# A hybrid approach for personal differential privacy preservation in homogeneous and heterogeneous health data sharing

**Pavan Kumar Vadrevu, Sri Krishna Adusumalli, Vamsi Krishna Mangalapalli**

## Abstract

*Personal Privacy Preserving Data Publication (PPPDP) always focuses the difficulty of revealing sensitive data when mining the useful information. Differential Privacy (DP) offers a way to extract the useful and required information about sensitive data without exposing the individual privacy. Design of new Differential privacy mechanisms from the scratch is very difficult. From the contemporary literature it is evident that ε-differential privacy model provides strongest and efficient privacy protection to the sensitive data without considering the attacker background knowledge. The existing methodologies reveal that ε-differential privacy addresses the problem of exposing relational and set-valued data in privacy preserving environment. In this paper we proposed two algorithms namely GenDP and cluster-basedDP by considering relational and set-valued data of health data in differential private disclosure. The GenDP generalizes the input data as homogeneous or heterogeneous based on the partitioning algorithm and classifier is defined to defense ε-differential privacy. It depends on a simple generalization of typical differential privacy known as Personalized Differential Privacy (PDP). The cluster-basedDP groups the data based on the similarity and defines the classifiers to handle the ε-differential privacy. Experimentation has done with COVID-19 health data, that results the proposed algorithms are scalable and efficient than available classification analysis.*

***Keywords:*** *Personal Privacy, Differential Privacy, Generic and Cluster based Differential Privacy*

## 1. Introduction

Differential privacy (DP) is a known promising method which guarantees unconditional privacy without the background information of the intruders. While delivering the consequences of results of statistical queries on sensitive information, including range queries, DP guarantees that with or without specific record in the dataset, the results of calculations are formally impossible to differentiate. In general, DP gives an intuitive interface to information evaluators (for the non- interactive information discharge model, finding an efficient method for some areas stays an open issue, which we won't examine in this paper). However, the interactive model requires a server that can generally answer information evaluators' inquiries in time. Information contributor can't be disconnected if they hold their sensitive information firmly. Considering the risk of information providers remaining online for a long time and the straightforwardness of cloud computing, information providers consider moving their delicate information to a cloud service provider (CSP), which makes the CSP respond each and every query from information evaluators. However, the CSP may not always be reliable and trustworthy. The privacy of sensitive information becomes questionable when CSP has been compromised.

**PINQ-style Global Privacy Budget:** PINQ is an implementation of interactive differential protection which guarantees, at runtime, that queries adhere to a worldwide privacy budget. Third-party customer code freely decides how sensitive data sets ought to be processed and queried. The run-time framework guarantees this doesn't break a predetermined privacy budget ε. PINQ[1] construct on a set of standard differentially private primitive queries, along with basic composition principles – mathematical properties appreciated by the meaning of differential privacy. One focal rule is that multiple queries (for example with differential privacy ε1 and ε2 respectively) have an additive impact (ε1 + ε2) on the overall differential privacy. Another focal idea is to track sensitivity of functions to quantify how much adjustment in the input may influence the estimation of the information. Together, these components permit the framework to track how much to deduct from global privacy budget on each invocation of a primitive query.

**Limitations of the Global Privacy Budget:** In a batch system where all calculations are depicted in advance as a monolithic program, a global budget is sensible. In an interactive framework, however, there are few restrictions to this style of accounting. Imagine a situation involving a huge informational collection of individuals – a cross - set of the population – containing a variety of data about medical and lifestyle.

Let us further assume that we aim for ε-differential privacy for some specified value of ε. On Monday the analyst selects all the individuals from the database who have a specific blood category, AB-negative, and develops an algorithm which extract data about them as a part of medical exploration. Since only 0.6% of the populations have this blood type the proportion of the database involved in this study is moderately little, yet the database is known to be large enough for it to be significant. As per the framework, Let the cost of this analysis be ε1. On Tuesday the analyst gets another assignment, to extract data about the lifestyle of smokers, towards promoting nicotine gum. This is an altogether bigger part of the database, perhaps covering Monday's research group. The analyst has ε − ε1 left to spend. If ε1 is large, the analyst has blown the budget by analysing the small group, despite the fact that that review didn't contact the information of the largest aspect of the populace. PINQ offers a path around this issue by adding nonstandard information database primitives. Here we would divide the information into AB− or not AB− and perform the two investigations in parallel, with cost being the limit of the expense of the two investigations. This leads to a batch-style reasoning and an unnatural programming style. But it also has another restriction. Imagine a scenario where the database is live – we get new information over time, or if information is consistently being added. A global budget drives us to be unnecessarily pessimistic about new up 'til now unexploited information.

Today, in the midst of fast adoption of electronic medical records and wearable's, the worldwide health care systems are methodically gathering longitudinal patient health data, e.g., diagnoses, medication, lab tests, techniques, demography, clinical notes, and so on. The patient medical data is produced by at least one experience in any medical services delivery frameworks [2]. Medical data is currently estimated in Exabyte's, and it will soon in zettabyte and in future in the range of yottabyte[3]. Although suitable in variety of circumstances, numerous conventional strategies for analysis don't automatically capture complex and hidden attributes from large-scale and perhaps unlabeled

information [4]. Practically speaking, numerous medical applications rely upon domain knowledge to build relevant attributes, some of which are additionally gathered based on supplemental information. This procedure isn't straightforward, time consuming and also may result in missing opportunities to discover new patterns and attributes.

## 2. Related work

Mohammed et al. [5] proposed a probabilistic based generalization to address the problem of dimensionality in high dimensionality data. Xiao et al. [6] defined the DPCube method which uses KD-tree partitioning algorithm for multi-dimensionality health data and it produced a differentially personal cell histogram by splitting the noisy cell data with Laplace noise. Though, the multi-dimensional partitioning results the increase of errors of estimation due to huge domain values and attributes in multi-dimensional data.

Su et al. [7] designed DP-SUBN , it is based on PrivBayes to yield two-way negligible tables of input attribute sets by overlapped covering design to increase the adaptability of Bayesian Network with reduced communication cost. Chen et al. [8] introduced a sampling-based method which is designed with threshold algorithm to address a systematic survey on pair-wise attributes and publish the values with junction tree algorithm. This approach works efficient on binary and non-binary data, but failed to maintain efficiency and errors while minimizing the resultant errors.

In Li et al. [9] the authors addressed the problem of privacy budget exhaustion, specifically observations of frequent datasets. However, their methodology failed in compressed sensing, in which sparse tree reconstruction is undesirable and vulnerabilities in privacy protection. These are well suited for multi-dimensional data, but failed to handle the high volume of data. These methods impose significant information for introducing an efficient differential private privacy method for publishing high-dimensional data.

From the contemporary literature the algorithms defined to defense the differential privacy is categorized into interactive and non-interactive methods. The interactive method allows the data miner to dump aggregate queries through secure and personal methodology and server provides responses as an answers. Majority of the existing methods uses interactive framework for differential privacy. In non-interactive method the server first anonymizes the metadata and published the same for user availability. In this manuscript we adopt the non-interactive method of differential privacy and publish the contingency tables or marginals of the metadata using cluster analysis based on the similarity of the attribute behaviors. Nevertheless, this method is suitable for high-dimensional metadata with a huge domain.

To provide a ε-differential privacy protection a non-interactive method is proposed with heterogeneous health data. The main goal of the data release or publishing is to protect the privacy of the critical element and also it is significant that efficacy of the published data is equally preserved. Cluster-based privacy preserving algorithm for metadata is proposed to effectively personalize for preserving information in data mining applications.

1. From the literature it is evident that differential privacy algorithm with simultaneous access of relational and set-valued data not available. Hence, the proposed differential privacy data algorithm preserves the information in cluster analysis with generalized methods. It is also evident that deterministic techniques failed to achieve the ε-differential privacy, because these are depends on the disseminated data. However, differential privacy data is published with the addition of ambiguity procedures.
2. The proposed cluster based method handles the different types of attributes and doesn't require pre-discretized for numerical attributes. While preserving ε-differential privacy, the proposed method determines the centroids based on the similarity of the attributes for obtaining accurate classification.
3. The ε-differential privacy defines strong privacy protection from the literature, but the efficacy of the data published by the various personal differential privacy algorithms has been neglected. Experiment results defines our cluster-based method provides better security compared to personal interactive algorithm for designing classifiers.

## 3. Differential Privacy-preserving in homogeneous and heterogeneous health data

The rapid advancement of Electronic health record system (EHRS) [10] increased the exceptional rate of health data collection. The sharing of these data is essential for several applications among varies parties, which includes policy development, decision making and data mining applications. But the important constraint is to provide the privacy for sensitive data of individual people. However, now a days majority of the privacy protection schemes depends on their own rules or policies and organizational specific guidelines.

For example, in fine-grained epidemiology [11] survey with more than 20K residents need detailed ages over 89. Some scenarios revel that the performance of computerized heath records will be influenced by the privacy rules and law enforcement interference. Many researchers [12]concluded that de-identified data is not sufficient for protection and thee privacy rules affect the biomedical research.The institute of medical committee recently defined that privacy guidelines not sufficiently defense privacy and affect the high quality research on health and privacy information. As a result the health records of the patients' are not protected well and the researchers unable to use them efficiently for their research. The health data publishing require technical efforts to effective use of health data and privacy preserving. In recent days the health data is defined as either relational data (example: demographics) or set valued data (example: codes and test values). The relational data [13] is defined as one value for every attribute like age, BMS(body mass index), gender, height and weight. Whereas in set-valued data multiple values for every attributes is assigned like predefined medical codes and laboratory test results.

In the recent days the medical research requires heterogeneous data [14]to be published simultaneously and the correlation between them should be preserved. However, this type of heterogeneous data access is not properly covered in the recent literature with privacy technology. From the contemporary literature it is evident that these all are addressed homogeneous data and failed

to defense the privacy attacks involved with heterogeneous data access.We propose an algorithm which handles the heterogeneous health data in a differently private manner by retaining the sensitive information with data mining tasks. This scenario further explored the privacy threats with COVID-19 heterogeneous health data.

Table 1: Sample attributes in COVID-19 Dataset

| | |
|---|---|
| Patient-ID | Aadhar Card |
| Name | Address |
| Age | Height |
| Sex | Weight |
| Date of birth | BMI |
| Mobile Number | Diagnostic Code |
| Temperature | Family name |
| Blood group | Reference |
| Class | Hospital code |

Example 1: Consider the COVID-19 patient data in table1, each row defines the patient information. The attribute Sex is categorical. Age, weight, height, Mobile number, BMI and Diagnostic code are numerical, and set-valued. Suppose if the data owner needs to publish table 1 for analysis of patients classification based on the class attribute where it consists of Y and N which indicates the decease status of the patient. If some records are too specific such that not many patients can match it, explores the data causes the re-identification of the patient details.For example Lokids et.al proved that in code of International Classification of Diseases Nine (ICD-9), set-valued data of one source is used to linkage patient identities by an adversaries. The privacy attack is easier for an opponent, when victim's relational and set-valued data is known. It is very difficult to predict the opponent knowledge. The opponent can identify the victim with partial or full knowledge on set-valued data.

Preserving privacy for individual is very difficult and suggested seven privacy principles to public health authorities, researchers and industries to follow and move forward to track, test the developed technologies to handled COVID-19 disease.

1. *Publish meaningful consent with transparent methodologies defined for collecting data like what data is collected and how long the data is live*. Data is collected from the participants with consent of their willingness. Encourage the voluntary participation of the people with clear and user-friendly information and ensure that the participants are interacting freely with the technology and making their own choices before participation. The participants should know the benefits of data collection, type of data collected and amount of time the data is live to use.
2. *Data is collected and made available for health of public* .To trace the people who are physically contact with the infected person the data is collected and the individual person's health purposes the data should be owned by the health authorities and it contains the type of data required for fighting the disease.

3. *Optimize the amount of data collected*. The amount of data collected by the health authorities for tracing should be minimal and time period identified is essential.
4. *Educate the people about the storage of data.* The data collected is under the control of individual person with choice of storing the data in device or cloud and provide the permissions to access it whenever it is required by others.
5. *Provide suitable protections to protect the data.* The data is protected with encryption, de-identification, and random identifiers or similar approaches from data misused and hacking attempts.
6. *Limit the data sharing and health status and don't allow without consent of the people*. The health data should not be shared to individual contacts or others without consent the individuals and violation of these pursuant to legal issues
7. *Clear or remove the data if it is not useful further in future.* The health authorities must remove or delete the data once it is not useful further in future. The authorities must inform to the individuals the status of their data once the tracing is completed.

These principles are considered to COVID-19 technicalclarifications that comprise the group and use of private data such as health data, precise geolocation data, proximity or adjacency data, and identifiable contacts.

## 4. Deferential Privacy

A privacy model that minimizes the chances in the identification of individual records and provides the maximize privacy is stated as Differential privacy (DP) [15]. The amount of information that can be revealed about some ones's data which is available in the database to the third party/adversary can be bounded using the principle of DP. Traditionally, these bounds in DP are denoted by epsilon (ε) and delta (δ), which specify the level of privacy cause to be by a randomized privacy preserving algorithm (M) over a specific database (D).

***Definition 1. (Differential privacy)***

A randomized function $RF$ with a well-defined probability density $PD$ satisfies $\epsilon -$ Differential privacy if, for any two neighboring datasets $DS1$ and $DS2$ that differ by only one record and for any $RE \in range(MR)$

$$PD\ (RF\ (DS1) = RE) \leq e^{\epsilon} * PD\ (RF\ (DS2) = RE)$$

Generally, Laplace mechanism is used to achieve the Differential privacy (DP). To generate Laplace distribution parameters are selected based on global sensitivity and privacy budget.

***Definition 2. (Global sensitivity)***

Let the database is mapped to a fixed-size vector of real numb using a function $fn$. The global sensitivity of $fn$ for all neighboring databases $DS1$ and $DS2$ is computed as

$$\Delta\ (fn) = max_{DS1,DS2} \| fn(DS1) - fn(DS2) \|$$

where||. || denotes the L1 norm.

Noise should be added to response a range query with $\epsilon - DP$. Because one record will contaminate the estimation of the appropriate response by just one, the global sensitivity $\Delta f$ in our scheme should be 1. Let $Lap(\lambda)$ indicate the Laplace probability distribution with a mean of zero and scale $\lambda$, where $\lambda = \Delta f/\epsilon$.We can add noise sampled from $Lap(\lambda)$ to an original response $re$ to achieve $\epsilon - DP$.

***Definition 3. (Laplace mechanism)*** Let $rq$ be an response to a range query, and let $\eta$ be a random variable such that $\eta \sim Lap(\delta f/\epsilon)$. The Laplace mechanism is defined as follows:

$$\overline{\mathrm{r}\acute{\mathrm{q}}} = \mathrm{r\,q} + \eta.$$

Sequential composition and parallel composition are the two main properties of Differential privacy.

The hypothesis confirms that adding up the precise quantity of noise to statistical queries, one can achieve positive results at the same time provided a quantifiable conception of privacy. According to the definition, it doesn't conform to the rules of syntax from the data rather it is formed by comparing results of a query on any database with or without any one individual: a query Q (a randomized function) is ε-differentially private if the difference in probability of any query outcome on a data-set only varies by a factor of e ε (approximately 1 + ε for small ε) whenever an individual is added or removed. A variety of query mechanisms are developed in literature which provides Differential privacy for quantifiable collection of statistical problems. A little state-of-art has focused private mechanisms on composition principles to design the system. These principles build more complex differentially private building blocks in principled way, so that the resulting programs are guaranteed to be differentially private by construction [16][17]. The initial point for the current state-of –art is PINQ [1].

### 4.1 Personalized Differential Privacy

Let's have the discussion on the main concept introduced in this paper i.e., Personalized or "big epsilon" differential privacy, and its analogous compositionality properties.

**Definition 3.1 (Personalised (Big Epsilon) Differential Privacy**).

The data sets $DS1$ and $DS2$ differ on record $re$, represented as $DS1 \sim^{re} DS2$, if $DS1$ can be obtained from $DS2$ by adding the record $re$, or vice-versa.

Let function from records to non-negative real number is represented as $\xi$. $\xi$ -differential privacy is achieved for a randomized query $RQ$ if for all records $re$, and all $DS1 \sim^{re} DS2$, and any set of outputs $OT \subseteq range(RQ)$, then

$$P\,re[RQ(DS1) \in OT] \leq P\,re[RQ(DS2) \in OT] * e^{\xi(re)}$$

Private personal privacy level can be achieved by individual records is permitted with Personalized differential privacy. This may turn out to be a valuable idea in its own right, but its primary purpose in

this work is as a generalization that allows a more fine-grained accounting in the development of classical differentially private mechanisms, and one which plays well with dynamic databases. The following proposition summarizes the relation to "small-epsilon" differential privacy:

- If RQ is $\xi$-differentially private, then RQ is $\lambda x.\xi - differentially$ private.
- If RQ is $\xi - differentially$ private, and $sub(range(\xi)) = \varepsilon$ then Q is $\varepsilon$ -differentially private.

In personalized differential privacy (PDP) by offering a simple generalization of differential privacy allows every individual to have personalized privacy budget. The definition upholds generalized versions of the composition principles whereupon frameworks like PINQ are based, and besides appreciate various properties which take into account less inefficient compositional principles. For instance, any query about the drinking habits for adults offers 0-differential protection for Adrian, aged 13, as it accomplishes for any records of individual which enter the database after the query has been made. From these standards we plan a framework, in the style of PINQ, called Provenance for Personalized Differential Privacy.

The framework preserves a personal budget for each record entering the system. Rather than using sensitivity, it tracks the provenance of each record in the system and utilizations the specific provenance to calculate how a query should influence the budgets of the individuals. As PINQ, the system is depicted as a theoretical proper model for which we demonstrate personalized differential security. This is significant because the correctness of this method isn't clear for two reasons. Initially, the individual budgets become highly sensitive and how we handle them is novel. All the more explicitly, if a query includes records that would break the budget of an individual they are quietly dropped from the database whereupon the query is determined. In the above example, Tuesday's analysis of smokers will automatically exclude information got from any ailing people when the expense of the query surpasses their budget.

Furthermore, it is important to limit the domain of computations over data sets to a class which ensures that the provenance of any inferred record is relative (zero or one record), otherwise the number of records which may get rejected because of a little change in the input may be too enormous to provide privacy guarantees. The methodology is appropriate for integration with different frameworks, since we assume the existence of basic primitives providing classical differential privacy. We have put into operation a prototype of this methodology which broadens the PINQ framework [18] with personalized budgets and the ability to input real information. We analyze the performance of our provenance-based implementation with PINQ to show that the runtime overhead isn't critical. We conclude with a conversation of related work, an outline of our contributions, and the current restrictions of the methodology [19] as well as directions for future work.

## 5. Generic Differential Privacy and Cluster Based Differential Privacy for Differential privacy

In this section we introduced two independent methodologies or algorithms for Differential privacy namely generic differential privacy (GenDP) and Cluster-Based Differential privacy (Cluster-basedDP) algorithms. The GenDP provides generalized privacy protection algorithm for homogeneous and heterogeneous data. The GenDP algorithm consider homogeneous data as single group and heterogeneous data or multi-dimensional data will be portioned using portioned algorithm before defining the classifiers for validating the accuracy. The cluster-based DP groups the data into clusters based on the similarity of the behavior or attribute values, later it defines the classifiers for validating the personal or differential privacy.

### 5.1 Generic Differential Privacy

A variety of partition-based models [24][25][26]are proposed in the recent literature to prevent the privacy attacks and this provides insufficient protection, because these are vulnerable to several privacy attacks. In this paper, a differential privacy (DP) model is proposed which guarantees privacy and protects data from all privacy attacks. This differential privacy (DP) model [27][28]does not depend on the opponent's background knowledge. The probability of any output (published data) is from identically distributed input datasets and which promises all these outputs are insensitive to any single personal data. This states that the personal privacy is not under risk, because disclosed data set is included. The GEnDp and Portioning algorithms are given in Algorithm1 and Algorithm2.

**Algorithm 1: DiffGen**

**Input:**COVID-19 data set $DS$, privacy budget $\varepsilon$, and number of specializations $s$

**Output:** Generalized data set $\widehat{DS}$

1. Initially, the highest value is assigned to every value of $DS$
2. The highest value is initially assigned to $Ct_i$
3. Forspecification of predictors, Set privacy budget as
$$\acute{\varepsilon} \leftarrow \frac{\acute{\varepsilon}}{2(\left|AT_n^{pr}\right| + 2s)}$$
4. Specify the split value for each$va_n \in \cup Ct_i$with probability $\alpha \exp\left(\frac{\acute{\varepsilon}}{2\Delta u} u(DS, va_n)\right)$;
5. Calculate the score for each candidate$\forall\ va \in \cup Ct_i$
6. for $i = 1$ to $s$ do
7. Select $va \in \cup\ Ct_i$ with probability $\alpha \exp\left(\frac{\acute{\varepsilon}}{2\Delta u} u(DS, va)\right)$;
8. Specialize $va$ on $DS$ and update $\cup Ct_i$;
9. for each new$va_n \in \cup Ct_i$, Determine the split value with probability $\alpha \exp\left(\frac{\acute{\varepsilon}}{2\Delta u} u(DS, va_n)\right)$;
10. Update score for$va \in \cup\ Ct_i$;
11. end for
12. each group with count $\left(co + lap\left(\frac{2}{\varepsilon}\right)\right)$ is returned, where $Lap(\cdot)$denotes the probability density function of Laplaciandistribution.

**Algorithm 1 Differential-private partition Algorithm**

**Input:** $DB_{t_i}$,$TH_D$,$TH_R$,$TH_L$,$q.\epsilon_i$

**Output:** Sub datasets$BU$;

1. **Initialization:** Set $sz = 0; i = 1; j = 1; BU = \phi$;
2. $\widehat{TH_D} = TH_D + ZN, \widehat{TH_R} = TH_R + \overline{ZN}, \triangleleft ZN, \overline{ZN} \sim Lap((q.\epsilon_i))$
3. $bu_j \rightarrow db_i; Min = Max = Current = db_i; sz++; i++;$
4. **while** $i \leq length(DB)$ **do**
5. **if** $Current \neq Null$ and $|Current - db_i| > \widehat{TH_R}$ **then**
6. **if** $bu_{j-1}.length > 1$ **then**
7. ▷Last bucket is not a single bin bucket
8. $last = BU.pop(); bu_j = lasy.pop();$
9. $BU \leftarrow last; BU \leftarrow bu_j; j++; bu_j \leftarrow db_i; BU \leftarrow bu_j;$
10. $j++; Current = x; i++; sz = 0;$
11. **else** ▷Last bucket is a single bin bucket
12. $bu_j \leftarrow x; BU \leftarrow bu_j;$
13. $j++; Current = x; i++; sz = 0;$
14. **elseif**
15. **else if** $sz == 1$ **then**
16. $BU \leftarrow bu_j; j++; bu_j \leftarrow db_i; j++;$
17. $Current = db_i; sz = 0; i++;$
18. **else if** $sz \geq 1$ **then**
19. $last = bu_j.pop(); BU \leftarrow bu_j; j++; bu_j \leftarrow last;$
20. $BU \leftarrow bu_j; j++; bu_j \leftarrow db_i; BU \leftarrow bu_j; j++;$
21. $Current = x; i++; sz = 0;$
22. **else if**
23. $Max = \max(Max, db_i); Min = \min(Min, db_i);$
24. **if** $|Max - Min| \leq TH_D$ and $sz \leq TB_S$ **then**
25. $bu_j \leftarrow db_i; Current = db_i; sz++; j++;$
26. **else**
27. $BU \leftarrow bu_j; Current = db_i; sz = 0; j++;$
28. **end if**
29. **end while**
30. **return** $BU$

### 5.2 Cluster-based Differential Data Method

In this section cluster-based differential method [20][21][22] is introduced and dataset is divided into subsets based on the similarity of the attributes with multi-dimensional properties. The similarity of the objects is strong in the same subset. To differentiate the cluster centers the data is divided into heaps, so that the distance from the cluster centers to the clusters of the samples is minimized. K-means clustering algorithm is adopted to group the behavior of the clusters.

Consider the set of samples containing $'n'$ data objects as$S = \{S_1, S_2, S_3, \ldots\ldots, S_i, \ldots\ldots, S_n\}$. Each data object contains $'m'$ features represented as$S_i = \{S_{i1}, S_{i2}, S_{i3}, \ldots\ldots, S_{im}\}$. The set is further divided into $'p'$ clusters as $S = \{cluster_1, cluster_2, cluster_3, \ldots\ldots, cluster_k\}$ and each cluster contains $'k'$ samples. $CL = \{CL_1, CL_2, CL_3, \ldots\ldots, CL_p\}$be the cluster centers where $p < n$. The Euclidean distance between two points is calculated as follows:

$$d(S_i, S_j) = \sqrt{\sum(S_{ix} - S_{jx})^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where $i = 1, 2, \ldots\ldots, n$and$x = 1, 2, \ldots\ldots, m$.

The sum of distances between the samples in the set divided by the total number of samples in the set is the average distance between he two points. Randomly set two samples from the set of samples. The average distance (AD) can be formulated as

$$AD = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} d(S_i, S_j)}{A_n^2} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

The density (DT) of the samples $S_i$ is defined as: Let $S_i$ be the center of the circle with the data objects and the number of objects in the circle is represented as $\propto * AD$, where be the coefficient adjustment of radius. If $d(S_i, S_j) \leq \propto * AD$ condition is true then the $cnt()$ function is cumulatively incremented by 1. 1 is the default value of the density of the sample.

$$DT(S_i) = \sum_{i=1}^{n} cnt(d(S_i, S_j) \leq \propto * AD) \ldots\ldots\ldots\ldots\ldots\ldots (3)$$

Where $i = 1, 2, \ldots\ldots, n$and $j = 1, 2, \ldots\ldots, n$.

Further the average density ($ADT$) of the samples $S$ is calculated as follows:

$$ADT = \frac{\sum_{i=1}^{n} DT(S_i)}{n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4)$$

A collection of data objects whose density is a certain multiple of the average density of the sample set $S$ is specified as high-density point set and defined as

$$HD = \{S_h\} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)$$

Where $S_h$ is a data object which belongs to dataset $S$ which satisfy the condition $DT(S_h) \geq \beta * ADT$. $\beta$ is the adjustment factor of density whose default value is 1.
The mean of sample set $S$ is denoted as center of sample set $S$ and represented as

$$S_{Center} = \frac{S}{n} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$$

The squared error sum $(ERR)$ of the cluster is represented as

$$ERR = \sum_{i=1}^{a}\sum_{j=1}^{b}\left|S_{ij} - CL_i\right|^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

In the above equation, $j^{th}$ data object of the $i^{th}$ cluster is represented as $S_{ij}$ and center of $i^{th}$ cluster is represented as $CL_i$

**Algorithm: Cluster-basedDP**

**Input:** A set of COVID-19 data,
**Output:**COVID-19 based data for validating the differential privacy.

1. By using the equations (1 – 3) , Calculate the density of each object in the sample set S
2. Based on the equations (4 - 6) , Measure the high-density point set HD and also center of sample set $S_{Center}$
3. According to the equation 6, compute the distance from HD to $S_{Center}$ and select $HD_i$ which satisfies the condition $\max(d(HD_i, S_{Center}))$ as the first cluster center $CL_1$ to join the set CL;
4. repeat
5. Select the data object $HD_j$, if the condition $\max(d(HD_i, S_{Center}) * d(HD_i, CL_1))$ is true as the second initial cluster center $CL_2$ is added to the set CL;
6. Until $|CL| = a$
7. Return metadata
8. //The center of the cluster is selected
9. By distance, data objects are divided into the nearest cluster in the sample set S
10. As per the equation 7, compute the squared error sum (ERR) of the cluster and specify whether the union is converged
11. If the converges occurs, the clustering algorithms stops. Otherwise, previous step is executed again and the cluster center is updated again
12. // Dividing the datasets

## 6. Experimental Study

This section explores the experimental study on the impact of differential privacy to validate the quality of the data using classification accuracy and test the scalability of the proposed methods to handle large multi-dimensional data. Comparison of the methods GenDP and cluster based algorithm has been done to validate the accuracy when the data is of homogeneous or partitioned into groups when it is heterogeneous using partition algorithm and in cluster based approach data is grouped into clusters based on the behavior and validate the data.

The experiments were carried out on an INTEL core i5 2.9GHz PC with 4 GB RAM with COVID-19 dataset. The COVID-19 dataset is collected from the repositories and research centers, which contains 12000 records of effected people. However, we have selected 13features which include name, age, sex, mobile number; address, aadhar number, marital status, temperature, religion, payment details, source of admission, family members and diagnostic code. Among these 13 attributes some are categorical and some are set-valued attributes.

The proposed methods GenDP and Cluster based approach are applied on COVID-19 data set to evaluate the performance. To evaluate the classification efficiency in GenDP method with the partitioned algorithm, the data is divided into training and testing sets. Firstly, the algorithm is applied to evaluate the partitioning of the data in training phase and calculate the Ucuti. Later it was applied in GenDP to test the testing data for producing generalized test set and construct the classifier to measure the classification Accuracy (CA) on the generalized records of test data. In cluster-based approach the data is grouped based on the similarity and define the classifiers to evaluate the heterogeneous behavior of the data.

The COVID-19 data set is applied to both the approaches GenDP and cluster-based approach, where the privacy budget is ε=0.1, 0.25, 0.5, 1, with 8 specializations h=8. For the dataset 60% is given for training and defining the classifier in GenDP and 40% is utilized for validating the accuracy. The proposed methods explore the good performance with CA as 89% and 94% for various privacy budgets. The GenDP exhibits better performance for homogeneous data and classification-based approach have the consistent behavior for both homogeneous and heterogeneous data under ε-differential privacy. The figure 1 shows the performance of the same.

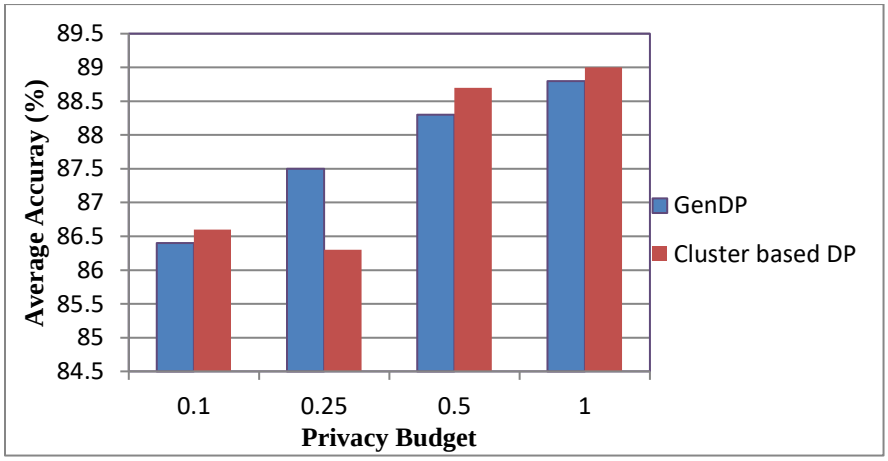


Figure 1: Classification Accuracy (CA) of GenDp and Cluster-basedDP for h=8

The Classification accuracy (CA) for the GenDP with privacy budget ε=0.1, 0.25, 0.5, 1, and the specializations ranging from 5 to 25. The Classification Accuracy for budget ε=1 is 84% and for budget ε=0.1 is 75%. The CA values for various specifications under different budget value have shown in figure2, where the number of specifications is increased then the accuracy is decreased.

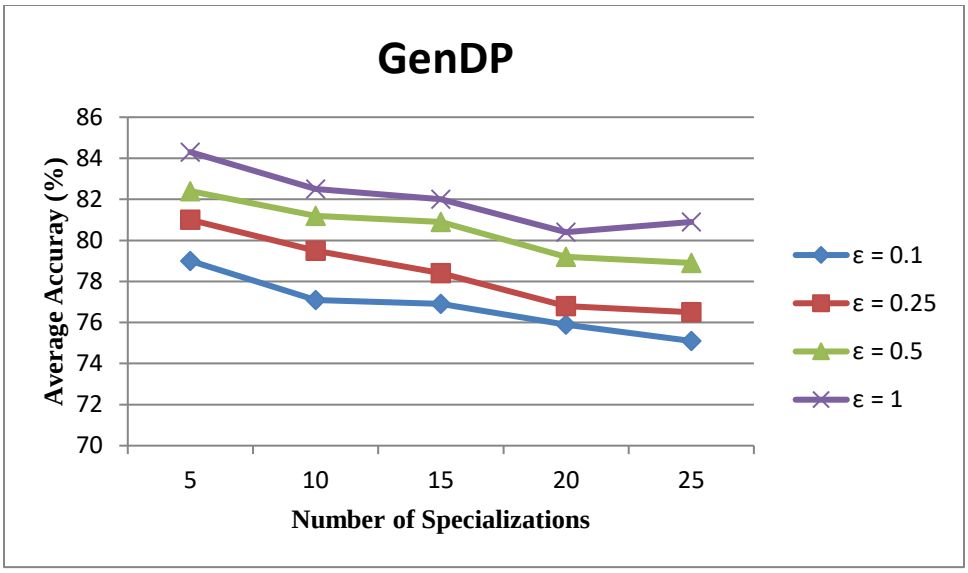


Figure 2: Classification Accuracy (CA) of GenDpfor various specializations and budgets

The partitioning algorithm is used to group the data into different buckets or groups based on the behavior of the data. The number of partitions for homogeneous data is 1 and when the number of dimensions or volume of data increased the accuracy will be decreased. The CA values for various partition count has shown in figure 2.

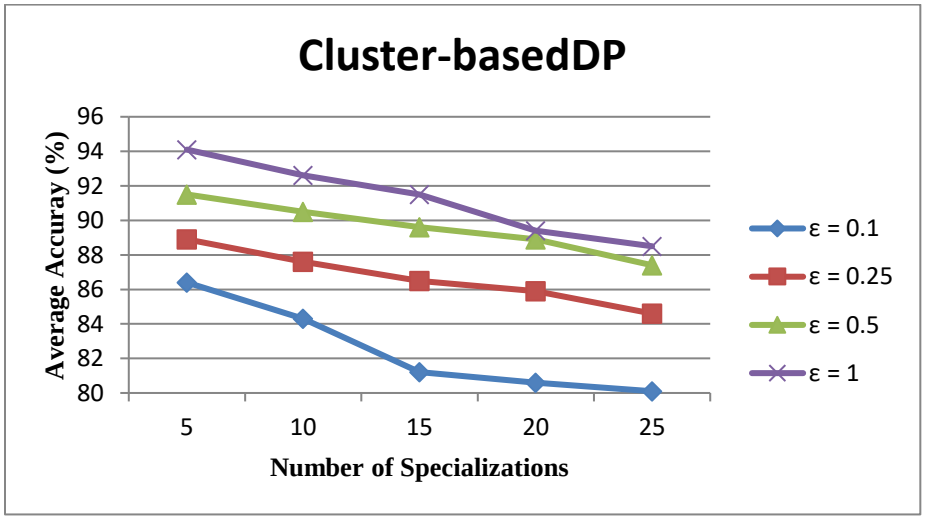


Figure 3: Classification Accuracy (CA) of Cluster-based DP for various specializations and budgets

The Classification accuracy (CA) for the cluster-basedDP with privacy budget ε=0.1, 0.25, 0.5, 1, and the specializations ranging from 5 to 25. The Classification Accuracy for budget ε=1 is 94% and for budget ε=0.1 is 86%. The CA values for various specifications under different budget value have shown in figure3, where the number of specifications is increased then the accuracy is decreased.The performance of the partition algorithm for variable partitions is shown in figure 4.

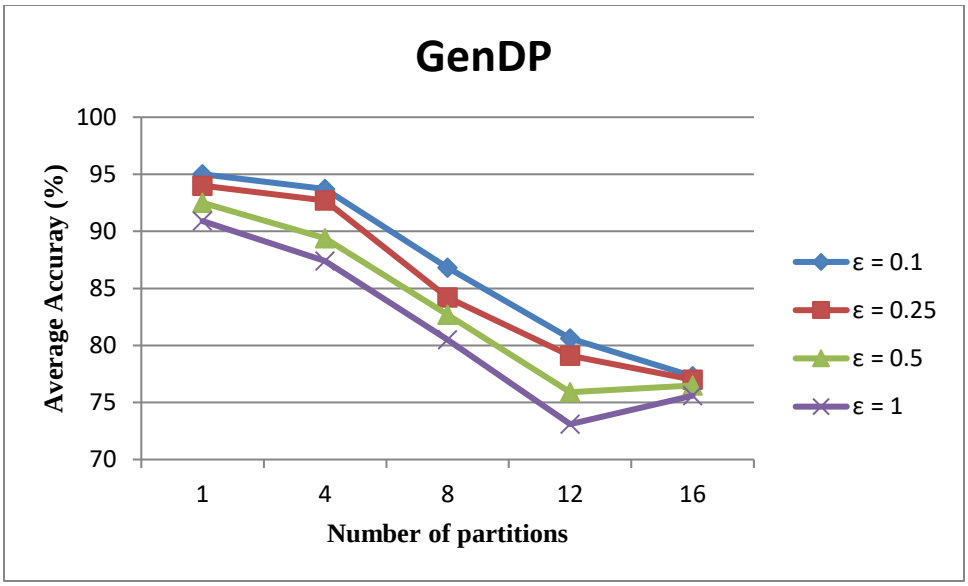


Figure 4: Classification Accuracy (CA) of GenDpfor various partitions

## Conclusion

This paper defined two new algorithms for addressing the ε-differential privacy in large homogeneous and heterogeneous data with multi-dimensionality. This algorithm classifies health data and disseminated to the advisories without revealing the sensitive information of the individuals. The typical generalized DP algorithm explores the privacy in homogeneous and heterogeneous and defines the classifiers based the groups for validating the individual or personal privacy. The clustering-based DP addresses the large multi-dimensional data and groups it based on the similarity for defining the classifiers to validate the ε-differential privacy. The methods are validate with COVID-19 dataset and proved these methods maintain the good classification accuracy for various specifications and budget values.

## Acknowledgements

We express thanks to the Information Technology Department and Management of Shri Vishnu Engineering College for Women (A), Bhimavaram for providing all the necessary resources to carry out this work.

## Author Affiliations

1. **Pavan Kumar Vadrevu**, Research Scholar, Department of Computer Science and Engineering, Centurion University of Technology and Management, Paralakhemundi, Orissa (e-mail: vadrevu.pavan@gmail.com)

2. **Sri Krishna Adusumalli**, Associate Professor, Department of Information Technology, Shri Vishnu Engineering College for Women, Bhimavaram , Andhra Pradesh (e-mail : srikrishna@svecw.edu.in)

3. **Vamsi Krishna Mangalapalli**, Professor, Department of CSE, Chaitanya Institute of Science and Technology, Kakinada. (e-mail: vamsimangalam@gmail.com)

## References

[1] F. D. McSherry. Privacy integrated queries: an extensible platform for privacy-preserving data analysis. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data, pages 19–30. ACM, 2004.

[2] Jamoom, E. W., Yang, N., &Hing, E. (2016). Adoption of certified electronic health record systems and electronic information sharing in physician offices: United states, 2013 and 2014. NCHS Data Brief, 236, 1–8.

[3] Fang, R., Pouyanfar, S., Yang, Y., Chen, S. C., &Iyengar, S. S. (2016). Computational health informatics in the big data age: A survey. ACM Computing Surveys, 49(1), 12:1–12:36. doi:10.1145/2932707.

[4] Miotto, R., Li, L., Kidd, B. A., & Dudley, J. T. (2016). Deep patient: An unsupervised representation to predict the future of patients from the electronic health records. Scientific Reports, 6, 26094. doi:10. 1038/srep26094.

[5] V. Bindschaedler, R. Shokri and C. A. Gunter, "Plausible deniability for privacy-preserving data synthesis," Proceedings of the VLDB Endowment, vol.10, no.5, pp. 481-492, 2017.

[6] Y. Xiao, J. Gardner and L. Xiong, “Dpcube: Releasing differentially private data cubes for health information,” in Proceedings of 2012 IEEE 28th International Conference on Data Engineering (ICDE’12), 2012, pp. 1305-1308

[7] S. Su, P. Tang, X. Cheng, R. Chen and Z. Wu, “Differentially private multi-party high-dimensional data publishing,” in Proceedings of IEEE 32nd International Conference on Data Engineering (ICDE) ,2016, pp. 205-216.

[8] R. Chen, Q. Xiao, Y. Zhang and J. Xu, “Differentially private highdimensional data publication via sampling-based inference,” in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 129-138.

[9] Y. D. Li, Z. Zhang, M. Winslett and Y. Yang, “Compressive mechanism: Utilizing sparse representation in differential privacy”, in Proceedings of The 10th Annual ACM Workshop on Privacy in The Electronic Society, ACM, 2011, pp. 177-182

[10] O’Keefe CM. Privacy and the use of health data—reducing disclosure risk. Electronic J Health Informatics 2008;3:e5:1–e5:9.

[11] Benitez K, Loukides G, Malin BA. Beyond safe harbor: automatic discovery of health information de-identification policy alternatives. The 1st ACM International Health Informatics Symposium; ACM, 2010:163–72.

[12] Baumer D, Earp JB, Payton FC. Privacy of medical records: IT implications of HIPAA. ACM ComputSoc (SIGCAS) 2000;30:40–7.

[13] McGraw D. Why the HIPAA privacy rules would not adequately protect personal health records: Center for Democracy and Technology (CDT) brief. http://www.cdt.org/brief/ why-hipaa-privacy-rules-would-not-adequately-protect-personal-health-records (accessed 20 Feb 2012).

[14] Institute of Medicine. Beyond the HIPAA privacy rule: enhancing privacy, improving health through research. Washington (DC): National Academies Press (US); 2009.

[15] M. Chamikara, P. Bertok, D. Liu, S. Camtepe, and I. Khalil, "An efficient and scalable privacy preserving algorithm for big data and data streams," Computers & Security, vol. 87, p. 101570, 2019.

[16] I. Roy, S. T. V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel. Airavat: Security and privacy for mapreduce. In NSDI, pages 297–312. USENIX Association, 2010.

[17] A. Haeberlen, B. C. Pierce, and A. Narayan. Differential privacy under fire. In USENIX Security Symposium, 2011.

[18] D. Mir, S. Muthukrishnan, A. Nikolov, and R. N. Wright. Panprivate algorithms via statistics on sketches. In Proceedings of the Thirtieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS '11, pages 37–48. ACM, 2011.

[19] J. T. Wittbold and D. M. Johnson. Information flow in nondeterministic systems. In IEEE Symposium on Security and Privacy, pages 144–161, 1990.

[20] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, A. Liu, Big da620 ta cleaning based on mobile edge computing in industrial sensor-cloud, IEEE Transactions on Industrial Informatics, 2019. doi:10.1109/TII. 2019.2938861.

[21] L. Qi, X. Zhang, W. Dou, Q. Ni, A distributed locality-sensitive hashingbased approach for cloud service recommendation from multi-source data, 625 IEEE Journal on Selected Areas in Communications, 35 (11) (2017) 2616– 2624.

[22] T. Wang, D. Zhao, S. Cai, W. Jia, A. Liu, Bidirectional prediction based underwater data collection protocol for end-edge-cloud orchestrated system, IEEE Transactions on Industrial Informatics, 2019. doi:10.1109/TII. 630 2019.2940745.

[23] W. Gong, L. Qi, Y. Xu, Privacy-aware multidimensional mobile service qual645 ity prediction and recommendation in distributed fog environment, Wireless Communications and Mobile Computing, 2018 (2018)

[24] Wong RCW, Fu AWC, Wang K, et al. Can the utility of anonymized data be used for privacy breaches? ACM Trans Knowledge Discov Data 2011;5:1–24.

[25] Friedman A, Schuster A. Data mining with differential privacy. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10); New York, NY, USA: ACM Press, 2010:493–502.

[26] Fung BCM, Wang K, Yu PS. Anonymizing classification data for privacy preservation. IEEE Trans Knowledge Data Eng 2007;19:711–25.

[27] McSherry F, Talwar K. Mechanism design via differential privacy. Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07); IEEE, 2007:94–103.

[28] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter intelligent monitoring in intensive care II: a public-access intensive care unit database. Crit Care Med 2011;39:952–60.

[29] Pavan Kumar Vadrevu, Sri Krishna Adusumalli, Vamsi Krishna Mangalapalli, A Survey on Personal Privacy Preserving Data Publication in IoT , International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-6C2, April 2019

[30] Pavan Kumar Vadrevu, Sri Krishna Adusumalli, Vamsi Krishna Mangalapalli, Motion Detection to Preserve Personal Privacy from Surveillance Data using Contrary Motion, International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8 Issue-6, March 2020