Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

# Generalized Canonical Discriminant Congruential Signcryption for Secure Communication with Big Data

[1]S. Sangeetha, [2]P. Suresh Babu
[1]Research Scholar, [2] Associate Professor
[1,2] Department of Computer Science, Bharathidasan College of Arts and Science, India
E-Mail: [1]sangee.siva2011@gmail.com [2]ptsuresh77@gmail.com

**Article Info**
**Page Number: 1540-1556**
**Publication Issue:**
**Vol. 71 No. 4 (2022)**

***Abstract***

Big data is a collection of huge data employed to examine and extract information from large datasets. With the generation of a large volume of data, it faces severe security risks and challenges such as data leakage, malicious use, etc. Many researchers carried out their research for performing secured data communication. But, the data confidentiality level was not improved by using existing cryptographic methods. Therefore, a generalized canonical linear discriminant-based Multiplicative congruential signcryption (GCLD-MCS) technique is introduced for secured data communication with higher data confidentiality and lesser communication overhead.

The GCLD-MCS technique performs two processes, namely data classification and secured communication. Initially, a number of data are collected from the big dataset. After that, generalized canonical statistic distributive linear discriminant analysis is carried out to examine the linear combination of data and to classify into a number of classes. After data classification, secured data transmission is performed using Multiplicative congruential Rabin cryptographic signcryption technique. Multiplicative congruential Rabin cryptographic signcryption includes three processes namely Multiplicative congruential Key Generation, Signcryption and Unsigncryption. Experimental evaluation of the proposed GCLD-MCS technique is carried out with respect to classification accuracy, data confidentiality level, data integrity, and communication overhead, with a different number of data. The discussed results indicates that the performance of GCLD-MCS technique increases communication security with higher data confidentiality rate, integrity, and minimum overhead than the other state-of-the-art methods.

**Keywords**:- Secured data communication, Generalized canonical statistic distributive linear discriminant analysis, Multiplicative congruential Rabin cryptographic signcryption.

**Article History**
**Article Received: 25** March 2022
**Revised: 30** April 2022
**Accepted: 15** June 2022
**Publication: 19** August 2022

## 1. Introduction

Nowadays, security and privacy issues of big data systems are noteworthy and require precise observation. However, the large volume of big data is susceptible to attacks in modern communication systems. Big data security events occur recurrently in recent years. Hence, Privacy-preserving techniques have been developed to guarantee the authorized individual to access the private information for enhancing the data security in communication systems.

A novel security scheme known as Lightweight Hybrid Scheme (LHS) was introduced in [1] based on Elliptic Curve Cryptography and provides the secure exchange mechanism. However, it failed to define a novel scheme for big data encryption and decryption that leads to a higher level of security and withstand attacks existence during the communication. The blockchain-Based Trusted Service Evaluation (BCSE) model was designed in [2] over Big Data. But, it failed to ensure the stability of the performance, especially in a large and increasing number of data samples.

In [3], different methods of data mining algorithms were considered to ensure security. However, the efficient machine learning method was not applied for minimizing the overhead of secure data transmission. A big data provenance model (BDPM) was developed in [4] to provide the entire data transformation through dissimilar components of a big data

system. But the performance of data confidentiality and integrity was not analyzed. A trusted third-party-aided searchable and verifiable data protection approach was introduced in [5] to enhance the integrity of uploaded or downloaded data at any time. But it failed to use the secure data sharing scheme with big data applications.

A Lossless Computational Secret Image Sharing (CSIS) technique was developed in [6] to convert the encrypted secret data against attacks. But the designed technique has morecomputational complexity and time requirement for secret reconstruction. A novel deep learning model and a secure inferencing approach were developed in [7] for protecting the privacy of the data. However, it failed to reduce communication costs. A new privacy-preserving and communication-efficient algorithm (DPCrowd) was designed in [8] for sharing differential privacy parameters. But the communication overhead was not minimized.

A generic Privacy-Preserving Auction Scheme (PPAS) was introduced in [9] for Big Data sharing Using Homomorphic Encryption. But it failed to integrate with other efficient cryptosystems to improve the performance of the security. A cloud-enabled IoT environment maintained by multifactor authentication and a lightweight cryptography approach was developed in [10] to protect the big data system. But it failed to attack detection during the big data communication in cloud servers.

### 1.1 Contributions

In order to overcome the existing issues, a novel GCLD-MCS technique is introduced with the following novel contributions

- To improve the security of data communication, a novel GCLD-MCS technique is introduced based on two major processes namely generalized canonical statistic distributive linear discriminant analysis and Multiplicative congruential Rabin cryptographic signcryption.
- First, the generalized canonical statistic distributive linear discriminant analysis is employed in the GCLD-MCS technique to change the unstructured dataset into a structured data format through the classification process. The generalized canonical correlation classifies the data into different classes with higher accuracy. Hotelling t-squared statistic distribution is also applied for minimizing the variance within the class and maximizing the variance between the classes. This process minimizes the communication overhead of secure data transmission.
- To increase data confidentiality and integrity, a Multiplicative congruential Rabin cryptographic signcryption is applied in the GCLD-MCS technique. The proposed signcryption technique performs Multiplicative congruential key generation, signcryption, and unsigncryption to improve the secure data communication process by avoiding the attacks.
- Finally, comprehensive experiment evaluations are carried out to estimate the performance of our GCLD-MCS technique and other cryptographic techniques along with the various metrics.

### 1.1 Organization of Paper

The rest of the paper is organized into five different sections as follows. Section 2 reviews the related works. Section 3 provides a brief description of the proposed GCLD-MCS technique with a neat architecture diagram. Section 4 describes the experimentation with the dataset description. In section 5, the performance results of the proposed technique and existing methods are discussed with different metrics. At last, Section 6 concludes the paper.

## 2. Related Works

Homomorphic encryption algorithms were designed in [11] for big data security analysis while preserving privacy. But the higher data confidentiality was not achieved. Attribute-Based Honey Encryption (ABHE) scheme was introduced in [12] for performing

encryption-decryption on different sizes of files. This encryption scheme failed to use more robust approaches to ensure optimum security of big data. A novel blockchain technology was introduced in [13] for big data security solutions by integrating fragmentation, encryption, and access control. But the higher data confidentiality level was not achieved.

Blockchain technology was introduced in [14] to handle the healthcare system by providing the solution. The designed scheme guaranteed security with a smart contract. But, the overhead was not reduced. An effective Secure Information Propagation (SIP) approach was developed in [15] for E-health Networks. The designed security analysis was not feasible to find that illegitimate user. A blockchain integrated IPFS (interplanetary file system) was developed in [16] to guarantee secure and privacy-aware E-Health verification. However, the designed system failed to consider the cryptographic technique to improve security and privacy.

A privacy-preserving approach using Blockchain technology was introduced in [17] that maintain the security, privacy, and integrity of the e-health data. But the overhead of e-health data distribution was not reduced. A blockchain-assisted secure system using Lamport Merkle Digital Signature (LMDS) was developed in [18] for medical IoT data transmission. The designed system increases data confidentiality but the overhead of secure communication was not minimized. An efficient Lightweight integrated Blockchain (ELIB) technique was introduced in [19] to improve the security as well as privacy of data communication. The designed technique minimizes the time but performance of confidentiality and integrity analysis were not performed. A Secure Communication and Authentication for IoT applications using blockchain was introduced in [20]. The designed method minimizes the communication cost but the higher data confidentiality rate was not achieved.

## 3. Proposed Methodology

Big Data is a vast term for implementation with the huge volume and complex data sets. While the data set is bigger in volume and the conventional methods are insufficient to extract the important and accurate data from large unstructured data. With the help of classification methods, unstructured data is turned into structured form as a result that any user accesses the required data easily. These classification techniques are applied over the big databases to offer data services from the huge volume of data sets. In addition, big data security is the major concern to perform the data point communication. A traditional security technique provides slow performance and is time-consuming in a Big Data context. Based on the above motivation, a novel technique called
GCLD-MCS is introduced.

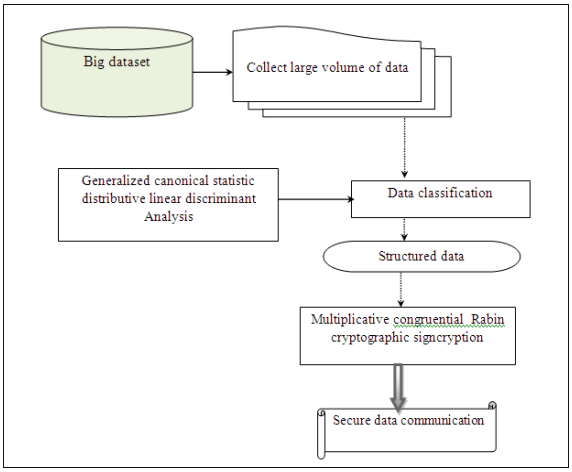


**Figure 1 Architecture of the Proposed GCLD-MCS Technique**

The proposed GCLD-MCS technique performs two major processes big data classification and secure communication. The classification process of the proposed GCLD-MCS technique uses the machine learning technique called generalized canonical statistic distributive linear discriminant Analysis. Then the security of communication is achieved by applying Multiplicative congruential Rabin cryptographic signcryption. With the use of the above-said two techniques, the architecture of the proposed GCLD-MCS technique is constructed as given above.

Figure 1 demonstrates the structural representation of the proposed GCLD-MCS technique that consists of two different main processes namely classification and secure data transmission with big data. Initially, the big dataset is considered. The proposed GCLD-MCS technique first performs the classification of the data using Fisher's linear discriminant analysis. After classifying the data, secure data transmission from sender to receiver is said to be performed using Multiplicative congruential Rabin cryptographic signcryption. These processes of the proposed GCLD-MCS technique are briefly described in the following subsections.

## 3.1 GENERALIZED CANONICAL STATISTIC DISTRIBUTIVE LINEAR DISCRIMINANT DATA CLASSIFICATION

The first process of the proposed GCLD-MCS technique is to perform the data classification using generalized canonical statistic distributive linear discriminant analysis. The proposed linear discriminant analysis is a machine learning technique that helps to find the linear combination of data that distinguishes two or more classes using generalized canonical correlation and Hotelling t-squared statistic distribution. This helps to improve the data classification. Applying the Hotelling t-squared statistic distribution to the Hotelling t-squared statistic distribution is mainly used for testing the differences in means for the multivariate data. Here the multivariate data involves more than two dependent variables (i.e. features), resulting in a single result is determined in a precise manner and the error rate being well controlled.

Let us consider the number of data in the big dataset,

$$D_i = \{D_1, D_2, D_3, \dots D_n\} \in DS \quad (1)$$

In (1),$D_i$ denotes a set of mean in the big dataset $DS$. A number of classes are initialized. The mean value of the class is calculated as given below,

$$m_c = \frac{1}{n} \sum_{i=1}^{n} D_i \ (2)$$

From (2), $m_c$ denotes a mean value of class and $n$ denotes the number of data. Therefore, the relationship between the mean and data is measured by applying a generalized canonical correlation as given below,

$$G_c = Cov\ (D_i, m_j) \qquad (3)$$

From the above mathematical equation (3), the covariance '$Cov$' is calculated by using data '$D_i$' and mean value '$m$'.

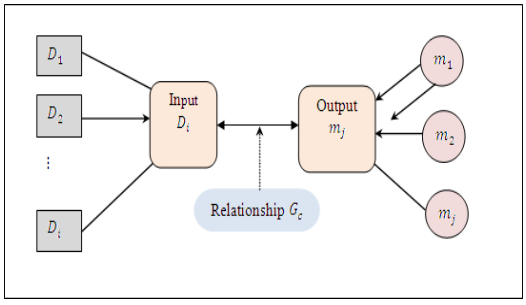


**Figure 2 Generalized Canonical Correlation**

The covariance between data and mean is mathematically determined as below,

$$Cov\ (D_i, m) = \frac{\sum(D_i - m_D)(C_j - m_c)}{n} \quad (4)$$

From the above mathematical formula (4), '$m_D$' signifies the mean of data '$D_i$' and '$m_c$' denotes the mean of the class '$C_j$'. By using (4) relationship between the data and output class is measured. Based on correlation measures, the data is classified into a particular class. Therefore, this process maximizes the variance between the classes and minimizes the variance within the class using Hotelling t-squared statistic distribution. The variance between and within the class is measured as follows,

$$Cov\ (w) = \sum\sum(D - m_c)(D - m_c)' \quad (5)$$

From (5),$Cov\ (w)$denotes a variance within the class and '$D$ denotes a data, $m_c$ represents a mean of the class. $(D - m_c)'$ denotes a transpose of a matrix. Similarly, the variance between the classes is formulated as follows,

$$Cov\ (b) = \sum\sum(m_i - m_j)(m_i - m_j)' \quad (6)$$

From (6),$Cov\ (b)$ $d$enotes a variance between the classes, '$m_i$' $m_j$ denotes a mean of the two classes. Therefore, the proposed technique maximizes the variance between the classes and minimizes the variance within the classes.

$$\arg\min Cov\ (w) \quad (7)$$

$$arg\ \max\ Cov\ (b) \quad (8)$$

From (7), (8), arg min and arg max function denote an argument of minimum and maximum function to minimize the variance within the class and maximize between the classes. In this way, the data is classified into different classes.

**Algorithm 1: Generalized canonical statistic distributive linear discriminant data classification**

**Input:** Big dataset, Number of data $D_i = D_1, D_2, D_3 \ldots D_n$
**Output:** Improve classification accuracy

**Begin**

1. **Define the number of classes** $C_j = C_1, C_2, \ldots C_m$
2. **Calculate the mean of classes '$m_c$'**
3. **For each data** $D_i$ in dataset
4. **For each** mean '$m_c$'
5. Measure correlation '$G_c$'
6. **end for**
7. **end for**
8. Categorizes the data into a particular class
9. Measure variation within class '$Cov\ (w)$'
10. Measure variation between class '$Cov\ (b)$'
11. Obtain ( classification results)

**End**

Algorithm 1 given above describes the step-by-step process of big data classification using a generalized canonical statistic distributive linear discriminant classifier. First, initializes the number of classes. Then the mean is estimated based on the number of data. Then the correlation between the mean of class and the data is measured. Based on the correlation measure, the data is classified correctly into a particular class. The result indicates that the variance within the class is minimized and maximizes the variance between the classes.

### 3.2 Multiplicative Congruential Rabin Cryptographic Signcryption

After the data classification, the proposed GCLD-MCS technique performs the secure data transmission from sender to receiver using a multiplicative congruential Rabin cryptographic signcryption scheme. The Multiplicative congruential Rabin cryptographic signcryption scheme consists of three major processes namely Key Generation, Signcryption, and Unsigncryption to increase the performance of secure data transmission.

- **Mltiplicative congruential Key Generation**

The proposed technique initially performs the key pair generation such as private and public keys. In a public key cryptographic, the multiplicative congruential Key Generation is used for the particular communication session. These keys are disabled after the session is completed.

This helps to avoid unauthorized access hence it improves the confidentiality level. The private key is kept secret whereas the public key is distributed for further processing. By applying the multiplicative congruential linear pseudorandom number generator, the amount of memory available is often strictly limited.

Let us consider the multiplicative linear congruential generator for randomly generating the two prime numbers $g$ and $h$

$$g = z_1.R_1 \ mod \ m_1 \quad (9)$$
$$h = z_2.R_2 \ mod \ m_2 \quad (10)$$

Where, $m_1$, $m_2$ are the prime numbers, $z_1$, $z_2$ denotes a multiplier is an element of primitive root modulo $m_1$, $m_2$. From (9) (10), $R_1$ $R_2$ are coprime to $m_1$, $m_2$. Therefore, the public key is generated as given below,

$$Q = [g * h] \quad (11)$$
$$P = [g, h] \quad (12)$$

Where , $Q$ indicates a public session key, '$P$' denotes a private key. In this way, the pair of session keys are generated.

- **Signcryption**

Once the pair of keys is generated, the proposed GCLD-MCS technique simultaneously performs encryption GCLD-MCS technique and digital signature generation.

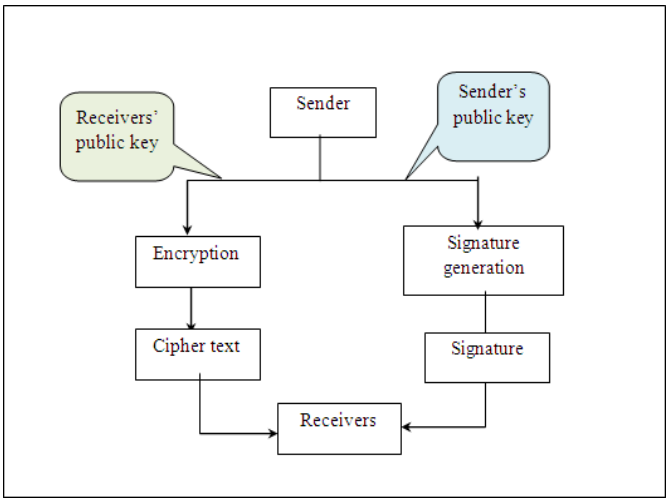


**Figure 3 Flow Process of Signcryption**

Figure 3 illustrates a block diagram of the flow process of signcryption to improve the security of data transactions. The signcryption consists of two major processes namely encryption and signature generation.

First, Encryption is performed to convert the input data i.e. plain text into unreadable form (i.e. ciphertext). Let us consider the classified data $CD_1$, $CD_2$, $CD_3$, ... . $CD_n$. The encryption process is performed as follows.
Let us consider the data $CD$ as plain text. The ciphertext of the input data is obtained as,

$$K \leftarrow CD^2 \ mod \ Q_r \quad (13)$$

Where $K$ denotes a ciphertext of original data '$CD$', $Q_r$ is a session public key of the receiver. The signature generation is performed simultaneously with the sender's private key. A valid digital signature is used for verifying the authenticity of original data and did not alter by any intruders. Therefore, the digital signature generation is performed with the secret private key. This digital signature is generated in the form of a hash value. A hash value is any function that is used to map the data of random size into data of fixed length. Let us consider the random number '$R$', the signature is generated as follows,

$$\beta_s = h\,(CD \| R) \quad (14)$$

From (14), '$\beta_s$' denotes a signature generated by the sender, (||) denotes a concatenation, $h$ is the cryptographic hash function. The encrypted data and signature is sent to the receiver.

- **Unsigncryption**

Finally, the proposed technique performs the unsigncryption that comprises signature verification and decryption at the receiver end.

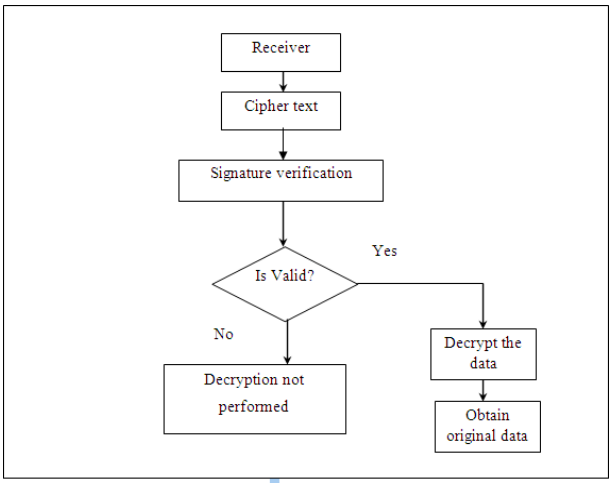


**Figure 4 Flow Process of Unsigncryption**

Figure 4 illustrates the flow process of unsigncryption to obtain original data. The ciphertext is given as input to the unsigncyption process. First, the signature is verified with the public key of the sender. First, compute,

$$\beta_s' = h\,(CD \| R) \qquad (15)$$

Where,$\beta_s' =$ indicates a signature at the receiver. Finally, verifies that the generated signature is $\beta_s' =$ matched with a signature generated at the sender '$\beta_s$'. When both the signature gets matched, it is said to be valid and then the receiver is said to be authorized. Then the authorized user received the original data. Otherwise, the signature is invalid and the receiver

is said to be the unauthorized user (i.e. attack) and did not get the original data. The decryption process is obtained as given below,

$$CD = (u_g.g.b_h - u_h.h.b_g) \; mod \; Q \qquad (16)$$

Where, $b_g = K_4^{\frac{1}{(g+1)}} \; mod \; g$, $b_h = K_4^{\frac{1}{(h+1)}} \; mod \; h$, $u_g.g + u_h.h = 1$

From (16),$CD$ denotes original data. In this way, the security of data communication from the sender to receiver is performed to improve data confidentiality. The algorithmic process of secure data transmission is described as follows

**Algorithm 2: Multiplicative congruential Rabin cryptographic signcryption**

**Input: D**ataset**,** Number of classified data $CD_1, CD_2, CD_3, \ldots . CD_n$
**Output:** increase the security of data transmission

**Begin**
**// Multiplicative congruential Session key pair generation**
**1. for data transmission**
**2. Generates the pair of key** $(Q, P)$ at a particular session
**3. end for**
**// Signcryption**
**4. For each data** $CD$
**5. Encrypt the data** with the public key of the receiver $K \leftarrow CD^2 \; mod \; Q_r$
**6. Obtain ciphertext** '$K$'
**7.** Generate the digital signature $\beta_s$
**8.** Send ciphertext and digital signature $t$o receiver
**9. End for**
**\\ Signature verification and decryption**

**10.** The receiver obtains the ciphertext '$K$' and digital signature $S$
**11.** Receiver generates signature $\beta_s = h\,(CD \| R)$
**12.** Verify the signature
**13. If** $(\beta_s = \beta_s{}'')$ **then**
**14.** Signature is valid
**15.** Decrypt the data ad obtain the original plaintext
**16. else**
**17.** The signature is not valid
**18.** Decryption is not performed
**19. end if**
**20.** Obtain secure data transaction
**End**

Algorithm 2 describes a step-by-step process of secure data transaction using Multiplicative congruential Rabin cryptographic signcryption. First, the classified data are taken as input. Then Multiplicative congruential key generation process said to be performed to generate the private key and public key of the sender as well as the receiver. Then the encryption and signature generation are carried out using the receiver's public key and sender's private key and sent to the receiver. On the receiver side, the signature verification is performed. If the two signatures get matched, then the receiver performs decryption and obtains the original data. Otherwise, the decryption is not performed. This helps to improve the security and data confidentiality level.

## 4. Experimental Scenario

In this section, an experimental evaluation of the proposed GCLD-MCS technique and existing LHS [1] and BCSE [2] are implemented in Java using WUSTL EHMS 2020 Dataset for Internet of Medical Things (IoMT) Cybersecurity Research taken from https://www.cse.wustl.edu/~jain/ehms/index.html. The WUSTL-EHMS-2020 dataset was created using a real-time Enhanced Healthcare Monitoring System (EHMS) testbed and it collects the network flow metrics and patients' biometrics. The patients' conditions are collected from the sensors attached to the patient's body and sent the data to the server. An attacker interrupts this data before they arrive at the server. Therefore, a GCLD-MCS is responsible for capturing the type of attacks (man-in-the-middle attacks: spoofing and data injection) and abnormalities detection during the patient's biometric data distribution. The spoofing attack breaks the patient's data confidentiality. The data injection attack alters the patient's biometric data, which breaks the data's integrity. This dataset consists of 44 features and 16000 instances where 35 features are network flow metrics, eight patients' biometric features, and one feature for the output label (0 or 1). The samples with the attacker are labeled as 1, while the rest as 0.

## 5. Performance Results and Discussion

In this section, the performance of the GCLD-MCS technique and existing methods LHS [1] and BCSE [2] are discussed with different metrics namely Classification accuracy, Data Confidentiality Rate, Data Integrity Rate, and Communication overhead. The performance of different metrics is analyzed with the help of a table and graphical representation.

### 5.1 Impacts of classification accuracy

The classification accuracy is defined as the ratio of a number of data that are correctly classified to the number of data. The formula for calculating the accuracy is given below,

$$Acc = [\frac{ncc}{n}] * 100 \quad (17)$$

From (17), $Acc$ represent a classification accuracy, '$ncc$' indicates the number of data correctly classified, $n$ denotes the total number of data taken as input. The classification accuracy is measured in terms of percentage (%).

**Table 1 Comparison of Classification Accuracy**

| Number of data | Classification accuracy (%) | | |
|---|---|---|---|
| | GCLD-MCS | LHS | BCSE |
| **1600** | 93.75 | 88.12 | 86.25 |
| **3200** | 93.12 | 89.06 | 85.93 |
| **4800** | 94.16 | 90 | 86.45 |
| **6400** | 93.9 | 89.37 | 86.09 |
| **8000** | 94.75 | 91.87 | 88.12 |
| **9600** | 94.19 | 90.1 | 88.02 |
| **11200** | 93.75 | 88.39 | 86.6 |
| **12800** | 92.57 | 89.84 | 87.89 |
| **14400** | 94.44 | 90.27 | 88.19 |
| **16000** | 93.12 | 89.37 | 87.5 |

Table 1 reports the performance of classification accuracy versus a number of data taken from 1600 to 16000. As exposed in the tabulated results, the classification accuracy using the GCLD-MCS technique and existing methods namely LHS [1] and BCSE [2] compared with each other methods, GCLD-MCS technique provides improved accuracy results when compared to existing methods. For example, with the number of data of 1600, the accuracy observed by using the GCLD-MCS technique is 93.75% and the accuracy of existing LHS and BCSE [2] was found to be 88.12%, 86.25% respectively. Followed by, nine different performance results are observed for each method. The average of ten comparison results indicates that the proposed GCLD-MCS technique enhances the classification accuracy by 5% and 8% when compared to [1] and [2] respectively.

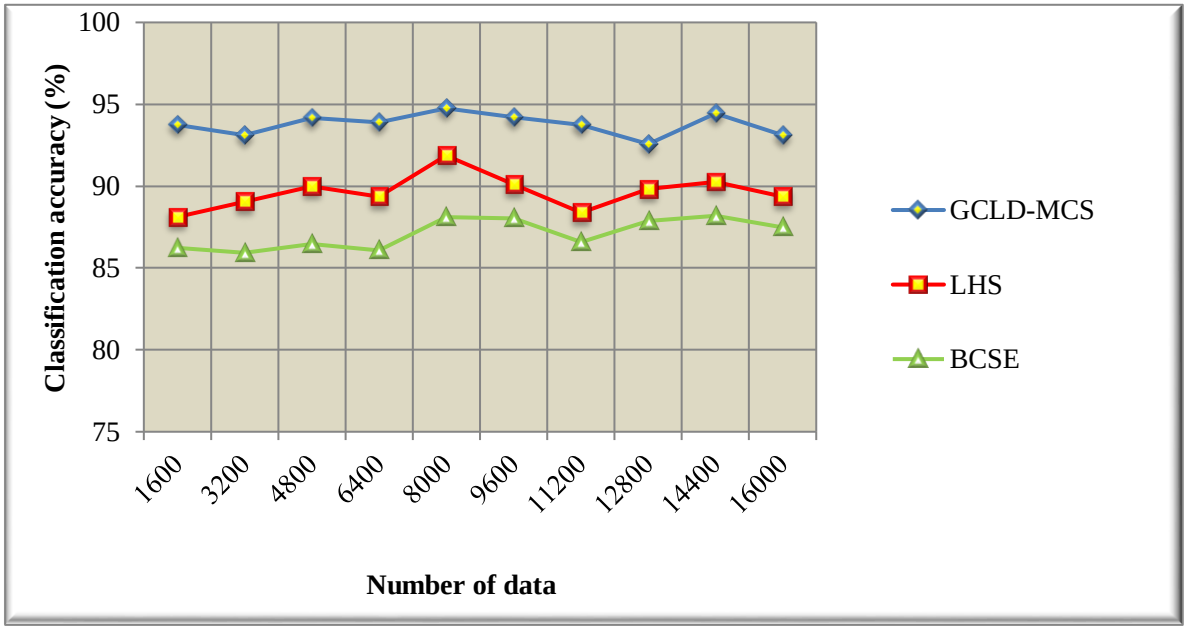


**Figure 5 Classification Accuracy Performance Comparision**

Figure 5 represents the graphical illustration of classification accuracy for ten different numbers of data. As exposed in figure 5, numbers of data are represented on the '$x$' axis, and the accuracy of different methods is observed at the '$y$' axis. The graphical illustration indicates that the classification accuracy of the GCLD-MCS technique is found to be improved than the other two existing methods. This is due to the application of the generalized canonical statistic distributive linear discriminant analysis to find the linear relationship between the data and the mean of a particular class by estimating the generalized canonical correlation. Based on the correlation analysis, the GCLD-MCS technique provides the final better accuracy results.

## 5.2 Impacts of Data Confidentiality Rate

The data confidentiality rate is measured as the ratio of the number of data that are received by the authorized receiver. The confidentiality rate is measured as given below,

$$C_{Rate} = [\frac{N_{AR}}{n}] * 100 \quad (18)$$

From (18), $C_{Rate}$represent a data confidentiality rate, '$N_{AR}$' indicates the data only received by the receiver, $n$ denotes the total number of data taken as input. The confidentiality rate is measured in terms of percentage (%).

**Table 2 Comparison of Data confidentiality rate**

| Number of data | Data confidentiality rate (%) | | |
|---|---|---|---|
| | GCLD-MCS | LHS | BCSE |
| **1600** | 93.5 | 87.5 | 85.62 |
| **3200** | 93.06 | 88.12 | 85.78 |
| **4800** | 94.06 | 89.79 | 85.83 |
| **6400** | 93.75 | 89.06 | 85.93 |
| **8000** | 94.62 | 90.62 | 87.5 |
| **9600** | 94.06 | 89.06 | 87.70 |
| **11200** | 92.85 | 87.94 | 86.33 |
| **12800** | 91.79 | 88.28 | 87.10 |
| **14400** | 94.30 | 89.58 | 88.02 |
| **16000** | 93.06 | 88.75 | 86.25 |

Table 2 indicates the performance results of the data confidentiality rate of the proposed GCLD-MCS technique with two state-of-the-art methods namely LHS [1] and BCSE [2]. The number of data taken from the dataset is 1600 to 16000. The proposed GCLD-MCS technique increases the performance of data confidentiality rate when compared to existing methods. This is proved through statistical assessment. Let us consider 1600 data samples taken as input in the first iteration. By applying the GCLD-MCS technique, 1496 data are correctly received by the authorized receiver, and hence the observed data confidentiality rate is 93.5% whereas the data confidentiality rate of existing LHS [1] and BCSE [2] are 87.5% and 85.62% respectively. Similarly, the nine varieties of iterations are performed for each method with a different number of input data. Finally, the performance of the GCLD-MCS technique is compared to the results of existing methods. The average of ten comparison results indicates that the performance of the confidentiality rate gets increased by 5% compared to [1] and 8% when compared to [2].

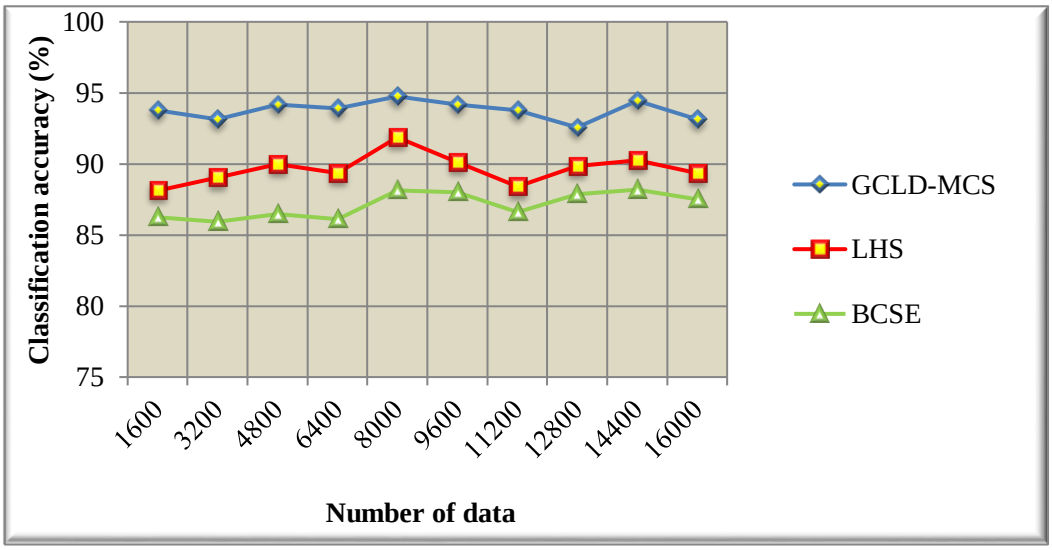


**Figure 6 Data Confidentiality Rate Performance Comparison**

Figure 6 depicts the performance comparison graph of data confidentiality rate versus a number of data samples using three methods GCLD-MCS technique, LHS [1], and BCSE [2]. The above graphical results demonstrate GCLD-MCS technique increases the data confidentiality rate when compared to existing methods. To discover the best performance, the GCLD-MCS technique uses Multiplicative congruential Rabin cryptographic signcryption. In signcryption, the GCLD-MCS technique performs the encryption and signature generation process. In the encryption process, the original plain text is converted into ciphertext. At the receiver, the signature verification is performed for only the authorized user who received the data and avoids unauthorized access.

### 5.3 Impact of Data Integrity Rate

Integrity rate is measured as the ratio of the number of data that are not altered or modified by any intruders (attacks) to the number of data transmitted. The data integrity rate is calculated as given below,

$$I_{rate} = [\frac{NNI}{n}] * 100 \quad (19)$$

From (19), $I_{rate}$ signifies a data integrity rate, $NNI$ denotes the number of data that are not altered or modified by others, '$n$' indicates the total number of data. The data integrity rate is measured in terms of percentage (%).

**Table 3 Comparison of Data Integrity Rate**

| Number of data | Data integrity rate (%) | | |
|---|---|---|---|
| | **GCLD-MCS** | **LHS** | **BCSE** |
| **1600** | 93.12 | 86.87 | 85.31 |
| **3200** | 92.96 | 87.81 | 85.62 |
| **4800** | 93.95 | 89.58 | 85.66 |
| **6400** | 93.43 | 88.75 | 85.62 |
| **8000** | 94.5 | 89.37 | 86.87 |
| **9600** | 94.01 | 88.85 | 87.5 |
| **11200** | 92.67 | 87.76 | 86.16 |
| **12800** | 91.64 | 88.16 | 86.95 |
| **14400** | 93.75 | 89.23 | 87.91 |
| **16000** | 92.93 | 88.5 | 86.12 |

Table 3 describes the performance of the data integrity rate of three different techniques GCLD-MCS technique, LHS [1], and BCSE [2] consistent with the number of data samples collected from the dataset.

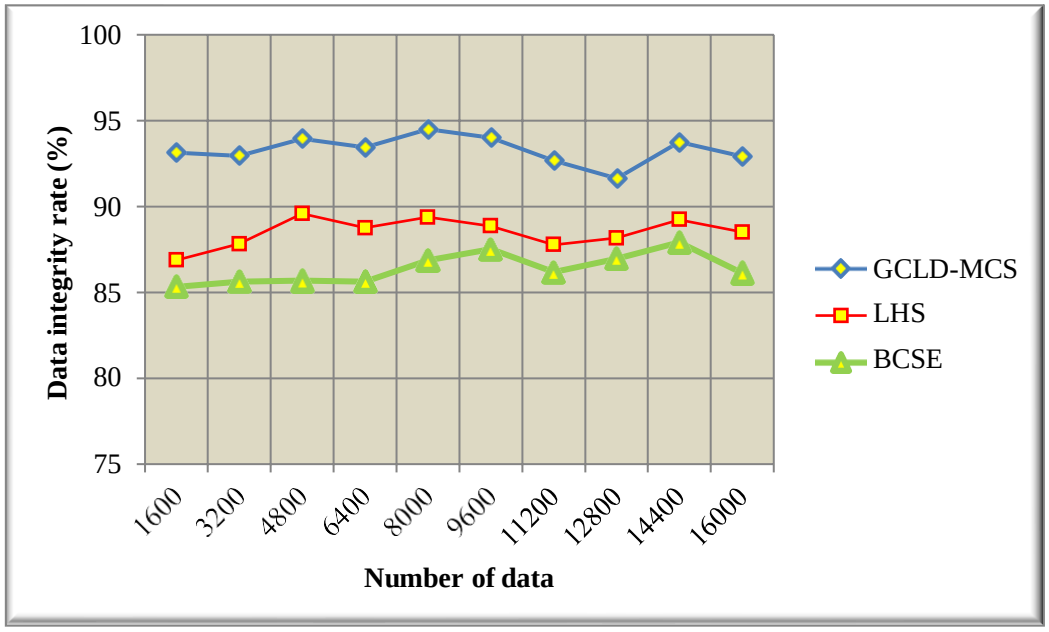


**Figure 7 Data Integrity Rate Performance Comparison**

For the experimental consideration, the number of data samples taken ranges from 1600 to 16000. Compared to existing methods, the GCLD-MCS provides improved performance of data integrity rate. Let us consider 1600 data samples for conducting the experiment. The observed results of data integrity rate using the GCLD-MCS technique is93.12%. Similarly, the data integrity rate of existing [1] [2] is observed as 86.87% and 85.31% in the first iteration. Likewise, the different performance results are obtained with respect to a number of data samples. The overall results indicate that the GCLD-MCS technique increases the overall performance of the data integrity rate by 5% when compared to [1] and 8% when compared to [2].

Figure 7 provides the performance comparison of the data integrity rate regarding the number of data samples from 1600 to 16000. As presented in figure 7, the horizontal axis provides a number of data while the vertical axis shows the data integrity rate. However, the proposed GCLD-MCS showed better performance over existing methods like LHS [1] and BCSE [2]. From the results, it is realized that the proposed GCLD-MCS showed better performance over the state of the arts. This is due to the application of signature generation and verification. At the receiver end, signature verification is performed. If the signature is valid, the receiver is said to be authorized. Then the authorized user received the original data. Otherwise, the receiver is said to be the unauthorized user (i.e. attack) and did not get the original data. The proposed GCLD-MCS technique finds the attack and minimizes the data alteration. This in turn increases the data integrity rate.

### 5.4 Impact of Communication overhead

The communication overhead is measured as the amount of time taken by the algorithm for secure data transmission from sender to receiver. The formula for calculating the communication overhead is expressed as given below,

$$CO = n * T(SDT) \quad (20)$$

Where $CO$ denotes a communication overhead, $n$ denotes the number of data, $T$ denotes a time for secure data transmission ($SDT$). It is measured in terms of milliseconds (ms).

**Table 4 Comparison of Communication Overhead**

| Number of data | communication overhead (ms) | | |
|---|---|---|---|
| | **GCLD-MCS** | **LHS** | **BCSE** |
| **1600** | 27.2 | 31.2 | 35.2 |
| **3200** | 35.2 | 38.4 | 41.6 |
| **4800** | 39.84 | 43.2 | 45.6 |
| **6400** | 43.52 | 48 | 51.2 |
| **8000** | 48.8 | 52 | 56 |
| **9600** | 50.88 | 55.68 | 57.6 |
| **11200** | 53.76 | 58.24 | 61.6 |
| **12800** | 58.88 | 62.72 | 66.56 |
| **14400** | 60.48 | 64.8 | 69.12 |
| **16000** | 64 | 67.2 | 72 |

The performance results of the communication overhead using three different methods GCLD-MCS technique, LHS [1], and BCSE [2] are shown in table 3. The observed communication overhead results indicate that the performance of the GCLD-MCS technique is relatively minimum than the existing methods. With the consideration of 1600 samples of data, the time taken to perform secure data transmission was found to be '$27.2ms$'. However, the time consumption of existing LHS [1] and BCSE [2] was found to be $31.2ms$' and $35.2ms$. The observed results designate that the GCLD-MCS technique decreases the communication overhead. After obtaining the ten results, the overall time consumption of the GCLD-MCS technique is compared to the existing results. The average of ten results demonstrates that the proposed GCLD-MCS technique reduces the time consumption of secure data communication by 8% and 14% as compared to the [1] and [2] respectively.

Figure 8 reveals the communication overhead involved in secure data transmission from the sender to the receiver. From Figure 8, it is obvious that the proposed method showed better performance over the state of the art. As presented in Figure 8, the horizontal axis provides data samples in numbers while the vertical axis shows communication overhead. As data samples are increased, the communication overhead is relatively increased. Therefore, the communication overhead is proportional to the number of data samples considered. However, the proposed GCLD-MCS technique showed better performance than existing methods. This is because the GCLD-MCS performs data classification using generalized canonical statistic distributive linear discriminant analysis before secure transmission. Moreover, the classified results are given to the signcryption for enhancing the security of data transmission with minimum overhead.

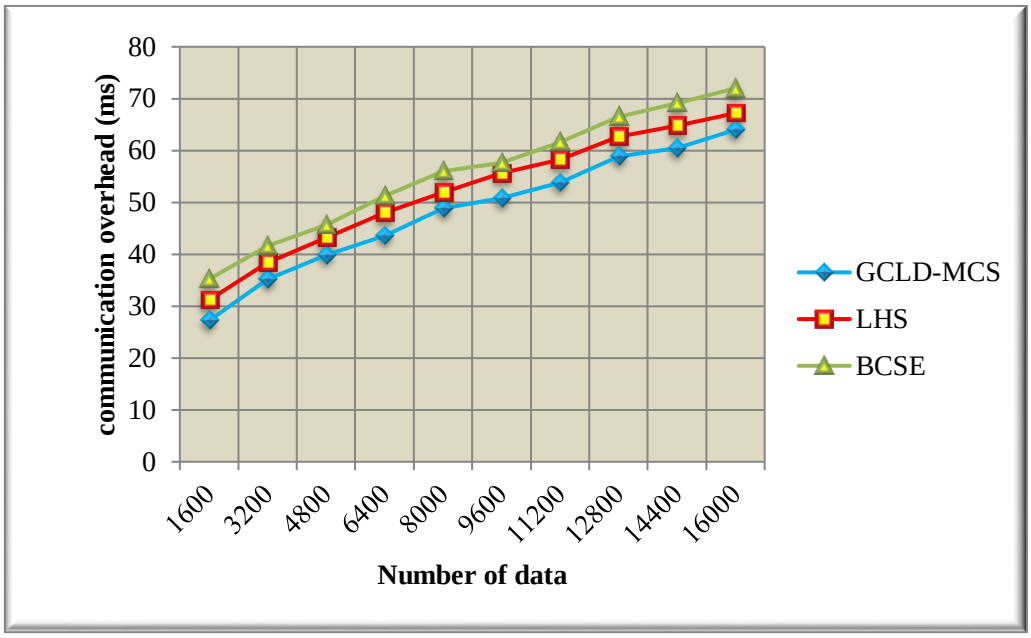


**Figure 8 Communication Overhead Performance Comparison**

## 6. Conclusion

In this paper, a novel technique called GCLD-MCS is used to deal with big data security. The technique is processed based on generalized canonical statistic distributive linear discriminant analysis and Multiplicative congruential Rabin cryptographic signcryption. First, the big data are taken as input, and perform the data classification to obtain the structured format to minimize the time consumption of the secure data communication. After the classification, the GCLD-MCS performs the Multiplicative congruential Rabin cryptographic signcryption to improve the data confidentiality and integrity. In this way, secure big data transmission is performed. The experimental assessments are carried out with respect to a number of data samples and compare the results of the proposed technique with two existing algorithms. The observed numerical results have confirmed that the proposed GCLD-MCS technique provides improved performance results in terms of classification accuracy, data confidentiality rate, data integrity rate, and communication overhead than the other cryptographic techniques.

## References

[1] Mallepalli Prasanna Kumari and Tumma Srinivasa Rao, "A lightweight hybrid scheme for security of big data", Materials Today: Proceedings, Elsevier, March 2021, Pages 1-15. https://doi.org/10.1016/j.matpr.2021.03.151

[2] Fengyin Li, Xinying Yu, Rui Ge, Yanli Wang, Yang Cui, and Huiyu Zhou, "BCSE: Blockchain-Based Trusted Service Evaluation Model over Big Data", Big Data Mining and Analytics , Volume 5, Issue 1, 2022, Pages 1 – 14. DOI: 10.26599/BDMA.2020.9020028

[3] Veselska Olga, Ziubina Ruslana, Finenko Yuriy , Nikodem Joanna, "Big Data Analysis Methods Based on Machine Learning to Ensure Information Security", Procedia Computer Science, Volume 192, 2021, Pages 2633-2640. https://doi.org/10.1016/j.procs.2021.09.033

[4] Yuanzhao Gao, Xingyuan Chen, Xuehui Du, "A Big Data Provenance Model for Data Security Supervision Based on PROV-DM Model", IEEE Access, Volume 8, 2020, Pages 38742 – 38752. DOI: 10.1109/ACCESS.2020.2975820

[5] Jian Shen, Member, Chen Wang, Anxi Wang, Sai Ji, and Yan Zhang, "A Searchable and Verifiable Data Protection Scheme for Scholarly Big Data", IEEE Transactions on Emerging

Topics in Computing , Volume 9, Issue 1, 2021, Pages 216 – 225. DOI: 10.1109/TETC.2018.2830368

[6] Parsa Sarosh, Shabir A.Parah, G.MohiuddinBhat, KhanMuhammad, "A Security Management Framework for Big Data in Smart Healthcare", Big Data Research, Elsevier, Volume 25, 2021, Pages 1-10. https://doi.org/10.1016/j.bdr.2021.100225

[7] Chin-Yu Sun, Allen C.-H. Wu, TingTing Hwang, "A novel privacy-preserving deep learning scheme without a cryptography component", Computers & Electrical Engineering, Elsevier, Volume 94, 2021, Pages 1-15. https://doi.org/10.1016/j.compeleceng.2021.107325

[8] Xuebin Ren, Chia-Mu Yu, Wei Yu, Xinyu Yang, Jun Zhao, Shusen Yang, "DPCrowd: Privacy-Preserving and Communication-Efficient Decentralized Statistical Estimation for Real-Time Crowdsourced Data", IEEE Internet of Things Journal, Volume 8, Issue 4, 2021, Pages 2775 – 2791. DOI: 10.1109/JIOT.2020.3020089

[9] Weichao Gao, Wei Yu, Fan Liang, William Grant Hatcher, Chao Lu, "Privacy-Preserving Auction for Big Data Trading Using Homomorphic Encryption", IEEE Transactions on Network Science and Engineering, Volume 7, Issue 2, 2020, Pages 776–791. DOI: 10.1109/TNSE.2018.2846736

[10] Saleh Atiewi1, Amer Al-Rahayfeh, Muder Almiani, Salman Yussof3, Omar Alfandi, Ahed Abugabah, And Yaser Jararweh, "Scalable and Secure Big Data IoT System Based on Multifactor Authentication and Lightweight Cryptography", IEEE Access , Volume 8, 2020, Pages 113498 – 113511. DOI: 10.1109/ACCESS.2020.3002815

[11] Rafik Hamza, Alzubair Hassan, Awad Ali , Mohammed Bakri Bashir, Samar M. Alqhtani, Tawfeeg Mohmmed Tawfeeg and Adil Yousif, "Towards Secure Big Data Analysis via Fully Homomorphic Encryption Algorithms", Entropy, Volume 24, Issue 4, 2022, Pages 1-17. https://doi.org/10.3390/e24040519

[12] Gayatri Kapil1, Alka Agrawal1, Abdulaziz Attaallah2, Abdullah Algarni, Rajeev Kumar and Raees Ahmad Khan, "Attribute based honey encryption algorithm for securing big data: Hadoop distributed file system perspective", Peer J Computer Science, Volume 6, 2020, Pages http://doi.org/10.7717/peerj-cs.259

[13] Hanan E. Alhazmi, Fathy E. Eassa; Suhelah M. Sandokji, "Towards Big Data Security Framework by Leveraging Fragmentation and Blockchain Technology", IEEE Access, Volume 10, 2022, Pages 10768 – 10782. DOI: 10.1109/ACCESS.2022.3144632

[14] Pratima Sharma, Malaya Dutta Borah and Suyel Namasudra, "Improving security of medical big data by using Blockchain technology", Computers & Electrical Engineering, Elsevier, Volume 96, Part A, December 2021, Pages 1-15. https://doi.org/10.1016/j.compeleceng.2021.107529

[15] Liping Zhang, Zhen Wei, Wei Ren, Xianghan Zheng, Kim-Kwang Raymond Choo, Neal N. Xiong, "SIP: An Efficient and Secure Information Propagation Scheme in E-Health Networks", IEEE Transactions on Network Science and Engineering ,Volume 8, Issue 2, 2021,Pages 1502 – 1516. DOI: 10.1109/TNSE.2021.3063174

[16] Sarath Sabu, H.M. Ramalingam, M Vishaka, H.R. Swapna, Swaraj Hegde, "Implementation of A Secure and Privacy-Aware E-Health Record and IoT Data Sharing using Blockchain", Global Transitions Proceedings, Elsevier, Volume 2 Issue 2, 2021, Pages 429-433. https://doi.org/10.1016/j.gltp.2021.08.033

[17] Shekha Chenthara ,Khandakar Ahmed,Hua Wang,Frank Whittaker,Zhenxiang Chen, "Healthchain: A novel framework on privacy preservation of electronic health records using blockchain technology", PLoS ONE, Volume 15, Issue 12, 2020, Pages 1-35. | https://doi.org/10.1371/journal.pone.0243043

[18] Jafar A.Alzubi, "Blockchain-based Lamport Merkle Digital Signature: Authentication tool in IoT healthcare", Computer Communications, Elsevier, Volume 170, 2021, Pages 200-208. https://doi.org/10.1016/j.comcom.2021.02.002

[19] Sachi Nandan Mohanty, K.C. Ramya, S. Sheeba Rani, Deepak Gupta, K. Shankar, S.K. Lakshmanaprabu, Ashish Khanna, "An efficient Lightweight integrated Blockchain (ELIB) model for IoT security and privacy", Future Generation Computer Systems, Elsevier, Volume 102, January 2020, Pages 1027-1037. https://doi.org/10.1016/j.future.2019.09.050

[20] LokendraVishwakarma, Debasis Das, "SCAB-IoTA: Secure communication and authentication for IoT applications using blockchain", Journal of Parallel and Distributed Computing, Elsevier, Volume 154, 2021, Pages 94-105. https://doi.org/10.1016/j.jpdc.2021.04.003