Received: October 30, 2017

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

A Combined Clustering and Geometric Data Perturbation Approach for Enriching Privacy Preservation of Healthcare Data in Hybrid Clouds

Vulapula Sridhar Reddy[1]* **Barige Thirumala Rao[1]**

[1]*Department of Computer Science and Engineering, Koneru Lakshmaiah University, Vijayawada, India*
* Corresponding author's Email: vsridhar0318@gmail.com

Abstract: In this paper, we plan a combined clustering and geometric data perturbation approach for improving privacy preservation of health care data in hybrid clouds. We will possibly plan an answer that can productively give protection to information put away in the cloud without presenting substantial overhead on both computation and communication. At first, the high-dimensional data are separated into various parts by utilizing the K-mean clustering method, each partition is considered as a cluster. At that point the mean estimate of each cluster processed; after that the contrast between the each cluster member and the mean of the cluster value is computed. In the next stage, the clustered information is perturbed by utilizing the Geometric Data Perturbation (GDP) algorithm which makes the values difficult to be recognized. These perturbed values are stored in the public cloud and the key parameters for randomizing and clustering is stored in the private cloud. Our approach would contribute in reduction of too much storage on private cloud on the off chance that we basically store the entire sensitive information on private clouds. The experimental results show that, the GDP algorithm has better privacy preserving compared with the other existing methods.

Keywords: Hybrid cloud, Privacy preserving, K-mean clustering, GDP (Geometric data perturbation).

1. Introduction

In recent year, cloud computing technology perfectly matches such "big data" challenges by providing nearly unlimited storage resources on demand [1, 2]. In health care, it is also gaining significant popularity by facilitating an inter-organizational medical data sharing environment [3]. The expansion of cloud computing services also empowers hospitals and institutions to effortlessly export their health care information to the cloud, which gives omnipresent access to the information and on-request data administration easily [4]. Currently, many cloud service providers (CSPs), including Box, Microsoft, Verizon and Dell, have announced their support for this business associate agreement [5, 6].

The cloud services also involve many security and privacy risks that lead to concerns among patients and medical workers [7-9] who are being particularly afraid of losing control over sensitive medical records while storing them on not fully trusted third-party servers [11-12]. Most existing solutions (e.g., [13-16]) employ encryption/ decryption techniques combined with access control and auditing system to provide security and privacy for data stored on a public cloud. Different approaches are used to outsource health care data. They are; High-dimensional sensitive healthcare attributes personalized protection at attribute level, Collusion resistance etc. [17, 18].

To satisfy the above practical requirements, the existing techniques provide a privacy preserving framework to outsource healthcare data to a hybrid cloud [19, 20]. The hybrid cloud is a network configuration that includes a combination of internal hardware, private cloud resources and public cloud capabilities. The hybrid cloud still suffers from the same privacy and security issues that plague the popular perception of public cloud platform providers. In addition, hybrid cloud as well as public cloud is a poor fit for the circumstances in which

data transport on both ends of the cloud is a mission critical operation that is sensitive to the delay from transporting data across a network and the latency in ping times.

In this work, our goal is to enrich a privacy preservation of health care data in the hybrid cloud without introducing a computation and communication overhead between the private cloud and the public cloud. Here, initially the K-mean clustering algorithm is used for partitioning the high-dimensional data, each partitioning is considered as a cluster. The GDP algorithm is used to perturb the clustering data, these perturbed values are hard to be recognized. The public cloud stores the perturbed data and the key parameters for randomizing technique and the clustering techniques are stored in a private cloud. Then the data retrieval technique is used to recover the original data from the perturbed data.

The rest of the segment of the paper is depicted in the section underneath. In section 3.1, the data privacy preservation is depicted, in section 3.2, the clustering technique is explained. The GDP algorithm is delineated in section 3.3. The test result and the conclusion are examined in section 4 and 5.

2. Related work

Numerous research works have previously existed in literature which was based on the RDF security based access control techniques and schemes. Some of the works are reviewed here.

Wei Wang *et al.* [21] have proposed a privacy-preserving framework to transit insensitive data to the commercial public cloud and the rest to trusted private cloud. Nonetheless, the encryption leads to large overhead when answering queries. Hui Zhang *et al* .[22] have introduced a hybrid cloud computing model in which users may adopt as a viable and cost saving method to make the best use of public cloud services. But in that infrastructure, workload factoring happens between a primary server and proxy servers. Jin Li *et al.* [23] have proposed the convergent encryption technique to encrypt the data before outsourcing. But this technique introduced significant cost when applied to healthcare data with high-dimensional sensitive attributes. Jingwei Li *et al.* [24] have aimed at tackling the challenge of privacy-preserving utilization of data in cloud computing. Furthermore, the cryptography method is a quite expensive, primitive and the data owner may not execute it easily.

Xuyun Zhang *et al.* [25] have proposed an efficient quasi-identifier index based approach to ensure privacy preservation and achieve high data utility over incremental and distributed data sets in the cloud. Nonetheless, this scheme only works for encrypted files and it suffers from the auditor statefulness and bounded usage, which may potentially bring in on-line burden to users when the keyed hashes are used up. Obviously, this is a costly technique. Kui Ren *et al.* [26] have proposed a secure cloud storage system supporting privacy-preserving public auditing. This scheme also suffers from data exploration problem. H. Liu *et al.* [27] have introduced a shared authority based privacy-preserving authentication protocol (SAPA) to address privacy issues for cloud storage. Their approaches also suffer from poor scalability and inefficiency because they are centralized and access all data frequently when an update occurs.

3. Proposed method

In this work, our goal is to design a solution that can efficiently provide privacy to data stored in the cloud without introducing large overhead on both computation and communication. Here, initially the high-dimensional dataset is divided into multiple partitions and each of the sensitive attributes is considered as a cluster. On the following stage, an optimal cluster head is selected by using a cluster head selection technique and values are computed based on the relation between each of the cluster members and the cluster head. In the next stage of our data privacy, the geometric data perturbation technique is applied to the previously computed values which make the values hard to be recognized. The perturbed values of each sensitive attribute is then stored in the public cloud whereas all the key parameters used for clustering and randomizing are kept in the private cloud. Our approach would contribute in reduction of too much storage in a private cloud if we simply store the entire sensitive information on private clouds. Other contributions include the reduction of communication overhead between private and public cloud and the delay introduced by communications between private and public cloud.

3.1 Data privacy preservation via hybrid cloud

Fig. 1 demonstrates the architecture of hybrid cloud. The real data comes from the private cloud and these real data is partitioned by using the K-mean clustering algorithm. In the next stage of our data privacy, the geometric data perturbation technique is applied to the previously computed values which make the values hard to be recognized. Then the perturbed values are stored in the public cloud and the key parameters used for clustering and

the GDP methods are stored in the private cloud. The database contains multiple sensitive pieces of information. It incorporates emergency information, medications, immunization, allergies and other health information. Here, every data contain an immense measure of information, in this way, the computation and the communication overhead have happened. Here, at first, the high dimensional dataset is separated into multiple groups based on their similarity function by utilizing the clustering procedure. The K-mean clustering calculation gave the preferable outcome over other existing techniques.

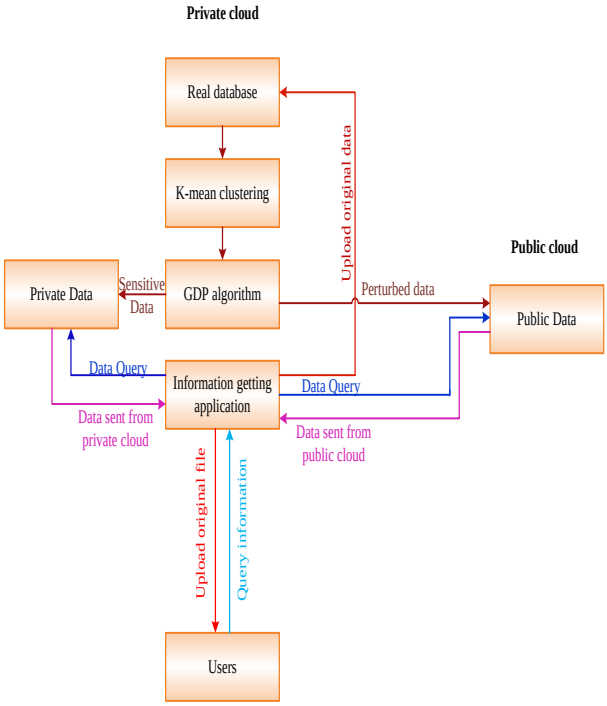

Figure.1 Hybrid cloud architecture

3.2 K- mean clustering

The K-mean clustering algorithm used to partition the n-number of items into k-groups. The protest partitioning depends on the similitude capacity of the object. In the clustering technique, accept the given arrangement of members q_1, q_2, ….,q_n each member has private information a_1, a_2, …, a_n. At first, consider the cluster value k, then apportioning the k data randomly to appraise the underlying centroids for the k-groups. At that point, consider that each information indicates and relegates it to the cluster which is nearest. Recalculate cluster center by discovering the mean of data points having a place with a similar cluster, then backpedal to the past two stages and rehash the procedure until the group part at no time in the future changes or a most extreme number of emphases is reached. At long last, the square error criterion is computed to isolate the k cluster comes about.

Input: Data Data $p= \{a_1, a_2,\ldots,a_n\} \subset D^r$, assume the number of clusters $k \in N$

Initialization: The random cluster centers are $m_1, m_2 \ldots m_k \in D^r$.

Repeat: Find closest cluster centers $\nabla \alpha \in P$
Update cluster centers $m_1, m_2 \ldots m_k$

$$m_j = \sum_{a_i \in M_j} a_i / \left| M_j \right| \tag{1}$$

Where, a_1 is the cluster point, M_j is the total number of points and m_j *is* the initial centroid of the points.

Until: no change in cluster centers

Output: The squared error criterion is reduced by a set of output k cluster centers $m_1, m_2 \ldots m_k$.

The squared error criterion is given by,

$$S_e = \sum_{j=1}^{k} \sum_{a_i = M_j} \left| a_i - m_j \right|^2 \tag{2}$$

Where, S_e is the squared error, m_j is the initial centroid of the cluster, a_i is the cluster point.

The original health care databases that are put away in arbitrary request appeared in Table 1. At first consider the cluster esteem k=3 and the underlying cluster mean is m_1 = 23, m_2 = 33, m_3 = 48, so the high-dimensional medicinal services information are changed over into three gatherings in light of their comparability work, K1= {23, 25, 28, 30}, K2 = {33, 32, 35, 36} and K3 = {48, 42, 46, 50}. On the following stage, we need to process the centroid of each cluster gathering, K1= (23, 25, 28, 30)/4= 26.5, K2 = (33, 32, 35, 36) /4= 34 and K3 = (48, 42, 46, 50)/4=46.5. Yet, we can't be certain that the cluster partitioning is correct or off-base. Along these lines, we need to ascertain the separation between the each point and its own particular cluster mean and the inverse group mean. Here the point 30 has less separation in the inverse group (K2) than its own cluster distance (K1). Presently the underlying grouping is changed, the cluster point 30 is changed

Table 1. Original health care database

No	Age	Sex	Height	Weight	Disease
1	23	M	162	53	Rabies
2	25	M	165	58	CP
3	32	M	172	67	Dengue fever
4	35	F	164	69	Rabies
5	28	M	167	59	Viral infection
6	30	F	156	72	I
7	36	F	180	75	CP
8	48	M	158	62	Bronchitis
9	33	M	163	76	Dengue fever
10	42	M	170	50	Viral infection
11	46	F	157	48	Dengue fever
12	50	M	182	68	CP

*M= Male, *F= Female, *CP= Chickenpox, *I= Indigstion

toward the next cluster group K2. In this way, now the grouping is K1= {23, 25, 28}, K2 = {30, 33, 32, 35, 36} and K3 = {48, 42, 46, 50}. Recalculate the cluster center until there is no adjustment in the group gathering.

3.3 Geometric data perturbation (GDP) technique

In our paper, a Geometric Data Perturbation (GDP) technique is used for randomizing the previously computed cluster values. The data utility and privacy guarantee are well preserved by the GDP algorithm. This GDP algorithm comprises of random geometric transformation, incorporates multiplicative transformation (*T*), translation transformation (*Vs*) and distance perturbation Ω.

$$GDP(P) = TP + Vs + \Omega \tag{3}$$

Where,
$$\Omega = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{4}$$

Where, *T* is the random projection matrix, *P* is the original data, *Vs* is the translation matrix, Ωis the random Guassian noise.

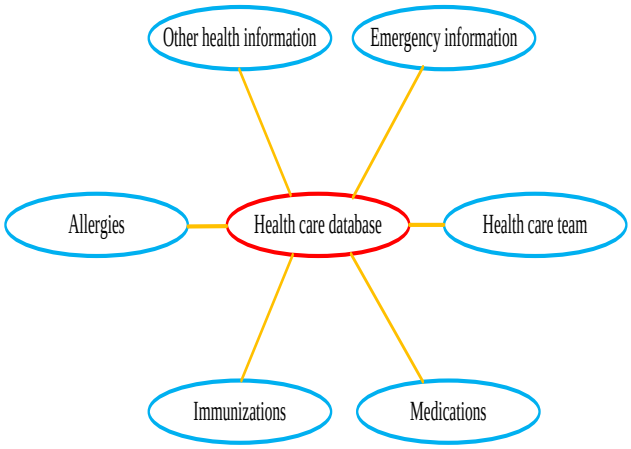

Figure.2 Health care database

Table 2. K-mean clustering for Table 1

No	Cluster groups	Age	Sex	Ht	Wt	Disease
1	K_1	23	M	162	53	Rabies
2		25	M	165	58	CP
3		28	M	167	59	Viral infection
4	K_2	30	F	156	72	Indigestio
5		3	M	163	76	Dengue fever
6		32	M	172	67	Dengue fever
7		35	F	164	69	Rabies
8		36	F	180	75	Chickenp
9	K_3	48	M	158	62	Bronchitis
10		42	M	170	50	Viral infection
11		46	F	157	48	Dengue fever
12		50	M	182	68	CP

*Ht= Hight, *Wt= Weight

Here, the health care service contains multiple pieces of information about the patients. It incorporates emergency information; medications, immunization, allergies and other health information are shown in Fig 2. Every data in the database has different attributes. The same area information is stored in the separate table.

The perturbation technique takes the data from any one of the tables to randomize the data. The GDP algorithm includes random translation perturbation, random rotation perturbation and noise addition to minimize the high dimensional dataset.

3.3.1. Random rotation transformation

To change the default, adjust the template as follows. The GDP algorithm only perturbs the

Table 3. Random rotation transformation values

Age	Sex	Ht	Wt	Disease
1748	Male	55679	9010	Rabies
1900	Male	81510	9860	Chickenpox
2128	Male	82498	10030	Viral infection
4980	Female	130260	25848	Indigestion
5478	Male	136105	27284	Dengue fever
5312	Male	143620	24053	Dengue fever
5810	Female	136940	24771	Rabies
5976	Female	150300	26925	Chicken pox
8928	Male	105386	14136	Bronchitis
4032	Male	113390	11400	Viral infection
8556	Female	104719	10944	Dengue fever
9300	Male	121394	15504	Chickenpox

numerical data. Along these lines, at first, the original data is changed over into numerical data by utilizing the American Standard Code for Information Interchange (ASCII). After numerical transformation, the values are accessible in the *P*. At that point the random matrix *T* is gotten, in light of the span of the numerical data esteem. In the event that the random matrix *T* is positive, we can rotate the matrix in against clockwise course otherwise rotate them in clockwise bearing. At that point multiply the real data matrix *P* and the random matrix *T*.

The random rotation transformation esteems are secured on the table 3. At first, consider any one of the attributes (age or height or weight) cluster values k_1 = {23, 25, 28}, k_2= {30, 33, 32, 35, 36} and k_3= {48, 42, 46, 50}. At that point the clustering data are rotated in 180^o clock astute bearing, k_1= {28, 25, 23}, k_2= {36, 35, 32, 33, 30} and k_3= {50, 46, 42, 48}. Then the rotated data is multiplied with the original dataset, so the rotation transformation esteems are k_1= {1748, 1900, and 2128}, k_2= {4980, 5478, 5312, 5810, 5976}, k_3= {8928, 4032, 8556, 9300}. A similar procedure is rehashed though the delicate properties in the health care data.

3.3.2. Translation transformation

In the translation transformation, one consistent number (on the off chance that it is sure or negative) is added to the original dataset. In the view of the original data *P*, the Gaussian noise Ω is introduced. At last, we add the product of original data *P* and the rotation matrix *T*, translation transformation matrix V_s and the random Gaussian noise Ω. At long last, the perturbed data *GDP* (*P*) is created. These perturbed data are stored in the public cloud and the perturbation techniques are put away in the private cloud.

The perturbed healthcare datasets are appearing in Table 4. After the random rotation transformation, the translation transformation matrix is shaped by including one steady number (c=3) to the original dataset. In the wake of finding the translation matrix, the random Gaussian noise is presented. At long last, every one of the qualities is (random rotation esteems, translation transformation esteems) added together to form the perturbed data. These perturbed data is difficult to be perceived by the client. The user does not see the original data without the data proprietor's consent. On the off chance that the clients require original data, they send the question demand to the information proprietor. The information proprietor offer authorization of the clients (restorative insurance agencies or associations and so forth.) to get to the information in view of their accreditations joins secret word, individual distinguishing proof number and so on. In the access control method, the systems already have the credential information about the patients. Initially, the patient's accreditation snippets of the data are given to the control panel. The control panel analyzes the information to an officially existing information list. In the event that the data in the access control system and the control panel is not equivalent, the user permission is denied to accept the data. If the data are equal the permission is granted to accept the data. Nevertheless, the control system does not give the permission to the users to accept all the data about the patient, only the users obtain certain data, all other data are perturbed.

3.4 Recovering data

Data recovering is very important in the information technology. In our system, the health care database owner gives permission to the three types of users to accept the original database. They are users, insurance companies and physician. But they do not get the all original data about the patients, they only get certain data based on the user's requirements and all other data are perturbed. When the data is queried, the request is sent to both the private cloud and the public cloud at the same time. Our system contains two databases; they are private data and the public data. The private data contains the original data and the perturbed data, the public cloud contains the perturbed data only.

Table 4. Perturbed data by using GDP algorithm

Age	Sex	Ht	Wt	Disease
1774	Male	55844	9066	Rabies
1928	Male	81678	9921	Chickenpox
2159	Male	82668	10092	Viral infection
5013	Female	130419	25923	Indigestion
5514	Male	136271	27363	Dengue fever
5347	Male	143795	24123	Dengue fever
5848	Female	137107	24843	Rabies
6015	Female	150483	27003	Chicken pox
8979	Male	105547	14201	Bronchitis
4077	Male	113563	11453	Viral infection
8605	Female	104879	10995	Dengue fever
9353	Male	121579	15575	Chickenpox

Initially, the user sends the request query to the system. This query request checks the private cloud data and takes the perturbed value of that. This perturbed value checks the public cloud data and retrieves all the attributes in that row. The GDP algorithm includes data perturbation phase and the data retrieval phase. Input: Original data *P*, its size *n* and delicate characteristic [*V*].

3.4.1. Data perturbed phase`

Intermediate result: Perturbed data set *P'*

Output: Clustering results *k* and *k'* of dataset P and $P^{'}$ respectively.

Steps:

a) Given the input dataset *P*, its tuple estimate n and the relating sensitive attribute $[V]_{n\times1}$
b) Sensitive attribute $[V]_{n\times1}$ is rotated in 180º clock-wise course, so the random rotation matrix $[T]_{n\times1}$ is generated.
c) The result of $[T]_{n\times1}$ and the $[V]_{n\times1}$ is obtained in step 3. The duplicated esteems will be, $[X]_{n\times1} = [T]_{n\times} \times [V]_{n\times1}$.
d) Compute the translation transformation matrix [*s*] as mean of sensitive attribute $[V]_{n\times1}$
e) Generate transformation $[Vs]_{n\times1}$ by applying the transformation matrices to $[V]_{n\times1}$.
f) Gaussian distribution Ω as a probability density function for Gaussian noise computed in eqn . 4
g) Now the perturbation data is *GDP* (*P*) = $[X]_{n\times1} + [Vs]_{n\times1} + \Omega$.
h) The perturbed dataset $P^{'}$ is created by supplanting attributes$[V]_{n\times1}$ in original dataset *p*with $[GDP\ (P)]_{n\times1}$
i) Apply k- means clustering algorithm with various estimation of [*k*] on original dataset [*P*] having sensitive attribute [*V*]
j) Apply *k* means clustering algorithm with various estimation of k on perturbed dataset *P'* having transformed sensitive attribute *GDP* (*P*)
k) Create cluster membership matrix of results from step 9 and step 10 and dissect.

3.4.2. Data retrieval phase

Input: Perturbed data $P^{'}$, sensitive attribute [*V*]

Intermediate result: Geometric data perturbation of sensitive attributes *GDP* (*P*).

Output: Original clustering data P of the perturbed data $P^{'}$

Steps:

a) Given the perturbed dataset $[P^{'}]$, its tuple estimate *n* and the relating sensitive attribute $[V]_{n\times1}$
b) Sensitive attribute $[V]_{n\times1}$ is rotated in 180º anti clock-wise direction, so the random rotation matrix $[T']_{n\times1}$ is generated.
c) TThe result of $[T']_{n\times1}$ and the $[V]_{n\times1}$ is obtained in step 3. The duplicated esteems will be, $[X^{'}]_{n\times1} = [T']_{n\times} \times [V]_{n\times1}$.
d) Compute the translation transformation matrix $[s^{'}]$ as mean of sensitive attribute $[V]_{n\times1}$
e) Generate transformation $[Vs^{'}]_{n\times1}$ by applying the transformation matrices to $[V]_{n\times1}$.
f) Compute Gaussian distribution $\Omega^{'}$
g) Now the result data is $GDP(P^{'}) = [X^{'}]_{n\times} + [Vs^{'}]_{n\times1} + \Omega^{'}$.
h) The original dataset *P*is created by supplanting attributes $[V]_{n\times1}$ in perturbed dataset $p^{'}$with $[GDP\ (P)]^{'}_{n\times1}$

4. Experimental results and discussion

In this section, the evaluation results of the proposed method are described and the performance of the proposed GDP algorithm is compared with the existing ElGamal's encryption algorithm. The GDP technique can be best utilized for perturbing for medical big data by clustering the whole data set utilizing K-means clustering and perturbing the

clustered data utilizing GDP strategy. The proposed GDP is more interpretable and can quickly manage the huge number of markers and the component hugeness can be surveyed in the midst of getting ready for insignificant additional computation. The privacy preserving in a hybrid cloud incorporates the data partitioning, random rotation transformation, and translation transformation to perturb the first data. The proposed approach would add to the lessening of an excess of capacity in private clouds in the event that we just store the whole sensitive data on private clouds. Different commitments incorporate the lessening of communication overhead between private and public cloud and the postponement presented by interchanges amongst private and public cloud.

4.1 Datasets description

For the examination purposes, we used the real time database and it has appeared in Table 1. This health care data contain 12 information about the patients. In reality health care dataset, each record fuses individual data and medicinal services checkup of a person. Here, the individual data about the patient is viewed as insensitive, while the social insurance checkup data about the patient is viewed as sensitive. Here, the perturbation is connected to the sensitive attributes (age, height, and weight) in the health care dataset. The table in the underneath areas (table 5) demonstrates the experimental results of Elgamal's encryption algorithm contrasted and the aftereffects of the proposed framework.

4.2 Privacy evaluation

In this area, the sensitive data about the patient is effectively impeded in light of the degree of privacy protection. The distinction between the first data values and the perturbed data values (known as Var) estimation is used to process the degree of privacy protection of settled data. The degree of privacy protection is calculated by,

$$\frac{Var(P-P')}{Var(P)}=\frac{\sigma^2(P-P')}{\sigma^2 P} \tag{5}$$

In the above equation, the variance function is represented as σ. Here dividing Var (X) is to enable the calculation results not to be impacted by the size range of the properties. Table 5 shows the privacy of the geometrical data transformation technique which includes four methods to protect the privacy, they are RFDP, SDP, TDP and RDP.

Table 5. Evaluating privacy of geometric data perturbation

Method	**Evaluating Privacy**
Reflecting Data Perturbation (RFDP)	0.7125
Scaling Data Perturbation (SDP)	0.7236
Translation Data Perturbation (TDP)	0.7596
Rotation Data Perturbation (RDP)	0.7498

4.3 Accuracy of perturbation results

To compute the accuracy of our proposed method, the same K-mean clustering algorithm is applied to the original data and the perturbed data. Table 6, show the accuracy of the existing KNN classifier algorithm and it is compared to our proposed Geometric data perturbation method. GDP algorithm is quicker to prepare and has less parameters additionally it is more interpretable as it plots the specimen vicinities and imagines the yield. From the table, obviously as the p esteems have expanded the accuracy is modified and consequently it is more productive than the officially existing frameworks and along these lines diminishes the many-sided quality of huge data to the best degree. The accuracy level of GDP algorithm is expanded when contrasted with the current KNN classification algorithm.

The accuracy of the perturbation result is calculated by,

$$Acc=\frac{1}{N}\sum_{i=1}^{k}(|Cluster_i(P)|-|Cluster_i(P')|) \tag{6}$$

In the above equation N is the number of original dataset, k is the number of clustering group, $|Cluster_i(P)|$ is the number of data in the original dataset, $|Cluster_i(P')|$ is the number of data in the perturbed dataset. In the geometric data perturbation algorithm, the accuracy of the perturbation result is more in light of the fact that the proposed approach used the translation, rotation, and reflection isometrics discover the accuracy of mining aftereffects of random reaction. So the proposed strategy has the preferable accuracy over the current perturbation techniques.

Table 6. Accuracy of perturbation results

KNN classifier [28]	Our proposed algorithm
66.23	65.89
66.56	71.25
73.33	74.59
67.82	79.58

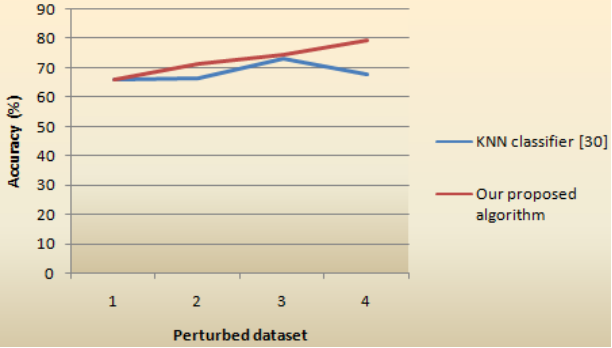

Figure.3 Accuracy comparisons for existing and proposed method

Table 7. Data privacy ratio and iteration results of our proposed algorithm.

Data Privacy Ratio and Iteration Results of Proposed Method	
Data Privacy Ratio	**Iteration Results**
10.588	0.7009
11.998	0.7129
16.875	0.7428
18.698	0.7596

Table 8. Time taken for encrypt the health care data of elgamal's encryption method is compared with our proposed method

S.No.	Record size	Time taken to encrypt the health care data (AES) (ms) [29]	Time taken to perturb the healthcare data (GDP algorithm) (ms)
1	3	112842	110882
2	5	122689	121987
3	7	148785	139887
4	8	168545	161558

4.4 Data privacy ratio and iteration results

In Table 7, the iteration result and the data privacy ratio of the proposed method is shown. Different perturbation methods are used for different privacy preserving recommendations. The proposed GDP algorithm has the best execution, on the grounds that, the proposed privacy preservation technique has a superior data privacy ratio, so it is significantly more effective than the current strategy. The GDP strategy has less iteration to perturb the first health care data when contrasted and the current technique and furthermore the GDP strategy is quicker than the current strategy, since it has less emphasis, while existing technique comprises of many rounds of encryption.

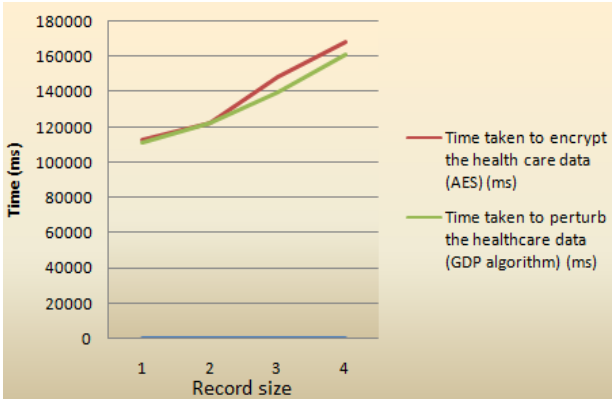

Figure.4 Time taken for encrypt the health care data of elgamal's encryption method is compared with our proposed method

4.5 Comparison analysis

In this section, we compare the performance of the proposed method with the existing Advanced Encryption Standard (AES) algorithm. Here, the proposed work contrasts the running time GDP algorithm and the current AES technique. We measure the time taken for perturbation technique by utilizing different quantities of records. The record estimate 3, 5, 7 and 8 is taken in the x-axis and the time utilization for encryption utilizing AES and perturbation utilizing GDP is going up against the y-axis.

Table 8 shows the time taken for encrypting the healthcare data by using AES algorithm methods and time taken to perturb the healthcare data by using the GDP algorithm. Here, we can see that the time taken for encrypting the data is expanded when the record size is bigger. For all the record sizes, the proposed GDP algorithm is ordinarily speedier than the current AES technique, in light of the fact that, the proposed strategy has no iteration to perturb the data however the current AES technique takes many rounds. Thus, the proposed strategy gives data privacy and it is more effective than the AES technique.

5. Conclusion

To upgrade the privacy preservation of health care data in the hybrid cloud, the existing methodologies have presented distinctive types of strategies. But, these strategies do not completely reduce the storage and computation overhead and the delay is also introduced. Some of the existing approaches are not suitable for particular operations (ex: healthcare data outsourcing and also suffer from data exploration, problem and also suffer from poor scalability and inefficiency since they are unified and get to all information frequently when an update happens. To overcome the above difficulties the proposed work implements a combined clustering and random data perturbation approach for enhancing the privacy preservation of health care data in the hybrid cloud. The proposed approach used to totally decrease the computation and communication overhead between the private and the public cloud and our approach doesn't encounter the evil impacts of poor scalability and inefficiency. Our experimental outcomes demonstrated that the proposed approach isn't just safeguarding the accuracy of the framework and furthermore gives the better privacy guarantee, contrasted with existing perturbation procedures. Here, translation, rotation, and reflection isometrics are used for the proposed approach, so the accuracy is a bigger number of times superior to the current methods.

Future work will examine the improved GDP system to accomplish a more elevated amount of privacy preservation. With the commitments of this paper, we intend to explore privacy-aware, productive booking of anonymized data sets in cloud by taking privacy preservation as a metric together with different measurements, for example, stockpiling and calculation later on.

References

[1] B. Fabian, T. Ermakova, and P. Junghanns, "Collaborative and secure sharing of healthcare data in multi-clouds", *Information Systems*, Vol. 48, No. 2, pp. 132-150, 2015.

[2] X. Zhang, W. Dou, J. Pei, S. Nepal, C. Yang, C. Liu, and J. Chen, "Proximity-Aware Local-Recoding Anonymization with MapReduce for Scalable Big Data Privacy Preservation in Cloud", *IEEE Transactions on Computers*, Vol. 64, No. 8, pp. 2293-2307, 2015.

[3] S. Shini, T. Thomas, and K. Chithraranjan, "Cloud Based Medical Image Exchange-Security Challenges", *Procedia Engineering*, Vol. 38, No. 12, pp. 3454-3461, 2012.

[4] L. Wang and C. Alexander, "Medical Applications and Healthcare Based on Cloud Computing", *International Journal of Cloud Computing and Services Science*, Vol. 2, No. 4, pp. 217-225, 2013.

[5] S. Rallapalli, "Improving Healthcare-Big Data Analytics for Electronic Health Records on Cloud", *Journal of Advances in Information Technology*, Vol. 7, No. 1, pp. 65-68, 2016.

[6] E. Achampong and C. Dzidonu, "Attribute-based Encryption for Electronic Health Records in a Cloud Computing Environment", *International Journal of Cloud-Computing and Super-Computing*, Vol. 2, No. 2, pp. 1-6, 2015.

[7] J. Yang, J. Li, and Y. Niu, "A hybrid solution for privacy preserving medical data sharing in the cloud environment", *Future Generation Computer Systems*, Vol. 43-44, No. 2, pp. 74-86, 2015.

[8] L. Huang, H. Chu, C. Lien, C. Hsiao, and T. Kao, "Privacy preservation and information security protection for patients' portable electronic health records", *Computers in Biology and Medicine*, Vol. 39, No. 9, pp. 743-750, 2009.

[9] Y. Kao, W. Lee, T. Hsu, C. Lin, H. Tsai, and T. Chen, "Data Perturbation Method Based on Contrast Mapping for Reversible Privacy-preserving Data Mining", *Journal of Medical and Biological Engineering*, Vol. 35, No. 6, pp. 789-794, 2015.

[10] J. Peng, "A New Model of Data Protection on Cloud Storage", *Journal of Networks*, Vol. 9, No. 3, pp. 217-225, 2014.

[11] C. Danwei, C. Linling, F. Xiaowei, H. Liwen, P.Su, and H. Ruoxiang, "Securing patient-centric personal health records sharing system in cloud computing", *China Communications*, Vol. 11, No. 13, pp. 121-127, 2014.

[12] H. Elmogazy and O. Bamasag, "Securing Healthcare Records in the Cloud Using Attribute-Based Encryption", *Computer and Information Science*, Vol. 9, No. 4, pp. 60, 2016.

[13] S. Worku, C. Xu, J. Zhao, and X. He, "Secure and efficient privacy-preserving public auditing

scheme for cloud storage", *Computers & Electrical Engineering*, Vol. 40, No. 5, pp. 1703-1713, 2014.

[14] V. V. Tejaswini, K. K. Sunitha, and S. S.K. Prashanth, "Privacy Preserving and Public Auditing Service for Data Storage in Cloud Computing", *Paripex - Indian Journal Of Research*, Vol. 2, No. 2, pp. 131-133, 2012.

[15] S. Nepal, R. Ranjan, and K. Choo, "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds", *IEEE Cloud Computing*, Vol. 2, No. 2, pp. 78-84, 2015.

[16] P. Suganthi, K. Kala, and C. Balasubramanian, "Using k-means clustering algorithm for handling data precision", In: *Proc.of the International Conf. on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, Kovilpatti, India, pp. 1-6, 2016.

[17] J. Dwivedi and Z. Wang, "An Optimised Key cryptography for Securing Cloud Data Sharing and Storage Environment", *International Journal of Science and Research*, Vol. 4, No. 12, pp. 924-926, 2015.

[18] K. Govinda and E. Sathiyamoorthy, "Privacy Preservation of a Group and Secure Data Storage in Cloud Environment", *Cybernetics and Information Technologies*, Vol. 15, No. 1, pp.46-54, 2015.

[19] D. Chandramohan, T. Vengattaraman, and P. Dhavachelvan, "A secure data privacy preservation for on-demand cloud service", *Journal of King Saud University - Engineering Sciences*, Vol. 29, No. 2, pp. 144-150, 2017

[20] X. Dai, Z. Wang, and Y. Zhang, "Data Security and Privacy Protection of Cloud Computing", *Advanced Materials Research*, Vol. 846-847, No. 12, pp. 1570-1573, 2013.

[21] W. Wang, L. Chen, and Q. Zhang, "Outsourcing high-dimensional healthcare data to cloud with personalized privacy preservation", *Computer Networks*, Vol. 88, No. 13, pp. 136-148, 2015.

[22] X. Huang and X. Du, "Achieving data privacy on hybrid cloud", *Security and Communication Networks*, Vol. 8, No. 18, pp. 3771-3781, 2015.

[23] H. Zhang, G. Jiang, K. Yoshihira, and H. Chen, "Proactive Workload Management in Hybrid Cloud Computing", *IEEE Transactions on Network and Service Management*, Vol. 11, No. 1, pp. 90-100, 2014.

[24] J. Li, Y. Li, X. Chen, P. Lee, and W. Lou, "A Hybrid Cloud Approach for Secure Authorized Deduplication", *IEEE Transactions on Parallel and Distributed Systems*, Vol. 26, No. 5, pp. 1206-1216, 2015.

[25] J. Li, J. Li, X. Chen, Z. Liu, and C. Jia, "Privacy-preserving data utilization in hybrid clouds", *Future Generation Computer Systems*, Vol. 30, No. 1, pp. 98-106, 2014.

[26] C. Wang, S. Chow, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Secure Cloud Storage", *IEEE Transactions on Computers*, Vol. 62, No. 2, pp. 362-375, 2013.

[27] X. Zhang, C. Liu, S. Nepal, and J. Chen, "An efficient quasi-identifier index based approach for privacy preservation over incremental data sets on cloud", *Journal of Computer and System Sciences*, Vol. 79, No. 5, pp. 542-555, 2013.

[28] H. Jalla and P. N. Girija, "Distance Based Transformation for Privacy Preserving Data Mining Using Hybrid Transformation", In: *Proc. Of the International Conf. on Computer Science and Information Technology*, Chennai, India, pp. 16-23, 2014.

[29] S. Balasubramaniam and V. Kavitha, "Geometric Data Perturbation-Based Personal Health Record Transactions in Cloud Computing", *The Scientific World Journal*, Vol. 2015, No.1, pp. 1-9, 2015.