Efficient Data Anonymization Approach to Preserve Privacy of Sensitive Data In Cloud Storage

Sahana Lokesh R[1]*, H R Ranganatha[2]*

Abstract

In cloud platforms, data privacy is a significant feature of stored data. The cloud contains a significant part in the medical field anywhere and it contains some sensitive data like the effects of the illness and the nature of the disease. Publishing and sharing sensitive data individually from the cloud infrastructure is an important task in the medical field. Hence, it is significant to preserve the information of the patients with more data privacy and high security. The African Vultures Optimization (AVOA) algorithm is the combination of a Genetic Algorithm and a simulated annealing method (GAVOSA) employed in this paper to protect patients' privacy. To calculate the fitness function, the optimization algorithm uses the generalized information loss and average equivalence value. The outcomes of the introduced method showed that the introduced technique can efficiently protect the medical databases' privacy.

Keywords: Genetic Algorithm-African Vultures Optimization algorithm-Simulated Annealing, cloud platform, Encryption, Privacy preservation, Data Security, Mondrian K-anonymity, Privacy.

DOI Number: 10.14704/Nq.2022.20.17.Nq880110 **Neuroquantology 2022; 20(17): 850-860**

1. Introduction

Many service areas like online data storage, infrastructure and application are distributed by cloud computing. It is the distribution of computing services by the Internet, and also counting data storage, networking, software, databases and servers. It is used to store, recover and manage information like medical records, educational records, individual data, financial transactions, and so on [9] Cloud computing has secured data like a persons' address, email, name, phone number, date of birth, password, health, medical records, treatment details, finance, zip code, etc., [28]. If the sensitive information is published, then the individuals' privacy is at risk. The Privacy-Preserving Data Publishing (PPDP) technique is used to publish the data while the data privacy is protected. Multiple anonymization methods, prototypes, algorithms and frameworks, are proposed for PPDP.

Data anonymization is utilized to preserve patient-sensitive information. It is the process of information purification that includes encrypting or removing the personal information in the dataset. Also, when the information is touching across boundaries then the data anonymization is used to reduce the risk of data loss [13]. In the privacy preservation field, k-anonymity is the most generally used method. K-anonymization contains a better performance in data privacy protection in the aspects of social network, data publication and location-based. PPDP does the main task of protecting both the privacy and utility of the data anonymization concurrently. In this research, the combination of the Mondrian-based k-anonymization approach and metaheuristic-based optimization approach are used to develop utility and privacy. In this paper, the optimization algorithm has proposed the combination of the simulated annealing approach (SA) , genetic algorithm (GA) [11] and African vulture's optimization algorithm

***Corresponding Author:-** Sahana Lokesh R, H R Ranganatha
Address:[1*]Research Scholar, Sapthagiri College of Engineering(VTU).
[2*]Professor & Head, Department of ISE, Sapthagiri College of Engineering, Bengaluru
Relevant conflicts of interest/financial disclosures: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

(AVOA) [28]. AVOA was inspired by the African vultures' lifestyle, navigation behaviors, and foraging. The GA is one of the powerful optimization algorithms that has shown its dependability in resolving huge compound real world problems and therefore used in PPDP [19]. Simulated Annealing (SA) is one of the easiest and most popular metaheuristic approaches because of its better performance and simplicity. Asymptotic convergence and statistical equilibrium are the two theoretical properties of basic SA [20] . GA–based encryption methods show better result in data encryption and the data encryption methods can be protected the anonymization information in the cloud. In the DNA-Genetic Algorithms' field [29], has done a great amount of research. Zang et al., proposed that DNA-GA is attracted through the biological membrane structure [29]. The performance of the proposed optimization algorithm, encryption approach and anonymization method are calculated with various datasets such as Breast Cancer Wisconsin (Diagnostic) Dataset, Polycystic ovary syndrome (PCOS), Lung Cancer Dataset, Cervical Cancer, Thyroid Disease Dataset based on Average Equivalence Value (C_A), Generalized Information Loss (GI_L) , throughput and Running time.

The remaining part of the article is discussed in the section below. Section 2 presents literature review of the proposed work. Background approaches are explained in Section 3. Section 4 discusses the proposed methodology. The experiments with performance metrics and dataset decryption are explained in section 5. Section 6 discusses the experimental result. Section 7 details the conclusion.

2. Literature Review

In [1] an anonymization method for PPDP is proposed. The purpose of this method is to findout the privacy issues in social networks' application-particular scenarios. Cloud computing faces many security problems, they are DoS attacks, vulnerable systems, security threats, data breaching, data privacy, data loss, service disruption, data integrity, multi-tenancy issues, malicious insiders, outside malicious attacks, and compromised authentication [18]. A slicing method is suggested in [13] for protecting the privacy of data. Slicing can handle high dimensional data. When compared with bucketization and generalization, the slicing contains numerous advantages. It protects more attributes connected with the sensitive attributes (SA) than bucketization. Compared to the generalization method, slicing protects an improved data utility.

An anonymization approach is presented in[19] that delivers both knowledge protection and privacy preservation. The anonymization method keeps statistical relations between data to knowledge protection through the data randomization; however knowledge is lost in most anonymization approaches. A general framework is used in [10]to calculate the clustering-based k-anonymity. The k-anonymity is the approach for distributing protection of privacy by making sure that the individual data cannot be found. A rough encrypt approach with the DBSCAN algorithm is introduced in [11] for clustering mobile SA data to develop the k-anonymizations' performance. This method is used to evaluate the clusters' purity and then it handles the unpredictability and develops the categorical cluster datas' performance. K-anonymity is introduced by [14] for privacy protection in the data collection before reaching data holders from the data owners, in a distributive and collaborative way. Though this method is possible for attacks, it is not secure like membership disclosure and attributes. The clustering method for the k-anonymization approach is presented in [17] to reduce the loss of information though at the same time promising the quality of data. One of the main challenges of this method is to reduce the loss of information during the process of anonymization. Using the k-anonymization technique, preservation of the privacy-based information publishing model is proposed [9]. This paper provides the result for protection of privacy against outsider and insider threats.

In[2] a GA-GWO-based k-anonymization approach is introduced. This method contains some evaluation metrics which are the Generalized Information Loss (GI_L), throughput, Average Equivalence Value (C_A), encryption time and decryption time. A ciphertext-policy attribute-based encryption (CP-ABE) approach has been introduced in [18] which allows a fine-grained encrypted IoT data access control on the

cloud. CP-ABE is considered the most assured method to deliver fine-grained and flexible access control. Based on the IoT systems CP-ABE is fairly suitable for secured cloud. In [28] a comprehensive privacy security protection framework is suggested. This paper also presents the cloud computings' risk for privacy security. In [6] the fusion of cryptography and signal processing as developing examples for protecting the users' privacy is proposed.

3. Background

This section details the encryption methods, optimization methods and anonymization methods used in the introduced technique.

3.1 DNA-GA based data encryption method

The DNA-GA encryption method contains random key generation, decryption and encryption. DNA computing is proposed into GA to develop the genetic algorithms' performance because of its natural connection with GA and DNA. Then the DNA computing is used to find the all possible results. The DNA-GA is attracted through the structure of the biological membrane. Thus, the DNA-GA algorithm overcomes the traditional genetic algorithms' demerits.[12], join the DNA-GA with other pattern identification approaches like multi-object optimization, clustering analysis and classification. DNA-GA is based on a multipoint find operation.

In genetic algorithm, mutation, crossover and replication are performed after the encryption stage. In the population of the chromosome, the redesigning operation is used to create the genetic materials like DNA, RNA and Genes that are moved to the activity and next iteration. In this step, the first chromosome number and length are determined. For each round, these values either vary or constant. After building the parent chromosomes, the crossover operation is performed. The two types of crossover approaches are used in this operation sequentially. In the first crossover operation, the parent chromosomes are chosen in the breeding pool. By exchanging the parents' heads, two new offsprings are produced. The mutation process, modifies the element strings after the crossover process and there are two methods. In the first method, 2 mutation points are explained among the first and last bits and also the data is converted into a binary vector. In the second method the complement of bits are taken and then the encrypted data is stored in the cloud. The encryption method is reversed to perform the decryption. Pseudocode for the DNA-GA based encryption method is shown in algorithm1.

Algorithm 1: Pseudo-code of DNA-GA-based encryption method
Read anonymized input text is saved in the Data
Using ASCII code conversion the binarize data is saved in the Data1
Redesign Data1
DNA bases are saved in the Data2
while (Round Number ≠ 0) **do**
// Encryption
Data 2 Encrypt with key
// GA
Redesign
Mutation
Crossover process
End while
Redesign
In the cloud to save the encrypted text file
In Data3 to save the brinarize encrypted text
while (Round Number ≠ 0) **do**
// Genetic Algorithm
Mutation
Crossover process
Redesign
// Decryption
Data 3 decrypt with key
End while
Redesign
Save file

3.2 Genetic Algorithm

The Genetic Algorithm (GA) contains three important processes which are selection, crossover and mutation. The biological evolution operation attract the GA. Natural selection is the important motivation of the genetic algorithm. The first method of this algorithm is about the initial population . Using the fitness function, the GA calculates the fitness value for each individual in the population. The GA calculates the number of multiple individuals and generates best solutions. The key elements of the GA is fitness function, chromosome representation, selection, mutation and crossover. GA is mainly utilized in problem optimization and machine learning methods. A genetic algorithm has proved as a powerful and reliable optimization method that can be applied to various real-world problems of important complexity.

3.3 African vulture's optimization algorithm (AVOA)

African vulture's optimization algorithm (AVOA) depends on the number of vultures in such an environment. In the metaheuristic algorithm, the N-vultures determine a similar number of population, and also this amount depends on the difficulty, the analyst wants to put into the African vultures optimization algorithm. Many vultures are physically separated into two groups in the natural environment. In this algorithm to separate the vultures into groups, the fitness function of the initial population is calculated first. Vultures provide the most important natural function that can be formulated which is the reason for dividing the groups in this algorithm. Each category of vultures has various capacities to discover the food and eat. During the preparation period, anti-hunger agreement, assumes that feeding the hungriest and weakest are the worst solution for the population. So the vultures try to present the greatest solution and stay away from the worst solution. Allowing the best vultures and strongest vultures are regarded as the two best solutions in the AVOA and the supplementary vultures also try the best approach. After forming the initial group, to evaluating fitness, the first and second-better solutions are chosen as the first and second-better vulture using Equation (1).

$$R(i) = \begin{cases} BestVulture_1 if p_i = L_1 \\ BestVulture_2 if p_i = L_2 \end{cases} \quad (1)$$

$$p_i = \frac{F_i}{\sum_{i=1}^{n} F_i} \quad (2)$$

Equation (1) shows that selecting the chosen vultures' probability is moved to the other vultures towards a better solution for every group is evaluated. $R(i)$is denoted as the best vultures. Before the search operation, the parameters $L_1$and L_2 are to be measured, with the values between 0 and 1 and some of both parameters are shown in Equation 1.

$$F = (2 \times rand_1 + 1) \times z \times \left(1 - \frac{iteration_i}{maxiterations}\right) + t \quad (3)$$

$$t = h \times \left(sin^w\left(\frac{\pi}{2} \times \frac{iteration_i}{maxiterations}\right) + cos\left(\frac{\pi}{2} \times \frac{iteration_i}{maxiterations}\right) - 1\right) \quad (4)$$

$F$is the satisfied vultures, max-iteration is the iterations' total number, then $the\ iteration_i$ denoted as the current iteration number and the z is a random number among-1 and 1. If the z value drops greater than 0, it means that the vulture is satisfied and if the z value drops less than 0, then it means the vulture is hungry. It can be seen in Equations (3) and (4).

$$p(i+1) = \begin{cases} Equation(6) if P_1 \geq rand_{P1} \\ Equation(8) if P_1 < rand_{P1} \end{cases} \quad (5)$$

$$p(i+1) = R(i) - D(i) \times F \quad (6)$$

$$D(i) = |X \times R(i) \times P(i)| \quad (7)$$

If $P_1 \geq rand_{P1}$, then update the position of the vulture using Equation (6).where $p(i+1)$ represents the vulture position vector, and using Equation (4) F is denoted as the amount of vulture being satisfied which is gained in the present iteration. $R(i)$ is denoted as the best vultures given in Equation (7), which is chosen through the use of Equation (1) in the present iteration.

$$p(i+1) = R(i) - F + rand_2 \times \left((ub - lb) \times rand_3 + lb\right) \quad (8)$$

If $P_1 < rand_{P1}$, then update the vultures' position updated using Equation (8). In the present iteration, Equation (1) is used to select the best vulture $R(i)$ in Equation (8). Using Equation (4), F is denoted as the rate of vulture satiation. Then $rand_2$ contains an irregular value among 0 and 1 then lb and up prove the lower bound and upper bound of the variables.

$$p(i+1) = \begin{cases} Equation(10) if P_2 \geq rand_{P2} \\ Equation(13) if P_2 < rand_{P2} \end{cases} \quad (9)$$

$$p(i+1) = D(i) \times (F + rand_4) - d(t) \quad (10)$$

$$d(t) = R(i) - P(i) \quad (11)$$

If $P_2 \geq rand_{P2}$ then update the vultures' position update using Equation (10). Then $rand_4$ contains a random value between 0 and 1.

$$S_1 = R(i) \times \left(\frac{rand_5 \times P(i)}{2\pi}\right) \times \cos(P(i))$$

$$S_2 = R(i) \times \left(\frac{rand_6 \times P(i)}{2\pi}\right) \times \sin\left(P(i)\right) \quad (12)$$

$$p(i+1) = R(i) - (S_1 + S_2) \quad (13)$$

If $P_2 < rand_{P2}$, then update the position of the vulture using Equation (13). Where *sin* and *cos* are functions of sine and cosine. Then $rand_5$ and $rand_6$ represent the random number between 0 and 1. Where S_1 and S_2 are used for Equation (13). Using Equation (13), update the vultures' location.

$$p(i+1) = \begin{cases} Equation(16) if P_3 \geq rand_{P3} \\ Equation(17) if P_3 < rand_{P3} \end{cases} \quad (14)$$

$$a_1 = BestVulture_1(i) - \frac{BestVulture_1(i) \times P(i)}{BestVulture_1(i) - P(i)^2} \times F$$

$$a_2 = BestVulture_2(i) - \frac{BeastVulture_2(i) \times P(i)}{BestVulture_2(i) - P(i)^2} \times F \quad (15)$$

In the current iteration, $BestVulture_1(i)$ represents the first groups' best vulture and the $BestVulture_2(i)$represents the second groups' best vulture. Then $P(i)$represents the vultures' present vector position. If $P_3 \geq rand_{P3}$, then the position of the vulture is updated using Equation (16).

$$p(i+1) = \frac{a_1 + a_2}{2} \quad (16)$$

In Equation (16) all vultures' aggregation is carried out. Equation (15) is used to get the values of a_1 anda_2. Then $p(i+1)$ represent the vulture positions' vector in the next iteration.

$$p(i+1) = R(i) - |d(t)| \times F \times Levy(d) \quad (17)$$

If $P_3 < rand_{P3}$, then update the position of the vulture using Equation (17).$d(t)$ is the vultures' distance, which is evaluated by Equation (11).

3.4 Simulated Annealing (SA)

The physical annealing operation of metalwork is the motivation of simulated annealing (SA). SA is highly utilized in real-life applications. This technique is commonly utilized in most optimization issues to get the best neighbouring solution. In metallurgy, SA is a single-solution based metaheuristic process for optimization motivated by the annealing. Simulated annealing is one of the famous metaheuristic approaches because of its better performance and simplicity. The capacity of SA to accept transitions that reduce the objective function is the simulated annealings' main characteristic. SA examines both impossible and possible programs by starting from a random feasible solution and an initial temperature. The simulated annealing (SA) algorithm is the most famous physical based algorithm.

3.5 K-anonymization

A k-anonymized database is the smallest of k-1 separate rows that distribute the same identifiers. K-anonymization contains two main operations and they are suppression and generalization. In the database, the attributes are categorized into two identifiers: Quasi-Identifiers (QI) and Sensitive Identifiers (SI). The examples of original records with SI and QI attributes are shown in Table 1.

3.5.1 Generalization

In the generalization operation, the broader category is used to change the discrete attribute. The main goal of this method is to enlarge uncertainty. The anonymized data set depends on the generalized process. An example of the generalization method is, if 'AGE' is 47, then it can be categorized as 40-50 or 45-50. The attributes' generalization is shown in Table 2.

3.5.2 Suppression

In the suppression process, the value of an attribute is exchanged with the special character '*'. The attributes' suppression is shown in Table 3.

3.5.3 Quasi Identifiers (QI)

QI is the attributes' sub-set, it is used to differentiate the person's identification. In the k-anonymization method, using generalization and suppression approaches in the QI database are usually made. Examples of quasi-identifiers are age, zip code, and gender.

3.5.4 Sensitive Identifiers (SI)

A sensitive identifier is an attribute that directly identifies personal records. Examples of SI are disease and income. These attributes are removed from the database, before publication.

Table 1. Patient records' sample

Name	Gender	Age	Zip code	Disease
Robert	Male	42	1115	Heart attack
Michel	Male	39	1158	Lupus
Sofia	Female	40	1562	Shingles
Lokesh	Male	56	1123	Anxiety

Table 2. Generalization

Name	Gender	Age	Zip code	Disease
Robert	Male	40-45	1115	Heart attack
Michel	Male	30-40	1158	Lupus
Sofia	Female	40-50	1562	Shingles
Lokesh	Male	55-60	1123	Anxiety

Table 3. Suppression

Name	Gender	Age	Zip code	Disease
Robert	Male	42	11**	Heart attack
Michel	Male	39	11**	Lupus
Sofia	Female	40	15**	Shingles
Lokesh	Male	56	11**	Anxiety

4. Proposed methodology

This section explains the introduced PPDP model with encryption and anonymization methods. In the introduced approach, the anonymized database through the data supplier using the GAVOSA using Mondrian k-anonymity approach for anonymization data. Before publishing in the cloud, the database anonymization is encoded and then creates the key. Using this key, the client can access the database anonymization by the process of decryption. The main aim of the introduced method is to defend the individuals' privacy while ensuring data utility. The proposed method PPDP block diagram is shown in Figure1.

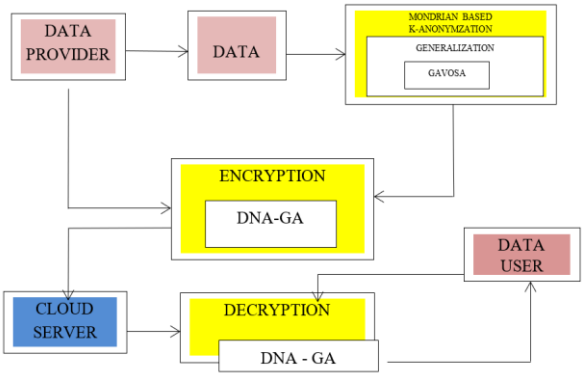

Figure 1. Flow Diagram of proposed PPDP approach

Each QI attribute represents the dimension in the Mondrian-based k-anonymization. In the private table, each tuple denotes a space point explained by QI. In dimension, tuples' planning is used to make a multidimensional space. In the region, all the points are general to a similar single value. The corresponding tuple points are changed with the general values. The median partitioning method is used to choose the splitting values for the attributes. It is tested for the acceptable cut in the selected splitting value. If there is an acceptable cut, then the space is divided into 2 areas. This method is repeated till all the areas are inspected. In all areas, there is no acceptable cut. The quasi identifier 'Gender' is categorized into female and male, and then the 'Age' is categorized into 3 levels. Figure 2 shows the taxonomy tree of gender and age. A similar generalization is achieved for all the documents. The utility and privacy are confirmed for each document based on the fitness value.

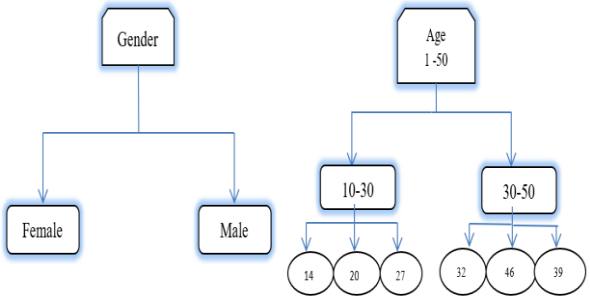

Figure 2. Taxonomy tree of gender and age

In the proposed method a hybrid model of African vulture's optimization, GA and simulated annealing are used to achieve the best convergence rate. To compare the other optimization algorithms, GAVOSA has less convergence time and cost-effective performance. GAVOSA includes three important steps and they are GA-based optimization, AVOA-based optimization and SA-based optimization.
In Equation (16), added the a_3 value of AVOA algorithm. Then GA is used to evaluate the a_3 value is shown in Equation (18).

$$p(t+1) = \frac{a_1 + a_2 + a_3}{3} \qquad (18)$$

A genetic algorithm has five phases and they are selection, initial population, mutation, simulated annealing, crossover and fitness function. With the group of the solution to the issue, the population is initialized. In terms of the fitness value, each fitness chromosomes are determined. Then chromosomes are selected for reproduction. Two chromosomes are selected in the selection process in terms of the fitness value. This is an important step for a_3 calculation. Compared to

the other chromosomes, the selected 'A' and 'B' are the two greatest solutions that have the best fitness value. The new offsprings are created in the crossover operation and these are joined to the population. Simulated annealing is used to improve the fitness value of the proposed method.Equation (19) shows the chromosomes' position.

$$[B_1, B_2] = X_A \otimes X_B \quad (19)$$

Where the chromosomes' position 'A' and 'B' are denoted as X_A and X_B. Then B_1 and B_2 are the offsprings. The population diversity is maintained by the mutation operation to produce two more chromosomes by using random numbers. During the mutation operation, the identified chromosomes are shown in Equation (20).

$$[D_1, D_2] = B_1 \circ B_2 \quad (20)$$

With a higher degree of usefulness, the best information is created and privacy can be maintained by the fitness values optimization. Using GI_L and C_A, the fitness value is calculated.The fitness value is calculated by using the greater degree of utility and privacy. The usefulness is determined with the GI_L and the privacy is determined with the C_A. However the smallest value of GI_L and C_A correspondingly show the maximum utility and privacy, and then the utility and privacy are inversely propositional to GI_L and C_A. The pseudo code of GAVOSA in terms of the anonymization method is shown in algorithm 2.

Algorithm 2: The Pseudo-code for Mondrian k-anonymization using GAVOSA

// Mondrian based k-anonymization
1. Cut dimension is chosen for the broadest attribute
2. To find the center of the chosen attribute as the cut point
3. The dataset is classified into two new subsets from the cut point

// African vultures optimization algorithm (AVOA)
4. Using AVOA tocalculate a_1,a_2

// Genetic Algorithm
5. Perform mutation, selection and crossover for evaluating a_3
6. The better QI number is evaluated by evaluating$(a_1 + a_2 + a_3)/3$
7. For the chosen QI to perform generalization.
8. If the subset is not fulfilling the k-anonymizations privacy models' condition so the above process is repeated

// Simulated Annealing (SA)
9. Simulated Annealing is used to improve the fitness value

Return the anonymized data

In the encryption stage, the symmetric key approach is used. Using the key, the encryption is performed after changing the binary data into the DNA sequence. Crossover, replication and mutation are performed in the genetic algorithm after the encryption stage. Cloud is used to save the encrypted data. In order to get the original file through performing the mutation, decryption and crossover are performed. If the encryption operation is reversed, then the decryption is performed. The cloud data is changed to a DNA order and reshaped in the decryption operation.

5. Experiments

This section details the evaluation matrix and database description.

5.1 Evaluation Metrics

The average equivalence value (C_A), running time, general information loss (GI_L), throughput and fitness value form the evaluation matrix. It is used to analyse the utility and privacy of the introduced approach.

5.1.1 Generalized information loss (GI_L)

GI_L is an effective metric to gain the populations' lowest fitness value. Equation (21) is used to calculate the generalized information loss.

$$GI_L = \frac{1}{Re \times qi} \times \sum_{i=1}^{qi} \sum_{j=1}^{Re} \frac{U_{ij} - L_{ij}}{U_i - L_i} \quad (21)$$

Here, U_i and L_i are represented the ith quasi identifiers' bounds. Re represents the number of records. U_{ij} and L_{ij} represent the generalized values, and interval. The qi represents the quasi identifier.

5.1.2 Average equivalence value metric (C_A)

C_A is an important metric to get the smallest value for forecasting the goal and it is represented in Equation (22).

$$C_A = \frac{Re}{|SA_{q \times qi}| \times k} \quad (22)$$

Where SA indicates the sensitive attribute.

5.1.3 Fitness Function (Fit)

The calculation of the fitness value is used to maintain privacy at each iteration and it is represented in Equation (23).

$$Fit = B * GI_L(A) + D * C_A(A) \quad (23)$$

Where the constant values are B and D. If the fitness value has not improved, the obtained parameters are applied to the simulated annealing method.

5.1.4 Encryption Throughput

The encryption throughput is calculated by utilizing the total plain text and encryption time and it is shown in Equation (24)

$$Encryption\ Throughput = \frac{All\ plain\ text(bytes)}{Time\ of\ the\ Encryption\ (sec)} \quad (24)$$

5.2 Database Description

Breast Cancer Wisconsin (Diagnostic) Data Set, Polycystic Ovary Syndrome (PCOS), Lung Cancer Data Set, Cervical Cancer, and Thyroid Disease Data Set have been used for the experimental calculation of the proposed method.

6. Experimental Results

The experimental result for the introduced PPDP method is discussed in this section.

6.1 The comparison of proposed GAVOSA algorithm with other existing methods

In order to compare the performance of the introduced method GAVOSA with the three traditional methods namely Chimp Optimization Algorithm (ChOA), African vulture's optimization Approach (AVOA), and GA-ChOA, five datasets are used to display the efficiency of the introduced approach. These are Breast Cancer Wisconsin (Diagnostic) Data Set, Polycystic ovary syndrome (PCOS), Lung Cancer Data Set, Cervical Cancer, Thyroid Disease Data Set. For all five datasets, the set of k-values is 5, 10, 15, 20 and 25.

The framework of the proposed GAVOSA approaches first anonymized the data and then encryption is done to improve the efficiency of the algorithms. Table 4 show different existing approaches such as ChOA, AVOA and GA-ChOA compared with the proposed approach to predict the algorithm efficiency with anonymization and without anonymization. The outcome shows that the proposed approaches with anonymization produces better throughput and efficiency than the efficiency without anonymization

Table 4 Comparative analysis with anonymization and without anonymization

Approaches	Anonymization Based Encryption Technique		Encryption Without Anonymization	
	GI_{Loss}	Throughput	GI_{Loss}	Throughput
ChOA	0.5245	4.635	0.8333	3.524
AVOA	0.4819	5.932	0.7666	4.821
GA-ChOA	0.3281	6.421	0.6488	5.319
GAVOSA	0.0634	9.423	0.2166	8.312

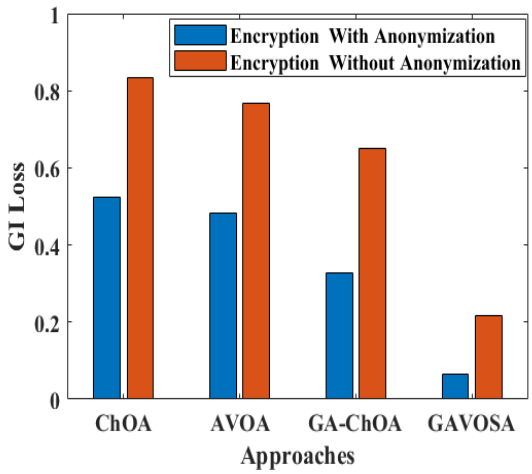

Figure 3. GI_{Loss} with anonymization and without anonymization

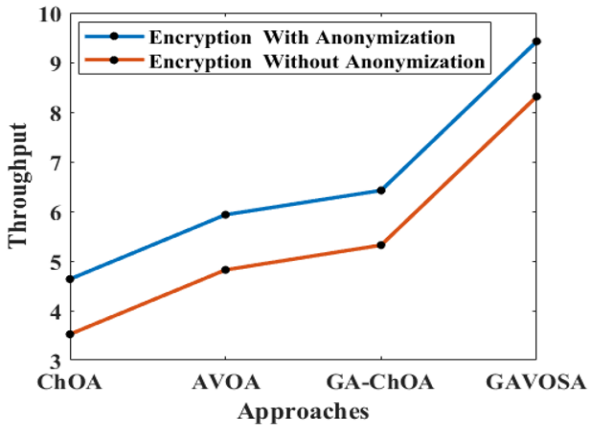

Figure 4. Throughput with anonymization and without anonymization

On analyzing the Figure 3 and 4 the generalized information loss of ChoA strategy is diminished by 37.06%, for AVOA it is diminished by 37.14%,

for GAChoA it is diminished by 49.43% and for GAVOSA it is diminished by 70.73%.

Table 5 depicts the GI_{Loss}, C_{Avg}, Fitness value, Encryption time and Decryption time of the Genetic algorithm with DNA-GA. From the table it is evident that the loss diminishes by 23.57%, C_{Avg} lessens by 11.68%, fitness reduces by 11.06%, the encryption time is lowered by 0.0089s, the decryption time is lowered by 0.0049s and also the total time is diminished by 0.35s.

Table 5 Comparison of all metrics with Genetic algorithm with DNA-GA

Approaches	GI_{Loss}	C_{Avg}	Fit	Encryption Time	Decryption Time	Total Time
GA	0.4205	0.4205	0.7841	0.2235	0.2895	1.87
DNA-GA	0.3214	0.3748	0.6974	0.2146	0.2846	1.52

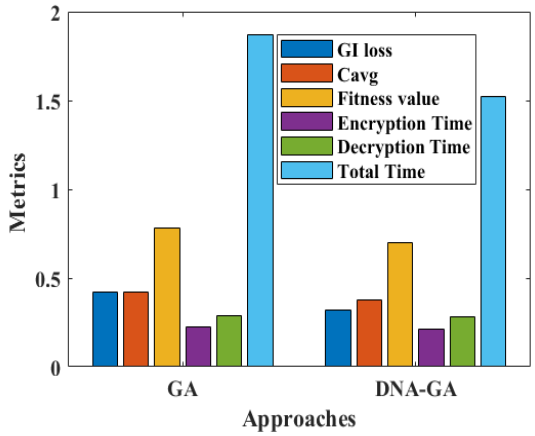

Figure 5. Performance of DNA-GA with GA

Figure 5 shows the performance of DNA-GA compared with genetic approach. While analyzing it is evident that DNA-GA produces better performance than GA

Table 6 depicts the GI_{loss}, C_{Avg}, Fitness value, Encryption time, Decryption time and total time of the SA versus CEN-SA. From the table it is evident that the loss diminishes by 1.80%, C_{Avg} lessens by 42.55%, fitness reduces by 42.1%, the encryption time is lowered by 0.0382s, the decryption time is lowered by 0.0024s and also the total time is diminished by 0.2s.

Table 6 Comparison of all metrics with SA with CEN-SA

Approaches	GI_{loss}	C_{Avg}	Fit	Encryption Time	Decryption Time	Total Time
SA	0.6876	0.3591	0.5237	0.2083	0.2601	1.37
CEN-SA	0.6752	0.2063	0.3031	0.1701	0.2361	1.05

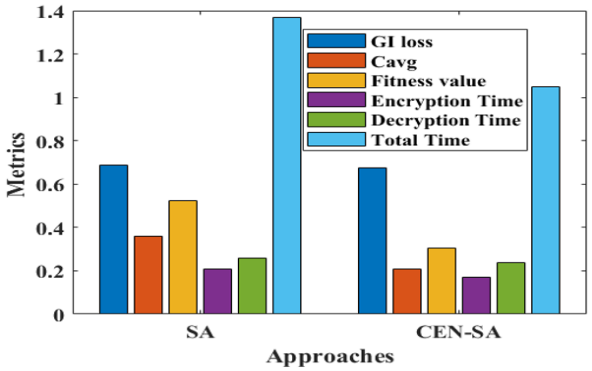

Figure 6. Performance of SA with CEN-SA

Figure 6 illustrates the performance of CEN-SA compared with genetic approach. While analyzing it is evident that CEN-SA produces better performance than SA interms of GI_{loss}, C_{Avg}, Fitness value, Encryption time, Decryption time and total time.

Table 7 depicts the GI_{loss}, C_{Avg}, Fitness value, Encryption time, Decryption time and total time of the AVA with GAVOSA. From the table it is evident that the loss upsurges by 13.57%, C_{Avg} upsurges by 13.57%, fitness reduces by 95.74%, the encryption time is lowered by 0.1443s, the decryption time is lowered by 0.152s and also the total time is diminished by 0.69s.

Table 7 Comparison of all metrics with AVA with GAVOSA

Approaches	GI_{loss}	C_{Avg}	Fit	Encryption Time	Decryption Time	Total Time
AVA	0.3214	0.3214	0.4379	0.1867	0.2459	1.18
GAVOSA	0.365	0.365	0.01865	0.0424	0.0936	0.49

Table 8 depicts the GI_{loss}, C_{Avg}, Fitness value, Encryption time, Decryption time and total time of the DNA-GA with GAVOSA. On analysing the table it is evident that the loss lessens by 49.57%, C_{Avg} diminishes by 2.61%, fitness reduces by 97.74%, the encryption time is lowered by 0.1722s, the decryption time is lowered by 0.191s and also the total time is diminished by 1.03s.

Table 8 Comparison of all metrics with DNA-GA with GAVOSA

Approaches	GI_{loss}	C_{Avg}	Fit	Encryption Time	Decryption Time	Total Time
DNA-GA	0.2214	0.3748	0.6974	0.2146	0.2846	1.52
GAVOSA	0.365	0.365	0.01865	0.0424	0.0936	0.49

Table 9 depicts the GI_{Loss}, C_{Avg}, Fitness value, Encryption time, Decryption time and total time of the CEN-SA with the proposed GAVOSA. On analyzing the table it is evident that the loss lessens by 45.57%, C_{Avg} diminishes by 76.93%, fitness reduces by 93.85%, the encryption time is lowered by 0.12s, the decryption time is lowered by 0.14s and also the total time is diminished by 0.56s.

Table 9 Comparison of all metrics with CEN-SA with GAVOSA

Approaches	GI_{Loss}	C_{Avg}	Fit	Encryption Time	Decryption Time	Total Time
CEN-SA	0.6752	0.2063	0.3031	0.1701	0.2361	1.05
GAVOSA	0.365	0.365	0.01865	0.0424	0.0936	0.49

On analyzing all the three approaches given in Table 7, 8 and 9 it has been determined that the suggested GAVOSA strategy yields superior results to the current methods.

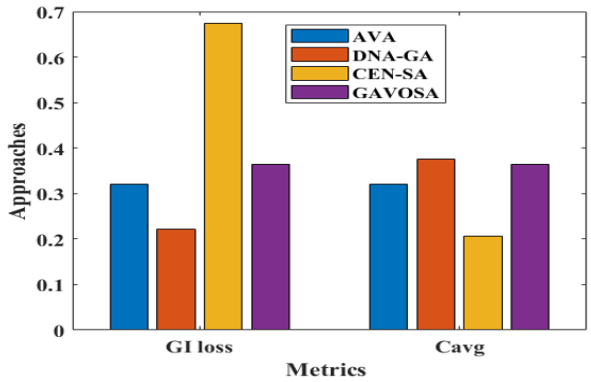

Figure 7. Performance of GAVOSA interms of GI_{Loss} and C_{Avg}

Figure 7 compares the GI_{Loss} and $C_{Avg\ of}$ the proposed GAVOSA and other three existing algorithms AVA, DNA-GA and CEN-SA. From the analysis it is proved that the proposed approach produces better outcome than the existing approach.

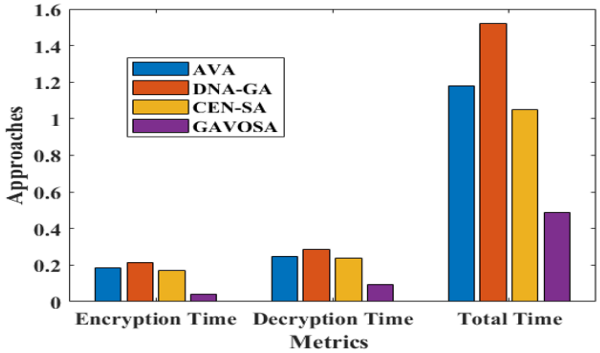

Figure 8. Performance of GAVOSA interms of Encryption time, Decryption time and total time

Figure 8 compares the encryption time, decryption time and total time of the proposed GAVOSA and other three existing algorithms AVA, DNA-GA and CEN-SA. Finally while taking the mean of all the approaches it is evident that the proposed GAVOSA approach produces better results than the existing approaches.

7. Conclusion

In this paper, the cost-effective data publishing method is introduced to protect the privacy and security of the medical databases. To anonymize the databases, the effective Mondrian based k-anonymization method is used with GAVOSA. In this paper SA is employed to update the position of the AVOA. To deliver the extra security for the anonymized data the DNA-GA based encryption method is used. Based on the Average Equivalence Value (C_A) and Generalized Information Loss (GI_L), to measure the utility and privacy. To calculate the proposed PPDP method by using the Breast Cancer Wisconsin Dataset, Polycystic Ovary Syndrome, Lung Cancer Dataset, Cervical Cancer, and Thyroid Disease Dataset. To analyze the proposed methods' security, the DNA-GA algorithms' encryption throughput is compared with other encryption methods. When compared to the current methods with developed privacy and utility, the proposed method gives better performance. The proposed method is giving better results by comparing with the average time taken with all other approaches. In future work we will enhance the security to secure the sensitive information from the third party attackers.

REFERENCES

Abdul Majeed;Sungchang Lee; (2021). Anonymization Techniques for Privacy Preserving Data Publishing: A Comprehensive Survey. IEEE Access, (), –. doi:10.1109/access.2020.3045700

BibalBenifa, J. V., &Venifa Mini, G. (2020). Privacy Based Data Publishing Model for Cloud Computing Environment. Wireless Personal Communications. doi:10.1007/s11277-020-07320-3.

Chen, Z. G., Kang, H. S., Yin, S. N., & Kim, S. R. (2016). An efficient privacy protection in mobility social network services with novel clustering-based anonymization. EURASIP journal on Wireless communications and networking, 2016(1), 1-9.

Kabir, M., Wang, H., &Bertino, E. (2011). Efficient systematic clustering method for k-anonymization. ActaInformatica, 48(1), 51-66.

Kundalwal, M. K., Chatterjee, K., & Singh, A. (2019). An improved privacy preservation technique in health-cloud. ICT Express, 5(3), 167–172.

Lagendijk, R. L.; Barni, M. (2013). Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation. IEEE Signal Processing Magazine, 30(1), 82–105. doi:10.1109/MSP.2012.2219653

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). Mondrian multidimensional K-anonymity, Proc.-International Conference on Data Engineering, 25. doi: 10.1109/ICDE.2006.101.

Li, J., Zhang, Y., Ning, J., Huang, X., Poh, G. S., & Wang, D. (2020). Attribute based encryption with privacy protection and accountability for CloudIoT. IEEE Transactions on Cloud Computing.

Lokesh, S., &Ranganatha H R. (2020). Review on privacy preservation methods in cloud computing. International Journal of Future Generation Communication and Networking, 13(3), 3990-4002.

M. ErcanNergiz; Chris Clifton (2007). Thoughts on k-anonymization. , 63(3), 622–645. doi:10.1016/j.datak.2007.03.009

McCall, J. (2005). Genetic algorithms for modelling and optimization, Journal of Comput. Appl. Math. 184(1), 205–222.

Mousa HM. (2016). DNA-Genetic Encryption Technique, International Journal of Computer Network and Information Security.8(7).1-9.10.5815/ijcnis. 2016. 07.01.

Murthy, Suntherasvaran; Abu Bakar, Asmidar; Abdul Rahim, Fiza; Ramli, Ramona (2019). [IEEE 2019 IEEE 5th Intl Conference on Big Data Security on Cloud (BigDataSecurity), IEEE Intl Conference on High Performance and Smart Computing, (HPSC) and IEEE Intl Conference on Intelligent Data and Security (IDS) - Washington, DC, USA (2019.5.27-2019.5.29)], doi:10.1109/BigDataSecurity-HPSC-IDS.2019.00063

Romanou, A. (2018). The necessity of the implementation of privacy by design in sectors where data protection concerns arise. Computer Law and Security Review, 34(1), 99–110.

Shah, M.A., Swaminathan& R., Baker, M. (2008). Privacy-preserving audit and extraction of digital contents. IACR Cryptol, 186.

Sun, PanJun (2020). Security and privacy protection in cloud computing: Discussions and challenges. Journal of Network and Computer Applications, 160(), 102642–. doi:10.1016/j.jnca.2020.102642

Tahir, M., Sardaraz, M., Mehmood, Z., & Muhammad, S. (2020). CryptoGA: a cryptosystem based on genetic algorithm for cloud data security. Cluster Computing. doi:10.1007/s10586-020-03157-4.

Tiancheng Li; Ninghui Li; Jian Zhang; Molloy, I. (2012). Slicing: A New Approach for Privacy Preserving Data Publishing. , 24(3), 561–574. doi:10.1109/tkde.2010.236

Weijia Yang; SanzhengQiao (2010). A novel anonymization algorithm: Privacy protection and knowledge preservation. , 37(1), 756–766. doi:10.1016/j.eswa.2009.05.097

Zang, Wenke; Ren, Liyan; Zhang, Wenqian; Liu, Xiyu (2017). A cloud model based DNA genetic algorithm for numerical optimization problems. Future Generation Computer Systems, (), S0167739X17304697–. doi:10.1016/j.future.2017.07.036

Zang, W., Ren, L., Zhang, W., & Liu, X. (2018). A cloud model based DNA genetic algorithm for numerical optimization problems. Future Generation Computer Systems, 81, 465–477. doi:10.1016/j.future.2017.07.036.

Zhan, Z.H., Liu, X.F., Gong, Y.J., Zhang, J., Chung, H.S.H., Li, Y. (2015). Cloud computing resource scheduling and a survey of its evolutionary approaches. ACM Comput. Surv. 47(4), 63.

Breast Cancer Wisconsin Data Set, https://www.kaggle.com/datasets/uciml/breast-cancer- wisconsin-data/code.

Polycystic ovary syndrome, https://www.kaggle.com/dataset/prasoonkottarathil/polycystic-ovary-syndrome-pcos.

Lung Cancer Data Set, https://archive.ics.uci.edu/ml/datasets/lung+cancer.

Cervical Cancer, https://www.openml.org/d/42912.

Thyroid Disease Data Set, https://www.kaggle.com/datasets/yasserhessein/thyroid- disease-dataset

Sahana Lokesh R. and H.R. Ranganatha (2022). An Enhanced Data Anonymization Approach for Privacy Preserving Data Publishing in Cloud Computing Based on Geneticza Chimp Optimization. International Journal of Information Security and Privacy/1096 16(1) DOI: 10.4018/IJISP.300326.

Zang, W., Ren, L., Zhang, W., & Liu, X. (2018). A cloud model based DNA genetic algorithm for numerical optimization problems. Future Generation Computer Systems, 81, 465–477. doi:10.1016/j.future.2017.07.03