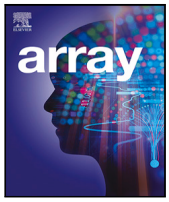
Array 20 (2023) 100324

Contents lists available at ScienceDirect

Array

journal homepage: www.elsevier.com/locate/array



Efficient perturbation techniques for preserving privacy of multivariate sensitive data

Mahbubur Rahman *, Mahit Kumar Paul, A.H.M. Sarowar Sattar

Department of Computer Science & Engineering, Rajshahi University of Engineering & Technology, Rajshahi-6204, Bangladesh

ARTICLE INFO

Keywords:
Data privacy
Information privacy
Perturbation
Data utility
Sensitive data

ABSTRACT

Cloud data is increasing significantly recently because of the advancement of technology which can contain individuals' sensitive information, such as medical diagnostics reports. While deriving knowledge from such sensitive data, different third party can get their hands on this sensitive information. Therefore, privacy preservation of such sensitive data has become a vital issue. Data perturbation is one of the most often used data mining approaches for safeguarding privacy. A significant challenge in data perturbation is balancing the privacy and utility of data. Securing an individual's privacy often entails the forfeiture of the data utility, and the contrary is true. Though there exist several approaches to deal with the trade-off between privacy and utility, researchers are always looking for new approaches. In order to address this critical issue, this paper proposes two data perturbation approaches namely NOS2R and NOS2R2. The proposed perturbation techniques are experimented with over ten benchmark UCI data set for analyzing privacy protection, information entropy, attack resistance, data utility, and classification error. The proposed approaches are compared with two existing approaches 3DRT and NRoReM. The thorough experimental analysis exhibits that the best-performing approach NOS2R2 offers 15.48% higher entropy and 15.53% more resistance against ICA attack compared to the best existing approach NRoReM. Furthermore, in terms of utility, the accuracy, f1-score, precision and recall of NOS2R2 perturbed data are 42.32%, 31.22%, 30.77% and 16.15% more close to the original data respectively than the NRoReM perturbed data.

1. Introduction

Because of the rapid growth of cloud technology and its storage ability, a big volume of data is being used constantly in this era of information technology. A substantial amount of data is transmitted and received on a daily basis via communication channels due to the rapid advancement of data gathering methods [1]. These data may contain many valuable and sensitive information such as commerce [1], education [2], banking [3], healthcare [3,4], lifestyle, habits and personal preferences [4]. People are often reluctant to give personally identifiable information. Such reluctance to provide personal information constrains the accessibility of vital data for knowledge discovery [5,6]. Data mining is the key technique of expertly identifying previously unknown patterns and information from data [7]. Many data mining techniques have made it possible to extract information from massive amounts of data [8]. Many companies and organizations utilize this data to make decisions in order to increase customer contentment [8]. The knowledge that is retrieved is crucial for modeling subsequent business operations. As a result, it is envisaged that during data mining operations, individual privacy should be protected and the data set's utility is anticipated to remain quite consistent with that of the actual data. Other big data repositories, such as crimes [9], epidemics [10], medical reports, customer information, and social physics [11] need the data to study human behavior and forecast events that are primarily centered on people. Therefore, protecting individuals' privacy while conducting analysis chores has become a crucial issue that necessitates forceful answers.

Let us think about the illustration in the Table 1 to highlight the privacy and utility scenario. It contains the customer ID, account number, amount of transaction and current balance of three customers of an organization. The customer ID, account number, amount of transaction and current balance of customers are highly sensitive and private. The organizations typically do not divulge any details about their clients. However, sometimes it may be important to evaluate this data to gain insight into the organization's and the customers' condition. The companies must study the customer information provided with a specific level of modification to achieve these goals of utility and privacy. Let us accomplish it by rotating in two dimensions at random angles $\theta_1 = 1.9°$ and $\theta_2 = 3.3°$ respectively. Since the 2-dimensional

* Corresponding author.
E-mail address: durlovrahman32@gmail.com (M. Rahman).

https://doi.org/10.1016/j.array.2023.100324
Received 14 May 2023; Received in revised form 21 September 2023; Accepted 30 September 2023
Available online 6 October 2023

Table 1
Customer information.

Customer ID	Account number	Amount of transaction	Current balance
8317	1325	8000	38211
9425	3026	10010	50000
1913	6022	13210	53250

Table 2
Rotated with smaller angles.

Customer ID	Account number	Amount of transaction	Current balance
8268.5	1600	5787.5	38 608.1
9319.5	3336.8	7115.2	50 493.3
1712.3	6082.1	10 122.8	53 922.1

Table 3
Rotated with larger angles.

Customer ID	Account number	Amount of transaction	Current balance
4605.1	7051.3	−26811.2	28 377.6
4089.3	9014.7	−35344.6	36 755.2
−3176.1	5462.3	−36187.4	41 238.7

rotation matrix has a dimension of 2×2, we analyzed two attributes at a time to fulfill matrix multiplication rule. So the first attribute pair is formed by customer ID and account number, whereas the second attribute pair is formed by amount of transaction and current balance. The 2-Dimensional rotation matrix by counterclockwise direction [12] is denoted by:

$$R_\theta = \begin{pmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{pmatrix}$$

The Table 2 represents the result after rotation. Then we use two larger random angles $\theta_1 = 47.8°$ and $\theta_2 = 55.2°$ which result is depicted in Table 3. Analyzing Tables 2 and 3, it is realized that the modified data in Table 2 is sufficiently similar to the original data. As a result, data mining algorithms find it quite simple to bring out information that closely resembles the actual data. However, it is difficult to hide the original data because the perturbed data is almost similar to the original one. On the other hand, it is observed that the revised data are very different from the original data in Table 3. Furthermore, some of them are negative. As a result, data mining algorithms have a great difficulty retrieving information unlike to the original data. However, in this scenario, approximating the information is difficult, which guarantees some privacy. As can be seen from Tables 2 and 3, one transformation fails to protect privacy but preserves the utility, while the other fails to keep up its utility but is able to safeguard privacy. A significant challenge in data mining is the trade-off between privacy and utility.

Privacy-preserving data mining (PPDM) is one of the well-known trends in privacy field. It is motivated by one of the information era's most important policy issues: the right to privacy. PPDM has been formed as a subject of study in which data mining techniques can be utilized to examine private confidential information about persons while protecting their privacy [5]. It is all about improving "traditional" data mining and making it more privacy conscious by employing unique privacy-preserving approaches that will create accurate data mining predictions, show fascinating data patterns, while also safeguarding privacy and data utility [13]. It has been extensively researched in the context of big data. In PPDM, the information that can be extracted from a modified data set by current data mining methods is virtually identical to what can be extracted from the actual data set [5,13]. The main objective of PPDM techniques is to ensure individual privacy while still considering data utility [1,5,6]. PPDM has two major concerns [5,14]. Firstly, the original data can only be accessed by the owner or authorized users [1] and only they can alter confidential raw data in a way that differs from the original database. Thus the privacy is protected. Second, such information ought to be eliminated from the data set altogether. So, the primary goal of data mining strategies based on PPDM is to develop methodologies that may change the original data while maintaining the privacy of the underlying information during the data mining process [5].

Privacy and data utility are two opposing factors in data mining. It is almost difficult and a huge challenge for any algorithm to be better than all other algorithms in terms of privacy, utility, complexity and so on [15]. Different research initiatives in the subject of PPDM are being examined to deal with this burgeoning challenge. There are different perturbation techniques, such as noise addition [16], condensation based approach [17], random projection-based approach [18] and many others. All these approaches have some limitations. The noise addition based perturbation approach generates random noise in the data set's particular columns [5]. There is another perturbation strategy which is based on randomization. Although this technique is straightforward, attacks based on reconstruction are a threat for this technique. On contrary, in another technique based on condensation, there is a privacy concern because it is highly likely that the perturbed data can be used to deduce the original data [19]. A few of the original data points are projected into a different space than the original multidimensional space in a random projection-based perturbation strategy [18]. Projection perturbation can preserve graphical fidelity because data set from different spaces can be merged more accurately than those from euclidean space. [20]. Consequently, the protection of sensitive details is mostly concentrated on various phases of a data mining task in other existing systems [5]. Several studies have also focused on the utility of data, such as in [5,21].

In this paper, two approaches namely NOS2R (Normalization, Scaling, Shearing and Reflection) based perturbation and NOS2R2 (Normalization, Scaling, Shearing, Reflection and Rotation) based perturbation have been proposed which maintain data utility and preserve privacy at the same time. Normalization is used to have a standard scale for every data point in the data matrix [5]. In this work, z-score is used as a normalization approach. Scaling is the technique of adjusting data. The coordinate points of the original data change as a result of scaling. It also changes the size of the data. Shearing deals with changing the geometric property along axes. The shape of the object is modified via shearing. In multi-dimensional space, rotation transformation preserves geometric attributes such as length, inner product, shapes such as hyper plane, and hyper curved surfaces. Data utility can be achieved by improving data perturbation parameters for scaling, shearing, reflection and rotation in a systematic way [3]. The privacy of the approach was then assessed in terms of privacy guarantee, attack resistance and an increase in entropy [3].

The remaining portion of this paper is organized as Section 2 summarizes the existing work which are related to this research work. Section 3 represents the overview of perturbation techniques. The metrics used for analyzing privacy and utility are discussed in Section 4. In Section 5, the proposed approaches NOS2R and NOS2R2 are described. Section 6 represents the experimental analysis and provides a comparative analysis of the performance. In the last section, our research work is concluded.

2. Related work

Individuals' information has become pervasive in today's world. In case of big data, there are a lot of potential as well as a lot of risk and on the other hand, a slew of security and privacy issues [22]. So, the researchers' mission of safeguarding privacy has been challenging. Several methodologies have been applied by various studies conducted over the last few years to discover solutions to this problem while maintaining utility and preserving privacy.

In [23], Chamikara et al. offers a perturbation algorithm called DISTPAB to preserve the privacy of horizontally partitioned data in a distributed context. This strategy reduces computational limitations by

dispersing the burden of privacy preservation across a dispersed ecosystem with resource-constrained devices and high-performance computers, taking advantage of the asymmetry of resources. But this work has a limitation i.e. it works only for horizontally partitioned data. So when some vertically partitioned data appears, this method cannot maintain the utility and privacy trade-off.

For sensitive data mining, a method called NRoReM is proposed by Paul et al. in [5]. This data perturbation approach consists of four stages. At first normalization, secondly geometric rotation, then linear regression, and finally scalar multiplication was followed. The exploratory study of privacy protection, attack resistance, information entropy analysis, data utility, and error analysis reveals that NRoReM protects individual privacy as well as data utility on a greater level. However, when scalar multiplication is used for better privacy, the size of the data set is altered, which sometimes results in a decrease in utility in some cases.

A method for protecting privacy is represented in [24] which is based on distance matrix. Its aim is to secure common comparable data that is disseminated among many organizations. In circumstances where the amount of data in each organization is limited and the data bias is substantial, the suggested method robustly integrates scattered data that is of the same quality as linked raw data. Furthermore, this proposed method can be used with data that has been corrupted by noise. Even after introducing noise, the statistics of the original data could still be calculated which is a threat and it necessitates a more robust method of privacy protection.

To alter and restore data, a method called RPCM is devised in [25] that leverages reversible data concealment mechanisms. According to the observations, when dealing with perturbed data, RPCM keeps almost similar knowledge as it does in original data and when the degree of data disruption grows, the risk of privacy leakage does not increase. In [26], a new data perturbation technique is proposed named Range Noise Perturbation (RNP). Data set is perturbed four times in this procedure. In the first step, it analyzes a group of attributes for perturbation; in the second, it selects a range noise value; and in the third step, it figures the range and determines the noise. The perturbed attributes are then used to update the data set. However, there is still a risk of privacy and utility trade-off because this method depends on the range noise value.

The paper [27] introduces a Bayesian-based PPDM approach for classification. This method is a data perturbation method which is algorithm-independent, implying that the perturbed data can be used directly by traditional classification methods. Experiments show that this strategy is effective at securing privacy while maintaining data utility. A semi-supervised privacy-preserving clustering approach is proposed in [28]. The method of this study learns a large margin nearest cluster metric using convex optimization utilizing a limited quantity of supervised data. The method then applies multiplicative perturbation on the original data based on the learned metric, which might change the distribution shape of the original data and therefore protect privacy information while assuring good data usability. Despite that, the inner-site sharing of the LMNC (Large Margin Nearest Cluster) metric within the distributed system, any site that acquires access to the modified data of others is endowed with the capacity to retrieve the corresponding original data.

Torra proposed a new approach microaggregation [29]. This method involves using a clustering algorithm to protect the data set, and then replacing each of the data with the cluster representative. Here, two parameters m_1 and m_2 are used for fuzzy partition and membership degrees respectively. So erroneous or faulty value of these parameters can still affect the balance between privacy and utility. The k-means technique is utilized in [30] to propose a Reversible Privacy-Preserving k-means Clustering (kRPP) algorithm that protects the clustering information of a data set. When the noise ratio or noise offset ratio is increased, the inserting noises to a cluster with a minimal total distance in the kRPP method has no significant effect on centroids movement. However, kmeans algorithm is not suitable idea for protecting privacy because this algorithm can worsen the attack resilience.

The authors of [31] suggested a data perturbation approach that follows min–max normalization. The technique starts with choosing the sensitive features and then applying min–max normalization to them. The altered attributes and the actual data are then blended in and some assessment is performed using this data. In this approach, min–max normalization has been used which is not suitable for the full feature data set. The security is jeopardized by the significant changes in the values' positions in the matrix. The authors of [8] suggested an approach which follows 3D rotation transformation. The technique is accomplished by rotating the attributes of data in a set of triplets. Then the angle of rotation is determined by thresholding to the variance of the attributes [32]. The balance between privacy and utility can be hampered because it uses a mixture of clockwise and anti-clockwise rotation for a better privacy at the expense of utility.

P^2R_0CAl is a privacy-preserving technique suggested by the authors of [4]. For balancing privacy and utility in data sets, this method makes use of rotation and condensation characteristics. In one version of this method, kmeans was used. This kmeans algorithm delivers better utility, but sometimes privacy suffers and attack resistance declines. A non-reversible perturbation method namely PABIDOT is proposed in [3] that uses optimal geometric transformation for privacy preservation of huge data. It uses multidimensional geometrical alternation, reflection, translation, and rotation to disrupt a data set, followed by randomized expansion and random tuple shift out.

For the sake of stream data mining privacy, two data perturbation methodologies are presented in [6] which use random projection, random translation, and two distinct additive noise structure in a coalition. Independent noise, which is produced for each individual entry, is one of the noise types. The other type of noise accumulates during the course of a stream's lifetime.

In [33], the authors proposed a new data perturbation algorithm, SEAL (Secure and Efficient data perturbation Algorithm utilizing Local differential privacy) which relies on Chebyshev interpolation and Laplacian noise. SEAL offers a fair balance of privacy and utility with high efficiency and scalability. In [34], the sensitive information of individuals is perturbed via a fuzzy data transformation technique. The technique uses a fuzzifier to transform linguistic data from crisp values so that the inference engine can compare them to the fuzzy rules which had already been stated. Then using a defuzzifier, perturbed data is then translated from the generated linguistic values to crisp values. In this approach, while perturbing the data, privacy is preserved, but the utility can be neglected.

Another approach namely RG+RP is proposed in [35] which consists of two stages. Each individual's data is perturbed in the first stage by running it through a nonlinear function called repeated Gompertz (RG); in the second stage, a specific random projection (RP) matrix is used to project the perturbed data to a lower dimension in a (almost) distance-preserving manner. Random projection avoids independent component analysis (ICA) attacks and ensures clustering accuracy, while maximum a posteriori (MAP) estimation attacks are mitigated by the nonlinear RG function. The recovery resistance of the proposed two-stage randomization method is evaluated in terms of MAP estimation assaults.

The authors of [36] discussed a method called value swapping. There exists no erroneous information as a result of value swapping. While retaining data set wise properties, methods like value swapping [36], condensation [19], random projection [35] and additive noise [6] sacrifice the secrecy of individual records by substituting sensitive information with less sensitive records.

Most of the existing methods still lacks in balancing the privacy and utility. Still a trade-off between utility and privacy has been noticed. In our work, the two proposed approaches follow a combination of normalization and some different transformations. When undergoing these various transformations, normalization aids in maintaining the

Euclidean distance between data points as closely as possible to the original data. And the combination of normalization and the transformations helps to preserve geometric features including length, inner product, hyper-plane forms, and hyper-curved surfaces. As our methods go through several stages of transformations, more data is being distorted. So the privacy of the data is being preserved. Thus our methods are able to provide significant degree of privacy and keep the utility at the same time.

3. Overview of perturbation technique

There are various number of innovative and effective approaches for preserving privacy. Each technique has advantages and disadvantages of its own [37]. For obtaining individual's privacy, several of the perturbation techniques employ many changes to the original data streams. Data privacy can be protected using these perturbation techniques and also the data utility is expected not to be overlooked. However, centralized and distributed scenarios can be initially classified as PPDM approaches [5]. This classification is determined by the position of computing for data mining performance. The perturbation method is well-suited to both centralized and distributed computing applications [38]. Perturbation techniques alter the original data before data mining activities. In most circumstances, individual records cannot be recovered, but aggregate distributions can [5]. All aggregated distributions are significant for data mining and analytical jobs. However, adversaries will have a difficult time extracting concealed information about individuals from the skewed published data [39]. There are two types of data perturbation techniques [40] — unidimensional and multidimensional data perturbation. Random noise addition is the most frequent unidimensional data perturbation technique, which is mathematically [41] expressed as Eq. (1).

$$x_i' = x_i + r_i \tag{1}$$

where x_i signifies the initial data value in a one-dimensional situation, x_i' represents the equivalent value of the processed data, and r_i means random noise from a distribution such as the Gaussian distribution. Unidimensional data perturbation strategies cannot give accurate mining results because data mining tasks necessitate knowledge from several correlated data dimensions [5,42]. On the contrary, multidimensional data perturbation approaches retain statistical information. When a value from the original multidimensional data is hidden, this is likewise true [42]. Multidimensional data perturbation methods change data across all dimensions. Multidimensional data perturbation procedures suffer from less information loss than uni-dimensional data perturbation strategies because a lot of statistical data is stored in various dimensions [5]. Random rotation transformation is described in [43] as Eq. (2).

$$g(X) = RX \tag{2}$$

Here R stands for random rotation matrix, X for original data attribute sets, and g(X) for changed attribute sets. The utility is maintained while the privacy of the user's data is increased through random rotation transformation.

4. Evaluation metrics

Perturbation approaches are commonly evaluated using two types of metrics: privacy metric and data utility metric. The most effective perturbation strategies aim to protect privacy and also maintain data utility. As described below, we have employed different criteria to quantify data utility and privacy in this study.

4.1. Privacy of data

The complexity of estimating the actual data items from the altered data is referred to as privacy [5]. We evaluated six different metrics to assess data privacy security in this study: Secrecy [5,44], Value Difference (VD) [5,45], RP [5,45], RK [5,45], CP [5,45], and CK [5,45].

4.1.1. Secrecy

The variance of actual and altered data values has traditionally been used to characterize the secrecy of a perturbation method [44]. Secrecy is defined as follows:

$$Secrecy = \frac{Variance(x_i - x_i')}{Variance(x_i)} \tag{3}$$

where x_i stands for a single original data attribute, while x_i' stands for the perturbed data attribute. This metric, secrecy measures privacy by comparing the initial data contents of an altered attribute to those which can be examined.

4.1.2. VD

VD uses the relative value difference in the Frobenius norm to express changes in original data items [45]. When the initial data values of a data matrix are altered, it is defined as the following equation in Frobenius norm [45].

$$VD = \frac{\|A - A'\|_F}{\|A\|_F} \tag{4}$$

where Frobenius norm is denoted by the subscript F, A denotes the actual data stream, and A' stands for the altered data stream.

4.1.3. RP

The ranks of the related data elements are altered when a data matrix is perturbed. The rank-based measure RP can be used to reflect mean changes in all attribute-specific ranks [45]. For example, we consider the data set $D_{m\times n}$, which has n attributes and m instances. RP is now defined as Eq. (5), where R_j^i represents the rank of the jth element of the original data matrix A_{ji} for the ith attribute and $R_j^{i'}$ represents the jth element's rank of the perturbed data stream A_{ji}' for the ith attribute.

$$RP = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}|R_j^i - R_j^{i'}|}{nm} \tag{5}$$

4.1.4. RK

Even after being perturbed, some data items can keep their distinct magnitude ranks for each of the attributes. RK can express the percentage of such data components [45]. It is depicted as Eq. (6).

$$RK = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} Rn_j^i}{nm} \tag{6}$$

Here Rn_j^i indicates if an element maintains its rank during the data perturbation process. Rn_j^i is characterized as the following equation:

$$Rn_j^i = \begin{cases} 1, & \text{if } R_j^i = R_j^{i'} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

4.1.5. CP

The rank of the average value of each attribute is predicted to vary following the perturbation of data items. CP [45] can indicate such variations in rank of the average value of attributes as Eq. (8).

$$CP = \frac{\sum_{i=1}^{n}|(R_{AV_i} - R'_{AV_i})|}{n} \tag{8}$$

where R_{AV_i} denotes the rank of the mean value of the ith attribute in ascending sequence. On contrary, R'_{AV_i} is rank of the mean value of the ith perturbed attribute in ascending sequence.

4.1.6. CK

After data perturbation, CK can reflect the percentage of qualities that can keep their respective ranks of the average value [45]. CK can be depicted as Eq. (9).

$$CK = \frac{\sum_{i=1}^{n} Cn^i}{n} \tag{9}$$

here Cn^i is characterized as the following equation:

$$Cn^i = \begin{cases} 1, & \text{if } R_{AV_i} = R'_{AV_i} \\ 0, & \text{otherwise} \end{cases} \tag{10}$$

4.2. Utility of data

Utility generally means the amount of important data about a data stream which is necessary for data mining activities [5]. We employed two utility metrics in this study, which are described next.

4.2.1. Accuracy

It represents how many instances are correctly predicted with respect to total observations. The percentage of appropriate results made by data mining algorithms is measured by accuracy [5,46]. Hence, the accuracy is defined as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

where TP is true positive, TN is true negative, FP is false positive and FN means false negative. The confusion matrix specifies all of these variables [47].

4.2.2. F1-score

The F1-score combines the precision and recall into a single metric by taking their harmonic mean. A balance between recall and precision can be established by F1-Score [5,46]. If precision P and recall R, then F1-Score can be defined as Eq. (12).

$$F1 - Score = 2 * \frac{P * R}{P + R} \tag{12}$$

where $P = \frac{TP}{TP+FP}$ and $R = \frac{TP}{TP+FN}$ are the precision and recall, respectively.

4.2.3. Precision

Precision is the probability that a data is relevant once the algorithm has returned it [48]. A model's total number of positive predictions divided by the proportion of true positives is known as precision [49]. Precision can be written in regards to true positives and false positives (FP) as follows [49]:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

4.2.4. Recall

Recall can be defined as the likelihood of retrieving a relevant data [48]. Recall is the ratio of true positives to total number of actual positives [49]. Recall can be written as follows [49]:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

5. Proposed approach

The proposed NOS2R approach consists of normalization, scaling, shearing and reflection. Then the NOS2R2 approach is the combination of NOS2R and 3D rotation transformation. The NOS2R algorithm is shown in Algorithm 1.

The proposed approach starts by using the data matrix $D_{m\times n}$ as input, where n represents the total number of attributes and m represents the total number of instances. At first, the data set is normalized to make it standardized. In NOS2R, The data set $D_{m\times n}$ is normalized using z-score normalization because it performs better than alternative normalization methods for the full feature data set [50]. Table 1 data set is normalized and the result is shown in Table 4. When rescaling an attribute, z-score normalization helps to re-scale it. For rescaling, the attribute's mean and standard deviation is used [5]. As a result, it generates a characteristic with a unit variance, zero mean, and good handling of outliers. It can be defined by the following Eq. (15) [50].

$$x_i' = \frac{x_i - \mu_i}{\sigma_i} \tag{15}$$

where the normalized form of the ith attribute is represented by x_i', x_i is original ith attribute. μ_i means the mean and σ_i represents the standard deviation of ith attribute.

Algorithm 1: NOS2R Perturbation

Input: $D_{m\times n}$
Output: $D_{8_{m\times n}}$

1 $D_{1_{m\times n}} \leftarrow zscore(D_{m\times n})$ ▷ Applying normalization;
2 $D_{2_{m\times n}} \leftarrow S \times D_{1_{m\times n}}$ ▷ Applying scaling transformation;
3 $D_{3_{m\times n}} \leftarrow Sh_x \times D_{2_{m\times n}}$ ▷ Applying shearing transformation along X axis;
4 $D_{4_{m\times n}} \leftarrow Sh_y \times D_{3_{m\times n}}$ ▷ Applying shearing transformation along Y axis;
5 $D_{5_{m\times n}} \leftarrow Sh_z \times D_{4_{m\times n}}$ ▷ Applying shearing transformation along Z axis;
6 $D_{6_{m\times n}} \leftarrow Rf_{xy} \times D_{5_{m\times n}}$ ▷ Applying reflection transformation along XY plane;
7 $D_{7_{m\times n}} \leftarrow Rf_{yz} \times D_{6_{m\times n}}$ ▷ Applying reflection transformation along YZ plane;
8 $D_{8_{m\times n}} \leftarrow Rf_{xz} \times D_{7_{m\times n}}$ ▷ Applying reflection transformation along XZ plane;

Table 4
Normalized data set.

Customer ID	Account number	Amount of transaction	Current balance
0.4353	−0.8968	−0.9159	−1.1301
0.7086	−0.1815	−0.1510	0.3597
−1.1439	1.0783	1.0669	0.7704

Table 5
Data set after scaling.

Customer ID	Account number	Amount of transaction	Current balance
0.4353	−0.8968	−0.9159	−1.1301
1.4172	−0.363	−0.302	0.7194
−3.4317	3.2349	3.2007	2.3112

Then scaling transformation is used to alter the data matrix $D_{1_{m\times n}}$ after normalization. Scaling is one of the fundamental transformations. The process of transformation entails changing and moving the current points. Scaling is applied to alter or modify an object's size. Scaling factors are used to make the shift [12]. The item size will either be increased or decreased depending on the scaling factor [51]. The data size increases if the scaling factor is greater than 1. It is known as magnification [12,51]. The data size is decreased when the scaling factor is less than 1 which is called reduction [12,51]. Thus the data distortion is materialized by the magnification or reduction of the data value. The general form of scaling transformation can be written as Eq. (16) [12]:

$$P' = S.(P) \tag{16}$$

For ease of understanding the scaling matrix [12] is given below which we have used for data distortion.

$$S = \begin{bmatrix} S_x & 0 & 0 \\ 0 & S_y & 0 \\ 0 & 0 & S_z \end{bmatrix}$$

Setting S_x=1, S_y=2 and S_z=3, the normalized data matrix $D_{1_{m\times n}}$ given in Table 4 is scaled using the scaling matrix S. The transformed data matrix $D_{2_{m\times n}}$ looks like Table 5.

After scaling, the data stream $D_{2_{m\times n}}$ is deformed using shearing transformation. Shearing changes the shape of the object. Shearing can be done in the three directions as follows [52]:

- Shearing in X-direction
- Shearing in Y-direction
- Shearing in Z-direction

Table 6
Shearing matrices.

Axis of shearing	Shearing matrices
X	$Sh_x = \begin{pmatrix} 1 & S_y & S_z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
Y	$Sh_y = \begin{pmatrix} 1 & 0 & 0 \\ S_x & 1 & S_z \\ 0 & 0 & 1 \end{pmatrix}$
Z	$Sh_z = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ S_x & S_y & 1 \end{pmatrix}$

Table 7
Data set after shearing along X axis.

Customer ID	Account number	Amount of transaction	Current balance
−6.3168	7.9004	7.9312	7.602
1.4172	−0.363	−0.302	0.7194
−3.4317	3.2349	3.2007	2.3112

In the case of shearing in the X-direction, the coordinate of X remains unchanged and the coordinate of Y and Z is changed. For shearing in the Y-direction, the coordinate of Y remains unchanged while the coordinate of X and Z is changed. Similarly, when shearing is done along the Z direction, the coordinate of Z remains unchanged while the coordinate of X and Y are changed. For the explanation let us see the shearing matrices as shown in Table 6 [52].

The data distortion takes place through shearing along X axis by the following equations [52] when P (x, y, z) is the initial data point and $P'(x', y', z')$ is the altered data point.

$$x' = x \tag{17}$$

$$y' = y + S_y * x \tag{18}$$

$$z' = z + S_x * x \tag{19}$$

Similarly, shearing along Y axis performs the data alteration as follows [52]:

$$x' = x + S_x * y \tag{20}$$

$$y' = y \tag{21}$$

$$z' = z + S_z * y \tag{22}$$

Again, data contortion is transpired through shearing along Z axis by the given equations [52]:

$$x' = x + S_x * z \tag{23}$$

$$y' = y + S_y * z \tag{24}$$

$$z' = z \tag{25}$$

The shearing transformation is done through step 3 to step 5 in the NOS2R algorithm. At first, $D_{2_{m\times n}}$ is transformed using Sh_x and $D_{3_{m\times n}}$ is generated. Then $D_{3_{m\times n}}$ is transformed using Sh_y and $D_{4_{m\times n}}$ is generated. After that, $D_{4_{m\times n}}$ is transformed using Sh_z and $D_{5_{m\times n}}$ is generated. The corresponding data matrices after shearing along X, Y and Z axis are given in Table 7, Table 8 and Table 9 respectively. To do so, let us use S_x=2, S_y=2.5 and S_z=3.

After shearing, reflection is done. The data value that is reflected always takes shape on the opposite side of the mirror. As the distance between a data point P and the mirror is same as the distance between the reflected point P' and the mirror, the reflection transformation can be written as Eq. (26) [12]. Thus reflection changes the coordinate points of the original data.

Table 8
Data set after shearing along Y axis.

Customer ID	Account number	Amount of transaction	Current balance
−6.3168	7.9004	7.9312	7.602
−21.5115	25.1425	25.1625	22.857
−3.4317	3.2349	3.2007	2.3112

Table 9
Data set after shearing along Z axis.

Customer ID	Account number	Amount of transaction	Current balance
−6.3168	7.9004	7.9312	7.602
−21.5115	25.1425	25.1625	22.857
−69.844	81.891	81.969	74.657

Table 10
Reflection matrices.

Reflection plane	Reflection matrices
XY	$Rf_{xy} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$
YZ	$Rf_{yz} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$
XZ	$Rf_{xz} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Table 11
After reflection transformation along three planes.

Customer ID	Account number	Amount of transaction	Current balance
6.3168	−7.9004	−7.9312	−7.602
21.5115	−25.1425	−25.1625	−22.857
69.844	−81.891	−81.969	−74.657

$$P' = Rf.(P) \tag{26}$$

The reflection matrices along planes are given in Table 10. The reflection transformation takes place through step 6 to step 8 of algorithm 1. The reflected data matrix of our example is shown in Table 11.

This resultant data matrix of reflection transformation is the perturbed data set using the NOS2R method. And finally the rotation transformation is applied on the resultant data matrix of Algorithm 1 according to Algorithm 2.

In this study, 3D rotation was accomplished using the double rotation matrices shown in Table 12 [8]. Therefore for rotating the data set using geometric multiplication, three attributes of the relevant data set must be executed simultaneously. As an execution unit, we employed three sequential attributes $(A_i, A_j, A_k) \in D_{8_{m\times n}}$ and termed it a triplet. Triplet production is simple if the number of attributes is like 3t in total, where t = 1, 2, 3, etc. When there are 3t + 1 number of attributes, the third attribute is combined with the first two to form a triplet. For example, if the final attribute is n, the attributes of the triplet are [n−2, n−1, n]. The triplet can be created by combining the attribute n−2 with n−1 and n, if there are a total of 3t+2 attributes. Then the attributes of the triplet are [n−2, n−1, n]. Here, t_1 = [*Customer ID, Account number, Amount of Transaction*] and t_2 = [*Account number, Amount of Transaction, Current Balance*] are the triplets used in our example. Due to the fact that Table 11 has 3t + 1 attributes where t = 1, Account number and Amount of Transaction are repeated to construct t_2 with the Current Balance.

Multiple rotation angles are produced once the initial triplet is created using the sequential first three attributes. By setting θ = 0.1 to 360, a 3D rotation matrix R_θ is produced. In line 6 of Algorithm 2, a rotation is done by the matrix multiplication of (A_i, A_j, A_k) and R_θ

Algorithm 2: NOS2R2 Perturbation

Input: $D_{8_{m\times n}}$
Output: $D_{9_{m\times n}}$

1 $k \leftarrow \lceil n/3 \rceil$ ▷ n is the number of attributes;
2 **for** *each axes pair* $q \in xy, yz, xz$ **do**
3 $P_k \leftarrow k$ triplets (A_i, A_j, A_k) in $D_{8_{m\times n}}$;
4 **for** *different values of* θ **do**
5 **for** *each selected triplet* P_k **do**
6 $(A_i^{'}, A_j^{'}, A_k^{'}) \leftarrow R_\theta \times (A_i, A_j, A_k)$
7 **end**
8 Variance:
$(v_1 = var(A_i - A_i^{'}), v_2 = var(A_j - A_j^{'}), v_3 = var(A_k - A_k^{'}))$;
9 $\theta_k \leftarrow$ SecurityRange
$(v_1 \geq \delta_1, v_2 \geq \delta_2, v_3 \geq \delta_3; \delta_1, \delta_2, \delta_3 > 0)$;
10 $(A_i^{'}, A_j^{'}, A_k^{'}) \leftarrow R_{\theta_k} \times (A_i, A_j, A_k)$
11 **end**
12 $\theta_m \leftarrow$ angle giving maximum variance;
13 **end**
14 $D_{9_{m\times n}} \leftarrow$ data set with maximum variance;

Table 12
3D rotation matrices.

Single rotation		Double rotation
Axis of rotation	Rotation matrices	
x	$R_x = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix}$	$R_{xy} = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ \sin^2\theta & \cos\theta & \sin\theta\cos\theta \\ \sin\theta\cos\theta & -\sin\theta & \cos^2\theta \end{pmatrix}$
y	$R_y = \begin{pmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{pmatrix}$	$R_{yz} = \begin{pmatrix} \cos^2\theta & -\sin\theta\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta & 0 \\ \sin\theta\cos\theta & -\sin^2\theta & \cos\theta \end{pmatrix}$
z	$R_z = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$R_{xz} = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta\cos\theta & \cos^2\theta & \sin\theta \\ -\sin^2\theta & \sin\theta\cos\theta & \cos\theta \end{pmatrix}$

Table 13
Final perturbed data set.

Customer ID	Account number	Amount of transaction	Current balance
−8.266	9.2536	9.2417	8.0817
59.422	−69.86	−69.93	−63.84
63.079	−73.73	−73.79	−67.03

which results in the triplet like (A_i', A_j', A_k'). The variances v_1, v_2, and v_3 are then calculated in line 8 by computing $var(A_i - A_i')$, $var(A_j - A_j')$ and $var(A_k - A_k')$ respectively. Step 9 of Algorithm 2 compares variances v_1, v_2, and v_3 to the security thresholds st $=[\delta_1, \delta_2, \delta_3]$ where $\delta_1, \delta_2, \delta_3 > 0$ and the inequalities are $v_1 > \delta_1$, $v_2 > \delta_2$ and $v_3 > \delta_3$. We applied the security thresholds to ensure that the rotated triplet's attributes fulfill a minimum variance. For other triplets, these security criteria might be different. Every triplet may differ in variance range because of this. Then the variances and the angles are selected for which the inequalities become true. The covariance matrix of the relevant triplet's components was then added to the resulting difference matrix's covariance matrix to calculate the overall variance. Then in line 12 of Algorithm 2, θ_m is calculated that yields the highest total variance. Then this θ_m was used to generate the final perturbed data set.

After applying rotation, the resultant data set is the final perturbed data set of NOS2R2 approach. The perturbed data set of our example is shown in Table 13 where we have used Z axis as rotation axis and $\theta = 38°$ as the angle that provides maximum variance.

The block diagrams representing NOS2R and NOS2R2 are given in Fig. 1 and Fig. 2 respectively.

6. Performance analysis

6.1. Data set

For the purpose of the evaluation, ten UCI data set were used. Table 14 provides a full summary of the data set that were used. Data sets with multiple dependent variables and characteristics were utilized for analysis activities. Only numerical attributes are included in the data set. The majority of the data set contains confidential information. A fair selection procedure was used.

6.2. Classifiers and experimental set-up

Various classification techniques are used by different classifiers [53]. We have used one benchmark classifier named Decision Tree (DT) [54]. A decision tree is a tree-based method where each path leading from the root is characterized by a data separating sequence up until a Boolean result is attained at the leaf node [54]. In these tree structures, the leaves stand in for class labels and the branches signify the combinations of features that result in those class labels [55]

We have used MATLAB R2020a as the implementation platform. The existing methods 3DRT and NRoReM and the proposed methods NOS2R and NOS2R2 are implemented using MATLAB R2020a. We also calculated the privacy metrics using MATLAB. After that,for the classification model, Weka 3.8.5 has been employed to generate data mining results. The selected measures are Accuracy, F1-score, Precision and recall values. The analysis tasks are carried out using 10-fold cross-validation. After calculating these utility and privacy metrics, we plotted those data using Microsoft Excel. The data values were plotted in different types of graphs such as: line graphs, dot plot and bar charts. We used Microsoft Excel to generate all these types of graphs. We used a machine with the following specifications: Intel(R) Core(TM) i5-6200U CPU @ 2.30 GHz 2.40 GHz; RAM: 8.00 GB; system type: 64-bit Operating System, x64-based processor.

6.3. Analysis of privacy

We have studied the privacy-preserving performance of NOS2R and NOS2R2 utilizing the metrics which were described in Section 4.1. Again an increase in information entropy and ICA attack analysis [41] was evaluated to measure its privacy-preserving capabilities. The proposed NOS2R and NOS2R2 approaches are mostly related to NRoReM [5] and 3-Dimensional Rotation Transformation (3DRT) [8]. That is why, in this work, NOS2R and NOS2R2 are compared to 3DRT and NRoReM.

6.3.1. Privacy protection

NOS2R and NOS2R2 both have the potential to protect individuals' privacy. The secrecy level for NOS2R and NOS2R2 are higher than that of both existing approaches which we have used in this study. Table 15 shows the secrecy level comparison.

The mean Friedman test ranks are provided in the final row of Table 15. A lower FMR value denotes a lower level of secrecy, whereas a higher value denotes a higher level of secrecy. According to the experiment's test statistics: chi-square χ^2 value of 28.08 and the probability value p is less than 0.00001. The p-value shows that the variations in secrecy levels between 3DRT, NRoReM, NOS2R and NOS2R2 are significant. We can conclude from the Friedman Mean Rank (FMR) data that NOS2R offers better secrecy than 3DRT and NRoReM. And NOS2R2 provides the greatest secrecy level. As reported by our study, both NOS2R and NOS2R2 offer better secrecy across all the data set compared to NRoReM and 3DRT. Even NOS2R2 provides higher secrecy than our proposed first method NOS2R. So our proposed both methods

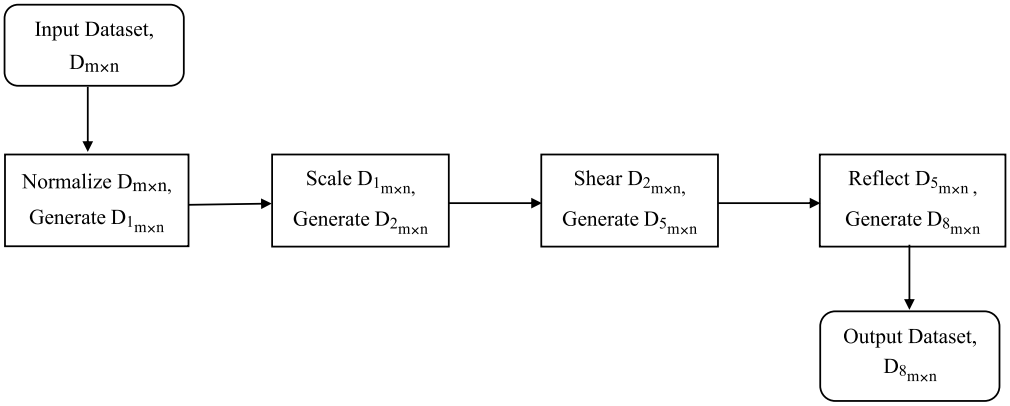


Fig. 1. Proposed four stage NOS2R approach.

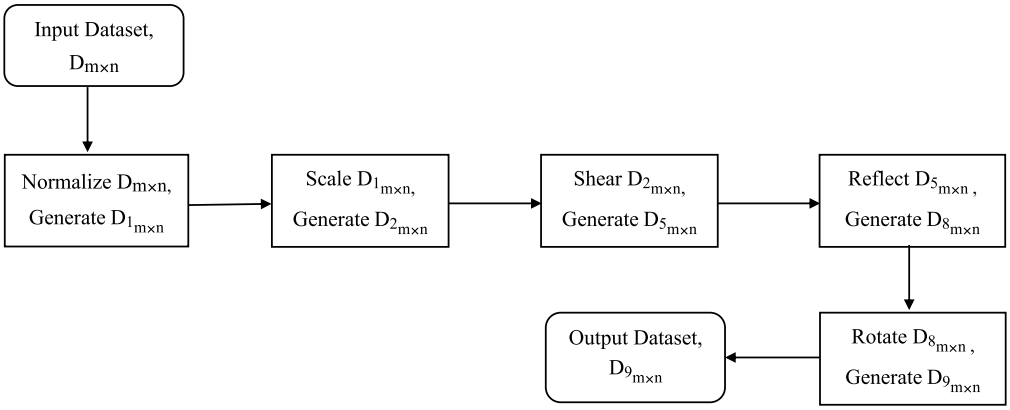


Fig. 2. Proposed five stage NOS2R2 approach.

Table 14
UCI data set used for Analysis.

SI No.	Original data set name	Names used in this paper	Number of attributes	Attribute type	Number of instances
1	Haberman's survival	HBRM	3+1	Integer	306
2	Blood Transfusion Service Center	TRNF	4+1	Real	748
3	Mammographic mass	MAMO	5+1	Integer	961
4	Breast Cancer Wisconsin (Original)	WSCN	10+1	Integer	699
5	Diabetic Retinopathy	DBTC	19+1	Integer, Real	1151
6	Breast Cancer Wisconsin (Diagnostic)	WDBC	31+1	Real	569
7	ionosphere	IONS	34+1	Integer, Real	351
8	SPECTF Heart	SPTF	44+1	Integer	267
9	spambase	SPMB	57+1	Integer, Real	4601
10	Connectionist Bench (Sonar, Mines vs Rocks)	SNAR	60+1	Real	208

work well than the existing 3DRT and NRoReM approach. And among all the methods NOS2R2 method is superior.

VD, RP, and CP should have higher values for improved privacy protection; however, the value of CK and RK should be lower [45].

Table 16 shows the value of VD, RP, RK, CP and CK. Here, it is seen that, at least two of the metrics show that the NOS2R strategy protects more privacy compared to 3DRT and NRoReM for 90% the data set and NOS2R2 works better for all the data set. Similarly, in case of

Table 15
Comparison of secrecy levels.

Data set	3DRT [8]	NRoReM [5]	NOS2R	NOS2R2
HBRM	0.4021	0.4801	0.7479	1.1486
TRNF	0.2277	0.2561	0.4037	0.4409
MAMO	2.4086	38.231	65.6515	171.8649
WSCN	1.3614	23.1025	38.7413	485.382
DBTC	0.1619	2.7813	5.2957	7.3771
WDBC	106.3386	689.458	4611.768	8154.139
IONS	0.1003	3.7854	4.1851	6.6131
SPTF	0.0184	0.0165	0.061	1.0957
SPMB	0.0176	0.0172	0.0178	0.0181
SNAR	643.8542	2178.789	30 666.1	83 809.57
FMR	**1.2**	**1.8**	**3**	**4**

Table 16
Comparison of VD, RP, RK, CP, and CK for privacy protection.

Data set	Methods	VD	RP	RK	CP	CK
HBRM	3DRT [8]	1.0017	116.0958	0.0033	1.33	0.33
	NRoReM [5]	1.0019	122.2506	0.0033	1.35	0.1
	NOS2R	1.0184	133.1618	0.0033	1.38	0
	NOS2R2	1.0238	134.4009	0.0033	1.45	0
TRNF	3DRT [8]	0.998	151.4586	0.2533	1	0.6
	NRoReM [5]	0.999	207.0529	0.0247	1	0.5
	NOS2R	1.0021	347.8389	0.000858	1	0.5
	NOS2R2	1.0036	348.4532	0.00066845	1	0.25
MAMO	3DRT [8]	1.0053	320.013	0.0012	1.4	0.2
	NRoReM [5]	1.0254	327.248	0.003	1.7	0.2
	NOS2R	1.0891	360.37	0.002	2.1	0.02
	NOS2R2	1.103	369.018	0.0002	2.2	0
WSCN	3DRT [8]	1.0916	168.8673	0.0035	3.2	0.2
	NRoReM [5]	16.602	191.214	0.0016	3.2	0.2
	NOS2R	50.7264	308.9672	0.00087847	3.2	0.2
	NOS2R2	54.9648	309.5909	0.00029283	3	0.3
DBTC	3DRT [8]	1.0025	442.4007	0.00686	5.3684	0.2105
	NRoReM [5]	1.0012	464.5204	0.0068	6.2045	0.1896
	NOS2R	1.0061	516.784	0.000777356	6.8421	0.0526
	NOS2R2	1.1323	520.6913	0.000274361	5.8947	0.0526
WDBC	3DRT [8]	1	128.0857	0.0067	9.8064	0
	NRoReM [5]	1	194.427	0.0028	9.7812	0
	NOS2R	1	267.0425	0.0011	9.4838	0
	NOS2R2	1	268.583	0.0009	11.8709	0
IONS	3DRT [8]	1.804	101.7	0.007	12.1176	0.068
	NRoReM [5]	4.8542	119.02	0.003	10.489	0.0314
	NOS2R	10.72	152.85	0.0015	11.82	0.33
	NOS2R2	12.454	152.193	0.00108	14.1705	0.0294
SPTF	3DRT [8]	0.9984	4.8869	0.1173	14	0
	NRoReM [5]	0.9984	6.0892	0.0478	13.7802	0
	NOS2R	1	11.3561	0.0016	13.6818	0.045
	NOS2R2	1.0257	37.1957	0.00074	13.8182	0
SPMB	3DRT [8]	1.006	1221.052	0.000938	20.6316	0
	NRoReM [5]	0.99	1203.478	0.0025	18.4623	0
	NOS2R	0.9983	1176.32	1	19.8596	0
	NOS2R2	1.0162	1484.6067	20.2632	13.8182	0
SNAR	3DRT [8]	2.4315	37.666	0.0147	20.4667	0.0167
	NRoReM [5]	17.2507	66.1094	0.0088	20.0629	0.0172
	NOS2R	23.8662	90.67	0.0025	20.733	0.0153
	NOS2R2	27.8277	100.304	0.0026	16.9	0.0333

three metrics, for 80% the data set, NOS2R protects more privacy than 3DRT and NRoReM; NOS2R2 offers more privacy for all the data set. Again, using four metrics among those five demonstrate that NOS2R and NOS2R2 provide higher level of privacy for 70% the data set and 90% the data set respectively than both existing approaches which we have used in this study. Moreover, NOS2R2 is compared against 3DRT, NRoReM and NOS2R.

It is seen that, at least two metrics demonstrate the superiority of NOS2R2 than others for all of the data set and that of for 90% of the data set for at least three and four metrics.

Table 17
Comparison of ICA attack resistance.

Data set	3DRT [8]	NRoReM [5]	NOS2R	NOS2R2
HBRM	158.2347	193.0289	216.5618	219.9357
TRNF	218.5934	218.5935	219.4241	219.5176
MAMO	279.268	289.265	296.3496	313.3135
WSCN	45.8631	52.6981	69.7644	73.8404
DBTC	180.815	184.2968	189.6886	190.5094
WDBC	156.301	156.301	156.3017	156.3018
IONS	43.0971	78.5236	114.4893	135.328
SPTF	285.6945	293.2979	302.2199	303.227351
SPMB	531.2711	531.8564	531.944	531.0337
SNAR	50.1018	68.1669	79.2819	77.2114
FMR	**1.15**	**2.05**	**3.2**	**3.6**

6.3.2. Attack resistance

There are various sorts of attacks which are addressed against data perturbation methods. One of the attacks is Independent Component Analysis (ICA), which aims to recreate the original data from the altered data [41]. The attack resistance of NOS2R and NOS2R2 is evaluated against an ICA attack because it is based on matrix multiplication. When there is a bigger difference between the altered and actual data, the ICA attack becomes more challenging.

Table 17 shows the values for ICA attack analysis. The values represent the standard deviation of the difference between the ICA-reconstructed data and the initial data. The last row shows the FMR values for the comparison of the existing and proposed approaches. Here the chi-square χ^2 value and the probability value, p are 22.35 and 0.000055 respectively, where p demonstrates the significance of the variations in ICA attack resistance among all the methods. According to our analysis, NOS2R offers 33% higher resistance against ICA attack than 3DRT and 11.5% higher resistance than NRoReM. And NOS2R2 is the most resistant method against ICA attack. NOS2R2 is 39.2% more resistant to ICA attack than 3DRT and 15.5% more resistant against ICA attack than NRoReM. So our proposed both methods work better than the existing 3DRT and NRoReM approach. And among all the methods NOS2R2 method is superior in case of attack resistance.

6.3.3. Increase in information entropy

Information entropy indicates the amount of information contained in an event. The average rate of information production from a random source of data is known as information entropy. Shannon's entropy can be used to calculate the information entropy of a data set. The formula of Shannon's entropy is given in Eq. (27) [3].

$$H(X) = -\sum_{i=1}^{n} P(x_i)\log_2 P(x_i) \tag{27}$$

where H(X), a discrete random variable which denotes the entropy of X, with $X = < x_1, x_2, \ldots, x_n >$. $P(x_i)$ represents the probability of occurrence of x_i and $\sum$ denotes the summation over X.

Using Eq. (28) [3], the average information entropy increase are determined.

$$AIE = \frac{\sum_{i=1}^{n}(H(x_i') - H(x_i'))}{n} \tag{28}$$

Here, $H(x_i)$ is the entropy for every attribute of the original data set. Similarly, $H(x_i')$ stands for every attribute's entropy of the altered data set. AIE represents an average increase in information entropy. When the AIE values are greater than 0, it means that the altered data set have more impurities than the original data set. Fig. 3 shows the AIE for 3DRT, NRoReM, NOS2R, and NOS2R2. The bar chart shows that, the entropy level is greater for NOS2R compared to NRoReM and 3DRT for 90% the data set. Also, NOS2R2 provides higher entropy than 3DRT, NRoReM and NOS2R for 90% of the data set. Our experimental analysis says that, NOS2R provides 6.94% higher entropy than 3DRT and 3.44% higher entropy than NRoReM. Furthermore, NOS2R2 provides 17.72% higher entropy than 3DRT and 15.48% higher entropy than NRoReM.

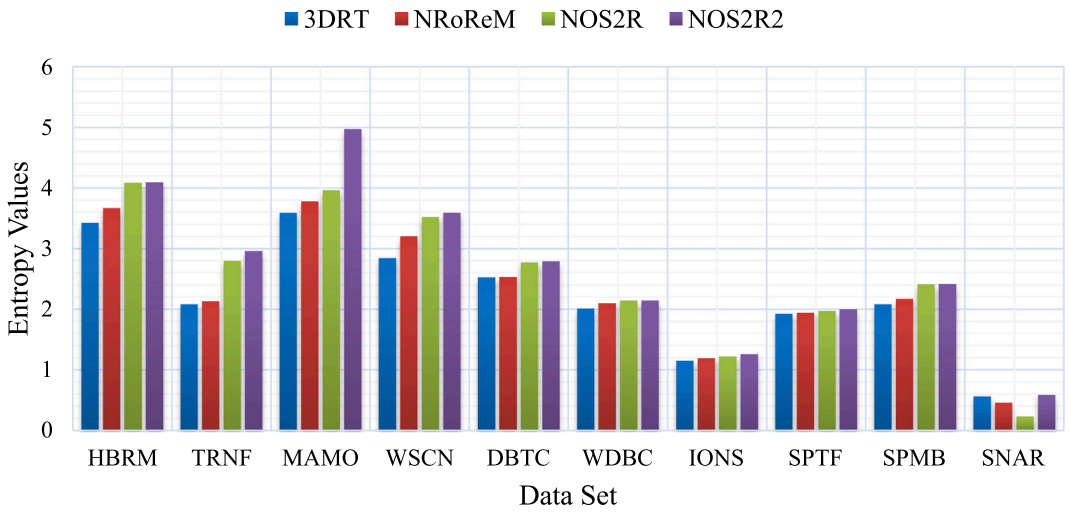


Fig. 3. Average increase in information entropy.

Table 18
Accuracy comparison.

Data set	Original	3DRT [8]	NRoReM [5]	NOS2R	NOS2R2
HBRM	69.23	72.12	67.26	69.68	69.26
TRNF	76.77	76.13	77.22	76.36	76.38
MAMO	83.33	79.78	80.73	82.27	82.76
WSCN	94.36	92.55	95.75	95.68	95.65
DBTC	61.13	58.82	59.89	59.59	59.34
WDBC	90.67	94.3	94.51	93.34	93.59
IONS	88.23	89.92	87.02	90.75	88.63
SPTF	61.36	64.07	63.48	59.52	62.97
SPMB	91.36	92.32	89.9	92.13	91.81
SNAR	73.68	74.28	72.12	71.42	72.85

Table 19
Absolute accuracy differences.

Data set	3DRT [8]	NRoReM [5]	NOS2R	NOS2R2
HBRM	2.89	1.97	0.45	0.03
TRNF	0.64	0.45	0.41	0.39
MAMO	3.55	2.6	1.06	0.57
WSCN	1.81	1.39	1.32	1.29
DBTC	2.31	1.24	1.54	1.79
WDBC	3.63	3.84	2.67	2.92
IONS	1.69	1.21	2.52	0.4
SPTF	2.71	2.12	1.84	1.61
SPMB	0.96	1.46	0.77	0.45
SNAR	0.6	1.56	2.26	0.83

6.4. Analysis of utility

From Table 18, it is seen that the accuracy of NOS2R and NOS2R2 are more similar to that of the original data than NRoReM and 3DRT. For over 70% data set, NOS2R outperforms NRoReM and 3DRT. And for 80% data set, NOS2R2 approach performs on a greater scale.

The absolute differences in classification accuracy shown in Table 19 demonstrate this analysis. Table 19 displays the differences between the accuracy found in the altered data set by 3DRT, NRoReM, NOS2R and NOS2R2 and that of the original data set.

From this accuracy difference, it can be said that the accuracy of NOS2R is 28.62% closer to the original data than 3DRT and 16.81% closer than NRoReM. As well, NOS2R2 is 50.5% closer to the original data than 3DRT and 42.32% closer than NRoReM.

It is also obvious from the analysis of Eq. (29) that the utility of NOS2R and NOS2R2 approach are maintained. The classification accuracy of the original data set, Ro, and the perturbed data set, Rp, are both expressed in Eq. (29). The value of e indicates the lowest level over which the utility of the classifiers can be maintained.

$$E = R_p - (1 - e)R_o \tag{29}$$

For this analysis, e is set to 0.03068. Then it is observed that the classifier is able to maintain the utility in case of NOS2R and NOS2R2; however fail in case of 3DRT and NRoReM for this lowest value of e. That is why e is set to 0.03068. In Fig. 4, this is depicted.

E_{min} values are plotted along the y-axis of the graph in Fig. 4 and are calculated as follows: At first, the minimum error returned for each strategy is identified. This is carried out for every data set. Next, a plot is made using the value of minimum error for every method. These E_{min} values are displayed against e which initializes at 0.1 since Fig. 4 shows that E_{min} is positive for greater values of e.

The utilities of the data set that the classifier, decision tree return, are maintained if Eq. (29) shows that the value of error E $\geq$ 0. As a result, we found that the utility of NOS2R and NOS2R2 can be maintained for $e = 0.03068$ but fail for 3DRT and NRoReM, as shown in Fig. 5. If we further decrease the value of e and set at 0.03067, then the utility for NOS2R2 is preserved but fail for 3DRT, NRoReM and NOS2R as well for this value of e. This is shown in Fig. 6.

The y-axis in Figs. 5 and 6 shows E_m, the minimum error. It has been found that NOS2R and NOS2R2 perturbation successfully preserves the utility of the entire data set for the lowest value of $e = 0.03068$. However, it turns out that 3DRT and NRoReM do not preserve 100% utility for the same value of e.

Besides accuracy, f1-score analysis is helpful to evaluate the effectiveness of classification algorithms on a larger scale. The bar chart in Fig. 7 shows that where analysis of accuracy compromises, the NOS2R approach provides 43.65% better result than 3DRT and 22% closer f1-score to original data than NRoReM. Again the NOS2R2 perturbation method outperforms all of the 3DRT, NRoReM and NOS2R approaches in terms of f1-score. NOS2R2 results in 50.25% closer f1-score to original data than 3DRT and 31.22% closer f1-score compared to NRoReM.

Furthermore, Precision is another important metric for utility evaluation. The bar graph in Fig. 8 demonstrates that, NOS2R algorithm yields precisions that are 48.03% and 21.89% better than 3DRT and NRoReM respectively. Similarly, NOS2R2 approach generates precisions which are 53.93% and 30.77% closer to original one in comparison to 3DRT and NRoReM respectively.

In addition to the other utility metrics, recall analysis is also of utmost significance. Fig. 9 illustrates the recall values for the existing and proposed methods. From Fig. 9, it can be found that the recall values for NOS2R is 21.4% and 16.15% closer to original one than that of 3DRT and NRoReM methods respectively. All of the 3DRT,

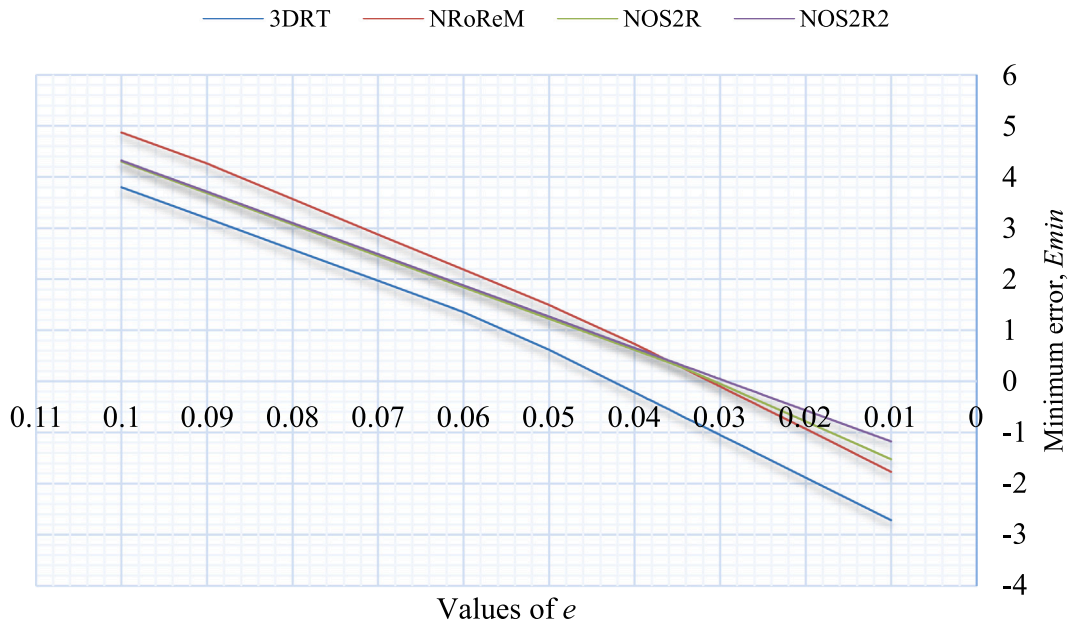


Fig. 4. Selecting the value of *e*.

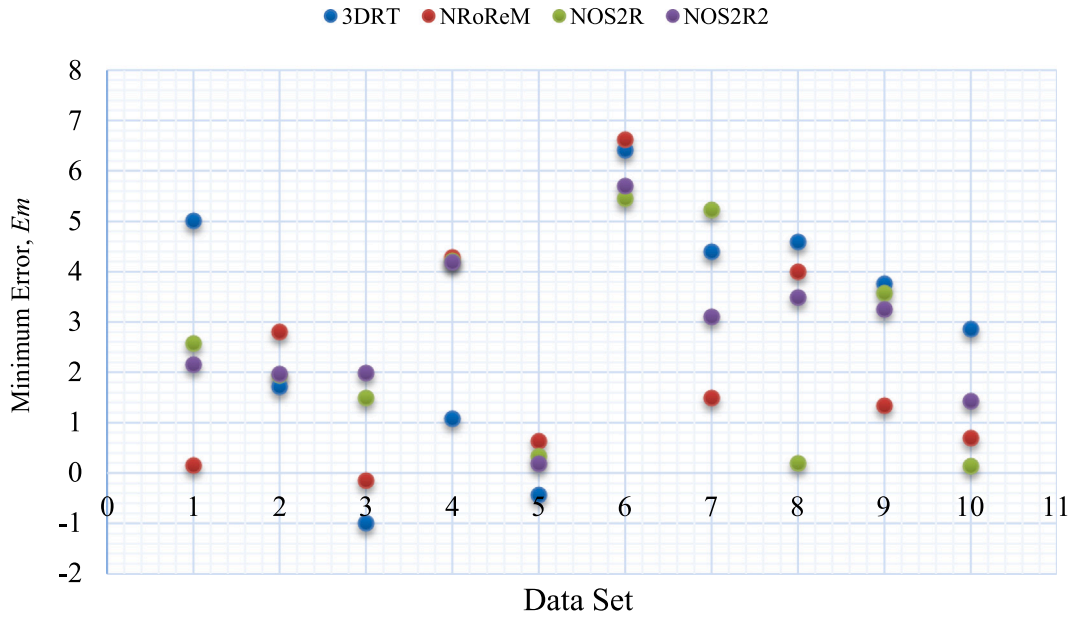


Fig. 5. Classification errors for e = 0.03068.

NRoReM, and NOS2R techniques are outperformed by the NOS2R2 perturbation method while evaluating recall values. NOS2R2 algorithm offers 39.85% and 35.82% better recall values than 3DRT and NRoReM respectively.

Hence, we see that the existing approaches are outperformed for the majority of the data set by NOS2R and NOS2R2. A comparative summary has been provided in the Table 20. Thus our methods provide better balance between privacy and utility. Our proposed methods work much better than the existing methods. But in case of big data there is an uncertainty regarding the amount of time these approaches will take.

6.5. Discussion and shortcomings

For most of the data set, NOS2R and NOS2R2 surpass 3DRT and NRoReM, but they fall short in a few instances. NOS2R is a four stage data perturbation technique and NOS2R2 is five stage data perturbation technique. The improved performance is a result of all phases of NOS2R and NOS2R2. Normalization helps to maintain a close relationship of euclidean distances of the data points. Normalization is followed by scaling transformation. Scaling can be used to modify an object's size. The coordinate values can be multiplied by scaling factors to achieve this. The original coordinate X can be multiplied by a scaling factor to get the modified coordinate X'. Then shearing transformation is done. Shearing also causes a change in object's shape and the data value and a new shape is appeared. Thus shearing play a vital role in preserving privacy as the data value is changed and a new shape is appeared. Again, geometric reflection and rotation are isometric transformations [56]. NOS2R uses scaling, shearing and reflection transformation. and NOS2R2 uses rotation along with these transformations. Rotation and reflection are strict isometric transformation. Scaling is not isometric transformation, but it still can maintain

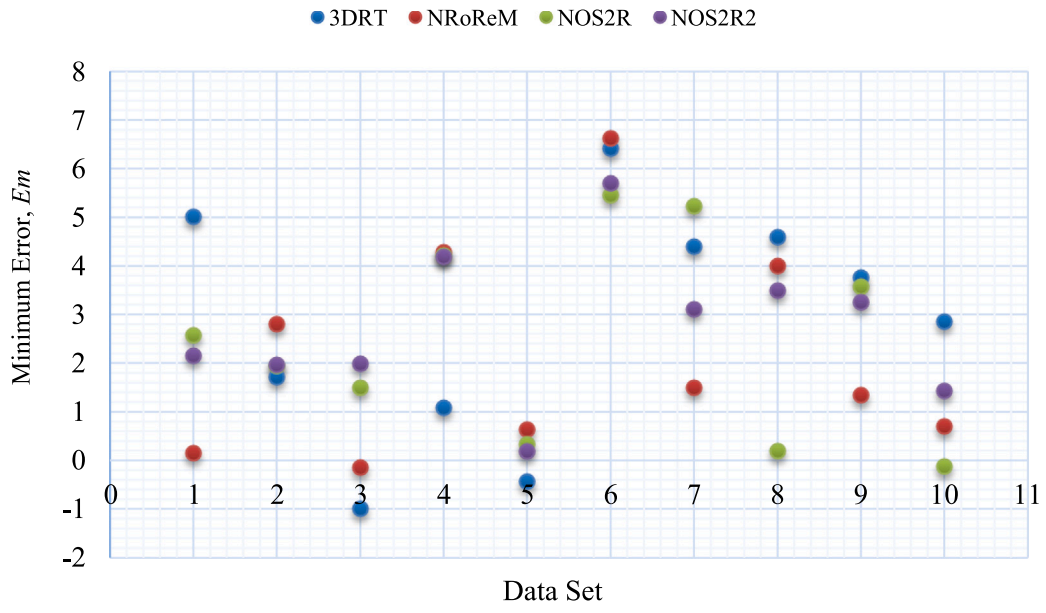


Fig. 6. Classification errors for e = 0.03067.

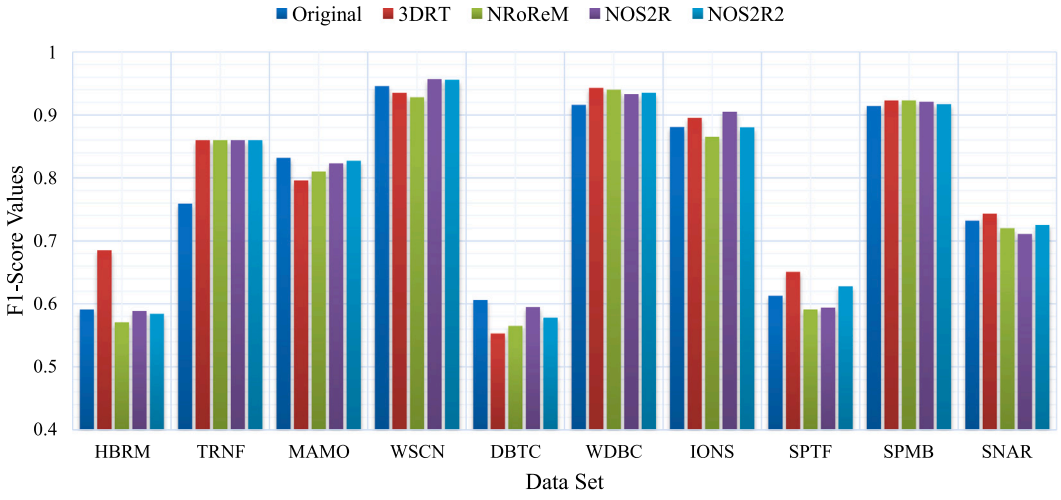


Fig. 7. F1-Score analysis.

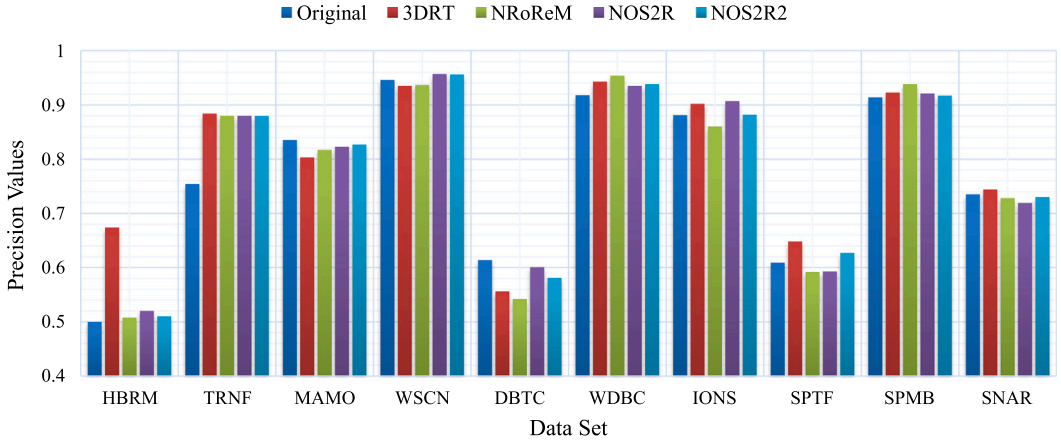


Fig. 8. Precision analysis.

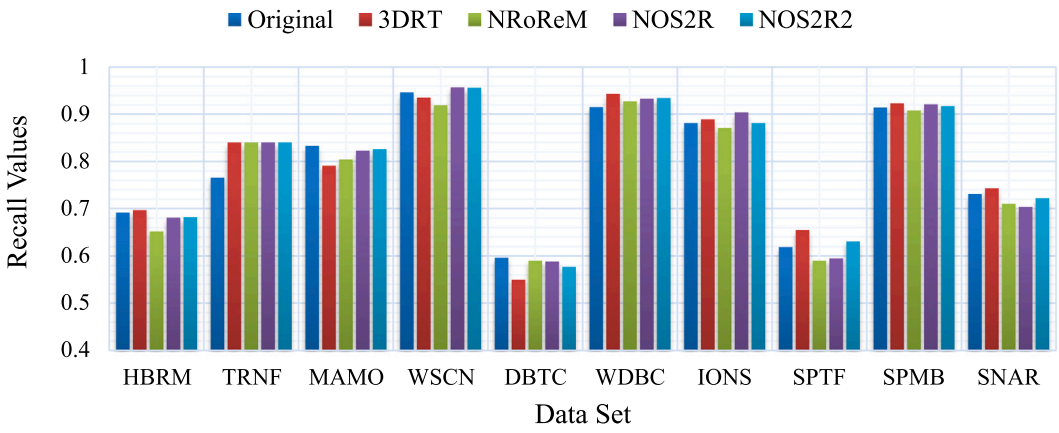


Fig. 9. Recall analysis.

Table 20
Summary of comparative analysis of existing methods and proposed methods.

Metric	3DRT [8]	NRoReM [5]	NOS2R	NOS2R2
Secrecy	FMR = 1.2	FMR = 1.8	FMR = 3	FMR = 4
Entropy	lesser than all other methods	3.6% higher than 3DRT; but lesser than proposed methods	6.94% higher than 3DRT 3.44% higher than NRoReM	17.72% higher than 3DRT 15.48% higher than NRoReM 11.4% higher than NOS2R
VD, RP, RK, CP, CK	higher privacy for 40% of the data set than NRoReM but lesser privacy for all of the other data set	greater privacy for 60% and 30% data set than 3DRT and NOS2R respectively; but lower than NOS2R2	higher for 70% data set than existing methods and only greater for 10% data set than NOS2R2	greater for 100% and 90% data set than the existing methods and NOS2R respectively.
ICA attack	FMR = 1.15	FMR = 2.05	FMR = 3.2	FMR = 3.6
Average absolute Accuracy difference	2.08	1.78	1.48	1.03
Absolute average F1-score difference	0.0394	0.0285	0.0222	0.0196
Absolute average Precision difference	0.0508	0.0338	0.0264	0.0234
Absolute average Recall difference	0.0271	0.0254	0.0213	0.0163

the similarity [56]. Thus scaling, rotation and reflection are used to maintain the utility. But in our work, to increase the level of privacy, a 3D rotation matrix with a mixture of clockwise and counterclockwise directions has been employed (for example, the z-axis is counterclockwise, the other two axes are clockwise). The rotated data points' spatial representation therefore varies in a number of ways which conflicts with maintaining utility. Several new study directions have been made possible by this work. To begin with, other transformations or methods can be applied with this work to compare the utility and privacy preservation. The proposed approaches might be used for other data mining tasks, such as regression, clustering, and so on. Currently, we have used numerical data to validate our methods. Adapting these methods to deal with categorical data can be potential avenue for future research. Again, investigating alternative attacks is another crucial topic for future research. ICA attack analysis has been employed in this work. Further attacks can be put under consideration for a better analysis.

7. Conclusion

The main goal of this research work is to perturb the data for preserving privacy along with maintaining the utility at the same time. Data perturbation techniques are applied to multivariate data set in order to change the original data items. The recent perturbation methods frequently address the trade-off between privacy and data utility. The proposed NOS2R and NOS2R2 approaches perturb all the features of the data set. The data items are standardized using a normalization process, which is followed by scaling, shearing, reflection and rotation. Thus the data elements are distorted. But the data set's geometrical shapes remain constant. This property facilitates in maintaining the original data set's utility and the composite transformation significantly

increases privacy to a significant level. Then the results of the proposed methods are compared with the other existing methods. And the proposed two methods provide better results than the existing methods. NOS2R and NOS2R2 perturbation approaches maintain data utility and preserve individual privacy on a larger scale for 90% data set. However, the work proposed in this paper can be well extended with feature reduction in multiparty clustering scenarios.

CRediT authorship contribution statement

Mahbubur Rahman: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization. **Mahit Kumar Paul:** Supervision, Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Writing – review & editing, Visualization. **A.H.M. Sarowar Sattar:** Supervision, Formal analysis, Validation, Investigation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Afrin A, Paul MK, Sattar AHMS. Privacy preserving data mining using non-negative matrix factorization and singular value decomposition. In: 2019 4th International conference on electrical information and communication technology. 2019, p. 1–6. http://dx.doi.org/10.1109/EICT48899.2019.9068846.

[2] Askinadze A, Conrad S. Respecting data privacy in educational data mining: An approach to the transparent handling of student data and dealing with the resulting missing value problem. In: 2018 IEEE 27th international conference on enabling technologies: Infrastructure for collaborative enterprises. 2018, p. 160–4. http://dx.doi.org/10.1109/WETICE.2018.00037.

[3] Chamikara M, Bertok P, Liu D, Camtepe S, Khalil I. Efficient privacy preservation of big data for accurate data mining. Inform Sci 2020;527:420–43. http://dx.doi.org/10.1016/j.ins.2019.05.053.

[4] Chamikara M, Bertok P, Liu D, Camtepe S, Khalil I. Efficient data perturbation for privacy preserving and accurate data stream mining. Pervasive Mob Comput 2018;48:1–19. http://dx.doi.org/10.1016/j.pmcj.2018.05.003.

[5] Paul MK, Islam MR, Sattar AHMS. An efficient perturbation approach for multivariate data in sensitive and reliable data mining. J Inf Secur Appl 2021;62:102954. http://dx.doi.org/10.1016/j.jisa.2021.102954.

[6] Denham B, Pears R, Naeem MA. Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining. Expert Syst Appl 2020;152:113380. http://dx.doi.org/10.1016/j.eswa.2020.113380.

[7] Salloum SA, Alshurideh M, Elnagar A, Shaalan K. Mining in educational data: Review and future directions. In: Hassanien A-E, Azar AT, Gaber T, Oliva D, Tolba FM, editors. Proceedings of the international conference on artificial intelligence and computer vision. Cham: Springer International Publishing; 2020, p. 92–102. http://dx.doi.org/10.1007/978-3-030-44289-7_9.

[8] Upadhyay S, Sharma C, Sharma P, Bharadwaj P, Seeja K. Privacy preserving data mining with 3-D rotation transformation. J King Saud Univ - Comput Inf Sci 2018;30(4):524–30. http://dx.doi.org/10.1016/j.jksuci.2016.11.009.

[9] Helbing D, Brockmann D, Chadefaux T, Donnay K, Blanke U, Woolley-Meza O, et al. Saving human lives: What complexity science and information systems can contribute. J Stat Phys 2015;158(3):735–81. http://dx.doi.org/10.1007/s10955-014-1024-9.

[10] Jalili M, Perc M. Information cascades in complex networks. J Complex Netw 2017;5(5):665–93. http://dx.doi.org/10.1093/comnet/cnx019.

[11] Capraro V, Perc M. Grand challenges in social physics: in pursuit of moral behavior. Front Phys 2018;6:107. http://dx.doi.org/10.3389/fphy.2018.00107.

[12] Xiang Z, Plastock RA. Schaum's outline of computer graphics 2/E. McGraw Hill Professional; 2000.

[13] Kreso I, Kapo A, Turulja L. Data mining privacy preserving: Research agenda. Wiley Interdiscip Rev Data Min Knowl Discov 2021;11(1):e1392. http://dx.doi.org/10.1002/widm.1392.

[14] Verykios VS, Bertino E, Fovino IN, Provenza LP, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving data mining. ACM Sigmod Rec 2004;33(1):50–7. http://dx.doi.org/10.1145/974121.974131.

[15] Malik MB, Ghazi MA, Ali R. Privacy preserving data mining techniques: Current scenario and future prospects. In: 2012 Third international conference on computer and communication technology. 2012, p. 26–32. http://dx.doi.org/10.1109/ICCCT.2012.15.

[16] Muralidhar K, Parsa R, Sarathy R. A general additive data perturbation method for database security. Manag Sci 1999;45(10):1399–415. http://dx.doi.org/10.1287/mnsc.45.10.1399.

[17] Gerhardt F, Oriti D, Wilson-Ewing E. Separate universe framework in group field theory condensate cosmology. Phys Rev D 2018;98(6):066011. http://dx.doi.org/10.1103/PhysRevD.98.066011.

[18] Sun Y, Guo Y, Tropp JA, Udell M. Tensor random projection for low memory dimension reduction. 2021, http://dx.doi.org/10.48550/arXiv.2105.00105, arXiv preprint arXiv:2105.00105.

[19] Chen K, Liu L. Geometric data perturbation for privacy preserving outsourced data mining. Knowl Inform Syst 2011;29(3):657–95. http://dx.doi.org/10.1007/s10115-010-0362-4.

[20] Liu K, Kargupta H, Ryan J. Random projection-based multiplicative data perturbation for privacy preserving distributed data mining. IEEE Trans Knowl Data Eng 2006;18(1):92–106. http://dx.doi.org/10.1109/TKDE.2006.14.

[21] Sattar AS, Li J, Liu J, Heatherly R, Malin B. A probabilistic approach to mitigate composition attacks on privacy in non-coordinated environments. Knowl-Based Syst 2014;67:361–72. http://dx.doi.org/10.1016/j.knosys.2014.04.019.

[22] Fang W, Wen XZ, Zheng Y, Zhou M. A survey of big data security and privacy preserving. IETE Tech Rev 2017;34(5):544–60. http://dx.doi.org/10.1080/02564602.2016.1215269.

[23] Chamikara M, Bertok P, Khalil I, Liu D, Camtepe S. Privacy preserving distributed machine learning with federated learning. Comput Commun 2021;171:112–25. http://dx.doi.org/10.1016/j.comcom.2021.02.014.

[24] Chang H, Ando H. Privacy-preserving data sharing by integrating perturbed distance matrices. SN Comput Sci 2020;1(3):1–10. http://dx.doi.org/10.1007/s42979-020-00127-w.

[25] Kao Y-H, Lee W-B, Hsu T-Y, Lin C-Y, Tsai H-F, Chen T-S. Data perturbation method based on contrast mapping for reversible privacy-preserving data mining. J Med Biol Eng 2015;35(6):789–94. http://dx.doi.org/10.1007/s40846-015-0088-6.

[26] Shan J, Lin Y, Zhu X. A new range noise perturbation method based on privacy preserving data mining. In: 2020 IEEE international conference on artificial intelligence and information systems. 2020, p. 131–6. http://dx.doi.org/10.1109/ICAIIS49377.2020.9194850.

[27] Li G. A new bayesian-based method for privacy-preserving data mining. In: International conference on intelligent and interactive systems and applications. Springer; 2017, p. 171–7. http://dx.doi.org/10.1007/978-3-319-69096-4_24.

[28] Huang M, Chen Y, Chen B-W, Liu J, Rho S, Ji W. A semi-supervised privacy-preserving clustering algorithm for healthcare. Peer-to-Peer Netw Appl 2016;9(5):864–75. http://dx.doi.org/10.1007/s12083-015-0356-9.

[29] Torra V. Fuzzy microaggregation for the transparency principle. J Appl Log 2017;23:70–80. http://dx.doi.org/10.1016/j.jal.2016.11.007, SI:Logical Methods in Artificial Intelligence.

[30] Lin C-Y. A reversible privacy-preserving clustering technique based on k-means algorithm. Appl Soft Comput 2020;87:105995. http://dx.doi.org/10.1016/j.asoc.2019.105995.

[31] Kiran A, Vasumathi D. Data mining: min–max normalization based data perturbation technique for privacy preservation. In: Proceedings of the third international conference on computational intelligence and informatics. Springer; 2020, p. 723–34. http://dx.doi.org/10.1007/978-981-15-1480-7_66.

[32] Oliveira S, Zaiane O. Data perturbation by rotation for privacy-preserving clustering. Technical Report TR04-17, Edmonton, AB, Canada: Department of Computer Science, University of Alberta; 2004, http://dx.doi.org/10.7939/R3344P.

[33] Chamikara M, Bertok P, Liu D, Camtepe S, Khalil I. An efficient and scalable privacy preserving algorithm for big data and data streams. Comput Secur 2019;87:101570. http://dx.doi.org/10.1016/j.cose.2019.101570.

[34] Shynu PG, Md. Shayan. H, Chowdhary CL. A fuzzy based data perturbation technique for privacy preserved data mining. In: 2020 International conference on emerging trends in information technology and engineering. 2020, p. 1–4. http://dx.doi.org/10.1109/ic-ETITE47903.2020.244.

[35] Lyu L, Bezdek JC, Law YW, He X, Palaniswami M. Privacy-preserving collaborative fuzzy clustering. Data Knowl Eng 2018;116:21–41. http://dx.doi.org/10.1016/j.datak.2018.05.002.

[36] Hasan AT, Jiang Q, Luo J, Li C, Chen L. An effective value swapping method for privacy preserving data publishing. Secur Commun Netw 2016;9(16):3219–28. http://dx.doi.org/10.1002/sec.1527.

[37] Siang DKK, Othman SH, Radzi RZRM. Comparative study on perturbation techniques in privacy preserving data mining on two numeric data set. Int J Innov Comput 2018;8(1). http://dx.doi.org/10.11113/ijic.v8n1.161.

[38] Prakash M, Singaravel G. An approach for prevention of privacy breach and information leakage in sensitive data mining. Comput Electr Eng 2015;45:134–40. http://dx.doi.org/10.1016/j.compeleceng.2015.01.016.

[39] Abitha N, Sarada G, Manikandan G, Sairam N. A cryptographic approach for achieving privacy in data mining. In: 2015 International conference on circuits, power and computing technologies. 2015, p. 1–5. http://dx.doi.org/10.1109/ICCPCT.2015.7159300.

[40] Zhang N, Zhao W. Privacy-preserving data mining systems. Computer 2007;40(4):52–8. http://dx.doi.org/10.1109/MC.2007.142.

[41] Okkalioglu BD, Okkalioglu M, Koc M, Polat H. A survey: Deriving private information from perturbed data. Artif Intell Rev 2015;44(4):547–69. http://dx.doi.org/10.1007/s10462-015-9439-5.

[42] Li X, Yan Z, Zhang P. A review on privacy-preserving data mining. In: 2014 IEEE International conference on computer and information technology. 2014, p. 769–74. http://dx.doi.org/10.1109/CIT.2014.135.

[43] Chen K, Liu L. A random rotation perturbation approach to privacy preserving data classification. Technical Report, Wright State University; 2005, 10.1.1.72.6442.

[44] Oliveira SRM, Zaiane OR. Privacy preserving clustering by data transformation. J Inform Data Manag 2010;1(1):37. http://dx.doi.org/10.5753/jidm.2010.940.

[45] Xu S, Zhang J, Han D, Wang J. Singular value decomposition based data distortion strategy for privacy protection. Knowl Inf Syst 2006;10(3):383–97. http://dx.doi.org/10.1007/s10115-006-0001-2.

[46] Tasnim N, Paul MK, Sattar AHMS. Identification of drop out students using educational data mining. In: 2019 International conference on electrical, computer and communication engineering. 2019, p. 1–5. http://dx.doi.org/10.1109/ECACE.2019.8679385.

[47] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern C (Appl Rev) 2012;42(4):463–84. http://dx.doi.org/10.1109/TSMCC.2011.2161285.

[48] Goutte C, Gaussier E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In: Losada DE, Fernández-Luna JM, editors. Advances in information retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg; 2005, p. 345–59. http://dx.doi.org/10.1007/978-3-540-31865-1_25.

[49] Ting KM. Precision and recall. In: Sammut C, Webb GI, editors. Encyclopedia of machine learning. Boston, MA: Springer US; 2010, p. 781. http://dx.doi.org/10.1007/978-0-387-30164-8_652.

[50] Singh D, Singh B. Investigating the impact of data normalization on classification performance. Appl Soft Comput 2020;97:105524. http://dx.doi.org/10.1016/j.asoc.2019.105524.

[51] Singhal A. Computer graphics. 2019, Gate Vidyalay. https://www.gatevidyalay.com/computer-graphics/.

[52] Computer Graphics - 3D shearing transformation. 2021, GeeksforGeeks. URL: https://www.geeksforgeeks.org/computer-graphics-3d-shearing-transformation/. Online. [Accessed 17 November 2022].

[53] Lessmann S, Baesens B, Seow H-V, Thomas LC. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European J Oper Res 2015;247(1):124–36. http://dx.doi.org/10.1016/j.ejor.2015.05.030.

[54] Charbuty B, Abdulazeez A. Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends 2021;2(01):20–8. http://dx.doi.org/10.38094/jastt20165.

[55] Yang Y, Farid SS, Thornhill NF. Prediction of biopharmaceutical facility fit issues using decision tree analysis. In: Kraslawski A, Turunen I, editors. 23rd European symposium on computer aided process engineering. Computer aided chemical engineering, vol. 32, Elsevier; 2013, p. 61–6. http://dx.doi.org/10.1016/B978-0-444-63234-0.50011-7, URL: https://www.sciencedirect.com/science/article/pii/B9780444632340500117.

[56] Liu J, Xu Y. Privacy preserving clustering by random response method of geometric transformation. In: 2009 Fourth international conference on internet computing for Science and engineering. 2009, p. 181–8. http://dx.doi.org/10.1109/ICICSE.2009.31.