

<sup>1</sup> Mr. Manjunath  
Singh. H

<sup>2</sup> Dr. Tanuja R

## Elucidation of Adaptive Long Short-Term Memory for Data Deduplication and Data Security Enhancement by Hashing Algorithm over Cloud Sector



**Abstract:** - The advanced big data technology and cloud computing methodology change the method of the user and efficiency while processing the data, where efficient storage and scalable computing are provided to the user by the cloud servers anytime and anywhere. Many cloud service providers have been attracted to the data deduplication approach that highly decreased storage costs. The bandwidth requirement of users has been reduced and data redundancies in cloud storage have been removed by using data deduplication techniques. Most of the general data deduplication models are affected by many privacy and security problems because of the outsourced data transmission techniques of cloud storage. Therefore, deduplication approaches have been implemented to handle specific privacy and security issues that leads to a wide range of trade-offs and solutions for cloud data. Hence, a new approach is introduced based on the adaptive network and hashing for secure data deduplication. Input attributes are collected initially using the standard dataset. The collected attributes are given to the Adaptive Long Short-Term Memory (ALSTM) model for data deduplication. Here the attributes in the implemented ALSTM are tuned by the Bald Eagle Search (BES) strategy. Further, the hashing function is included to encode the deduplication data to improve security. Finally, the attained result from the implemented framework is compared with the standard data deduplication method for analyzing the deduplication efficiency and security of the proposed model.

**Keywords:** Data Deduplication; Data Security Enhancement; Cloud Sector; Hashing Algorithm; Adaptive Long Short-Term Memory

### I. INTRODUCTION

Cloud storage provides flexible services over the on-demand data outsourcing paradigm with attractive features like ubiquitous data access from any location, eliminating the requirement of the storage management agreement, individual maintenance, and elusion of capital expenditures on software and hardware [1]. In the cloud, data hosting produces higher security issues to user data because data availability is increasing based on the amount of cloud users [2]. Hence, a huge amount of memory space is required for storing these data [3]. Data compression techniques are generally applied to reduce the size of data on a big scale, but the disturbance of these approaches is high and repeated data also causes some storage issues in the cloud. Subsequently, the storage space issues are effectively solved by the data deduplication approaches that increase the storage space efficiency. Usually, chunks and files are denoted with respect to hash values that are fingerprints, which can be contrasted with the remaining hash functions to find whether the file they denote is a duplicate [4]. Data deduplication is broadly adopted for cloud storage servers that maximize the bandwidth and reduce the amount of unused Random Access Memory (RAM) [5].

The advanced hash function-based data duplication model is needed to be employed in the cloud environment before outsourcing the data [6]. However, data encryption and deduplication are incompatible [7]. Hence, the images are recognized with respect to their visual content, which can be accomplished by the new perceptual hash function named as “Local Binary Pattern (LBP)”, which provides a secure imagery deduplication technique with the combination of clustering and hashing. Traditional data deduplication approaches in cloud data storage are affected by information integrity, information security, and unauthorized user access [8]. Generally, data security and redundancy challenges come from the storage of information and the economical dispersal of data among huge amount of users [9]. Traditional data deduplication approaches increase bandwidth consumption and storage requirements [10]. To improve the data deduplication effectiveness with respect to similarity and location-related indexing, a deep learning method is suggested in this work. The hash-based deep learning scheme detects the match cases for finding similar data to remove data redundancy in the cloud. The accuracy of finding deduplicated data using the deep learning model is greater.

The major contributions of the presented deep learning-assisted hash function-based data deduplication model are elucidated as follows.

- To implement a new hash function-based deep learning model for performing data deduplication in the cloud environment to decrease the storage requirements and cost. It also maintains data security by using hash functions for data storage.

<sup>1</sup> Research Scholar, Department of Computer Science, UVCE, Bangalore, India. Email: mansh24.singh@gmail.com

<sup>2</sup> Associate Professor, Department of Computer Science, UVCE, Bangalore, India. Email: tanujar.uvce@gmail.com

Copyright © JES 2024 on-line: journal.esrgroups.org

- To suggest a BES strategy for tuning hyperparameters from LSTM to improve the detection performance of duplicate files.
- To develop A-LSTM for performing duplicated data detection by matching the files from the cloud. It improves the detection efficiency in large-scale data management systems.
- To store the deduplicated data in the cloud system by removing duplicate files using hash functions to protect the information from third parties. The users access the data with ownership permission to give greater security, confidentiality, and integrity to the data.
- To validate the performance of the developed data deduplication scheme using various metrics among previously developed models.

The organization of the developed hash function with deep learning-based data deduplication model is summarized as follows. Unit II gives the literature review with its problem statement. Unit III provides the architectural view and dataset collection of the data deduplication model. Unit IV illustrates a brief description of the data deduplication scheme with an explanation of deep learning and hash functions. Experimental validation is explained in Unit V and the summary of the proposed framework is given in Unit VI.

## II. LITERATURE SURVEY

### A. Related Works

In 2024, Ghassabi *et al.* [11] have introduced a secure data deduplication mechanism for textual data. The developed client-side and cloud-side deduplication strategy accomplished higher compression rates. This model maintained higher security in the cloud environments. It significantly reduced the storage requirements and improved the efficiency in managing large-scale data.

In 2024, Periasamy *et al.* [12] have implemented a lightweight fuzzy-based “Convolutional Neural Network (CNN)” for data deduplication, especially in cloud servers. The data has been initially classified into highly sensitive and normal sensitive. The hash function-related data deduplication approach has been used for maintaining confidential information. The ideal keys were created with respect to security level using a fuzzy-based tuna search algorithm. The accuracy of this approach was higher than previous schemes.

In 2022, Ma *et al.* [13] have presented a deduplication method in server-side with dynamic ownership management. Unauthorized cloud users have been prevented from using pre-verified access control methods to download data. The communication overhead has been mitigated by this model and it provides higher security.

In 2017, Abusaimh *et al.* [14] have recommended a hybrid data deduplication method at the file level for cloud storage. The duplicated chunks were effectively identified by this approach. The execution time of this model was lower and it meets the satisfactory requirements of security.

In 2024, Davea *et al.* [15] have suggested a randomized encryption algorithm for deduplication in the cloud storage system. The Merkle hash tree has been used for encrypting the data. This approach has been highly suitable in real-world environments.

### B. Problem statement

High data availability is achieved in a large distributed storage system by maintaining the replication factor, which is a minimum number of replica data. The storage cost, storage requirements and computation energy are decreased by removing the duplicate data, which is above the value of the replication factor. But, these data matching techniques have lower efficacy and efficiency; hence data deduplication methods are suggested for removing the redundant data to improve the performance of the cloud environment. The challenges while performing data deduplication are given as follows.

- Most of the deep learning methodologies are affected by generalization issues and overfitting issues that affect the data deduplication performance.
- Labeling errors produced during data deduplication are higher in conventional approaches.
- A huge amount of data is produced by the users in cloud systems, and handling this massive amount of data is a challenging issue in traditional models.
- Existing approaches only focused on reducing storage costs and requirements, data security and data integrity are the challenging issues while storing and transferring data in the cloud.
- The time required for performing data deduplication in the conventional approach is high.

Therefore, advanced deep learning-assisted data deduplication methodologies are designed to provide better performance in cloud storage systems, but these techniques are affected by several issues that are illustrated in Table 1.

**Table.1** Benefits and disadvantages of previous data deduplication methodologies in cloud system

| Author [citation]            | Methodology                        | Features  | Challenges  |
|------------------------------|------------------------------------|---|---|
| Ghassabi <i>et al.</i> [11]  | Encryption                         | <ul style="list-style-type: none"> <li>• It achieves higher compression rates and it protects the confidentiality of the textual data.</li> <li>• It significantly reduces the storage requirements.</li> </ul> | <ul style="list-style-type: none"> <li>• The accuracy of this approach is poor.</li> <li>• The computational overhead of the implemented scheme is high.</li> </ul> |
| Periasamy <i>et al.</i> [12] | CNN                                | <ul style="list-style-type: none"> <li>• It maintains the confidential information.</li> <li>• It requires much less computational time and it achieves higher accuracy.</li> </ul>                             | <ul style="list-style-type: none"> <li>• It needs to encourage the cloud storage system for easy transmission and reception.</li> </ul>                             |
| Ma <i>et al.</i> [13]        | Data ownership-based deduplication | <ul style="list-style-type: none"> <li>• It highly mitigates the communication overhead.</li> <li>• It efficiently restricts unauthorized users to protect sensitive data.</li> </ul>                           | <ul style="list-style-type: none"> <li>• The energy consumption of the system is slightly higher.</li> </ul>  |
| Abusaimeh <i>et al.</i> [14] | Chunk-level data deduplication     | <ul style="list-style-type: none"> <li>• It has a lower execution time.</li> <li>• The efficiency of the system is increased by reducing the data size.</li> </ul>  | <ul style="list-style-type: none"> <li>• Processing smaller chunks of data has more computational overhead.</li> </ul>  |
| Davea <i>et al.</i> [15]     | Merkle hash tree                   | <ul style="list-style-type: none"> <li>• It decreases the storage and communication costs.</li> <li>• It provides higher confidentiality.</li> </ul>  | <ul style="list-style-type: none"> <li>• The energy consumption of the model is higher.</li> </ul>  |

### III. DEVELOPMENT OF DATA DEDUPLICATION WITH DATA SECURITY MODEL IN CLOUD SECTOR: ARCHITECTURAL ILLUSTRATION AND DATASET COLLECTION

#### A. Architectural Description of Proposed Model

Data deduplication is an efficient approach for decreasing the consumption of storage spaces by removing identical files and data blocks from storing in the cloud. Traditional data deduplication approaches suffer from computational overhead because of the huge amount of data. The bandwidth requirements for storing massive amounts of data in the cloud are also greater. Traditional data deduplication approaches do not address the security challenges in the cloud. Traditional deep learning techniques do not address the issue of data confidentiality preservation and effective load balancing in the cloud. The accuracy of the deduplication process using traditional machine learning models is low and handling a huge volume of data during training is also challenging. Therefore, an optimization-aided deep learning model is developed for performing the data deduplication process. Security enhancement is also considered in this work using the hash function to provide higher data security and integrity. The structural illustration of the presented data deduplication framework is given in Fig. 1.

A new data deduplication model is suggested in this work to detect redundant files to decrease the storage and memory requirements of the cloud system. Additional security of data is also provided by this developed model using hash functions. The input attributes such as file name, hashtag, file size, file location, block name, data pattern, last modified date, and file size are given to the input of the deep learning mechanism for performing the data deduplication process. The A-LSTM model is developed in this work for finding similar files from the cloud. The detection of duplicate file performance is enhanced via the optimization of hyperparameters such as learning rate, hidden neurons, and epochs from LSTM. The duplicate files are removed and new information is saved with a hash function in the cloud. Here, ownership permission is needed to access the data after storing it into the cloud. Hence, the confidentiality and security of the data is improved by this hash function. The effectiveness of the presented data deduplication framework is verified with the prior works to show the effectiveness of the model.

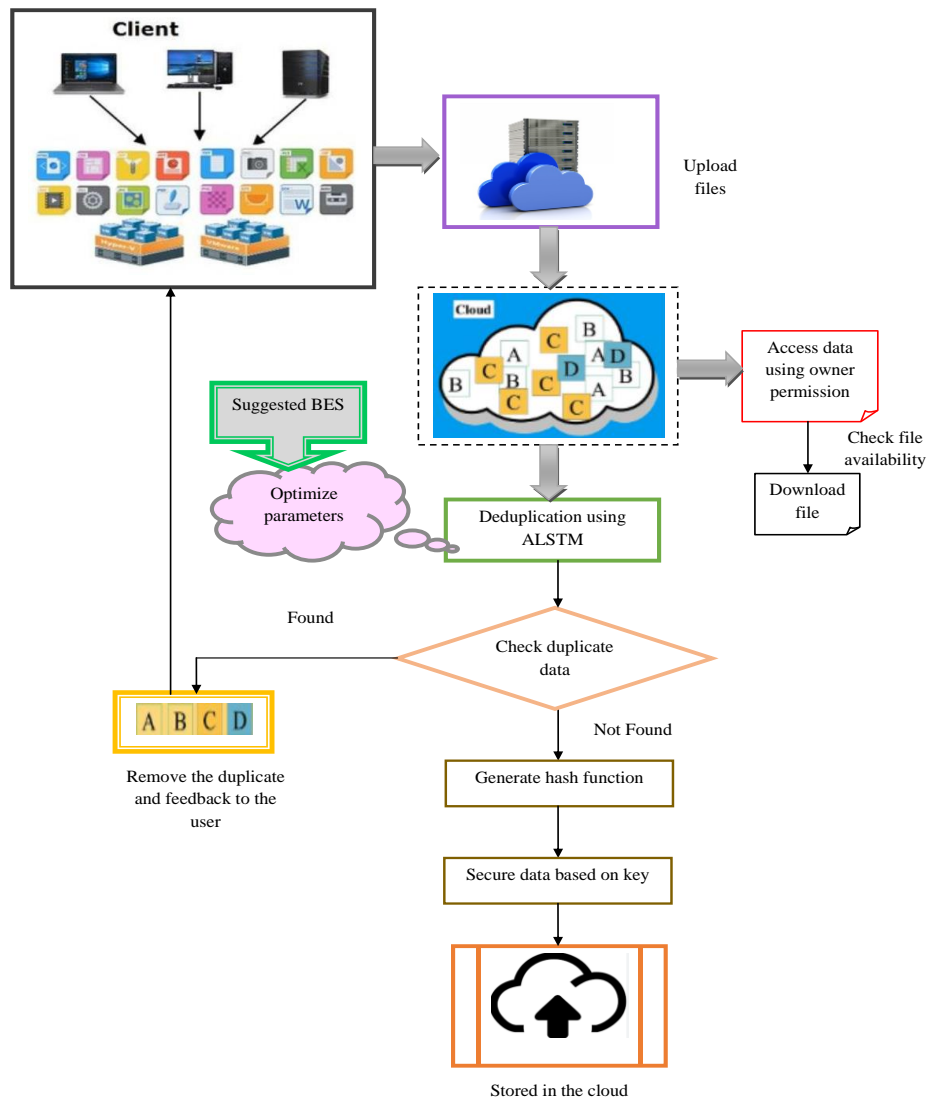


Fig. 1. Developed deep learning with hash function-based data deduplication model

B. Description of Deduplicate Data

The data required for performing data deduplication with data security enhancement is garnered from the publically available databases. Kaggle source is utilized for attaining the required information for the data deduplication process.

**Dataset 1 (Cholesterol):** From the website of “https://www.kaggle.com/mathurinache/cholesterol, the cholesterol data is collected with access date: 2024-05-28”. This directory includes 4 databases that concern heart disease identification datasets. All the attributes present in this dataset are numeric-valued. In this dataset, 76 raw attributes are there, from which only 14 are used.

**Dataset 2 “(Predict Diabetes from Medical Records)”:** The Kaggle source is utilized for garnering the data, which is taken from the website of “https://www.kaggle.com/paultimothymooney/predict-diabetes-from-medical-records/data: with access date 2024-05-28”. This dataset includes several medical predictor attributes named as independent and one target variable named as dependent.

C. Elucidation of Bald Eagle Search Algorithm

The BES algorithm is utilized in the developed framework for optimally tuning the variables from the LSTM structure to improve data deduplication performance. The optimal solution updating process using BES is described as follows.

The BES mimics the bald eagle's behaviour during hunting in the search space. The important stages of this algorithm are search space preference, searching inside the search space and swooping. The best area is preferred based on the food in the search stage and then it hunts the prey within the preferred search space. The select phase is mathematically represented in Eq. (1).

$$S_{new,j} = S_{bst} + \mathcal{G} * Rnd(S_{mean} - S_j) \tag{1}$$

The change in the position is controlled by the parameter  $\mathcal{G}$ , which is taken in the interval among  $[1.5,2]$ , the random attribute in the interval among  $[0,1]$  is denoted as  $Rnd$ , the search space to be currently preferred is indicated as  $S_{bst}$ , and the mean position based on the previously selected search space is indicated by the term  $S_{mean}$ .

The bald eagles search the prey and it moves in various directions. This prey-searching process is carried out in a selected search space. This search strategy is mathematically represented in Eq. (2).

$$S_{new,j} = S_j + u(j) * (S_j - S_{j+1}) + v(j) * (S_j - S_{mean}) \tag{2}$$

Here, the terms  $u(j)$  and  $v(j)$  represent the x-axis and y-axis spiral coordinates. The best position for swopping and hunting is selected in the spiral position.

The bald eagles swing from the appropriate location in the swopping stage based on the target prey. Here, all the bald eagles are moved towards the best location, which is mathematically indicated in Eq. (3).

$$S_{new,j} = Rnd * S_{bst} + u1(j) * (S_j - e1 * S_{mean}) + v1(j) * (S_j - e2 * S_{bst}) \tag{3}$$

Here, the terms  $u1(j)$  and  $v1(j)$  represent the new spiral coordinates on the x and y-axis. Moreover,  $e1$  and  $e2$  are the random parameters used in the range of  $[1,2]$ . The most appropriate solution is produced by increasing the movement intensity towards the centre and best points.

|   |   |
|---|---|
| <b>Algorithm 1: BES strategy</b>                          |   |
| Input: Randomly initialize the point $S_j$                |   |
| Output: Optimize hidden neurons, Learning rate and epochs |   |
| Evaluate fitness function for the initial point $S_j$     |   |
| While (termination condition not satisfied)               |   |
| Select space strategy                                     |   |
|   | For each point $j$ in the population                                |
|   | Update the new place based on the fitness function using Eq. (1)    |
|   | $S_{new} = S_{bst}$   |
|   | End for   |
| Searching in the selected space                           |   |
|   | Update the new place within the selected search space using Eq. (2) |
|   | $S_{new} = S_{bst}$   |
|   | End for   |
| Swooping stage  |   |
|   | Update the new place for hunting using Eq. (3)                      |
|   | $S_{new} = S_{bst}$   |
| End while   |   |
| Obtain best solution                                      |   |

#### IV. BRIEF ELUCIDATION OF DEEP LEARNING-BASED DATA DEDUPLICATION AND HASHING FUNCTION-BASED DATA SECURITY IN THE CLOUD SECTOR

##### A. Description of LSTM

LSTM is used in the proposed data deduplication model for checking duplicate files or chunks before storing them in the cloud. The network description of LSTM is given as follows.

LSTM [16] is mainly developed for solving the vanishing gradient issues, the input features are effectively learned using the memory cell and then the data is processed and stored in the long-term and short-term memory. The information entered and left in the memory cell is controlled by the gating functions in LSTM. Based on the hidden state value from the previous cell state, the forget gate resets the information that is not used in the current memory state. The new information to be included to the cell state is selected via the write gate. The information to be included in the cell state is manifested by the input gate at the time of writing. The reading process of information

is carried out by the output gate and it finally gives the hidden state vectors. The gate functions are described as follows.

$$F = \xi(W_{yF}y_r + W_{hF}h_{r-1} + W_{mF}m_{r-1} + B_F) \quad (4)$$

$$T = \tanh(W_{yT}y_r + W_{hT}h_{r-1} + W_{mT}m_{r-1} + B_T) \quad (5)$$

$$I = \xi(W_{yI}y_r + W_{hI}h_{r-1} + W_{mI}m_{r-1} + B_I) \quad (6)$$

$$O = \xi(W_{yO}y_r + W_{hO}h_{r-1} + W_{mO}m_{r-1} + B_O) \quad (7)$$

Here, the term  $m_{r-1}$  denotes the previous cell state memory, the input applied is indicated as  $y_r$ , and the value from the previous hidden state is represented as  $h_{r-1}$ . The “write gate, forget gate, output gate and input gate” are indicated by the terms  $T$ ,  $F$ ,  $O$  and  $I$ . The weight matrices and biases are signified by  $W$  and  $B$ , correspondingly. The activation function used is denoted by the term  $\xi$ .

For controlling the information, the gates play an important role and the expression for updating the new cell state is given in Eq. (8).

$$m_r = F * m_{r-1} + I * T \quad (8)$$

The activation function is updated in the output gate based on the output state, which is described in Eq. (9).

$$h_r = O * \tanh(m_r) \quad (9)$$

By forgetting the value of the previous cell state, a new cell state value is updated.

#### B. Data Deduplication using A-LSTM

The A-LSTM network is implemented in this proposed deduplication framework for checking the deduplicated data before storing it in the cloud. From the collected dataset, the input parameters such as file name, hashtag, file size, file location, block name, data pattern, last modified date and file size are given as the input of the A-LSTM network. The data deduplication procedure is carried out on the basis of the input attributes from the data. The labeled classes are classified using the A-LSTM model. The deduplicated features are detected by this A-LSTM and avoid these duplicate copies at the time of feature conversion. The exact matching data are removed before storing it in the cloud to decrease the storage requirements. Data deduplication using A-LSTM highly decreases the memory overhead and computational overhead while processing massive amounts of data. The memory bandwidth requirements are also reduced by using this developed A-LSTM. The data deduplication performance is further improved by optimizing the parameters such as “hidden neurons, learning rate, and epochs from LSTM” with the support of BES. The searching ability of this BES is high and hence better optimal solutions are obtained in the search space. Therefore, the hyperparameters are optimally tuned in better intervals to improve the data deduplication performance. The precision value gets maximized by this parameter optimization and the fitness function is given in Eq. (10).

$$Obj = \arg \min_{\{L_{x^*}^{LSTM}, H_{y^*}^{LSTM}, E_{z^*}^{LSTM}\}} \left( \frac{1}{Pr n} \right) \quad (10)$$

The fitness function of optimizing hyperparameters from LSTM is indicated as  $Obj$ , the optimally tuned learning rate in the interval between  $[0.01, 0.99]$  is signified as  $L_{x^*}^{LSTM}$ , the selected hidden neurons in the range between  $[5, 255]$  from LSTM is denoted as  $H_{y^*}^{LSTM}$  and the optimized epochs from LSTM in the interval of  $[5, 50]$  is indicated by  $E_{z^*}^{LSTM}$ . The matching performance for removing duplicated copies using A-LSTM is better and it is evaluated through the precision measure  $Pr n$ . It is calculated using the expression given in Eq. (11).

$$Pr n = \frac{X_{ps}}{X_{ps} + L_{ps}} \quad (11)$$

The false positive observation is signified as  $L_{ps}$  and the true positive observation is signified as  $X_{ps}$ . The block schematic representation of the data deduplication process using A-LSTM is given in Fig. 2.

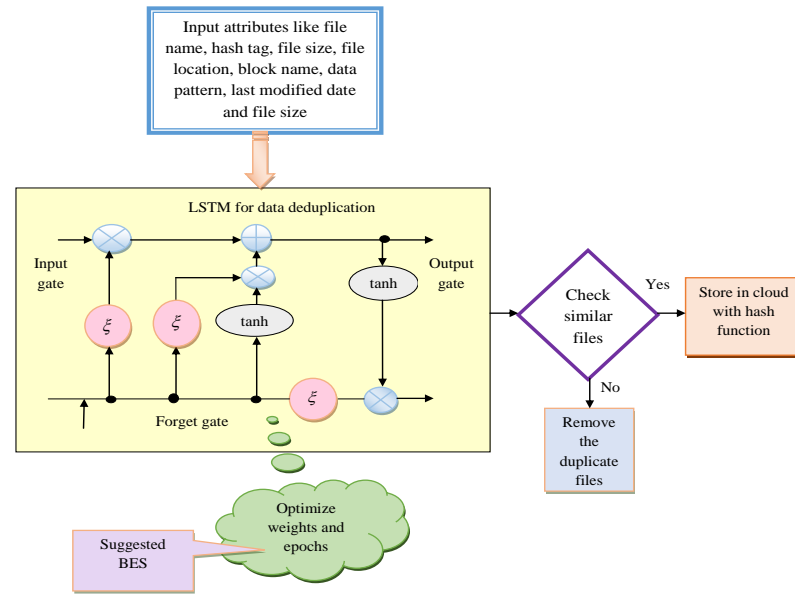


Fig. 2. Data deduplication in the cloud using A-LSTM

### C. Hashing Function-based Data Security

The duplicated data is effectively detected by the A-LSTM method, and the security of the data is to be ensured before storing it in the cloud. The confidentiality of the data is ensured on the client side using the hash function in this model. Hash function-based data storage protects the data from third parties to provide better integrity, security, and confidentiality for the ownership. The return values of the hash variables are identified through the hash codes, hash sums, and hash values. The isolation of data is made possible by these hash values, which is helpful for storing data with reduced computational overhead and the user wants to access the data, it easily provides the required data to the users without any complexities. Several key values are used for protecting the data from leakage and it provides higher security over the server side and client side. Data owners use the key values when downloading data, in this way, the data has been protected from third parties. To provide additional security, encryption algorithms are suggested before storing the information in the cloud environment. The overall processing steps of the proposed model are described as below.

- The data owner wants to store the file in the cloud.
- Data deduplication is performed using A-LSTM to detect duplicate copies to decrease the storage needs in the cloud.
- The hash function is applied over the detected deduplicated data; it removes the duplicate copies and stores the new data with some hash values. The use of hash values is helpful in identifying the files very easily for downloading the user and it provides higher security over the data. It avoids data leakage and protects the data from third parties.
- The detected duplicate content of the data is not saved in the cloud. The original information is accessed by the user with the owner's permission.
- Therefore, the strong trust between the cloud service provider and the cloud user is retained by this approach.

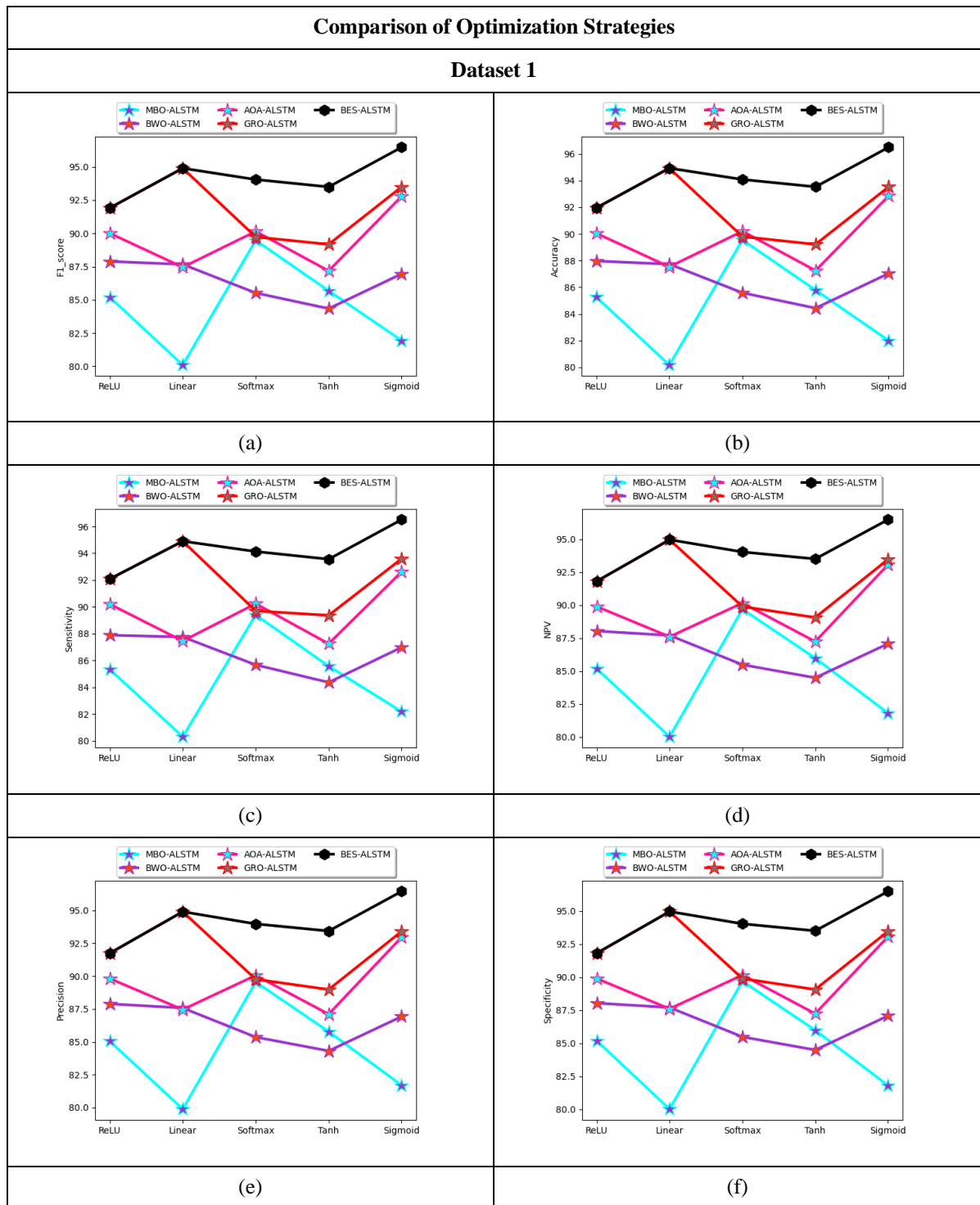
## V. RESULTS AND DISCUSSION

### A. Implementation Setup

Python software has been used for validating the performance of the proposed data deduplication scheme, and the experimental evaluation has been done to show the deduplication performance in the cloud by assuming the "number of population, chromosome length and maximum iteration count as 10, 4 and 50". The algorithms like "Migrating Birds Optimization (MBO) [17], Beluga Whale Optimization (BWO) [18], Arithmetic Optimization Algorithm (AOA) [19], and Gold Rush Optimization (GRO) [20]" and previously developed techniques like CNN [12], Convolutional Autoencoder [21], Gated Recurrent Unit [22], and Attention-based LSTM [23] were considered for the evaluation of duplication efficacy.

**B. Performance Analysis Using Positive Measures**

The positive measures used for validating the data deduplication performance among conventional optimistic strategies are provided in Fig. 3 and previously suggested techniques, which are shown in Fig. 4. The activation functions are varied for analyzing the performance because the effectiveness is clearly identified through the variation in activation functions. The accuracy rate of the recommended deduplication detection framework is approximately 97.2%, which is greater when compared to other optimization algorithms and techniques for considering the sigmoid activation function in dataset 1. The precision and f1-score of the presented BES-ALSTM model is above 95%, which demonstrates the similar file detection effectiveness of our proposed framework is more impressive than previous methods.



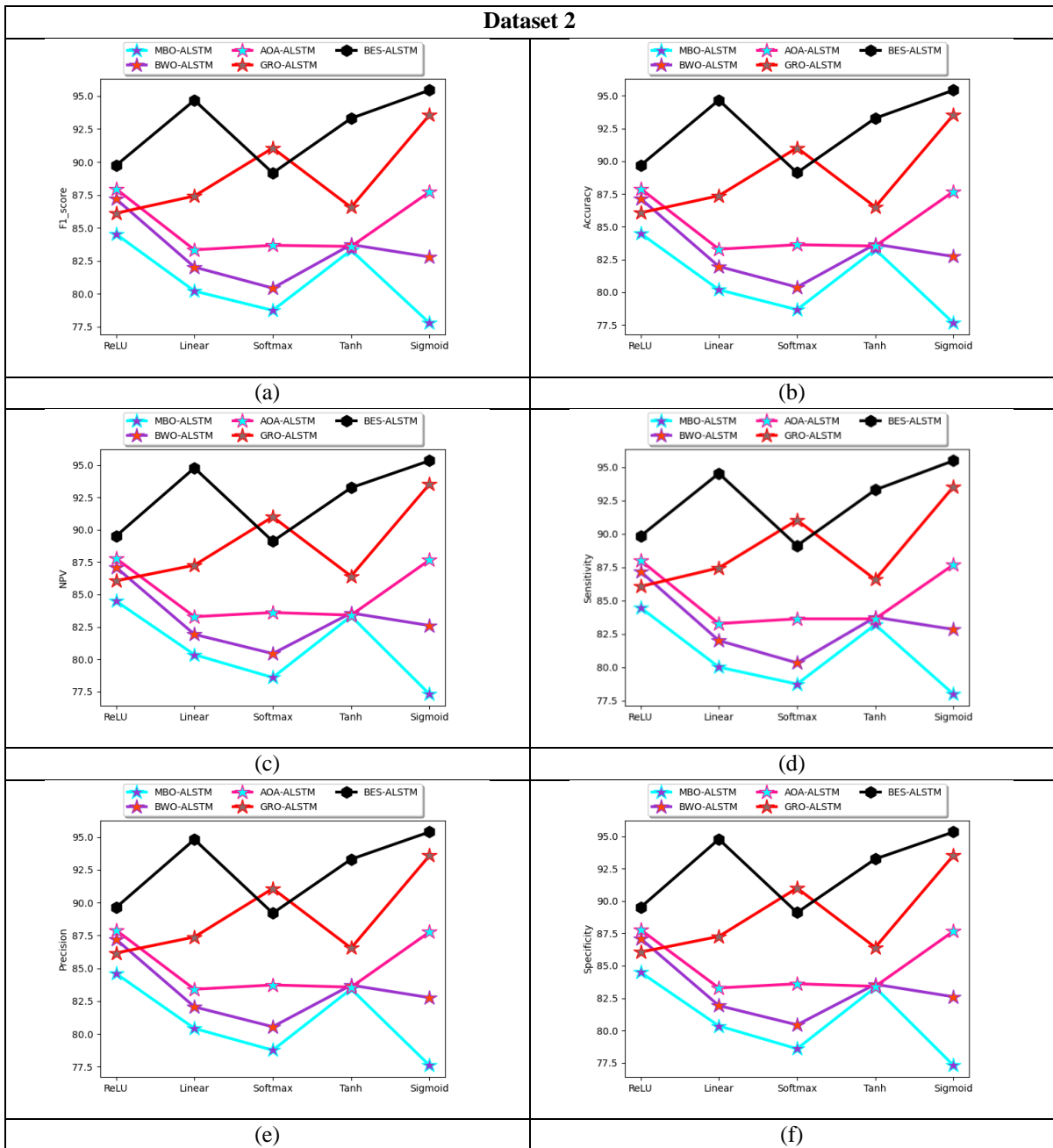
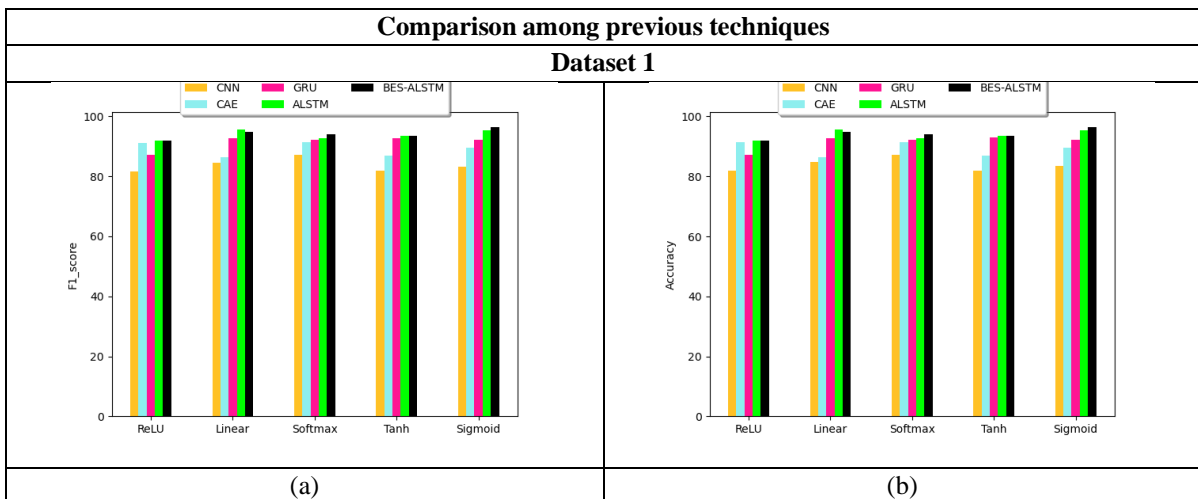
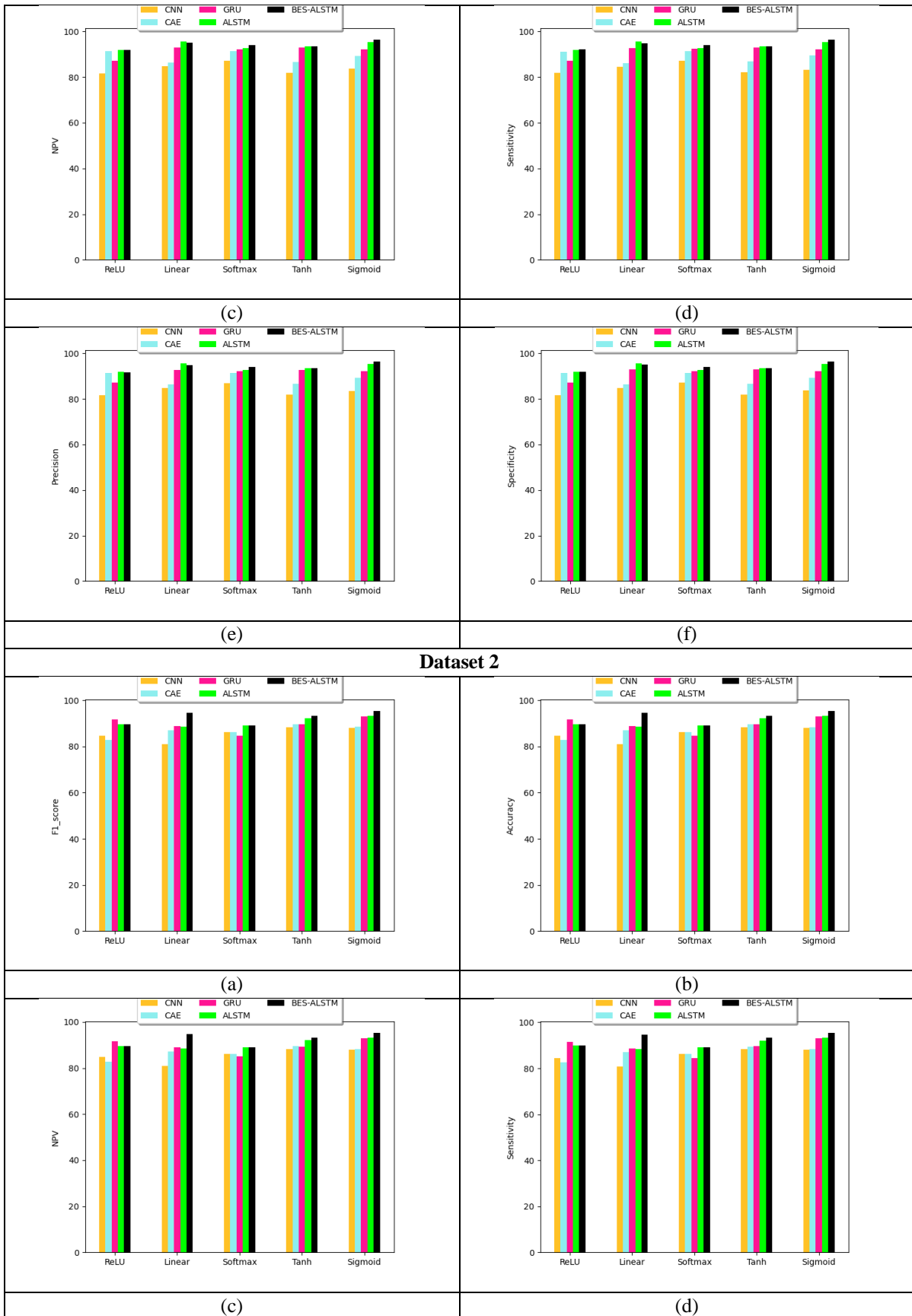


Fig. 3. Data deduplication efficiency of the proposed framework when compared to optimization methods regarding (a) F1-score (b) Accuracy (c) NPV (d) Sensitivity (e) Precision (f) Specificity





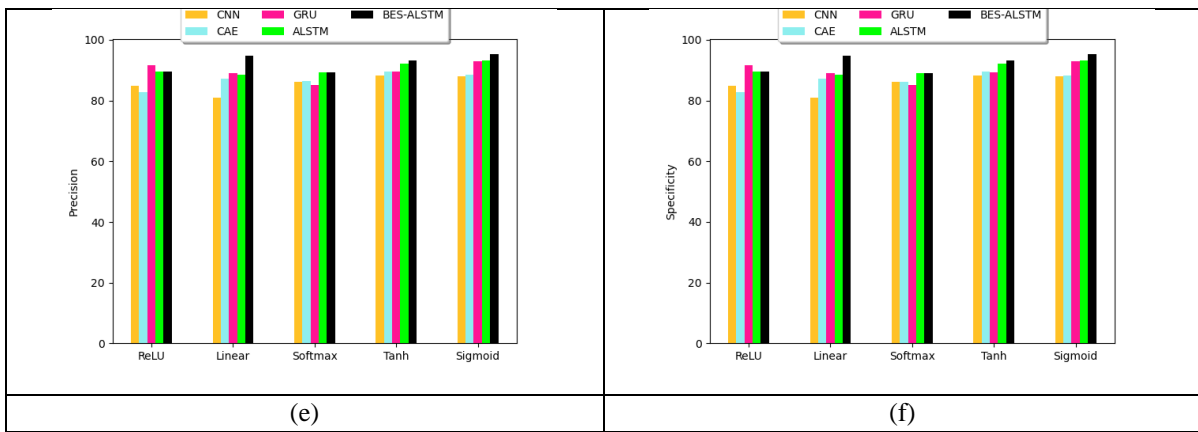


Fig. 4. Data deduplication efficiency the recommended framework over various techniques in terms of (a) F1-score (b) Accuracy (c) NPV (d) Sensitivity (e) Precision (f) Specificity

C. Convergence and ROC Validation

The convergence and ROC validation of the offered data deduplication framework is depicted in Fig. 5. This analysis is done over two datasets, convergence ability is analyzed by varying the number of iterations and the ROC validation is done according to true and false positive measures. Convergence examination is done over optimization algorithms and ROC validation is done over previous deduplication methods. The convergence plots clearly show the optimized deep learning network provides a better matching ability over huge amounts of data. The data deduplication detection efficiency is better than the previous model that is clearly shown from the ROC analysis.

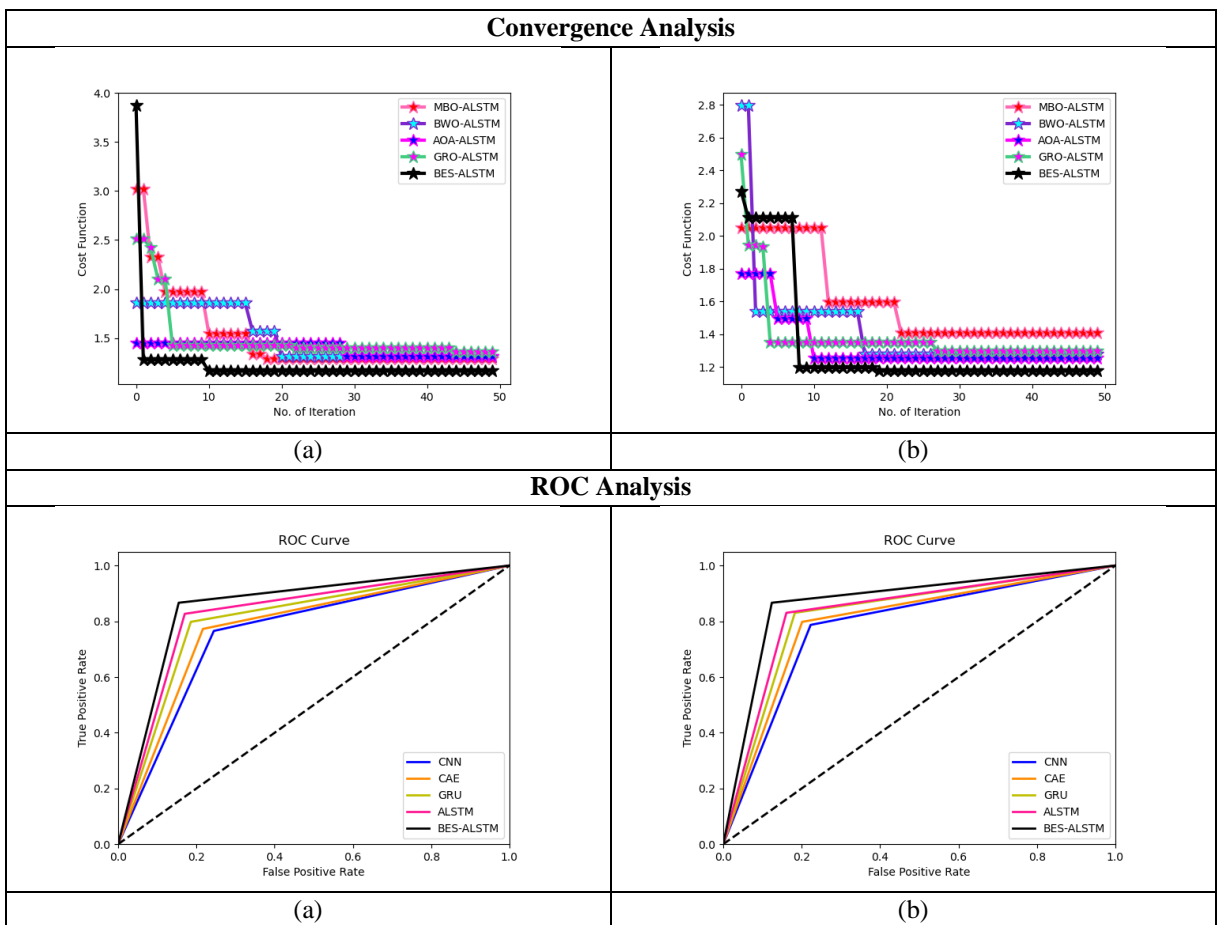


Fig. 5. Convergence and ROC performance of the suggested framework when compared to optimization methods for (a) Dataset 1 (b) Dataset 2

D. Numerical Validation over Negative Measures

The negative measures analysis of the proposed deep learning-assisted data deduplication framework is provided in Table II. The false negative and positive values show that the error rate of our proposed approach is highly minimized than other models. The reduced error rate shows that the deduplicated data detection effectiveness of the

presented model is greater than other heuristic strategies and previously developed methods. The statistical validation of the implemented scheme is provided in Table III. Various heuristic methods are suggested for analyzing the statistical measures, where the minimized value of statistical measures proves that the optimization performance of the developed scheme is high hence it provides better data deduplication performance than others.

**Table.2.** Numerical analysis of developed data deduplication model using Hash function.

| <b>Comparison of Optimization Algorithms</b> |                |                |                |                |           |
|--|----------------|----------------|----------------|----------------|-----------|
| Term   | MBO-ALSTM [17] | BWO-ALSTM [18] | AOA-ALSTM [19] | GRO-ALSTM [20] | BES-ALSTM |
| Dataset 1                                    |                |                |                |                |           |
| FPR  | 18.18483       | 12.91818       | 6.922822       | 6.52534        | 3.511096  |
| FNR  | 17.81281       | 13.01577       | 7.380074       | 6.407246       | 3.488762  |
| FDR  | 18.3061        | 13.07409       | 7.037037       | 6.595246       | 3.55347   |
| Dataset 2                                    |                |                |                |                |           |
| FPR  | 22.69988       | 17.39654       | 12.33427       | 6.468461       | 4.660506  |
| FNR  | 22.0231        | 17.16448       | 12.30585       | 6.451613       | 4.500199  |
| FDR  | 22.39398       | 17.2304        | 12.23595       | 6.414343       | 4.614161  |
| <b>Comparison of Previous Techniques</b>     |                |                |                |                |           |
| Terms  | CNN [12]       | CAE [21]       | GRU [22]       | ALSTM [23]     | BES-ALSTM |
| Dataset 1                                    |                |                |                |                |           |
| FPR  | 10.66578       | 7.784034       | 4.57105        | 3.511096       | 16.36303  |
| FNR  | 10.36565       | 7.849715       | 4.562227       | 3.488762       | 16.83999  |
| FDR  | 10.75484       | 7.880617       | 4.626215       | 3.55347        | 16.61621  |
| Dataset 2                                    |                |                |                |                |           |
| FPR  | 11.97268       | 11.65127       | 7.151466       | 6.870229       | 4.660506  |
| FNR  | 11.90761       | 11.54918       | 6.889685       | 6.770211       | 4.500199  |
| FDR  | 11.87251       | 11.54918       | 7.074722       | 6.807325       | 4.614161  |

**Table.3.** Statistical analysis of developed data deduplication model using Hash function

| Terms              | MBO-ALSTM [17] | BWO-ALSTM [18] | AOA-ALSTM [19] | GRO-ALSTM [20] | BES-ALSTM |
|--------------------|----------------|----------------|----------------|----------------|-----------|
| Dataset 1          |                |                |                |                |           |
| Median             | 1.291002       | 1.308576       | 1.447223       | 1.396947       | 1.163595  |
| Best               | 1.291002       | 1.308576       | 1.309974       | 1.362429       | 1.163595  |
| Standard deviation | 0.418984       | 0.25111        | 0.06774        | 0.284949       | 0.378328  |
| Mean               | 1.515644       | 1.505048       | 1.389578       | 1.496611       | 1.238196  |
| Worst              | 3.022512       | 1.856512       | 1.447223       | 2.511806       | 3.868808  |
| Dataset 2          |                |                |                |                |           |
| Best               | 1.407753       | 1.275603       | 1.255245       | 1.297731       | 1.174578  |
| Median             | 1.407753       | 1.275603       | 1.255245       | 1.352228       | 1.174578  |
| Standard deviation | 0.261809       | 0.305441       | 0.163276       | 0.21682        | 0.349192  |
| Worst              | 2.046492       | 2.793803       | 1.771056       | 2.494707       | 2.2717    |
| Mean               | 1.59865        | 1.414588       | 1.330843       | 1.385186       | 1.33242   |

## VI. CONCLUSION

A new data deduplication model in the cloud has been implemented using deep learning with a hash function to reduce the storage and memory requirements. The cost requirements for storing the data are decreased by the removal of similar files saved in the cloud. The input attributes related to the collected data were given as input to A-LSTM; it matches the similar features of the data already saved in the cloud to find the deduplicated data. The data deduplication performance has been maximized by optimizing the attributes from LSTM. The detected duplicate files were removed and new data were saved in the cloud with a hash function. The ownership permission for accessing the data has provided higher security and confidentiality over the stored data. The data deduplication effectiveness of the presented model was ensured with conventional frameworks to prove the detection efficiency, and the results proved that the precision rate is 95.8% for considering dataset 1 at the sigmoid activation function. The data deduplication performance is greatly enhanced with reduced error by the proposed model and it ensures data security. In future work, advanced data encryption algorithms will be recommended for providing additional security to the data to protect it from malicious activities.

## REFERENCES

- [1] C. Hema & Fausto Pedro Garcia Marquez, "Storage Enhancement in the Cloud Using Machine Learning Technique and Novel Hash Algorithm for Cloud Data Security," *Proceedings of the Sixteenth International Conference on Management Science and Engineering Management*, Vol. 1, 2022.
- [2] Widad Elouataoui, Imane El Alaoui, Saida El Mendili, and Youssef Gahi, "An End-to-End Big Data Deduplication Framework based on Online Continuous Learning," *International Journal of Advanced Computer Science and Applications*, Vol. 13, No. 9, 2022.
- [3] S. Li, C. Xu, Y. Zhang, Y. Du and K. Chen, "Blockchain-Based Transparent Integrity Auditing and Encrypted Deduplication for Cloud Storage," in *IEEE Transactions on Services Computing*, vol. 16, no. 1, pp. 134-146, 1 Jan.-Feb. 2023.
- [4] Amdewar Godavari, Chapram Sudhakar, T. Ramesh, "Hybrid deduplication system with content-based cache for cloud environment," *Journal of King Saud University - Computer and Information Sciences*, Vol. 36, Issue 5, pp. 102030, June 2024.
- [5] Mohd Akbar, Irshad Ahmad, Mohsina Mirza, Manavver Ali & Praveen Barmavatu, "Enhanced authentication for deduplication of big data on cloud storage system using machine learning approach," *Cluster Computing*, 2023.
- [6] D. Viji and S. Revathy, "Hash-Indexing Block-Based Deduplication Algorithm for Reducing Storage in the Cloud," *Computer Systems Science and Engineering*, Vol. 48, Number 1, 2024.
- [7] M. Song, Z. Hua, Y. Zheng, H. Huang and X. Jia, "Blockchain-Based Deduplication and Integrity Auditing Over Encrypted Cloud Storage," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 6, pp. 4928-4945, Nov.-Dec. 2023.
- [8] Rupali S. Patil, Amina Kotwal, Swati S. Patil, "Efficient Iot-Based Cloud Computing Framework For Secure Data Storage Using Machine Learning Algorithm," *Journal of Theoretical and Applied Information Technology*, Vol.101, 31st May 2023.
- [9] Ch. Prathima, Naresh Babu Muppalaneni & K. G. Kharade, "Deduplication of IoT Data in Cloud Storage," *Machine Learning and Internet of Things for Societal Issues*, pp. 147-157, 2022.
- [10] X. Yu, H. Bai, Z. Yan and R. Zhang, "VeriDedup: A Verifiable Cloud Data Deduplication Scheme With Integrity and Duplication Proof," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 1, pp. 680-694, 1 Jan.-Feb. 2023.
- [11] Kiana Ghassabi, and Peyman Pahlevani, "DEDUCT: A Secure Deduplication of Textual Data in Cloud Environments," *IEEE Access*, PP.99, 2024.
- [12] J. K. Periasamy, L. Selvam, M. Anuradha and R. Kennady, "A fuzzy optimal lightweight convolutional neural network for deduplication detection in cloud server," *Iranian Journal of Fuzzy Systems*, Vol. 21, Number 1, pp. 33-49, 2024.
- [13] Xuewei Ma, Wenyuan Yang, Yuesheng Zhu, Zhiqiang Bai, "A Secure and Efficient Data Deduplication Scheme with Dynamic Ownership Management in Cloud Computing," *arXiv:2208.09030v1 [cs.CR]*, 18 Aug 2022.
- [14] Hesham Abusaimh, Omar Isaid, "Hybrid Data Deduplication Technique in Cloud Computing for Cloud Storage," *Journal of Theoretical and Applied Information Technology*, Vol.95. No 24, 2017.
- [15] Jay Davea, Prithvi Hegde, Hitaishi Desai, Anshul Kanodia, Raj Srivastava, Kushagra Singh, "RESIST : Randomized Encryption for Deduplicated Cloud Storage System," *Computer science and engineering*, 2024.
- [16] Mudzfirah Abdul Halim, Azizi Abdullah, Khairul Akram Zainol Ariffin, "Recurrent Neural Network for Malware Detection," *Int. J. Advance Soft Compu. Appl*, Vol.11, No. 1, March 2019.
- [17] Erkan Ulker, Vahit Tongur, "Migrating birds optimization (MBO) algorithm to solve knapsack problem," *Procedia Computer Science*, Vol. 111, pp. 71-76, 2017.
- [18] Changting Zhong, Gang Li, Zeng Meng, "Beluga whale optimization: A novel nature-inspired metaheuristic algorithm," *Knowledge-Based Systems*, Vol. 251, pp. 109215, 5 September 2022.
- [19] Laith Abualigah, Ali Diabat, Seyedali Mirjalili, Mohamed Abd Elaziz, Amir H. Gandomi, "The Arithmetic Optimization Algorithm," *Computer Methods in Applied Mechanics and Engineering*, Vol. 376, pp. 113609, 1 April 2021.
- [20] Kamran Zolfi, "Gold rush optimizer: A new population-based metaheuristic algorithm," *Operations Research and Decisions*, Wroclaw University of Science and Technology, Faculty of Management, vol. 33(1), pp. 113-150, 2023.
- [21] Qi, Hui, et al. "Secure Deduplication Method Based on Autoencoder." *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking*, IEEE, 2020.

- [22] N. Mageshkumar, L. Lakshmanan, "Intelligent data deduplication with Deep Transfer Learning Enabled Classification Model for Cloud-based Healthcare System," *Expert Systems with Applications*, Vol. 215, 1 vol. 119257, April 2023.
- [23] Anesa Al-Najeh, Sana Abouljam, "PCF-LSTM for Dynamic Data Portability with CSHHC-Based Deduplication in Distributed Cloud," *International Research Journal of Advanced Engineering and Science*, Vol. 9, Issue 1, pp. 61-66, 2024.