high latency. Precise deduplication reduces redundant data storage, which enhances data processing and retrieval efficiency. This system functions more effectively, lowering the burden on storage by precisely detecting duplicates. The graphical visualization of performance assessments among algorithms is specified in Fig. 7 and analyses among different deduplication techniques are provided in Fig. 8. The accuracy of the proposed EPU-FSA-ADCL-AM model is 94.2%, which is better than traditional algorithms. Thus, for reducing duplicated data in the cloud system and for maintaining the integrity of unique data, accuracy analysis in a data deduplication model is essential.



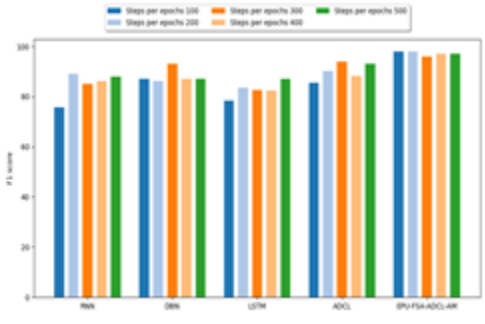


(e)

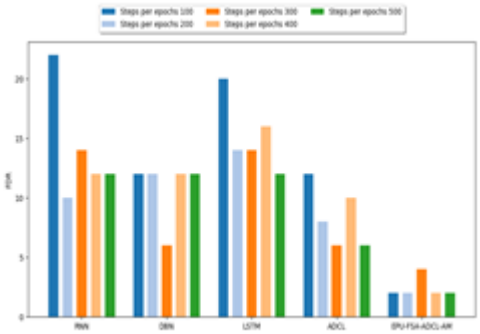
Fig. 7. Performance analysis of the proposed data deduplication model by varying algorithms regarding metrics like (a) Accuracy, (b) FI-score, (c) FDR, (d) FOR, and (e) Precision



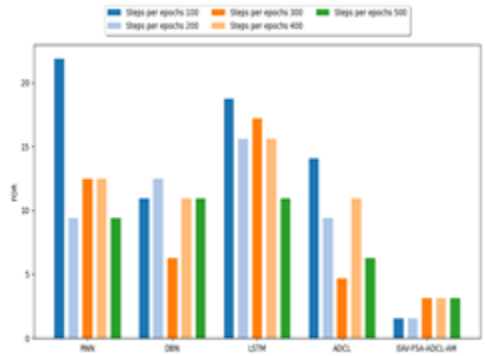
(a)



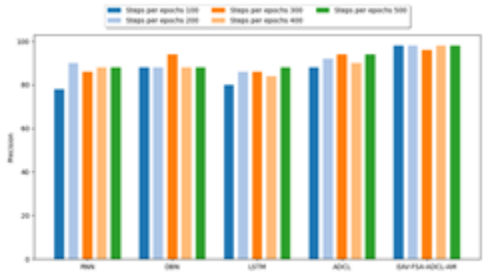
(b)



(c)



(d)



(e)

Fig. 8. Performance analysis of the proposed data deduplication model by varying classifiers regarding metrics like (a) Accuracy, (b) FI-score, (c) FDR, (d) FOR, and (e) Precision

E. Accuracy of Proposed Deduplication Model Analysis by Varying K-Fold Value

It is a statistical method used to assess the stability of the data deduplication model. It is commonly used to compare the performance of different models or algorithms to select the best one. It offers a reliable technique to assess a model's effectiveness. The numerical values of K-fold evaluation among numerous techniques and techniques are provided in Table II. By systematically training and testing the proposed model on different subsets of the data, K-fold analysis provides a more reliable estimate of how this model performs better to ensure that the model accurately detects the duplicated data. At a k-fold value of 5, the designed scheme obtained an accuracy of 98.2%, which is higher than other traditional models. Thus, the deduplication accuracy of the proposed EPU-FSA-ADCL-AM model is better than other deduplication methods.

TABLE II. ACCURACY ANALYSIS OF THE PROPOSED MODEL VARYING K-FOLD VALUES

Among Algorithms					
K-fold values	BWO-ADCL-AM [27]	GEO-ADCL-AM [28]	MRA-ADCL-AM [29]	FSA-ADCL-AM [26]	EPU-FSA - ADCL-AM
1	75.4386	86.84211	86.84211	88.59649	97.36842
2	83.33333	83.33333	85.96491	93.85965	96.49123
3	70.17544	84.21053	80.70175	91.22807	98.24561
4	77.19298	85.08772	85.96491	92.10526	95.61404
5	69.29825	84.21053	85.08772	88.59649	98.24561
Among Techniques					
K-fold values	RNN [35]	DBN [36]	LSTM [30]	ADCL	EPU-FSA - ADCL-AM
1	84.21053	90.35088	88.59649	88.59649	97.36842
2	86.84211	93.85965	86.84211	93.85965	96.49123
3	88.59649	87.7193	85.08772	93.85965	98.24561
4	82.45614	89.47368	82.45614	89.47368	95.61404
5	89.47368	92.98246	87.7193	96.49123	98.24561

F. State-of-the-Art Methods Comparison of Proposed Deduplication Model

The proposed data deduplication model is assessed with existing methods. The numerical values of these assessments are provided in Table III. Based on these analyses, the precision rate of the designed scheme is boosted with 28.9%, 6.5%, 19.5% and 8.8% than BF, PRE, LSTM, and ADCL-AM. State-of-the-art comparison helps to identify the most successful deduplication technique, which reduces redundancy and optimizes the usage of storage in cloud services. Thus, from these analyses, the proposed EPU-FSA-ADCL-AM model achieved high efficiency in identifying deduplicated data and provided faster data handling and retrieval times.

TABLE III. STATE-OF-THE-ART METHODS ANALYSIS OF THE PROPOSED DEDUPLICATION MODEL

TERMS	BF [4]	PRE [2]	LSTM [30]	ADCL-AM	EPU-FSA -ADCL-AM
Accuracy	75.4386	91.22807	84.21053	90.35088	97.36842
FI-score	73.07692	90.19608	82	89.10891	97.0297
Precision	76	92	82	90	98
FDR	24	8	18	10	2
FOR	25	9.375	14.0625	9.375	3.125

G. Entity Extraction Analysis of Proposed Model

The resultant analysis of entity extraction is specified in Fig. 9. In this proposed model, entity extraction is performed for grouping similar entities to form the structured data. By accurately extracting the entities, the

accuracy of the designed EPU-FSA-HECC-based deduplication system is significantly improved. In this entity extraction analysis, the outcome of the EPU-FSA-HECC model is assessed with other algorithms to obtain the high reliability of the proposed deduplication model. The correlation coefficient of the EPU-FSA-HECC model is 51.2% enhanced than other existing deduplication models.

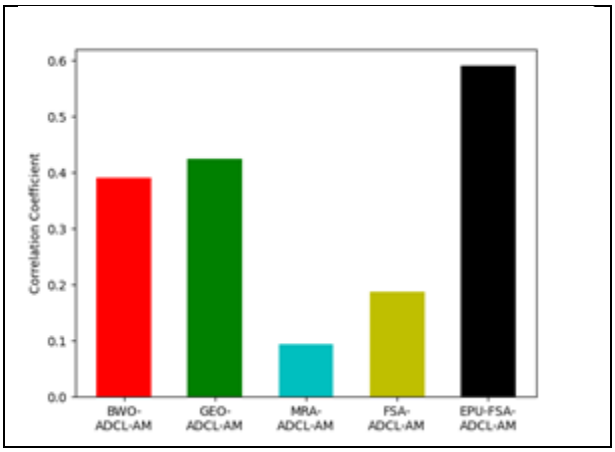


Fig. 9. Entity extraction analysis of the proposed model

H. CPA and KPA Attack Analysis of Proposed Encryption Model

Attacks like CPA and KPA analyses among different algorithms are provided in Fig. 10. It helps to identify and classify key elements within unstructured data. The analyses are carried out among different cases of attacks. It assists in ensuring that sensitive information is appropriately recognized and maintained in the setting of encrypted data. The correlation coefficient of the proposed EPU-FSA-HECC is decreased with 33.6% than DES, 30% than AES, 12.5% than RSA, and 10.2% than HECC. A lower correlation coefficient indicates that the extracted entities are less redundant and more distinct from one another. Thus, in cloud computing environments, the competence of the designed EPU-FSA-HECC model is enhanced for efficient deduplication process.

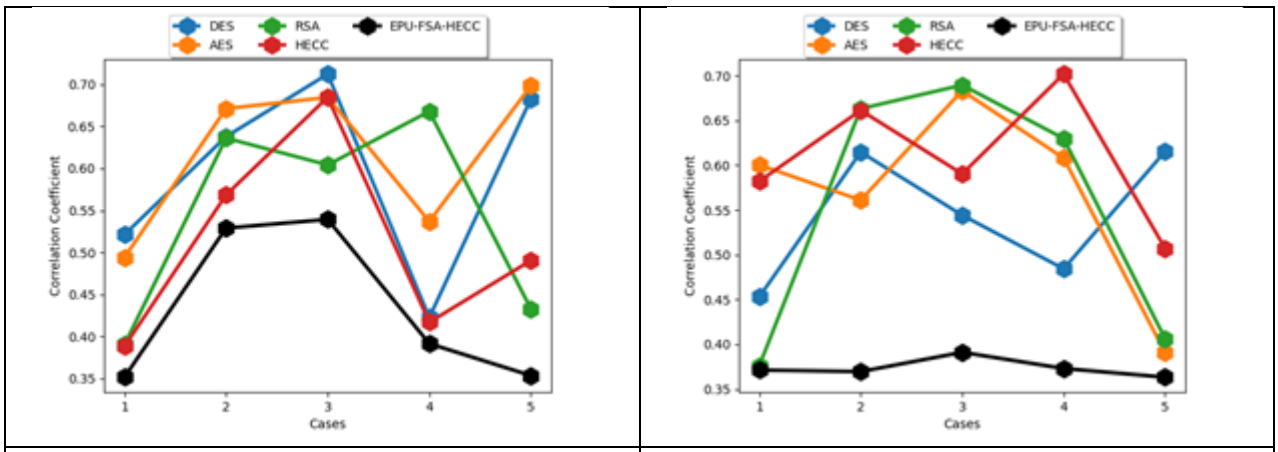
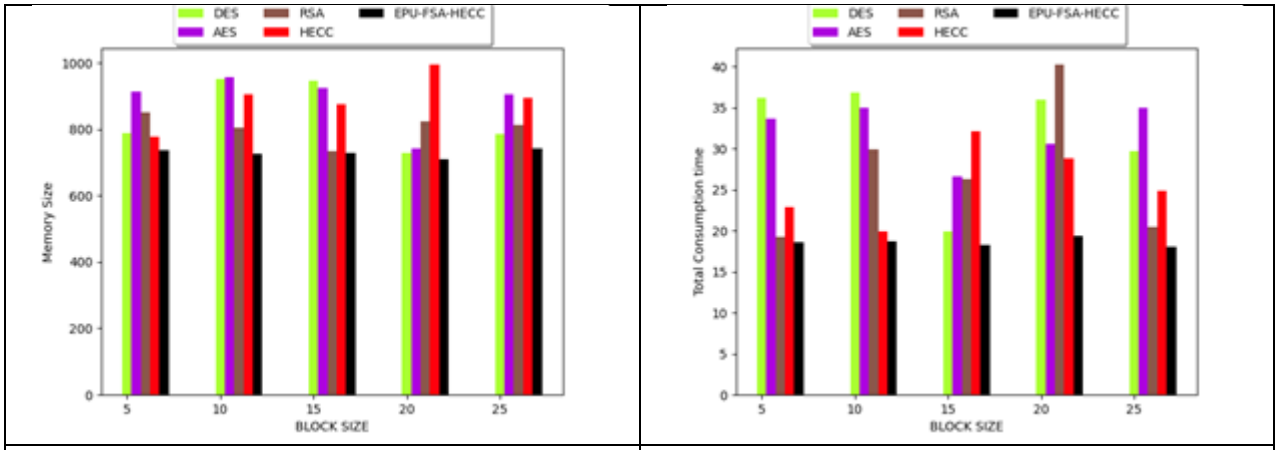


Fig. 10. Attacks analysis of the proposed encryption model regarding (a) CPA attack and (b) KPA attack

I. Computation Time and Memory Size Analysis of Proposed Encryption Model

Computation time and memory size are critical factors that significantly impact the efficiency of cloud platforms. The time complexity of the algorithms used for entity extraction is crucial and utilizing cryptographic methods significantly reduces computation time. By more effectively storing frequently requested data, encryption techniques aid in lowering memory use. The graphical representations of computation time and memory size analysis are provided in Fig. 11. The effectiveness of data processing algorithms is impacted by the block size selection. At, block 10, the memory size of the proposed EPU-FSA-HECC model is decreased by

2.5%, 8.2%, 6.02%, and 1.2% than DES, AES, RSA, and HECC. Varying block sizes improve the performance of cryptographic methods. Thus, by varying block sizes and analyzing their impact on computation time and memory size, the proposed EPU-FSA-HECC model leads to enhanced reliability in cloud platforms.

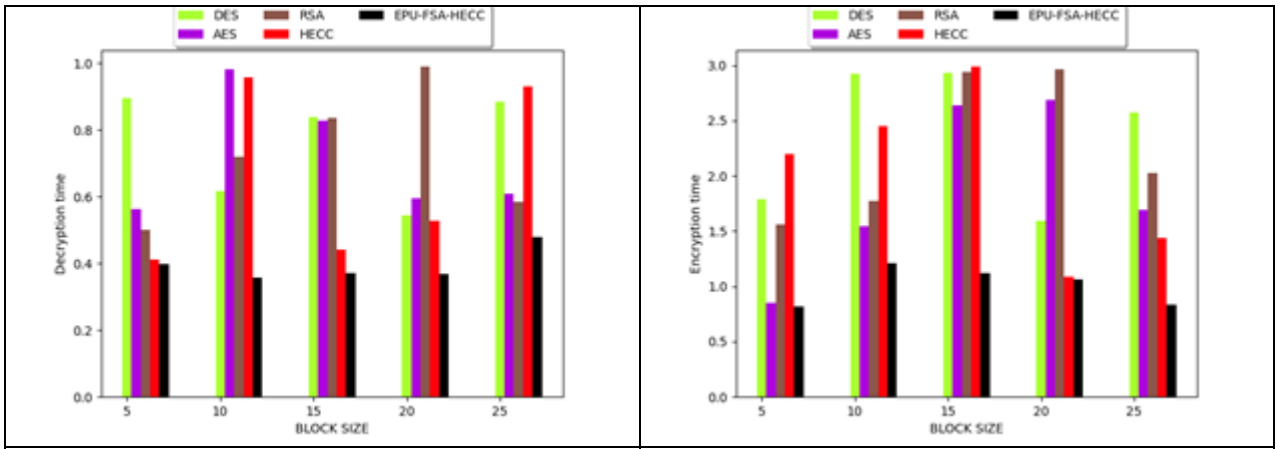


(a) (b)

Fig. 11. Analysis of the proposed encryption model regarding (a) Memory size and (b) Total computation time

J. Encryption and Decryption Time Analysis of Proposed Encryption Model

Fig. 12, shows the encryption and decryption analysis of the suggested framework. Analyzing the time needed to encrypt and decode data using different cryptographic methods like DES, AES, RSA, and HECC is graphically represented. The size of the encryption key greatly affects the time needed for encryption and decryption. While larger keys offer stronger security, they also result in longer processing times. Varying block sizes significantly influence encryption and decryption times. At block size 5, the decrypted time of designed EPU-FSA-HECC model is decreased by 55.5% than DES, 32.2% than AES, 16.6% than RSA, and 6.9% than HECC. While larger blocks improve the performance with security issues, smaller blocks result in higher overhead. Thus, it is very essential to conduct testing and analysis to determine the best block size. The encryption time of the designed EPU-FSA-HECC is reduced with 55.5% than DES, 13.3% than AES, 18.7% than RSA, and 48% than HECC. Thus, from these analyses, the encryption and decryption capability of the proposed EPU-FSA-HECC is enhanced than other cryptography techniques.



(a) (b)

Fig. 12. Analysis of the proposed encryption model regarding (a) Decryption time and (b) Encryption time

VI. Conclusion

A secured data deduplication model was developed using a deep network for the identification and elimination of duplicated copies of data in cloud storage. This proposed model helps to ensure that unique data was stored, which reduces the duplication issues by boosting the data flexibility. Different attributes were

collected from the standard database and from these attributes the entities were extracted with the aid of the ADCL-AM model. The stability of the entity extraction task was boosted by fine-tuning the parameters from the deep network with the aid of the EPU-FSA strategy. From this generated entity, data patterns were produced for the deduplication process. In this proposed model, the ADCL-AM model was optimized using EPU-FSA to find the duplicated data from the data patterns. The security issues raised in this model were diminished by implementing the HECC-based cryptography mechanism. Integrating the key-based deduplication provides an optimal security solution. Thus, the accuracy rate of the suggested framework is boosted with 29% than BF, 6.73% than PRE, 15.6% than LSTM, and 7.7% than ADCL-AM. To boost the flexibility, a data compression technique will be implemented to reduce the storage cost in cloud services. Thus, the future deduplication model eventually contributed to developing secure data exchange over the network with the support of individual secret keys during the cryptography mechanism.

References

[1] Liu, X., Lu, T., He, X., Yang, X. and Niu, S., "Verifiable attribute-based keyword search over encrypted cloud data supporting data deduplication", IEEE Access, vol.8, pp.52062-52074, 2020.

[2] Xixun Yu, Hui Bai, Zheng Yan, Rui Zhang, "VeriDedup: A Verifiable Cloud Data Deduplication Scheme With Integrity and Duplication Proof", IEEE Access, vol.20, pp.680-694, 2022.

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

[4] Ebinazer, S.E. and Savarimuthu, N., "An efficient secure data deduplication method using radix trie with bloom filter (SDD-RT-BF) in cloud environment", Peer-to-Peer Networking and Applications, vol.14, pp.2443-2451, 2021.

[5] Azad, P., Navimipour, N.J., Rahmani, A.M. and Sharifi, A., "The role of structured and unstructured data managing mechanisms in the Internet of things", Cluster computing, vol.23, pp.1185-1198, 2020.

[6] Dr. R. Vijaya Kumar Reddy, G. Venugopal, Girijaswi Rajanala, V. Sambasivarao,N. Harshavardhan, T. Ajay, "Transforming Unstructured data to Structured data using Map Reduce and HBase", International Journal of Emerging Trends in Engineering Research, vol.8, no.9, 2020.

[7] Sivarama krishnan N, Vandana V, Vishali, Dharshana S G, Subramaniyaswamy V, Umamakeswari A, "Conversion Of Unstructured Data To Structured Data With A Profile Handling Application" International Journal of Mechanical Engineering and Technology, vol.8, no.8, pp.623-630, 2017.

[8] Anujna M., Ushadevi A., "Converting and Deploying an Unstructured Data using Pattern Matching", American Journal of Intelligent Systems, vol.7, pp.54-59, 2017.

[9] Zhang, G., Yang, Z., Xie, H. and Liu, W., "A securely authorized deduplication scheme for cloud data based on blockchain", Information Processing & Management, vol.58, no.102510, 2021.

[10] Javadpour, A., Abadi, A.M.H., Rezaei, S., Zomorodian, M. and Rostami, A.S., "Improving load balancing for data-duplication in big data cloud computing networks", Cluster Computing, vol.25, pp.2613-2631, 2022.

[11] PG, S., RK, N., Menon, V.G., P, V., Abbasi, M. and Khosravi, M.R., "A secure data deduplication system for integrated cloud-edge networks", Journal of Cloud Computing, vol.9, no.61, 2020.

[12] Premkamal, P.K., Pasupuleti, S.K., Singh, A.K. and Alphonse, P.J.A., "Enhanced attribute-based access control with secure deduplication for big data storage in cloud", Peer-to-Peer Networking and Applications, vol.14, pp.102-120, 2021.

[13] Bai, J., Yu, J. and Gao, X., "Secure auditing and deduplication for encrypted cloud data supporting ownership modification", Soft Computing, vol.24, pp.12197-12214, 2020.

[14] Yin, J., Tang, Y., Deng, S., Zheng, B. and Zomaya, A.Y., "Muse: A multi-tierd and sla-driven deduplication framework for cloud storage systems", IEEE Transactions on Computers, vol.70, pp.759-774, 2020.

[15] [15] Shen, W., Su, Y. and Hao, R., "Lightweight cloud storage auditing with deduplication supporting strong privacy protection", IEEE Access, vol.8, pp.44359-44372, 2020.

[16] Yang, X., Lu, R., Shao, J., Tang, X. and Ghorbani, A.A., "Achieving efficient secure deduplication with user-defined access control in cloud", IEEE Transactions on Dependable and Secure Computing, vol.19, pp.591-606, 2020.

[17] Periasamy, J.K. and Latha, B., "Efficient hash function–based duplication detection algorithm for data Deduplication deduction and reduction", Concurrency and Computation: Practice and Experience, vol.33, no.e5213, 2021.

[18] Yao, F., Pu, C. and Zhang, Z., "Task duplication-based scheduling algorithm for budget-constrained workflows in cloud computing", IEEE Access, vol.9, pp.37262-37272, 2021.

[19] Xiong, J., Zhang, Y., Lin, L., Shen, J., Li, X. and Lin, M., "ms-PoSW: A multi-server aided proof of shared ownership scheme for secure deduplication in cloud", Concurrency and Computation: Practice and Experience, vol.32, no.e4252, 2020.

[20] Tang, X., Zhou, L., Hu, B. and Wu, H., "Aggregation-Based Tag Deduplication for Cloud Storage with Resistance against Side Channel Attack", Security and Communication Networks, no.6686281, 2021.

[21] Kan, G., Jin, C., Zhu, H., Xu, Y. and Liu, N., "An identity-based proxy re-encryption for data deduplication in cloud", Journal of systems architecture, vol.121, no.102332, 2021.

[22] Li, S., Xu, C., Zhang, Y., Du, Y. and Chen, K., "Blockchain-based transparent integrity auditing and encrypted deduplication for cloud storage", IEEE Transactions on Services Computing, vol.16, pp.134-146, 2022.

[23] Elkana Ebinazer, S., Savarimuthu, N. and Mary Saira Bhanu, S., "ESKEA: enhanced symmetric key encryption algorithm based secure data storage in cloud networks with data deduplication", Wireless Personal Communications, vol.117, pp.3309-3325, 2021.

[24] Premkamal, P.K., Pasupuleti, S.K., Singh, A.K. and Alphonse, P.J.A., "Enhanced attribute based access control with secure deduplication for big data storage in cloud", Peer-to-Peer Networking and Applications, vol.14, pp.102-120, 2021.

[25] Saharan, S., Somani, G., Gupta, G., Verma, R., Gaur, M.S. and Buyya, R., "QuickDedup: Efficient VM deduplication in cloud computing environments", Journal of Parallel and Distributed Computing, vol.139, pp.18-31, 2020.

[26] Wang Zhiheng And Liu Jianhua, "Flamingo Search Algorithm: A New Swarm Intelligence Optimization Algorithm", IEEE access, vol. 9, 2021.

[27] Hayyolalam, Vahideh, and Ali Asghar Pourhaji Kazem, "Black widow optimization algorithm: a novel meta-heuristic approach for solving engineering optimization problems," Engineering Applications of Artificial Intelligence, vol.87, pp.103249, 2020.

[28] Mohammadi-Balani, Abdolkarim, Mahmoud Dehghan Nayeri, Adel Azar, and Mohammadreza Taghizadeh-Yazdi, "Golden eagle optimizer: A nature-inspired metaheuristic algorithm." Computers & Industrial Engineering, vol.152, pp.107050, 2021.

[29] Desuky, Abeer S., Mehmet Akif Cifci, Samina Kausar, Sadiq Hussain, and Lamiaa M. El Bakrawy, "Mud Ring Algorithm: A new meta-heuristic optimization algorithm for solving mathematical and engineering challenges," IEEE Access, vol.10, pp.50448-50466, 2022.

[30] Ravikanth, M., Korra, S., Mamidisetti, G., Goutham, M., & Bhaskar, T, "An efficient learning based approach for automatic record deduplication with benchmark datasets," Scientific Reports, vol.14, issue.1, 16254, 2024.

[31] Zeebaree, S.R, "DES encryption and decryption algorithm implementation based on FPGA. Indones. J. Electr. Eng. Comput. Sci, vol.18, issue.2, pp.774-781, 2020.

[32] Abdullah, A. M, "Advanced encryption standard (AES) algorithm to encrypt and decrypt data," Cryptography and Network Security, vol.16, issue.1, 11, 2017.

[33] Meneses, Fausto, Walter Fuertes, José Sancho, Santiago Salvador, Daniela Flores, Hernán Aules, Fidel Castro, Jenny Torres, Alba Miranda, and Danilo Nuela, "RSA encryption algorithm optimization to improve performance and security level of network messages," vol.6, no. 8, 2016.

[34] Hafsa, A., Sghaier, A, Malek, J, & Machhout. M, "Image encryption method based on improved ECC and modified AES algorithm," Multimedia Tools and Applications, vol.80, pp.19769-19801, 2021.

[35] Kim, Joo-Chang, and Kyungyong Chung, "Recurrent neural network-based multimodal deep learning for estimating missing values in healthcare," Applied Sciences, vol.12, no. 15 pp. 7477, 2022.

[36] Kukkar, A., Mohana, R., Kumar, Y., Nayyar, A., Bilal, M. and Kwak, K.S, "Duplicate bug report detection and classification system based on deep learning technique," IEEE Access, 8, pp.200749-200763, 2020.

[37] Qiu, Jiayu, Bin Wang, and Changjun Zhou, "Forecasting stock prices with long-short term memory neural network based on attention mechanism," PloS one 15, no. 1, e0227222, 2020.

[38] Devarajan, Malathi, and N. Sasikaladevi, "An hyper elliptic curve based efficient signcryption scheme for user authentication," Journal of Intelligent & Fuzzy Systems 39, no. 6, pp.8487-8498, 2020.