Multi-Level Privacy Preservation Using Rotation Perturbation

R.PraveenaPriyadarsini [#1], Dr.M.L.Valarmathi [*2], Dr.S.Sivakumari [#3]

[1,3]Dept. of Computer Science & Engg.,
Avinashilingam University for Women, Coimbatore
[1]praveena.priya04@gmail.com
[3]hodcseau@gmail.com
[2] Dept. of Computer Science & Engg.,
Government College of Technology,
Coimbatore, India
[2] ml_valarmathi@rediffmail.com

Abstract— As the amount of data available and shared in the electronic media increases, the threat for its privacy and security also increases. Every organization publishes its data to many recipients for various reasons .Thus preserving the privacy of data during the process of data mining is the aim of Privacy Preserving Data Mining [PPDM]. Multi-trust level is a scenario in privacy preserving data mining where different versions of privacy preserved data are distributed to the users based on their trust level. This work presents three trigonometric based rotation perturbation algorithms for privacy preservation at multi-trust level. These algorithms are applied on three bench marked datasets to generate multiple sequential dataset versions at various privacy levels. The perturbed datasets are evaluated on utility, distortion rate and the ability to prevent linking attacks. The results show that the perturbed datasets have utility comparable to the original datasets and linking attacks are prevented.

Keywords: Privacy preserving data mining, Multi-Trust Level scenario, Rotation perturbation, utility, distortion rate, linking attacks

I .INTRODUCTION

The volume of data available on the internet has increased in the recent years. This makes the personal or organizational information contained in these data susceptible to breach of privacy. When these data are published or used for research and other purpose there is a risk that they may expose the private or sensitive information about the person or organization .Privacy Preserving Data Mining (PPDM) is an important field of research which tries to protect the sensitive and private information during the process of data mining [1]. There are many techniques used for privacy preservation like anonymization, suppression and generalization. A few directions in PPDM are privacy preserving data publishing, changing the data mining results to prevent privacy breach, query auditing and challenges of high dimensionality. A scenario in privacy preserving data publishing is Multi-Level Perturbation (MLP) where the objective is to perturb the dataset in such a way that sequential release of various versions of dataset at different privacy level can be released. The problem in this MLP is that the perturbed versions can be linked together to form the original data. Li and Chen [12] have discussed about the dimension of privacy preservation namely Multi Level Trust (MLT). In this scenario, diversity and linking attacks are unique challenges for privacy preservation. Hence in a Multi-Trust Level and MLP privacy preservation environment, where multiple users receive datasets with various privacy levels ,diversity and linking attack posses a great challenge to the data owner. In order to address this challenge, the present work propose three trigonometric based rotation perturbation algorithms for producing multiple versions of the dataset sequentially to be published to the users based on their trust level .The trust level of the user is derived by the owner of the dataset. Rotation perturbation is a privacy preservation technique that perturbs all the attributes of the dataset using uniform transformation function that transforms all the data in the dataset. Also, they maintain the geometric property of the data.

In the literature[16,23],rotation perturbation has be used to produce a single perturbed dataset ready for publishing without considering the Multi Trust Level scenario. Also, in the work [16,23] in the literature, datasets are normalized for the base level of perturbation and rotation function is then applied . The objective of the present work is to produce multiple releases of dataset that are perturbed at various privacy levels to be published to receivers with various trust levels using trigonometric rotation perturbation. Also, the produced datasets should have properties such that they cannot be linked together to obtain the original data. In the present work, the datasets are perturbed in two levels; one is the base level and the next is the derived level without normalizing the original dataset such that the value of metric variance of the attributes as well as the dataset is positive.

The proposed perturbation algorithms are as follows:

- Trigonometric single layer rotation perturbation
- Trigonometric double layer rotation perturbation
- Trigonometric cross rotation perturbation

This paper is organized as follows: Section 2 gives the Literature Survey, Section 3 explains the proposed trigonometric rotation perturbation algorithms, Section 4 gives the dataset description, Section 5 explains the evaluation metrics used in this work, Results and its analysis are explained in Section 6, Discussions on the obtained results are given in Section 7 and Section 8 gives the conclusions and the future work of this paper.

II. LITERATURE SURVEY

Data issued for research purpose tend to reveal sensitive information an about the individual or organization. Publishing data in an environment such that sensitive information in the dataset is preserved is termed as Privacy Preserving Data Publishing [PPDP]. Rakesh Aggarwal and Ramakrishna Srikant [2] state that data sets contain attributes that might be very sensitive and when these values are revealed to the third party, they expose personal information or confidential information. They have used randomization technique, where the sensitive information is replaced by some random value based on probability distribution. Various challenges and techniques in privacy preserving data publishing have been discussed by Fung at al [3]. In the Framework for high-accuracy privacy-preserving mining [FRAPP],a generalized matrix was designed to perturb data for PPDM. The randomized perturbed datasets obtained from this framework was tested for its utility on association and classification algorithms [4]. The privacy preservation using randomization techniques can be simply reconstructed by any attacker by using Expectation Maximization algorithm. An amplification of randomization technique which counters the privacy breaches that occur due to randomization was proposed by Alexandre.et.al., [5]. Aggarwal.et.al, [6] has discussed about expectation maximization algorithm, which provides privacy preservation as well as prevents information loss . Xiaokui Xiao and Yufei Tao[7] have proposed a technique called Anatomy which separates the quasi-identifiers (QI) and sensitive attributes (SA) into two separate tables and combine these attributes using grouping mechanisms to protect the privacy. They have proved that this is a better technique than the generalization. QI attributes are the set of attributes that could contain information to indentify the tuples of the datasets. Motwani and xu [8] have devised a technique to identify QI attributes and mask them. They had applied their techniques to a stream database. Lindell and Pinkas[9] have developed a protocol for preserving privacy to secure multiparty computation environment. The protocol devised by them allows two parties to run data mining algorithm on the union of the data bases such that unnecessary information is not revealed to either of the party. Keyvanpour [10] has discussed about the various techniques used in PPDM. Random rotation perturbation technique which preserves the Euclidean distance and inner product of data in a multi-dimensional space was proposed by Xiao.et.al., [12] for multilevel perturbation that aims at releasing multiple versions of datasets anonymized at different privacy levels. They have also proved that collusion / union of these sequential releases of datasets are useless. Charu C.Aggarwal [13] has proved that dataset with large number of features are prone to inference attacks. Thus, high dimensionality is also a problem in privacy preserving data mining. Zhijan Zham and Wen hang Du[14] have proposed multi group randomized response techniques which partitions data and assigns them to a group. Each group is randomized separately. They have used the accuracy of decision tree classifier to evaluate the effectiveness of this method. Keke chen and Lig Lin [15] have proved that the geometric property of dataset are preserved by geometric rotation. They have evaluated the quality of multidimensional perturbation and proved that their approach of random rotation perturbation provides high privacy guarantee and also maintains a zero loss of accuracy. Aggarwal and Yu [17] have developed a new condensation approach to privacy preserving data mining that maps the original dataset into a privacy preserved dataset. This multidimensional perturbation approach preserves the privacy of the datasets and all the characteristics and correlation of the original dataset. Li.et.al, has proposed a framework where the most trusted data miner is given a less perturbed data. They have proved that their framework would counter diversity attack [11]. S.M.Oliveira and O. R.Zaïane [22] have used rotation perturbation to preserve the privacy of data. Also variance and security metrics were used to validate their work. The rotation based perturbation method has good utility and the user chooses the security level of the rotation.

III.PROPOSED TRIGONOMETRIC ROTATION PERTURBATION ALGORITHMS

The following definitions are assumed for rotation perturbation in the present work

Definition 1:

Single layer rotation perturbation: Let D_{mxn} be the data matrix with m columns and n rows, the single layer rotation(SLR) function f(SLR) transforms the data matrix D to D’ such that variance (m’)>0 and security (m’)>0 where m’ $\in$ D’.

Definition 2:
Double /Cross rotation perturbation: The data matrix D'$_{mxn}$ obtained applying function f(SLR) on D$_{mxn}$ is rotated again using either function of Double Layer Rotation(DLR)-f(DLR) or function of Cross Layer Rotation(CLR) f(CLR) to produce D"$_{mxn}$ data matrix such that variance (m")>0 and security (m")>0 where m''∈ D"$_{mxn}$ .Also, the following theorem has been referred from literature [15]:
Theorem 1: SVM classifiers using radial basis, linear kernels and Perceptron classifier are strictly invariant to rotation perturbation.
Theorem 2: KNN classifies are strictly invariant to rotation and translation perturbation.
Rotation perturbation produces multi dimensional robust perturbation on all the attributes of dataset. Also, it maintains the geometric properties of the original data such that perturbed copy produces the same result on data mining rotation invariant classification algorithms. The datasets are rotated by the proposed algorithms and sequential releases of dataset at various privacy levels are generated. The flow of the proposed work is given in Fig. 1.

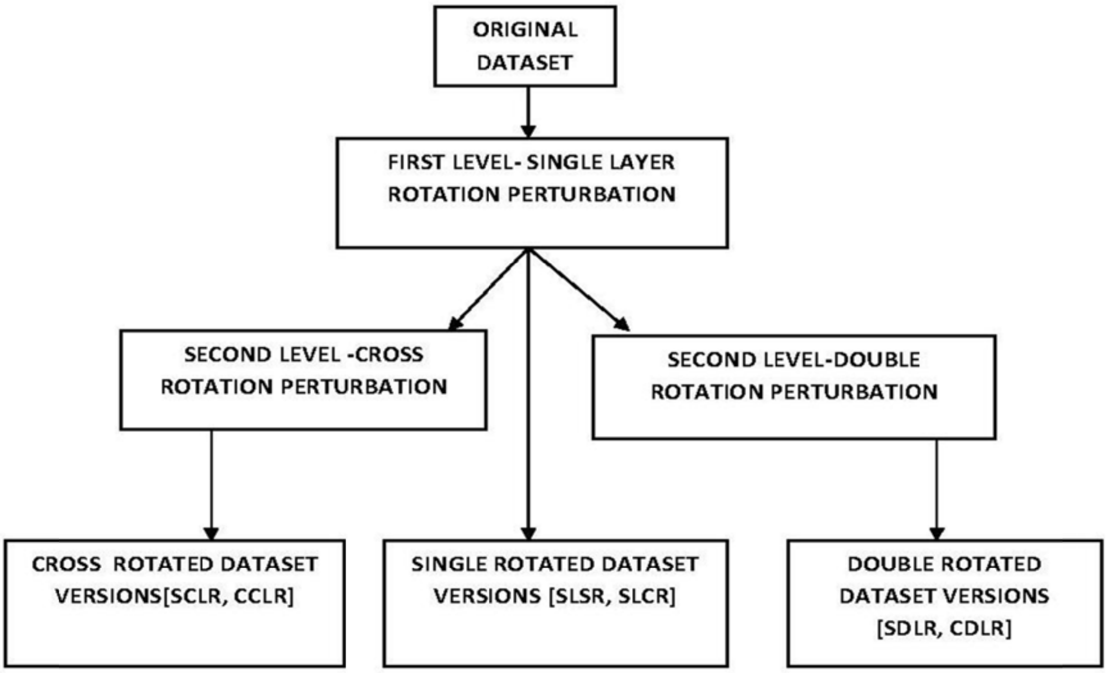

ORIGINAL DATASET
FIRST LEVEL- SINGLE LAYER ROTATION PERTURBATION
SECOND LEVEL -CROSS ROTATION PERTURBATION
SECOND LEVEL-DOUBLE ROTATION PERTURBATION
CROSS ROTATED DATASET VERSIONS[SCLR, CCLR]
SINGLE ROTATED DATASET VERSIONS [SLSR, SLCR]
DOUBLE ROTATED DATASET VERSIONS [SDLR, CDLR]

Fig 1: Model Flow Diagram

A. Trigonometric single layer rotation perturbation

The trigonometric functions can work only on numeric values. Hence, this perturbation technique is suited for numeric datasets. The datasets which have categorical values have to be converted into numerical values for applying rotation perturbation. Rotation perturbation is performed by applying a function that will perturb the data to a higher dimensional plane. In the proposed single layer rotation perturbation algorithm, sine and cosine values of the original data is used to distort the dataset. The working of trigonometric rotation perturbation is given below:

$$\text{Let the original dataset R} = \begin{matrix} r_{1,1} & r_{1,2} & \cdots & r_{1,m} \\ r_{2,1} & r_{2,2} & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ r_{n,1} & r_{n,2} & \cdots & r_{n,m} \end{matrix} \quad \text{----------------- (1)}$$

The data in the dataset R is transformed using the equation (2) or (3) in single layer rotation algorithm

$$\text{The rotation function } f(r) = \sin(\mathrm{r}) \times \mathrm{r} \quad \text{----------------- (2)}$$
$$\text{The rotation function } f(r) = cos(\mathrm{r}) \times r \quad \text{----------------- (3)}$$

Where r ∈ R

Perturbed dataset (R^1) = $$\begin{matrix} r'_{1,1} & r'_{1,2} & \cdots & r'_{1,m} \\ r'_{2,1} & r'_{2,2} & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ r'_{n,1} & r'_{n,2} & \cdots & r'_{n,m} \end{matrix}$$ ----------------- (4)

The single layered Sine /Cosine rotation perturbation is obtained using the equations (2) or (3).The proposed Sine/cosine single layer perturbation algorithm generates the perturbed version of the original dataset R namely Single Layer Sine Rotation perturbation (SLSR) using the equation (2) and Single Layer Cosine Rotation perturbation (SLCR) such that equation (5) and (6) will hold true on both the rotated datasets

Let r ∈ R, r' ∈ D'
Let r ∈ R, r' ∈ D' where D' = Sine rotated dataset/cosine rotated dataset then

$\text{SLSR} \ \cup \ \text{SLCR} \ \neq \ \text{R}$ -------------- (5)
$Variance\ (r')\ \&\ Security\ (r') > 0$ -------------- (6)

Since the rotation perturbation perturbs all the attributes of the dataset a unified measure like CK Value [18], security and variance metrics [22] are used to measure the privacy of the attribute in the datasets. The utility of these single layer perturbed datasets are studied using the rotation invariant classifiers like linear, RBF kernel of SVM classifier [16] KNN and Perceptron classifiers [15]. The algorithm for the Single Layer Rotation Perturbation is given in Fig 2.

```
Single layer rotation SLR Algorithm
Input $D_{mxn}$
Output $D'_{mxn}$
Let ($a_1$, $a_2$, ……….$a_n$) be the attribute of dataset D
If $D'_{mxn}$=Single layer sine rotation perturbation (SLSR) then
for ($a_i$= to $a_{i=n}$){for ($a_{n=1}$ to $a_{n=m}$){
$a'_1$=$a_n$* sin ($a_n$) }}
IF variance ($a'_1$)>0 && Security ($a'_1$)>0 then
SSLR=$a'_1$ U $a'_2$…….. U $a'_n$
Return (SSLR)
single layer cosine rotation perturbation
if $D'_{mxn}$ = single layer cosine rotation perturbation [SLCR]then
for ($a_i$= to $a_{i=n}$){for ($a_{n=1}$ to $a_{n=n}$){
a'=$a_n$* cos ($a_n$) }}
IF variance ($a'_1$)>0 && Security ($a'_1$)>0 then
CSLR=$a_1'$ U $a_2'$ ………..U $a_n'$
Return (CSLR)
}
```

Fig. 2: Algorithm Single Layer Rotation Perturbation

By applying the above algorithm the sequential versions of perturbed datasets obtained for the three datasets Adult, Wisconsin Breast Cancer, Ecoli dataset are given in Table I:

TABLE I: The Sequential Versions of Perturbed Datasets Obtained For the three datasets Adult, Wisconsin Breast Cancer and Ecoli

Datasets / **Perturbation Algorithms**	**Adult**	**Wisconsin Breast Cancer**	**Ecoli**
Single Layer Sine Rotation (SLSR)	Adult-Sin Single Rotated Dataset (ASLSR)	BC- Sin Single Rotated Dataset (BCSLSR)	Ecoli- Sin Single Rotated Dataset (ECSLSR)
Single Layer Cosine Rotation (SLCR)	Adult-Cos Single Rotated Dataset (ASLCR)	BC- Cos Single Rotated Dataset (BCSLCR)	Ecoli- Cos Single Rotated Dataset (ECSLCR)

B. Trigonometric Double layer rotation perturbation Algorithm

In the proposed double layer rotation perturbation algorithm, two levels of perturbation are performed. The SLSR and SLCR datasets of the first level perturbation is given as input to this algorithm. The second level of distortion is performed by multiplying the SLSR dataset with sine values of the corresponding data as shown below:

Let $r_1,r_2,\ldots r_n$ be the attribute such that $r_n \in$ D' where D' = Single Layer Sine Rotated dataset (SLSR) or Single Layer Cosine Rotated dataset (SLCR) . The double distorted Sine Double Rotation Perturbation (SDRP) and Cosine Double Rotation Perturbation (CDRP) are generated as given in equations (7) and (8) such that it satisfies the condition in equations (9) and (10) respectively. The security and variance values of the rotated attributes and datasets are calculated and only those attributes and datasets which have positive variance and security values are used for publishing.

The rotation function $f(r) = \mathrm{r}' \times \sin(\mathrm{r}')$ Where r'∈ SLSR -------- (7)
The rotation function $f(r) = r' \times \cos(r')$ Where r' ∈ SLCR -------- (8)

Such that Equations (7) and (8) holds true on Equations (9) and (10) respectively

If $r' \in \mathrm{SDLR} / \mathrm{CDLR}, r \in \mathrm{SLSR} / \mathrm{SLCR}$
Then $\frac{r'}{r} \neq r$ ---------(9)
$Variance\ (r')\ \&\ Security\ (r') > 0$ ---------(10)

This perturbation algorithm also distorts the values of the dataset more than the first level rotation. Also two versions of the dataset cannot be combined to obtain the third version as given in Equations (11) and (12)

$\mathrm{SLSR} \cup \mathrm{SDLR} \neq \text{Original Dataset R, SLSR}$ --------- (11)
$\mathrm{SLCR} \cup \mathrm{CDLR} \neq \mathrm{R, SLCR}$ --------- (12)

The algorithm for Double rotation Perturbation is given in Fig:3.

```
**Double layer rotation perturbation algorithm**
Input $D'_{mxn}$ (SLSR)
Output $D''_{mxn}$ (SDLR)
Let ($a'_{1}$, $a'_{2}$, ……….$a'_{n}$) € $D'_{mxn}$
If $D''_{mxn}$= Sine Double Level Rotation Perturbation (SDLR) then
for ($a'_{i=1}$ to $a'_{i=n}$ ){for  ($a'_{n=1}$ to $a'_{n=m}$){
$a''_{in}$= $a'_{in}$ * sin ($a'_{in}$) }}
IF  variance ($a''_{1}$)>0 && Security ($a''_{1}$)>0 then
$D''$ (SDRP) = $a''_{1}$ U $a_2''$ U $a_3''$ U……… $a_n''$
Return (SDLR)
}
else if $D'_{mxn\ =}$ Cosine Double Level Rotation Perturbation (CDLR) then
for ($a'_{i=1}$ to $a'_{i=n}$ ){
for  ($a'_{n=1}$ to $a'_{n=m}$){
$a''_{in}$= cos ($a'_{in}$) * $a'_{in}$
}
IF  variance ($a''_{1}$)>0 && Security ($a''_{1}$)>0 then
$D''$ (CDLR) = $a''_{1}$ U $a_2''$ U $a_3''$ U……… $a_n''$
}
Return (CDLR)
```

Figure 3: Algorithm for Double rotation Perturbation

By applying the above algorithm the perturbed sequential release of datasets are obtained for the three datasets used for experimentation are shown in Table II.

TABLE II: The Sequential Versions of Perturbed Datasets Obtained using Double Level Rotation Perturbation Algorithm

Datasets / **Perturbation Algorithms**	**Adult**	**Wisconsin Breast Cancer**	**Ecoli**
Sine Double Level Rotation (SDLR)	Adult-Sin Double Rotated Dataset (ASDLR)	BC-Sin Double Rotated Dataset (BSDLR)	Ecoli- Sin Double Rotated Dataset (ECSDLR)
Cosine Double Level Rotation (CDLR)	Adult-Cos Double Rotated Dataset (ACDLR)	BC- Cos Double Rotated Dataset (BCDLR)	Ecoli- Cos Double Rotated Dataset (ECCDLR)

C. Trigonometric Cross Layer Rotation Perturbation Algorithm

To increase the privacy of the dataset and also to prevent linking attack in a multi trust level environment, multi-dimensional trigonometric cross perturbation algorithm is proposed. In the cross perturbation algorithm the distorted datasets from the first level of distortion is taken as input. As a second level of distortion, Sine distorted single level perturbed data is multiplied with the Cosine value of the data and Cosine distorted data in the dataset CSLR is perturbed by multiplying it with the corresponding Sine value of the data as given in equations(13) and (14) respectively. Such that Equation (15) and (16) true

$$\mathrm{SCLR} = f(r) = r' \times \cos(r') \qquad \text{--------- (13)}$$

$$\mathrm{CCLR} = f(r) = r' \times \sin(r') \qquad \text{--------- (14)}$$

Where r' ∈ SLCR/SLSR

If r' ∈ CCLR/SCLR, r ∈SLCR/SLSR

Then $\frac{r'}{r} \neq r$ --------- (15)

$Variance\ (r')\&\ Security\ (r') > 0$ --------- (16)

Thus, dataset perturbed using this algorithm cannot be linked together with the previous level distorted dataset to obtain any new result. The algorithm for Cross Rotation Perturbation is given in Fig. 4.

```
Cross layer rotation perturbation algorithm
Input $D'_{mxn}$ (SLSR / SLCR)
Output $D''_{mxn}$ (CLSR / CLCR)
Let $(a'_1, a'_2, \ldots\ldots a'_n)$ € $D^1_{mxn}$
$D^1_{mxn}$ = Sine Cross Level Rotation Perturbation (SCLR)  then
for ($a'_{i=1}$ to $a'_{i=n}$ ){for  ($a'_{n=1}$ to $a'_{n=m}$){
$a''_{in} = a'_{in} * \cos(a'_{in})$ }}
IF  var ($a''$)>0 && Security( $a''$ ) >0 then
$D''_{mxn} = a''_1$ U $a_2''$ U $a_3''$ U……… $a_n''$
return (SCLR)
if $D''_{mxn}$ = Cosine Cross Level Rotation Perturbation (CCLR)   then
for ($a'_{i=1}$ to $a'_{i=n}$ ){for  ($a'_{n=1}$ to $a'_{n=m}$)
{
$a''_{in} = \sin(a'_{in}) * a'_{in}$}}
IF  var ($a''$)>0 && Security( $a''$ ) >0 then
$D''_{mxn} = a''_1$ U $a_2''$ U $a_3''$ U……… $a_n''$
Return CCLR
}
```

Fig. 4: Cross layer rotation perturbation algorithm

As given in Equation (15) one version of dataset cannot be obtained from another. The datasets generated using the above three algorithms can be counter linking attack since combining two versions of the perturbed dataset will not yield new information as given in Equations(17)and(18)

$\mathrm{SCLR}\ \cup \mathrm{SDLR}\ \neq \mathrm{SLSR}, \text{Original Dataset R}$ --------- (17)

$\mathrm{CCLR}\ \cup \mathrm{CDLR}\ \neq \mathrm{SLCR}, \text{Original Dataset R}$ --------- (18)

By applying the above algorithm the perturbed datasets obtained from the three datasets Adult, Wisconsin Breast Cancer, Ecoli dataset are shown in Table III.

TABLE III: The Sequential Versions of Perturbed Datasets Obtained using Cross Level Rotation Perturbation Algorithm

Datasets **Perturbation Algorithms**	**Adult**	**Wisconsin Breast Cancer**	**Ecoli**
Sine Cross Level Rotation (SCLR)	Adult-Sin Cross Level Rotated Dataset (ASCLR)	BC- Sin Cross Level Rotated Dataset (BCSCLR)	Ecoli- Sin Cross Level Rotated Dataset (ECSCLR)
Cosine Cross Level Rotation (CCLR)	Adult-Cos Cross Level Rotated Dataset (ACCLR)	BC- Cos Cross Level Rotated Dataset (BCCCLR)	Ecoli- Cos Cross Level Rotated Dataset (ECCCLR)

The utility and distortion of the data values of the dataset are the compared with the original dataset.

IV. DATASET DESCRIPTION

Three datasets Adult, Ecoli, Wisconsin Breast Cancer (WBC) datasets obtained from UCI Machine Learning Repository [19] have been used in the present work .Since, only numeric values can be used for rotation perturbation, Adult dataset which has categorical attributes was converted to numeric for experimentation.

A. Adult dataset

The Adult data set has 32,561 records where each record contains information about a person. There are 15 attributes including one class attribute, that has two categorical values, ">50K" and "<=50K". The description of the dataset is shown in TABLE IV.

TABLE IV **-**Adult dataset description

DATASET	**ADULT**
Attribute Characteristics	Categorical, Integer
Number of Instances	48842
Number of Attributes	14
Missing Values	Yes
No. of classes	2

Only those attributes which could have personal information are considered for this work.

B. Wisconsin Breast Cancer (WBC) dataset

The WBC data set has 10 numerical non-class attributes and one categorical class attribute. The description of the dataset is shown in TABLE V.

TABLE V**:** WBC dataset description

DATASET	**WBC**
Attribute Characteristics:	Integer
Number of Instances	699
Number of Attributes	10
Missing Values	Yes
No. of classes	2

C. Ecoli dataset

The Ecoli data set has 8 numerical non-class attributes and one categorical class attribute. The class attribute has 8 values. The description of the dataset is shown in TABLE VI.

TABLE VI**:** Ecoli dataset description

DATASET	**ECOLI**
Attribute Characteristics	Integer
Number of Instances	336
Number of Attributes	09
Missing Values	No
No. of classes	8

V.EVALUATION METRICS

The experiments were conducted using WEKA [23] and Rapid miner [24] software. In this work, two types of evaluation are performed on the perturbed datasets, one for checking the utility and the second to measure the distortion in the rank of the attributes and privacy after perturbation. Utility is assessed using the classification accuracy of the various versions of datasets on the rotation invariant classifiers SVM-Linear, SVM-RBF [16], KNN and Perceptron classification algorithms.

A. Classification accuracy is defined as the ability of the classifier to classify the given dataset and the equation for the same is given in (19).

$$\text{Classification Accuracy} = \frac{No.of\ tuples\ correctly\ classified\ accuracy}{Total\ No.of\ tuples\ in\ the\ dataset} \qquad \text{------- (19)}$$

A. *CK-Value* is defined as the data distortion level of the perturbed datasets which is a measure for comparing the rank of the attribute ranked using Info-Gain ranker method before and after perturbation.

CK [18] gives the measure of the attributes that keep their ranks after the distortion. Hence, it is calculated as given in equation (20):

$$CK = \frac{\sum_{i=1}^{m} Ck^{i}}{m} \qquad \text{-------(20)}$$

$$\text{CK} = \begin{cases} \text{CK=1} & \text{if the rank original attribute = Rank of perturbed attribute} \\ \text{CK= 0} & \end{cases}$$

B. *Variance*: Variance [22] measures the variation in the values of the attribute and the dataset after perturbation and is calculated using equations (21) and (22) respectively:

$$\text{Variance of attribute=}\ e = \frac{1}{\text{N}} \sum_{\text{i}=1}^{\text{n}} (x_i - x'_i) \quad \text{--- (21)}$$

Where N=no. of tuples in the dataset, n=number of values in attribute, x_i=original value of the attribute, $x_i^{'}$ =Perturbed value of the attribute.

$$\text{Variance of the dataset (D) =}\ \sum_{\text{m}=1}^{\text{n}} \left(\frac{a_m}{\text{N}}\right) \quad \text{-------- (22)}$$

Where N= no. of attribute in the dataset, a_m = Variance of attributes in the dataset

C. *Security*: The security [22] gives the security level of the dataset D after perturbation and is calculated using equation (23)

$$\text{Security} = \frac{\text{var(x−y)}}{\text{var(x)}} \quad \text{-------- (23)}$$

Where x=original attribute value,
y=perturbed attribute value.

VI. EXPERIMENTAL RESULTS AND ANALYSIS

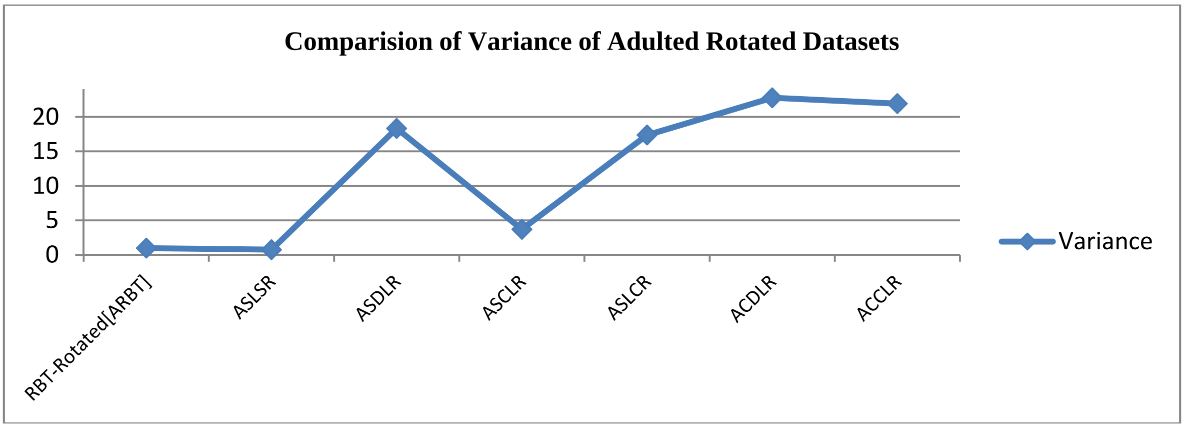
The perturbed datasets are generated from three bench marked dataset namely Adult, Wisconsin Breast Cancer and Ecoli datasets using three Trigonometric Rotation perturbation algorithms and are evaluated based on the following metrics: (i) utility of the dataset using accuracy on rotation invariant classifiers (ii) The distortion of the attributes using CK-Value [19] which compares the ranks of the attributes before and after perturbation based on information gain ranker method [21] (iii) Privacy of perturbed datasets using metrics Variance and Security.

The classification accuracy of the Adult dataset on the KNN, SVM-RBF, SVM-Linear and Perceptron algorithms are shown in the TABLE VII:

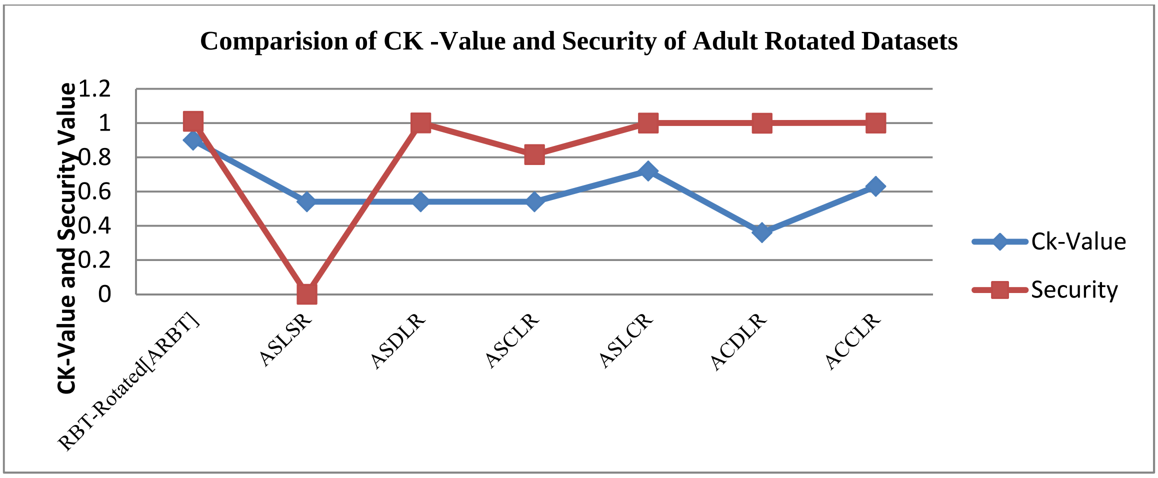
TABLE VII**:** Comparison of accuracies of Adult Rotated Datasets.

Adult Dataset-Rotated	**KNN**	**SVM-RADIAL**	**Perceptron**	**SVM-linear**
Adult-original	85.8722	77.40% +/- 0.97%	67.51% +/- 5.81%	75.68% +/- 0.27%
RBT-Rotated [ARBT]	85.0737	77.88% +/- 3.06%	69.65% +/- 6.06%	75.68% +/- 0.27%
ASLSR	85.3194	81.63% +/- 1.36%	49.56% +/- 11.42%	75.68% +/- 0.27%
ASDLR	85.8722	81.57% +/- 2.46%	66.77% +/- 7.15%	75.68% +/- 0.27%
ASCLR	85.1351	79.92% +/- 1.77%	64.80% +/- 4.85%	75.68% +/- 0.27%
ASLCR	84.2752	77.40% +/- 1.53	65.73% +/- 4.33%	75.68% +/- 0.27%
ACDLR	86.1794	79.73% +/- 1.15%	67.81% +/- 4.89%	75.68% +/- 0.27%
ACCLR	85.1966	77.64% +/- 1.73%	64.80% +/- 4.85%	76.78% +/- 1.50%

The TABLE VII shows that all the rotated Adult dataset versions ASLSR, ASDLR, ASCLR, ASLCR, ACDLR, ACCLR are rotated using the three trigonometric rotation perturbation algorithms. When the accuracy on SVM-RBF kernel is compared with RBT Perturbed dataset [15] ASDLR yields the highest accuracy. When the accuracy of trigonometric distorted datasets on SVM-LINEAR kernel is compared with the original and RBT perturbed dataset, all the trigonometric distorted datasets have the same accuracy value as that of the original dataset. On Perceptron algorithm all the datasets except ASLSR give the same accuracy as that of the original dataset. On KNN algorithm all the rotated datasets give the same accuracy. The Comparisons of Adult rotated dataset on the privacy metrics Variance, CK value and Security is shown in Graph 1 and 2.



Graph 1**:** Comparisons of Variance of Adult Rotated dataset versions.



Graph 2 **:** Comparisons of Ck-Value and Security of Rotated Adult dataset versions.

It is inferred from the graph1 that the variance of all the cosine rotated datasets are higher than the sine rotated and RBT rotated dataset, while the CK value and the security of all the rotated dataset are almost the same. Lower the CK value the more distorted is the dataset. The results indicate that ACDLR and ACCLR datasets are more distorted than the other perturbed versions.

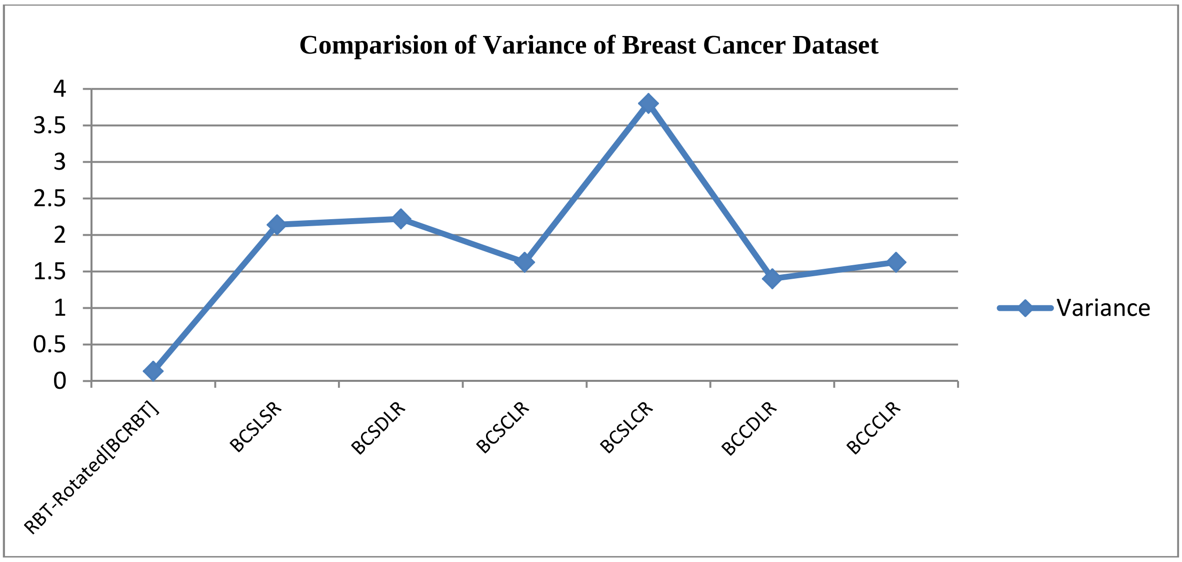
The classification accuracy of the Wisconsin Breast Cancer rotated datasets on the SVM-RBF, SVM-Linear, KNN and Perceptron algorithms are compared and shown in the TABLE VIII

TABLE VIII: Comparison of accuracies of Wisconsin Breast Cancer Rotated Datasets

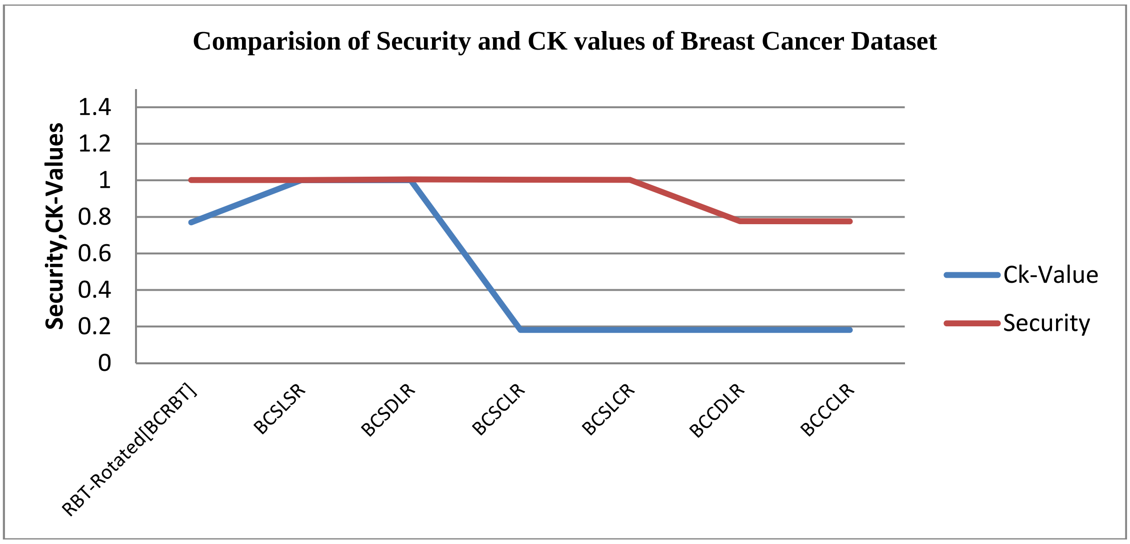
Breast-cancer Dataset-Rotated	**KNN**	**SVM-RADIAL**	**Perceptron**	**SVM-linear**
BC-original	95.1359	77.40% +/- 0.97%	34.48% +/- 0.68%	34.48% +/- 0.68%
RBT-Rotated[BCRBT]	87.1245	77.88% +/- 3.06%	34.48% +/- 0.68%	34.48% +/- 0.68%
BCSLSR	93.4192	40.05% +/- 16.66%	39.90% +/- 16.24%	84.99% +/- 5.45%
BCSDLR	94.134	81.57% +/- 2.46%	87.13% +/- 3.48%	87.12% +/- 3.00%
BCSCLR	94.4206	79.92% +/- 1.77%	37.33% +/- 9.00%	34.48% +/- 0.61%
BCSLCR	94.1345	77.40% +/- 1.53	91.13% +/- 3.12%	89.99% +/- 2.30%
BCCDLR	93.5622	81.08% +/- 1.08%	34.48% +/- 0.68%	34.48% +/- 0.68%
BCCCLR	94.4206	79.30% +/- 2.11%	34.48% +/- 0.68%	34.48% +/- 0.68%

The above TABLE VIII shows that BCSLR dataset gives the lowest accuracy on SVM RBF classifier. On SVM Linear algorithm BCSLSR, BCSDLR, BCSLCR version gives the highest accuracy than the original dataset. On KNN classifier all the proposed rotated dataset have ± 1% accuracy as that of the original dataset. On Perceptron algorithm except BCSDLR dataset all others have a higher accuracy than the original dataset.

Variance, Security and CK Values of rotated breast cancer datasets are shown in Graphs 3 and 4,



Graph.3**:** Comparisons of Variance of breast-cancer dataset versions



Graph.4: Comparisons of Ck-Value and Security of breast-cancer dataset versions

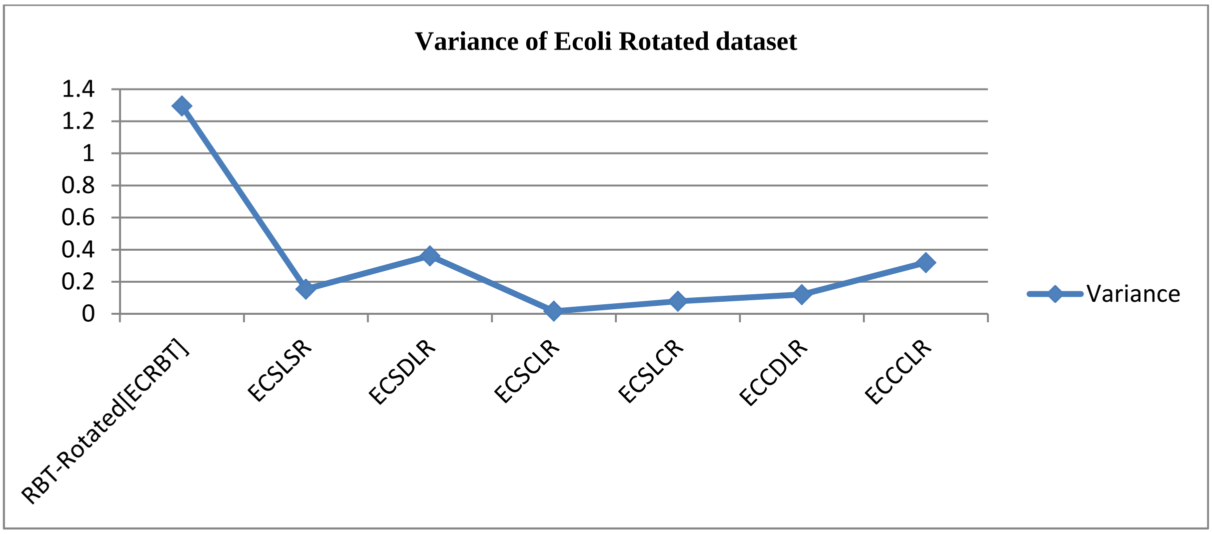
Graph 3 shows that the variances of all the proposed trigonometric rotated datasets are higher than RBT rotated dataset. Graph 4 indicates that CK values of BCSCLR, BCSLCR, BCCDLR and BCCCLR are lower indicating that these are the highest distorted dataset. The security of BCCDLR and BCCCLR are lowest among the distorted dataset.

The classification accuracy of rotated Ecoli-dataset on SVM-RBF, SVM-Linear, KNN and Perceptron algorithms are shown in TABLE IX:

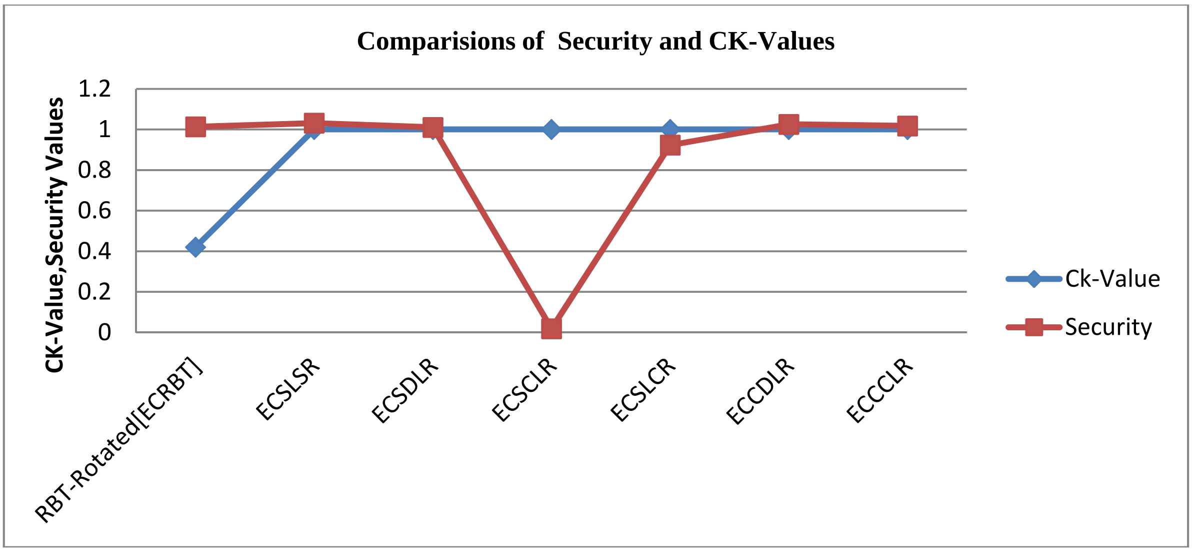
TABLE IX: Comparison of Classification Accuracy of Ecoli Dataset Versions

ECOLI- Dataset	**KNN**	**SVM-RBF**	**SVM-LINEAR**	**Perceptron**
ECOLI-original	80.3571	64.09%	65.03% +/- 2.24%	35.00%
RBT-Rotated [ECRBT]	80.6548	66.42% +/- 2.92%	66.42% +/- 2.92%	35.00%
ECSLSR	80.9524	65.03% +/- 2.24%	65.03% +/- 2.24%	35.00%
ECSDLR	83.0357	65.03% +/- 2.24%	65.03% +/- 2.24%	35.00%
ECSCLR	80.9524	65.03% +/- 2.24%	65.03% +/- 2.24%	35.00%
ECSLCR	78.869	65.03% +/- 2.24%	65.03% +/- 2.24%	35.00%
ECCDLR	77.9762	65.03% +/- 2.24%	65.03% +/- 2.24%	35.00%
ECCCLR	80.3571	65.03% +/- 2.24%	65.03% +/- 2.24%	35.00%

The TABLE IX shows that SVM-RBF classification accuracy on all the rotated datasets have an accuracy difference of ± 1% with the original dataset. On SVM linear and Perceptron classifier all the rotated datasets have the same accuracy as that of the original dataset. All the proposed rotated datasets except ECSLCR and ECCDLR give lower accuracy of about 3% from the original dataset on KNN classifier. The Graphs 5 and 6 show the comparison of all the privacy preserved versions of Ecoli dataset on privacy metrics, variance, CK value and security.



Graph 5: Comparison of Variance Ecoli dataset versions.



Graph **6**: Comparison of Ck-Value and Security metrics in Ecoli dataset versions.

The Graph 5 shows that the variance values of all the proposed datasets are lower compared to RBT rotated dataset. The graph 6 shows that except RBT rotated dataset all the other datasets show high CK value indicating that there is not much distortion in the information gain values of the attributes. The ECSCLR and ECSLCR datasets have lower security values than the other rotated datasets.

A. *Attack Analysis of Trigonometric Rotation Perturbation Algorithms*

The most prevalent attack in multi level perturbation environment is linking attack. The datasets rotated using the proposed algorithms obey the following equation 24 and 25, hence they cannot be rotated to the previous versions.

$\cos^{-1}(\mathrm{SLCR}) \neq m_{i\times j}$ -------------(24)

$\sin^{-1}(\mathrm{SLSR}) \neq m_{i\times j}$ -------------(25)

where $m_{i\times j} \in$ original dataset

Thus, the attacker cannot reverse the perturbed matrix back to the original matrix. Also, in a multi-level trust environment, each recipient is given a perturbed dataset with the level of perturbation based on their trust level. In linking attack two recipients of various trust levels combine their information to obtain more information than what is intended to them. In this proposed rotation perturbation the rotated datasets obeys the following equations:

$\mathrm{SLSR} \cup \mathrm{SLCR} \neq m_{i\times j}$ ------------- (26)

$\mathrm{SDLR} \cup \mathrm{SLSR} \neq m_{i\times j}$ ------------- (27)

$\mathrm{SCLR} \cup \mathrm{SDLR} \neq m_{i\times j}, \mathrm{SLSR}$ ------------- (28)

$\mathrm{CDLR} \cup \mathrm{SLCR} \neq m_{i\times j}$ ------------- (29)

$\mathrm{CDLR} \cup \mathrm{CCLR} \neq m_{i\times j}, \mathrm{SLCR}$ ------------- (30)

They prevent linking attack. As, the derived dataset are rotated on various Ө values from the base datasets, they cannot be linked or combined with the other versions to obtain the original dataset. Also the attacker who has various variants of perturbed datasets cannot reverse the transformation process since the variance values of each perturbed datasets are different. As the security and variance value of the perturbed dataset are different, linking and reversal back to the original dataset cannot be performed.

VII.RESULTS AND DISCUSSIONS

When the classification accuracy of the proposed rotated adult datasets are compared with RBF-rotated and original adult dataset on the accuracy of SVM-RBF classifier, ASLSR and ASDLR datasets have an increase of 4% accuracy than the original dataset. On SVM-Linear classifier all the other proposed rotated datasets have the same accuracy as the original dataset. Except for ASLSR dataset, all other rotated adult datasets give the same accuracy on Perception classifier. On KNN algorithm all the proposed rotated adult datasets have the same accuracy as that of the original dataset. The variances of adult rotated datasets are very high indicating that rotating them back to the original dataset is difficult. Also, all the versions have different variance values indicating that they cannot be linked back to the original dataset. The security values of the Adult rotated datasets are comparable with RBT rotated dataset. The CK value of all the adult rotated datasets are lower than that of RBT rotated datasets, indicating that the proposed rotated datasets are more distorted than the RBT rotated adult dataset. On Wisconsin Breast Cancer (WBC) dataset, the proposed rotated BCSSLR and BCSCLR dataset gives high accuracy on SVM linear classifier than the original dataset. On KNN Classifier all the proposed rotated WBC datasets gives accuracy of 1 or 2% lesser than the original dataset. On Perceptron classifier, BCSDLR and BCSCLR give high accuracy than the original dataset. On SVM-RBF classifier, BCSSLR dataset gives lesser accuracy than the original dataset. All the other versions have the accuracy of the original dataset. When the variance values of the WBC rotated datasets are compared, BCSCLR has the highest variance value and each version also has an unique variance value. BCCDLR and BCCCLR have the lowest security value. Except for BCSSLR and BCSDLR versions all the attributes in the other distorted versions retain the information gain value even after distortion. When the accuracy of SVM-Linear and Perceptron classifier on Ecoli-rotated datasets are compared, they give the same accuracy as that of the original dataset. On SVM-RBF classifier all the proposed dataset have an increase of 1% or 2% than the original dataset. When the KNN classifier accuracy on the rotated datasets are compared, they give an accuracy of ±3% than the original dataset. The entire attribute set regains their information gain value which is indicated by the CK values of the rotated dataset. The variance and security value of the rotated dataset is comparable with RBT rotated dataset in Ecoli rotated datasets and also unique preventing linking attack.

VIII.CONCLUSIONS AND FUTURE WORK

This work has presented three trigonometric rotation perturbation algorithms for sequential release of data in multi-trust level privacy preserving data mining. The proposed algorithms rotated the values of the datasets to some random value in order to preserve privacy. As a multidimensional perturbation, transformations are applied on all the attributes of the dataset. The level of privacy preservation applied depends upon the trust level of the user. All the proposed techniques are found to be having good accuracy measure and thus, maintain the utility of the dataset. Linking attacks is prevented by the proposed method since two versions of dataset cannot be combined to get more information. As the proposed rotated dataset has a different variance value, they cannot be rotated back to the original dataset. The result also indicates that the rank value of the attributes do not change much after perturbation. As a future work, random projection rotation can be combined with rotation for enhancing the privacy preservation at Multi-trust level.

REFERENCES:

[1] C.Aggarwal and P.Yu, “Privacy-Preserving Data Mining:, Models and Algorithms”. *Springer*, 2008.

[2] R.Agrawal, and R.Srikant, “Privacy preserving data mining,” in *Proc. ACM SIGMOD Int’l Conf. on Management of Data*, 2000.

[3] **C.**Benjamin and M.Fung,” Privacy-Preserving Data Publishing: A Survey of Recent Developments” in *Proc. ACM SIGMOD Int’l Conf. on Management of Data*, 2000.

[4] S.Agrawal, and J.R.Haritsa, “A framework for high-accuracy privacy-preserving mining.” In *Proc. of IEEE Intl. Conf. on Data Eng. (ICDE)* (2005), pp. 193–204.

[5] A.Evfimievski, J.Gehrke, and R.Srikant, “Limiting privacy breaches in privacy preserving data mining,” in *Proc. ACM Symposium on Principles of Database Systems*, 2003.

[6] D.Agrawal, and C.C.Aggarwal, “On the design and quantification of privacy preserving data mining algorithms,” in *Proc. of the 20th ACM Symposium on Principles of Database Systems*, Santa Barbara, California, M ay 2001, pp. 247–255.

[7] X. Xiao and T.Tao, ” Anatomy: Simple and effective privacy preservation”. In *VLDB*, pages 139–150, 2006.

[8] R.Motwani and Y.Xu, "Efficient Algorithms for Masking and Finding Quasi-Identifiers", in *Very Large Data Bases (VLDB) Conference*, Vienna, Austria., 2007.

[9] Y.Lindell,and B.Pinkas, “Privacy preserving data mining”. *Journal of Cryptology 15*, 3 (2000).

[10] Mohammad Reza Keyvanpour,” Classification and Evaluation the Privacy Preserving Data Mining Techniques by using a Data Modification–based Framework”, *International Journal on Computer Science. and Engineering* (IJCSE),2008

[11] Y.Li, and M.Chen, “Enabling multi-level trust in privacy preserving data mining,” EECS Department, *University of California*, Berkeley, Tech. Rep. UCB/EECS-2008-156, Dec 2011.

[12] X.Xiao, Y.Tao, and M.Chen, "Optimal random perturbation at multiple privacy levels," in *Proc. Int'l Conf. on Very Large Data Bases*, 2012

[13] C.Aggarwal, "Privacy and the Dimensionality Curse," *Privacy-Preserving Data Mining*, pp. 433–460, 2008.

[14] W.Du, and Z. Zhan, "Using randomized response techniques for privacy-preserving data mining". In *Proceedings of the 9th ACM SIGKDD*. ACM, New York. 2003

[15] Keke Chin and Ling Liu, "A Random Rotation Perturbation Approach to Privacy Preserving Data Classification," *Science direct* 2012.

[16] David Meyera , Friedrich Leisch and Kurt Hornikb "The support vector machine under test" *Neurocomputing* 55 (2003) 169 – 186.

[17] C.C. Aggarwal, and P.S.Yu, "A condensation approach to privacy preserving data mining," in *Proc. Int'l Conf. On Extending Database Technology (EDBT)*, 2004.

[18] S.T. Xu, J.Zhang, D Han, and J.Wang, "A singular Value Decomposition Based Data Distortion Strategy for Privacy Protection,". *Knowledge and Information Systems* (KAIS) journal,2006.

[19] A. Frank, and A. Asuncion, (2010): *UCI Machine Learning Repository*[http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science

[20] http://en.wikipedia.org/wiki/List_of_trigonometry_topics.

[21] J. Han, and M. Kamber, 2001. "*Data Mining*". Morgan Kaufmann Publishers.

[22] R.M.Oliveira, and O.R. Zaïane, "Achieving Privacy Preservation When Sharing Data for Clustering", *Proceedings of the International Workshop on Secure Data Management in a Connected World* (2004) 67-82

[23] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten (2009): "The WEKA Data Mining Software: An Update"; *SIGKDD Explorations*, Volume 11, Issue 1.

[24] Rapid-I RapidMiner. available from *http://rapid- i.com* (re-trieved: December'09).