A Non-Reversible Privacy Preservation Model for Outsourced High-Dimensional Healthcare Data

Syed Usama Khalid Bukhari [1], Anum Qureshi [2], Adeel Anjum [2], and Munam Ali Shah [2]

[1]University of Derby
[2]Affiliation not available

October 30, 2023

Abstract

Privacy preservation of high-dimensional healthcare data is an emerging problem. Privacy breaches are becoming more common than before and affecting thousands of people. Every individual has sensitive and personal information which needs protection and security. Uploading and storing data directly to the cloud without taking any precautions can lead to serious privacy breaches. It's a serious struggle to publish a large amount of sensitive data while minimizing privacy concerns. This leads us to make crucial decisions for the privacy of outsourced high-dimensional healthcare data. Many types of privacy preservation techniques have been presented to secure high-dimensional data while keeping its utility and privacy at the same time but every technique has its pros and cons. In this paper, a novel privacy preservation NRPP model for high-dimensional data is proposed. The model uses a privacy-preserving generative technique for releasing sensitive data, which is deferentially private. The contribution of this paper is twofold. First, a state-of-the-art anonymization model for high-dimensional healthcare data is proposed using a generative technique. Second, achieved privacy is evaluated using the concept of differential privacy. The experiment shows that the proposed model performs better in terms of utility.

A Non-Reversible Privacy Preservation Model for Outsourced High-Dimensional Healthcare Data

Anum Qureshi, Syed Usama Khalid Bukhari, Adeel Anjum, Munam Ali Shah

Abstract

Privacy preservation of high-dimensional healthcare data is an emerging problem. Privacy breaches are becoming more common than before and affecting thousands of people. Every individual has sensitive and personal information which needs protection and security. Uploading and storing data directly to the cloud without taking any precautions can lead to serious privacy breaches. It's a serious struggle to publish a large amount of sensitive data while minimizing privacy concerns. This leads us to make crucial decisions for the privacy of outsourced high-dimensional healthcare data. Many types of privacy preservation techniques have been presented to secure high-dimensional data while keeping its utility and privacy at the same time but every technique has its pros and cons. In this paper, a novel privacy preservation NRPP model for high-dimensional data is proposed. The model uses a privacy-preserving generative technique for releasing sensitive data, which is deferentially private. The contribution of this paper is twofold. First, a state-of-the-art anonymization model for high-dimensional healthcare data is proposed using a generative technique. Second, achieved privacy is evaluated using the concept of differential privacy. The experiment shows that the proposed model performs better in terms of utility.

Index Terms

Outsourced data; Cloud computing; GAN, NRPP, Classification, Differential Privacy

I. Introduction

THIS is a digital era of immense competition among advancing technologies. There is a hidden and non-hidden race of emerging computing strategies to save and steal the information of any entity or organization of the world. These days privacy preservation of high-dimensional healthcare data has been studied extensively due to the increase in the number of sensitive information. Everybody wants integrity, confidentiality, and availability at the same time which is the basic need in this computational world. Sharing of health care data is very common these days for the sake of analysis and validation [9]. A large amount of health care data is being processed by many pharmaceutical researchers for research purposes. Health care data has many dimensions which increase the risks of leakage of sensitive information. Because healthcare data in its raw form consists of many sensitive attributes. Many health care companies and organizations mask their data via different privacy preservation techniques for the sake of piracy. The hardest thing to do these days is to prevent your sensitive information from various attacks by intruders. In every new second, a new powerful attack is being launched against existing privacy preservation techniques and strategies.

Outsourcing of high-dimensional healthcare data on the cloud is becoming more common in this era. Cloud computing, has now been deployed everywhere and becoming a vital necessity of this generation because of its salient features such as availability. But on the other hand, it is full of risks concerning information breaches. Many organizations, medical or non-medical institutions, government official departments, and enterprises are now widely being using outsourcing and cloud computing to share, store, and exchange their data. Sharing and exchanging of high-dimensional data healthcare is more risky than past. Information that is digitized needs protection. Data of individuals is now being collected and processed by various organizations with the assurance of data privacy. But there is always a commutation between utility and privacy. Existing strategies are not fruitful for the protection of outsourced high-dimensional data. Because the high-dimensionality of data results in a huge degradation of utility of anonymized data [43]. Increase in dimensions of data means an increase in the number of attributes in data. Which directly links with the availability of maximum information of an individual to adversarial attacks. Because the adversary may have some prior or background knowledge that he can associate with his target can reveal the identity of the target. So, while data publishing protection of an individual's identity is the main concern. To protect the identity of an individual quasi-identifiers and sensitive information should be anonymized in a way that record linking attacks or attribute disclosure attacks can not be performed.

The objective of this paper is to provide an ingenious and inventive outsourced storage scheme that will not only guarantee the user's confidentiality, integrity and availability of data but also a conventional operational mechanism. This paper presents an efficient privacy preservation model that deals with the curse of high-dimensionality of healthcare data and provides defense against possible adversarial attacks. Because the adversary may have some prior or background knowledge that he can associate with his target can reveal the identity of the target. So, while data publishing protection of an individual's identity is the main concern. To protect the identity of an individual quasi-identifiers and sensitive information should be anonymized in a way that record linking attack or attribute disclosure attack can not be performed.

So, the objective of this research is to provide an ingenious and inventive outsourced storage scheme that will not only guarantee the user's confidentiality integrity and availability of data in a meantime but also a conventional operational

mechanism. To design an efficient privacy preservation model that not only will deal with the curse of high-dimensionality of data also will provide defense against possible adversarial attacks.

A. *Motivational Example*

Privacy preservation of high-dimensional data is the main curse of this era. Increase in the number of attributes, the chance of attack also increases. Due to the availability of background knowledge, the adversary can disclose the identity of the target.

Table I: Data Set

ID	Age	Sex	Race	Marital status	Edu-cation	Work hrs	Job	Fam	Class
1	23	F	B	Single	Grad	46	Doctor	4	0
2	24	M	W	Married	Uni	24	Clerk	6	0
3	30	M	W	Single	Grad	26	Teacher	7	1
4	34	F	B	Married	Uni	48	Engineer	5	1
5	35	M	B	Single	Uni	46	Writer	4	0
6	36	M	W	Single	Grad	51	Doctor	6	1
7	40	F	W	Married	Uni	47	Doctor	3	1
8	41	F	B	Married	Grad	25	Clerk	6	0
9	42	M	W	Single	Uni	45	Engineer	4	1

In [43] the authors presented an anonymization technique to preserve the privacy of high dimensional data using vertical fragmentation. In vertical fragmentation, attributes are split into multiple non-overlapping fragments. However, their technique is prone to attribute disclosure attacks. As attributes from multiple fragments can be combined to perform a record linkage attack. We are considering the Tables II and III that are the vertical fragments of Table I. These tables are anonymized and vertical partitioning of attributes is done using vertical fragmentation [43]. Before launching any type of attack the adversary gathers some background knowledge, for the sake of collecting sensitive information about the victim [34]. So, in Tables II and III[43] adversary can easily reveal the identity of the victim with the help of some basic knowledge of attributes that is related to his target and can get the record of that individual. Suppose the adversary gets information about an individual who is a female and she has done her bachelor's and currently working as a writer. The adversary might know her that she is above age 40. By using this available background knowledge adversary can easily identify the target from the fragmented tables (II, III) [43].

Table II: Anonymized Fragment 1

Seq.no	ID	Age	Fam	Race	Work hrs	Class
1	1	[26,40]	[2,5]	W	[46,48]	Y
2	6	[26,40]	[2,5]	W	[46,48]	Y
3	8	[26,40]	[2,5]	W	[46,48]	Y
4	2	[26,43]	[5,6]	A	[55,60]	N
5	7	[26,43]	[5,6]	A	[55,60]	N

Table III: Anonymized Fragment 2

Seq.no	ID	Sex	Education	Marital status	Job	Class
1	1	M	Grad school	Married	Engineer	Y
2	6	M	Grad school	Married	Engineer	Y
3	8	M	Grad school	Married	Engineer	Y
4	2	F	University	Single	Writer	N
5	7	F	University	Single	Writer	N

From records 4 and 5, in the above tables (II, III) [43] the adversary can identify the target record with the probability of $1/2$. If he can successfully identify one of the record's identities, he can also disclose the identity of the other record. Therefore, an efficient anonymization technique is required to protect the privacy of high-dimensional data.

B. Challenges

As dimensionality is the main curse of this era in data mining, classification and privacy which unfortunately have not been yet handled properly [5] [29]. Their are several challenges related to high dimensional data, some of them are as follows:

- As high-dimensional data has many features which means several attributes that can be linked to disclose the individual's identity. As with the advancement of new technologies security and privacy threats are associated to outsourced high dimensional data also emerging [13]. Many types of attacks are being launched by intruders using prior or background knowledge on the published datasets to gather personal and sensitive information of individuals. No matter how powerful a security strategy is applied for the preservation of sensitive content, the adversary can no longer be prevented to get intrude in it. As, if one's content gets hacked by the adversary and he lost access to it and the adversary grabbed the sensitive information, then he would get through penalties to get his sensitive information back. As we all know there is nothing perfect in this digital computational world.
- Datasets for high-dimensional data are often unregulated, which might make them more difficult to use. Furthermore, large datasets can contain noise and uncertainties. It can be difficult to process and implement appropriate data mining techniques to such noisy data. There is no analytical methodology that can provide insight into such situations, even for a small fraction of them. As a result, algorithms often become problem and data-specific. As a result, there is no generalized approach.
- With the increase in dimensions, the number of possible cluster combinations multiplies exponentially, making clustering non-deterministic polynomial-time hard (NP-hard), and there are no effective strategies for dealing with these difficult issues [46].
- While the performance of current computers continues to increase, and cheaper parallel and cloud computing facilities become accessible, but the challenges of high-dimensional data processing is persistent. Innovative technologies are still in desperate need of development [37]. To solve problems with high-dimensional data, a paradigm shift and a nontraditional style of thinking may be required.

*C. **Contributions***

Our research model is unique in a way that it can also handle security and utility of high dimensional data that is itself a challenge of this digital era [49]. It will also prevent many types of attacks like attribute or record linkage attack.

- We proposed the NRPP model which extends the basic concept of generative adversarial network (GAN). Our proposed model preserves the sensitive tabular data by converting it into a non-reversible and de-identified image.
- Our proposed approach ensures the prevention of any kind of unauthorized access that aims to do disclosure, disruption, modification, inspection and destruction of sensitive information or wants to use the information for illegal practices.
- For the privacy protection analysis of our proposed model, we have used the concept of differential privacy [8] [31].
- We have evaluated the performance of our NRPP model on two real-world datasets.Our simulation results indicate that our NRPP model is effective in maintaining the semantic meaning and privacy of textual content.

The rest of the paper is organized as follows.In Section 2 we reviewed the related works and done the comparative analysis of existing techniques. Preliminaries are described in Section 3. The proposed model is presented in Section 4 and in section 5 we have done the protection analysis. In section 6 experiments are conducted and results are scrutinized. Section 7 concludes the paper.

II. RELATED WORK

Protection of high-dimensional healthcare data against antagonists and nemesis is a prolonged battle that is never going to end in this world of emerging technologies and developments. Many data privacy protection techniques and terminologies have been launched to protect the confidentiality and integrity of individuals' sensitive data; we have briefly reviewed some of them in comparison to our proposed approach. Warner [44] was the first to propose the Randomized Response. This strategy was established in the statistics community to obtain confidential information from persons in such a way that data processors are unaware of which of two alternative questions the respondent has replied.

This further used by Agrawal and Srikant [4] and Kargupta [20] as they introduced a random perturbation scheme for privacy-preserving data mining and explained how the rebuilt distributions may be employed for data mining. A random variable is added to the value of a sensitive attribute in their randomization system. If x is the value of a sensitive attribute, for example, the database will store x+n rather than x, where n is a random noise picked from a specified distribution. It is demonstrated that given a random noise distribution, the original data distribution may be reconstructed. The randomization technique is not much efficient in terms of concealing sensitive information as the reconstruction of original data is still possible as well as it reduces the utility of data.

In [39] a k-anonymity privacy protection model has been proposed in which each record of an individual in an equivalence class is indistinguishable from other at least k-1 records. If the adversary has some background knowledge about its target he can reveal the identity of the target by performing linking attacks. This model is also prone to homogeneity attacks. [21] They developed an anonymization strategy for obscuring sensitive information about individuals from the records of their owners. K-anonymity is employed for data concealment and generalization, however it does not preserve an individual's sensitivity due to background information and homogeneity attacks of k-Anonymity. The k-anonymity approaches primarily focus on a universal approach that maintains the same level of anonymity for all individuals without taking into account their specific needs. As a result, some people may receive insufficient protection, while others may be subjected to excessive privacy controls.

Machanavajjhala et al [28] observed that when sensitive data have insufficient diversity, the user can identify sensitive values with high degree of confidence, and presented the l-diversity method, which is more feasible than K-anonymity but still prone to many adversarial attacks.

In [23] they developed a privacy notion called t-closeness to distribute the sensitive attributes in equivalence class to prevent attribute disclosure. Many existing works are presented on the base approach of k-anonymity and t-closeness to anonymize quasi-identifiers and protect sensitive properties in a static database. Syntactic anonymization, on the other hand, is difficult to implement to high-dimensional continuous data due to the difficulty in defining quasi-identifiers and sensitive attributes [40]. In this paper [43] they proposed a secured scheme to avoid any record linkage attack while publishing the high dimensional data using vertical fragmentation. But their technique is prone to attribute disclosure attacks.

[30] They present a new privacy model named as LKC-privacy model to ensure privacy of centralized and distributed scenarios for high dimensional health care data but high information loss and utility decreases. [14] Their proposed algorithm is productive for both privacy and utility of high dimensional data but this not fruitful for outsourced data publishing. [49] They introduced a data anonymization approach for big data and presented a concept of horizontal partition that is not fruitful for curing the privacy preservation of high dimensional data. [48] They introduced a data anonymization approach for big data and presented a concept of horizontal partition that is not fruitful for curing the privacy preservation of high dimensional data.

On the other hand, in [27] this paper they proposed the idea of vertical partition by whom [10] is inspired as they worked on the vertical partitioning of the data and in extension fragmentation of data is done to aim the prevention of attribute and QID disclosure. But still by combining several attributes from multiple fragments linkage attack can be performed. [35] They proposed an image disguise approach for privacy preservation deep learning. They have used blockwise permutation which is prone to re-identification attacks. [51] In this article, a deep learning privacy preserve model is proposed to train data and to share data among the users but their results show that this model is not fruitful for several numbers of participants because their model cannot afford too much load and failed to train effectively.

In this [25] a scheme is proposed in which the classifier owner outsourced users' data on the cloud and maintains its confidentiality and adopts a legitimate third party for security protocols and services. Furthermore, most of the existing works addressed the privacy-preserving classification in online scenario [22], [16]. In this scenario, a classifier administrator has a trained classifier model, and the end-users demand to analyze their data with it [26]. But if that third party got compromised the whole system can be occupied by an adversary and computational cost was very high for this scheme. [51] [50] [2] [38].

Differential private data sharing has now become very common[1]. Because differential privacy promises to keep the confidentiality of any sensitive information [12].

Table IV: Comparative analysis of some Anonymization Techniques

Ref no	**Technique**	**Advantages**	**Limitations**
[30]	Distributed and centralized Anonymization	Privacy in a centralized and distributed scenario	Prone to identity disclosure
[24]	Slicing	Prevent membership attack	Data mix-up due to permutation, high time consumption
[48][39]	k-Anonymity	Privacy preservation of big data	Prone to homogeneity and background knowledge attacks, identity disclosure
[15]	t-closeness	Promotes intra-group diversity, prevents attribute disclosure and skewness attack	On the increase of size or variety of the data, the chance of re-identification attack also increases.
[48][19]	l-Diversity	Promotes intra-group diversity	Attribute disclosure, probabilistic inference attack and homogeneity attacks are possible
[44]	Randomization	It is helpful for masking sensitive information of individuals	It considers all records equally and the data utility decreases. Not feasible for multiple attributes
[7]	Horizontal fragmentation	Managing dimensionality	Prone to attribute disclosure and record linkage attacks
[35]	Blockwise Permutation	Provides disguise to identification	Prone to re-identification attacks
[43][49]	Vertical fragmentation	Managing dimensionality, Privacy preservation of high-dimensional data	Prone to attribute disclosure and record linkage attack
[41][1]	Differential privacy	Prevent adversarial attacks by adding noise. Most suitable for big data	Data utility is reduced

Background knowledge is not considered in differential privacy [11] [40] [31]. Noise is added to the results according to the data type by mathematical calculations. It is most suitable for big data. However, data utility is reduced. As there is always a trade-off between utility and privacy. If there is a risk of privacy leakage more noise is added to the sensitive data. The entire amount of information that has been leaked as a result of the published representations is not limited [3]. However, existing approaches that accomplish differential privacy typically require local platforms to contribute to the backward propagation process, making them difficult to implement on lightweight platforms [22].

GANs (Generative Adversarial Networks) are recently proposed by the Deep Learning community that is still being extensively developed [6] [45]. GANs are used to produce samples that seem similar to those in the training set, rather than to categorize data into different categories (ideally with the same distribution) [35]. In these proposed approaches to instigate an effective attack against collaborative deep learning GANs is not a big deal [47]. As a result of the attack, any user acting as an insider can deduce sensitive data from the device of a victim. The attacker merely executes the collaborative learning algorithm on the victim's device to recreate sensitive information [18]. In the Table IV we have further evaluated some of the existing techniques.

So, we have presented the N-RPP model that is unique as compared to all of the above-mentioned privacy and utility maintaining concepts. In this case, we have thoroughly and briefly studied the existing techniques and stratagems to avert the adversary from any disclosure of delicate information and introduced a cost-effective approach that will efficiently sustain the privacy preservation of outsourced deep learning scenarios. Our main focus in this paper is to maintain privacy and classification accuracy and to overcome the information loss rate. The main purpose of our proposed N-RPP model is to sustain privacy and

utility at the same time.

III. PRELIMINARIES

In the wake of our assumptions and negotiations, fundamental concepts and definitions used in this research article are explained shortly in the following:

A. *Fundamental Conception and Definitions*

Any type of data either raw or high dimensional data consists of individuals 'specified information is composed of several attributes. Due to different categories of attributes with respect to their impartiality they are classified into different types some of them are as follows:

- Quasi-Identifiers: Are those attributes (e.g., zip code, age, gender) that can be linked with the external knowledge to identify the individual in the data [42].
- Sensitive Attribute: Are those attributes (e.g., income, disease, job) that are sensitive and confidential cannot be revealed [48].
- Explicit Attributes: Are those attributes (name, license number ID number etc.) that helps in revealing the identity of an individual completely, as these are the identifiers that clearly identifies the person [10]. Normally, before publishing or outsourcing of data all explicit and sensitive attributes are removed so that the identity of a person cannot be revealed easily. All generalization and anonymization terminologies are applied on the Quasi-identifiers to make it more confidential. But we are applying our terminology not only to QI attributes but to all attributes such as sensitive and explicit attributes which makes us unique until now that we will further discuss in detail.
- **Definition 1** (Equivalence class): This is a type of class, consists of set of generic records that have similar values for QI attributes [43].

- **Definition 2** (K-anonymity): The data-set table that satisfies K-anonymity if it consists of at least k-1 indistinguishable records with respect to QI attributes.So,the identification probability in the table T will be 1/k [43].

- **Definition 3** (Differential Privacy): An algorithm A is ϵ-differential private if for any subset of outputs O and for all datasets, $D1$ and $D2$ differing in at most one element. Where $A(D1)$ and $A(D2)$ are the outputs of an algorithm for input datasets $D1$ and $D2$ respectively and Pr is the randomness of the noise in the algorithm. Here epsilon is called the privacy budget. Note that usage of too much noise can damage the data utility. So the value of the privacy budget epsilon should be small [8].

$$Pr(A(D1) \in O)/Pr(A(D2) \in O) \leq e^{\epsilon} \tag{1}$$

- **Definition 4** (Deep Learning): Deep Learning Basically, deep learning involves training of an artificial neural network which may consists of several layers depend upon how deep you want your network to learn about your input [35]. Generally deep learning system models consist of multiple layers so that more complex and detailed functions and relations can be understood with the help of layers of network. Some common examples of deep networks are CNN(convolution neural network), RNN(recurrent neural network) and MLP(multiple layer perceptron)etc [1] [32]. There are some major and usual functions e.g.,

$$RELU function f(x) = max(0, 1) \tag{2}$$

$$sigmoid function f(x) = (1 + a - x) - 1 \tag{3}$$

  and hyperbolic function which we are also using in this paper. The modernization of this mechanism has attracted intensive research work in this field [36].
- **Definition 5** (Deep learning classification): In classification task, an instance vector x(x1.....xn) is classify into discrete category set of class. In this paper, we are using binary classification.As, we are classifying our output images into 0 and 1 class [18] [11] [33].

IV. PROPOSED APPROACH

Advances in information technologies need of data anonymization is approaching and everybody wants to gain access of high-quality data without any loss of information in a mean time. Use of outsourced cloud computing is also in practice. So, we have developed an efficient approach that is combination of deep learning and data privacy preservation terminology. This section describes the general framework and model of our proposed approach that is capable of preventing any type of disclosure to an individual's identity and sensitive information.

A. *General framework*

A deep neural network is an efficient source to deal with massive and intricate data architectures. The use of outsourced data publishing is also in practice. Many high-profile officialdom and organizations outsource their data to the public cloud service providers to get benefits from their outperformed GPU computation resources. With the advancement of information and technologies need for data anonymization is approaching. Outsourcing data publishing may lead to some serious security risks because cloud service providers cannot be trusted completely. For example, an adversary may analyze the outsourced data and exploit it for its benefits and may launch attacks to harm the resources of data. So, to avoid this a proper security mechanism should be deployed as shown in Figure 1.

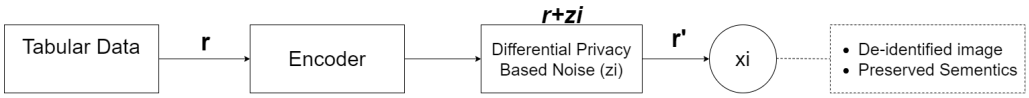

Figure 1: Framework of proposed NRPP model

In our proposed approach we are doing concealment of high dimensional data in a non-reversible image format that will prevent any type of adversary to take sensitive information that we made available on the cloud by outsourcing it. As our main motive is to maintain privacy and utility at the same time. Our proposed approach's methodology is consisting of many phases. As a first step, we cleaned and normalized the data set. We have used a label encoder for assigning numeric values to each attribute present in the data set. We then use the Min-max scaling term for normalizing the data set between 0 and 1.From the data set each row that is composed of several attributes is fed to the model and converted into an image as shown in equation(3.4) below. As all the attributes are anonymized into an image hence they can not be revealed by the adversary.

$$E_G = M(r) \sim [\Sigma a_n(r)] \tag{4}$$

$$M(r) \leftarrow r' \tag{5}$$

For classification, we have considered income as a target label based on which we have made two classes; class 1 for people who have a salary greater than 50,000 and class-0 for having a salary less than 50,000. As shown in Table 1. After that, we calculate the metrics i.e the F1-score and error rate to justify the performance of our proposed model.

B. *NRPP Model*

This section describes the architecture of our deep learning algorithm by which we are generating non-reversible images of individual healthcare records. Our deep learning algorithm has two phases. The workflow of our proposed NRPP model is shown in Figure 2.

Algorithm 1 Generation of non-reversible images from tabular data

Input r = $[r_1, r_2, ...r_n]$, r_1 = $[a_1, a_2.....a_n]$
for each Row $r \leftarrow [r_1, r_2, ...r_n]$ where r_1 = $[a_1, a_2.....a_n]$ **do**
    $E_G = M(r) \sim [\Sigma a_n(r)]$
    $M(r) \leftarrow r'$ //$r' = r + z_i$, where z_i is optimal noise according to $P_b(\epsilon' \leq P_b)$.
end for
Input $\leftarrow$ F_vec, //here r' is referred as F_vec(feature vector).
for each F_vec **do**
    F_vec $\rightarrow x_\iota$ // x_ι is a 2D image
    F= sigmoid(x_ι)
    if F ≤ 0.5 **then**
        $x_\iota \rightarrow class\_0$
    else
        $x_\iota \rightarrow class\_1$
    end if
end for

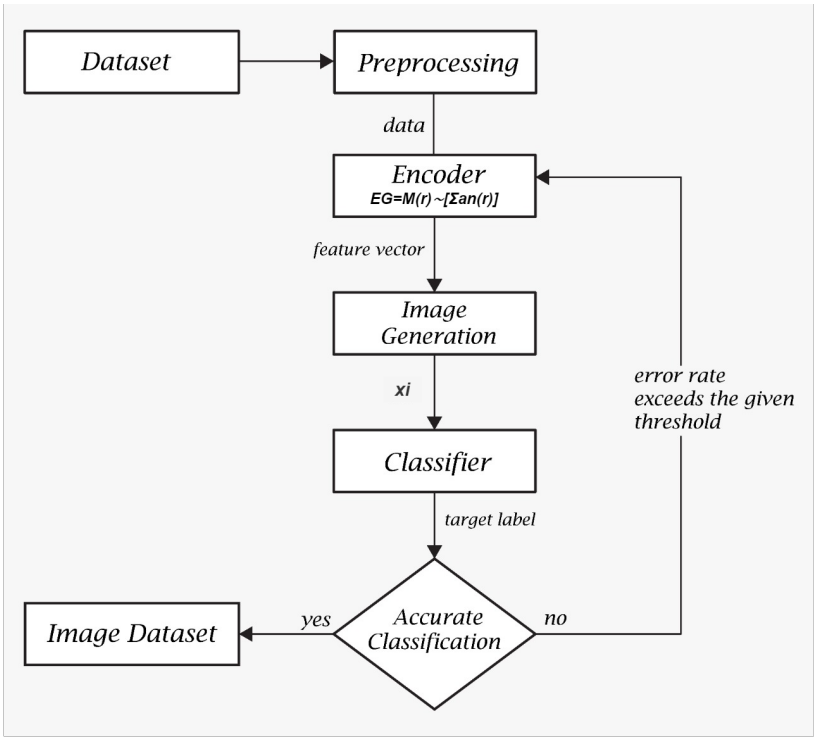

Figure 2: Workflow of proposed approach

1) **Phase 1: Generation of De-identified Content**
Relational data is full of information and publishing it without proper anonymization will result in a privacy violation. This will help the attacker to figure out user's information stored in the database, as well as which record is linked to it. Furthermore, releasing tabular content could result in the disclosure of the original content. For example, if an attacker has prior knowledge of the training mechanism, he can easily reverse engineer the input using a GAN attack algorithm. Hence, it's crucial to preserve the tabular data before publishing it. Differential privacy is an effective strategy for protecting the privacy of users' sensitive data by adding random noise, i.e., Laplacian noise, to the performance of the algorithm. This process is known as output perturbation, and it has been shown that it can guarantee differential privacy under certain conditions. The perturbation mechanism's key concept is to apply noise to an algorithm's output to guarantee differential privacy. There are two significant advantages of adding optimal noise to the tabular form of data. At first, the risk of re-identification of learned representation is reduced and the attacker can be prevented from inferring the sensitive content that exists in the database. Secondly, the attacker would find it difficult to retrieve the raw textual data. The Laplacian mechanism is a widely used method of adding noise to maintain differential privacy. Particularly, the laplacian mechanism perturbs the input with the addition of noise to it as follows:

$$r' = r + z_i \sim Lap(b), b = \frac{\Delta}{\epsilon}, n = 1, ..., d \tag{6}$$

where ϵ is the privacy budget, Δ is the L-sensitivity of the tabular content r, n are the dimensions of r and z_i is the noise vector.
Proposed approach consists of three blocks. Phase 1 has two blocks and Phase 2 contains one block. Output of the previous block becomes input for the next block: First two blocks of Phase 1 are mentioned below:

- **Block one:** This block is called a linear block. Where linear layer takes features' set (equal to number of columns) from each row of the dataset and passes them to dense layers; it gives a 784-feature vector as an output. We are using the rectified linear as activation function.

$$RELU function f(x) = max(0, 1) \tag{7}$$

:

- **Block Two:** It takes that 784 feature vector (generated by Block one) as an input and passed them to convolutional transpose layers to convert the vector into a 2-dimensional array (an image of 36x36). As shown in Figure 3.

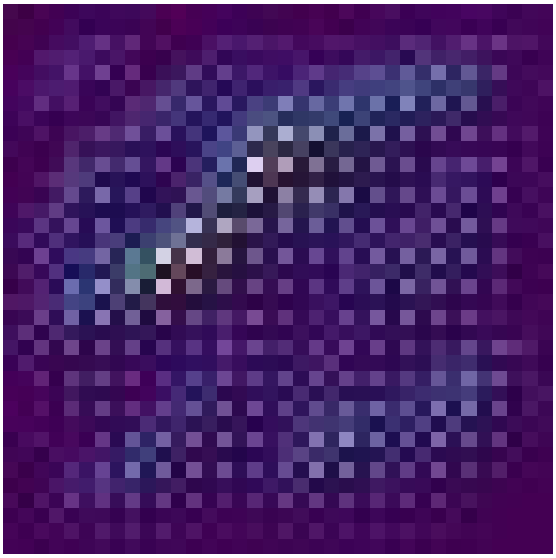

Figure 3: Image36*36

2) **Phase 2: Preservation of Semantic Meaning**
Normally, the latent representation is perturbed by adding noise to it. It prevents the adversary from the reconstruction of textual data and differential privacy is satisfied. However, adding a lot of noise can prevent the adversary to get sensitive information but it may destroy data for further use as the semantic meaning can no longer remain in it. It is very important to maintain privacy and utility at the same time. As only just give protection to sensitive data is not a good practice computationally. To preserve semantic meaning is important because it maintains the utility of data to use it further in other tasks that depend on it e.g classification. As classification is one of the most common tasks. To preserve semantic meaning for the sake of utility and privacy of data, an optimal amount of noise can be added to the data whether it is in textual form or image format. However, we have approached this challenge of maintaining semantic meaning and as well as privacy in a different way. We have added optimal noise according to our privacy budget P_b and converted all of the textual data into a non-reversible image format row by row. A large privacy budget can lead to large privacy limits. Thus, we add constraint $\epsilon < P_b$ where P_b is the pre-defined constraint [17]. Hence, the de-identified dataset of images is created that is not only diferentially private but also can be classified by any machine learning classifier. As we have done binary classification so, we have used the sigmoid function for our classifier thus we have equation 8:

$$\hat{w} = sigmoid(r', \theta_s) \tag{8}$$

θ_s weights are associated with the sigmoid function and $\hat{w}$ is the target label we have set for the binary classification.

For phase 2, our deep learning Algorithm has one block which is briefly explained below:

- **Block Three:** Block three is a convolutional neural network. It takes an input image of a size 36x36, and perform binary classification. In last layer we are using the sigmoid as activation function. Hence the output returns with the probability of 0 and 1. Which shows that image either belongs to class 0 or class 1 as shown in figures 4 and 5. In case the classification is done incorrectly the error is passed to Block 2 of phase 1. Using the error the Block 2 adjusts its weights and regenerates the image. We pass rows one by one in this hybrid multi phase neural network and train it until it learns how to generate easily classifiable images. Hence every row will generate an image that is easily classifiable. We run the Epoch to check the error rate and classification accuracy.

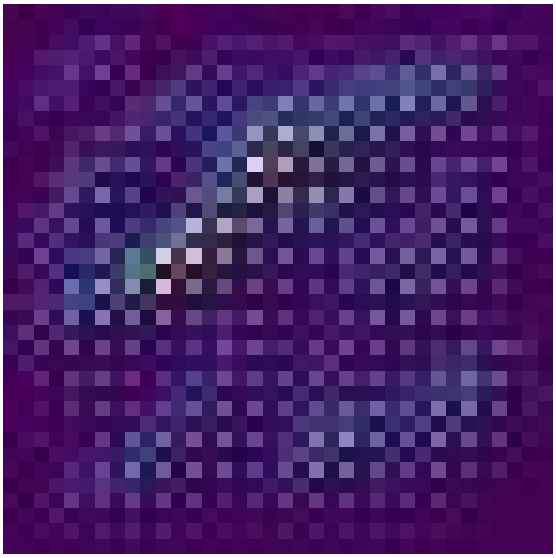

Figure 4: Class 0 Image

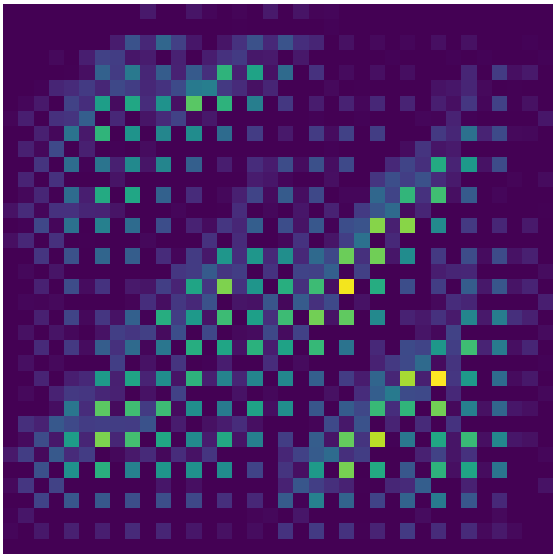

Figure 5: Class 1 Image

V. Protection Analysis

In this section, we have focused on the protection analysis of the proposed NRPP model. In our approach, we have anonymized high-dimensional relational data into non-reversible image format using deep learning.

In our work, noise involved for the anonymization of high dimensional data into an image format is ϵ'-differential private where $\epsilon' \leq P_b$. Here P_b is the optimal privacy budget. We demonstrate the protection assurance of our final non-reversible image x_i for each record. Theoretical findings indicate that NRPP model eliminates the possibility of disclosing the presence of any content.

Lemma1: If dataset D_1 satisfies differential privacy then the D_1 also satisfies $\epsilon' \leq P_b$

Proof: Suppose $r = [r_1, \ldots r_n], r_1 = [a_1, a_2 \ldots .. a_n]$ from the data set each row r combine with the noise vector z_i that is an optimal noise according to our pre-defined privacy budget P_b is fed to the model M and converted into a non-reversible image x_i that is de-identified and satisfies the differential privacy.

Theorem1: Let, $\epsilon' \leq P_b$ be the optimal value for the privacy budget ϵ in order to preserve the semantic meaning and private attributes. Suppose r be the original tabular content (row) for the set of attributes, $[a_1, \ldots a_n]$. In addition, Δ represents the L-sensitivity for the final anonymized output i.e image. If each element $z_i(m), m = 1, \ldots . . n$ in noise vector z_i is randomly selected from Lap($\frac{\Delta}{\epsilon}$), $\Delta = 2d$ the final output representation $r' = r + z_i$ gives image i.e x_i satisfies ϵ'-differential privacy.

Proof: As we know that $\epsilon' \leq P_b$ is the optimal value for the privacy budget. Then each element in z_i is distributed as Lap($\frac{\Delta}{\epsilon}$). Lap($\frac{\Delta}{\epsilon}$) is equal to arbitrarily picking each $z_i(m)$ from Lap($\frac{\Delta}{\epsilon}$) distribution, whose probability density function is

$$P_r(z_i(m)) = (\frac{\epsilon'}{2\Delta})^{e^{-\frac{\epsilon'|z_i(m)|}{\Delta}}} \tag{9}$$

Let D_1 and D_2 be any two datasets differing by only one record without any loss of information we assume that representation of our tabular content r is converted to r'. The L –sensitivity is equals to $||r - r'||1 \leq \Delta$ Then we have :

$$\frac{P_r[r+z_i=i|D_1]}{P_r[r'+z'_i=i|D2]} = \frac{\Pi_m \in (1,\ldots..,d) P_r(i-r(m))}{\Pi_m \in (1,\ldots..,d) P_r(i-r'(m))} \tag{10}$$

$$= \frac{\Pi_m \in (1,\ldots..,d) P_r(z_i(m))}{\Pi_m \in (1,\ldots..,d) P_r(z'_i(m))} \tag{11}$$

$$= e^{-\frac{\epsilon' \sum_m |z_i(m)|}{\Delta}} / e^{-\frac{\epsilon' \sum_m |z'_i(m)|}{\Delta}} \tag{12}$$

$$= e^{-\frac{\epsilon' \sum_m |z'_i(m)| - |z_i(m)|}{\Delta}} \leq e^{-\frac{\epsilon' \sum_m |z'_i(m)| - |z_i(m)|}{\Delta}} \tag{13}$$

$$= e^{\frac{\epsilon' ||z'_i - z_i||_1}{\Delta}} \tag{14}$$

Where z_i and z'_i are corresponding noise vectors w.r.t ϵ when inputs are D_1 and D_2 respectively. Since we have $z_i = i - r$ and $z'_i = i - r'$, we can write:

$$= ||z'_i - z_i||_1 = ||(i - r') - (i - r)||1 \tag{15}$$

$$||r' - r||_1 \in \Delta \tag{16}$$

This follows from the definition of L-sensitivity. We can rephrase the equation 10:

$$\frac{P_r[r+z_i=i|D_1]}{P_r[r'+z'_i=i|D2]} \leq e^{\frac{\epsilon' ||z'_i - z_i||_1}{\Delta}} \leq e^{\frac{\epsilon' \Delta}{\Delta}} = e^{\epsilon'} \tag{17}$$

VI. Experiments

In this section, we utilize two real world datasets for experiments to illustrate the efficiency of the NRPP model regarding privacy and utility.

A. *Datasets*

We used two datasets from the machine learning repository named Musk and Census. There are 72874 records in the Musk dataset, each having 168 numerical attributes. The task associated with it is to figure out whether the molecule belongs to musk or not. Census is the second dataset, which contains records collected from various precise population surveys. There are 199523 records and 40 categorical attributes in this dataset. Task associated to it is to predict income of a person is less than 50k or more than 50k. We did pre-processing and remove all the null characters of both datasets individually. After cleaning each dataset, we randomly chose 3000 records from each dataset to avoid the negative impact of imbalanced record values and then save the CSV files of each dataset separately to proceed further.

B. *Evaluation metrics*

The goal of our NRPP model is to keep as much data utility as feasible while protecting confidential information. The following two metrics are frequently used in the literature to determine the proficiency of an anonymization strategy.

1) ***Classification accuracy****:* Classification accuracy is a metric used to measure a classification model efficiency by dividing the number of correct predictions by the total number of predictions. It is the most commonly used statistic for testing the classifier and measuring its' ability to predict unknown vector classes. It is simple to calculate and understand. Normally for binary classification, F-measure is used. The F-measure is a metric for determining how accurate a model is on a given dataset. It is used to examine binary classification algorithms that categorize instances as either "positive" or "negative." This is also defined as the mean of the model's precision and recall. F-measure is a popular metric for assessing data mining algorithms as well as a variety of machine learning mechanisms, particularly in processing of languages.It ranges from 0 to 1 called as F1-score. F-measure is calculated by using formula given in the equation 18

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{18}$$

Where Precision is the proportion of true positive examples among those labeled as positive by the model and Recall is the proportion of positive instances rated as positive out of the total number of positive examples. With the increase in the value of the F-measure, the classification accuracy of a model also increases.

2) ***Loss****:* A loss function is also known as a error function. Which considers the likelihood or uncertainty of a prognosis based on how much it differs from the true value. It's a total of each sample's errors in the training or validation. This provides a more balanced view of the model's performance. Most commonly for binary classification binary cross-entropy loss is used which calculates the error rate between 0 and 1.

C. ***Simulation results***

We have evaluated the performance of our proposed NRPP model on the above-mentioned evaluation metrics. We have considered the Census dataset for the execution of the first experiment. As we are anonymizing each record of dataset consists of several attributes into a de-identified image with the help of our deep learning model. So, the identity of an individual can not be disclosed. To test, we used the Resent-18 pre-trained model. We passed the generated images dataset of Census dataset from deep learning classifier and metrics like classification accuracy and error rate are calculated. We get a 0.9 plus F1-score and error rate is 0.08 approximately. As shown in Figure 6

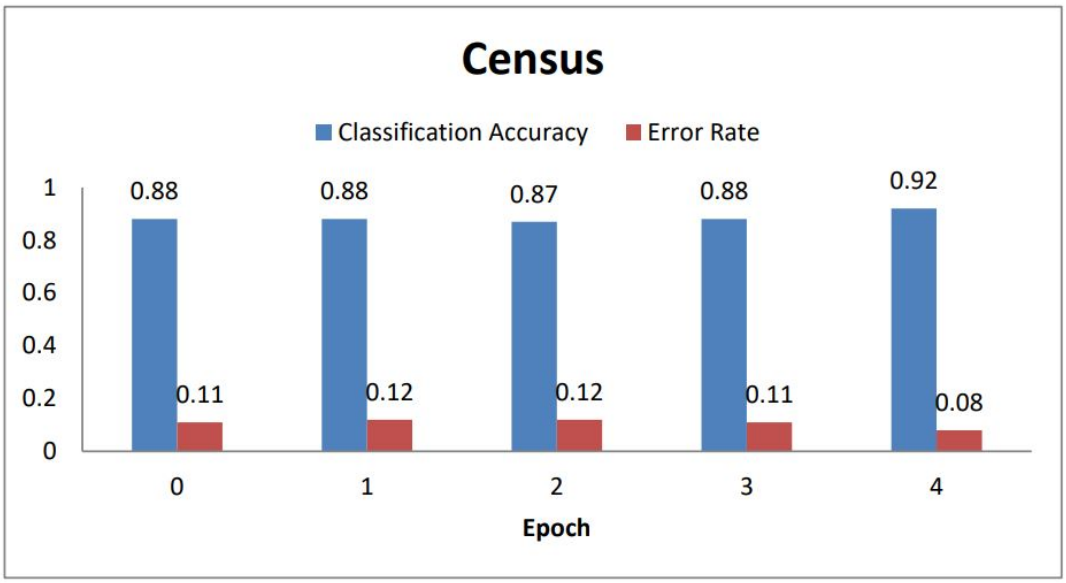

Figure 6: Classification accuracy of Census dataset

Figure 7 shows the result of the Musk dataset that we used to conduct the second experiment. It can be seen clearly that our model performed well as the classification accuracy of generated images' dataset of Musk is gradually increasing while the error rate is decreasing. Hence, we obtain a 0.92 plus F1−score, with a minimum error rate i.e 0.07 approximately while running just 5 epochs. Which indicates that the maximum utility metric is preserved.

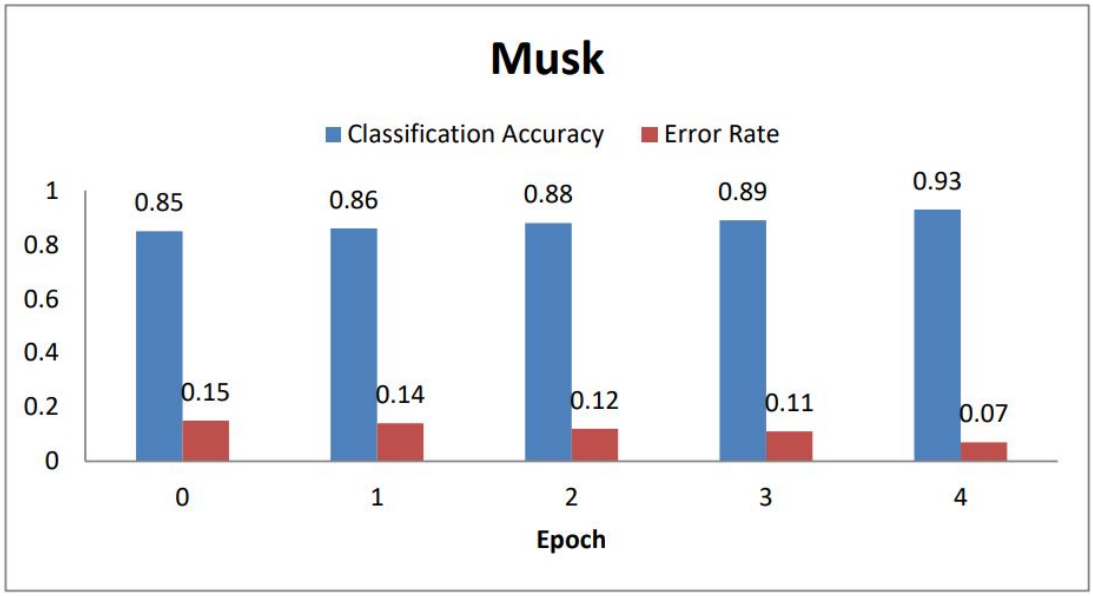

Figure 7: Classification accuracy of Musk dataset

Figure 8 and 9 represents the execution time of Census and Musk datasets respectively. It can be seen clearly that execution time is gradually decreasing on each epoch. The reason for this decrease is the tendency of our approach as with minimum learning rate it gives outstanding results. From figure 8 it can be seen that the execution time of Census dataset on Epoch 0 is 45 seconds which gradually decreases to 38 seconds on Epoch 4. Similarly, from figure 9 it is visible that the execution time of the Musk dataset on Epoch 0 is 59 seconds which further decreases to 55 seconds on Epoch 4.

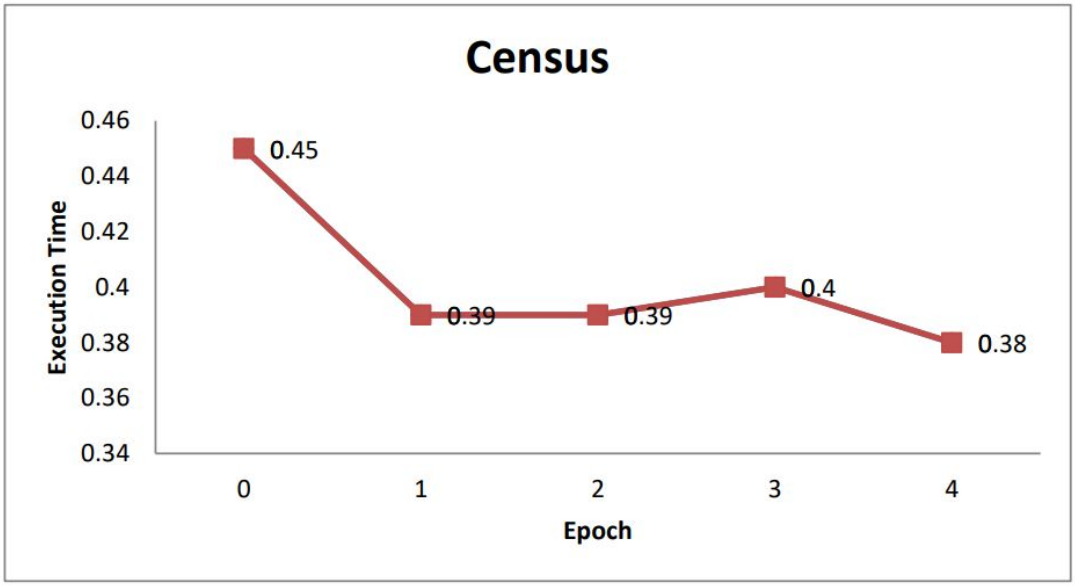

Figure 8: Execution time of Census dataset

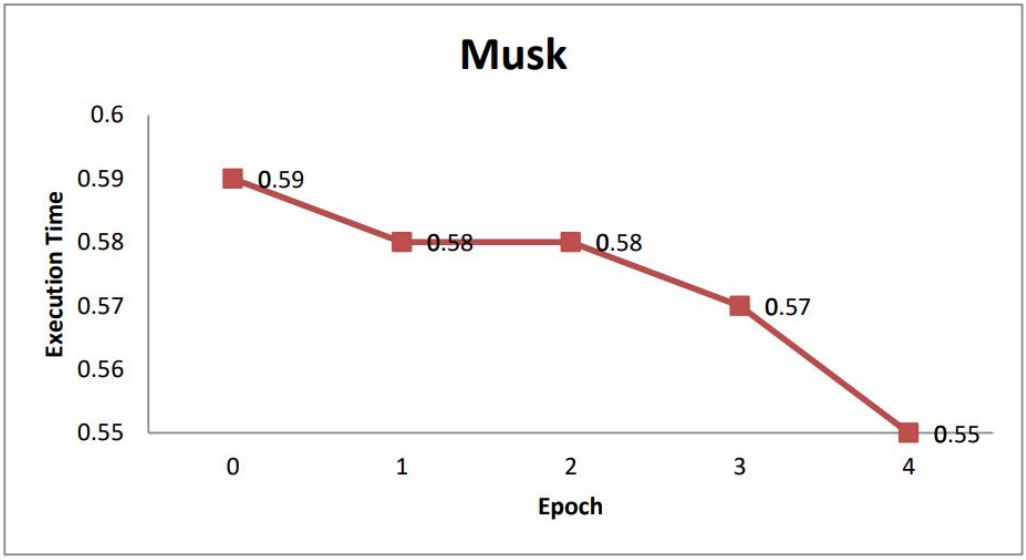

Figure 9: Execution time of Musk dataset

Our proposed approach is unique in terms of providing privacy and utility. As we are not using the traditional methods to protect sensitive and non-sensitive information. Because we are not publishing the data in the anonymized tabular format. Unlike we are converting the tabular data into a de-identified image format. Hence, it can not be compared with existing techniques in graphical result format.But for the sake of performance analysis, we compare the results of our NRPP model with [43] in terms of classification accuracy, loss rate and execution time for both datasets as shown in tables V and VI. This shows that our NRPP Model gives excellent results for classification accuracy, loss rate, and execution time. Moreover, the generated image dataset can be classified by any other machine learning algorithm which makes our proposed NRPP model unique and distinct.

Table V: Comparative analysis of approaches for Musk dataset:

Techniques	Musk		
	F-measure	Loss	Execution Time(mins)
Vertical Fragmentation [43] (k=5,frag=8)	0.76	0.31	6.32
NRPP Model	0.93	0.07	3.30

Table VI: Comparative analysis of approaches for Census dataset:

Techniques	Census		
	F-measure	Loss	Execution Time(mins)
Vertical Fragmentation [43] (k=5,frag=8)	0.79	0.21	17.45
NRPP Model	0.92	0.08	2.45

VII. Conclusion

Outsourcing of data is a common trend now which is one of the main causes of privacy violation. High-dimensional data is also the main curse these days. As data with many dimensions also increase the chances of adversarial attacks. Many privacy preservation techniques have been developed for the privacy and utility of big and high-dimensional data but still, the personal information of individuals is prone to many disclosure attacks. To deal with these issues we proposed a unique solution for publishing and outsourcing high-dimensional data as compared to traditional existing methods. We aimed in this approach to sustain the privacy and utility of outsourced high-dimensional data. The proposed scheme ensures the preservation of semantic meaning and de-identification of the data. We have developed an incentive NRPP model which resembles GAN's generator that converts the tabular data into a non-reversible image format and performs classification accuracy without losing any information. Our proposed methodology is entirely different and unique because we are publishing tabular data as de-identified image format which is non-reversible deferentially private for any adversary to get any sensitive information of a legitimate user even if he has some background knowledge. The empirical study of our experiments indicates that on minimum learning our proposed NRPP model outputs 0.9 plus F1- score with a minimum error rate. This shows that our proposed model provided the best and excellent results in terms of privacy and utility while publishing high dimensional that is the best practice for the desired trade-off while publishing sensitive data.

REFERENCES

[1] Martin Abadi et al. "Deep learning with differential privacy". In: *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*. 2016, pp. 308–318.

[2] Yashar Abed and Meena Chavan. "The challenges of institutional distance: Data privacy issues in cloud computing". In: *Science, Technology and Society* 24.1 (2019), pp. 161–181.

[3] Gergely Acs et al. "Differentially private mixture of generative neural networks". In: *IEEE Transactions on Knowledge and Data Engineering* 31.6 (2018), pp. 1109–1121.

[4] Rakesh Agrawal and Ramakrishnan Srikant. "Privacy-preserving data mining". In: *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. 2000, pp. 439–450.

[5] Sayyada Hajera Begum and Farha Nausheen. "A comparative analysis of differential privacy vs other privacy mechanisms for big data". In: *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE. 2018, pp. 512–516.

[6] Ghazaleh Beigi et al. "I am not what i write: Privacy preserving text representation learning". In: *arXiv preprint arXiv:1907.03189* (2019).

[7] Sabri Boughorbel, Fethi Jarray, and Mohammed El-Anbari. "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric". In: *PloS one* 12.6 (2017), e0177678.

[8] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. "Differentially private empirical risk minimization." In: *Journal of Machine Learning Research* 12.3 (2011).

[9] Jingxue Chen, Gao Liu, and Yining Liu. "Lightweight privacy-preserving raw data publishing scheme". In: *IEEE Transactions on Emerging Topics in Computing* (2020).

[10] Shekha Chenthara et al. "Security and privacy-preserving challenges of e-health solutions in cloud computing". In: *IEEE access* 7 (2019), pp. 74361–74382.

[11] Wenliang Du, Yunghsiang S Han, and Shigang Chen. "Privacy-preserving multivariate statistical analysis: Linear regression and classification". In: *Proceedings of the 2004 SIAM international conference on data mining*. SIAM. 2004, pp. 222–233.

[12] Cynthia Dwork, Aaron Roth, et al. "The algorithmic foundations of differential privacy." In: *Foundations and Trends in Theoretical Computer Science* 9.3-4 (2014), pp. 211–407.

[13] Benjamin CM Fung et al. "Privacy-preserving data publishing: A survey of recent developments". In: *ACM Computing Surveys (Csur)* 42.4 (2010), pp. 1–53.

[14] Benjamin CM Fung et al. "Service-oriented architecture for high-dimensional private data mashup". In: *IEEE Transactions on Services Computing* 5.3 (2011), pp. 373–386.

[15] Anjana Gosain and Nikita Chugh. "Privacy preservation in big data". In: *International Journal of Computer Applications* 100.17 (2014).

[16] Jihun Hamm et al. "Crowd-ml: A privacy-preserving learning framework for a crowd of smart devices". In: *2015 IEEE 35th International Conference on Distributed Computing Systems*. IEEE. 2015, pp. 11–20.

[17] Changhee Han et al. "Combining noise-to-image and image-to-image GANs: Brain MR image augmentation for tumor detection". In: *IEEE Access* 7 (2019), pp. 156966–156977.

[18] Briland Hitaj, Giuseppe Ateniese, and Fernando Perez-Cruz. "Deep models under the GAN: information leakage from collaborative deep learning". In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. 2017, pp. 603–618.

[19] Priyank Jain, Manasi Gyanchandani, and Nilay Khare. "Big data privacy: a technological perspective and review". In: *Journal of Big Data* 3.1 (2016), pp. 1–25.

[20] Hillol Kargupta et al. "On the privacy preserving properties of random data perturbation techniques". In: *Third IEEE international conference on data mining*. IEEE. 2003, pp. 99–106.

[21] Slava Kisilevich et al. "Efficient multidimensional suppression for k-anonymity". In: *IEEE Transactions on Knowledge and Data Engineering* 22.3 (2009), pp. 334–347.

[22] Meng Li et al. "Privynet: A flexible framework for privacy-preserving deep neural network training". In: *arXiv preprint arXiv:1709.06161* (2017).

[23] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity". In: *2007 IEEE 23rd International Conference on Data Engineering*. IEEE. 2007, pp. 106–115.

[24] Tiancheng Li et al. "Slicing: A new approach for privacy preserving data publishing". In: *IEEE transactions on knowledge and data engineering* 24.3 (2010), pp. 561–574.

[25] Tong Li et al. "Communication-efficient outsourced privacy-preserving classification service using trusted processor". In: *Information Sciences* 505 (2019), pp. 473–486.

[26] Xingxin Li et al. "On the soundness and security of privacy-preserving SVM for outsourcing data classification". In: *IEEE Transactions on Dependable and Secure Computing* 15.5 (2017), pp. 906–912.

[27] Entao Luo et al. "Privacyprotector: Privacy-protected patient data collection in IoT-based healthcare systems". In: *IEEE Communications Magazine* 56.2 (2018), pp. 163–168.

[28] Ashwin Machanavajjhala et al. “l-diversity: Privacy beyond k-anonymity”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1 (2007), 3–es.

[29] Antonio Maratea, Alfredo Petrosino, and Mario Manzo. “Adjusted F-measure and kernel scaling for imbalanced data learning”. In: *Information Sciences* 257 (2014), pp. 331–341.

[30] Noman Mohammed et al. “Centralized and distributed anonymization for high-dimensional healthcare data”. In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 4.4 (2010), pp. 1–33.

[31] Noman Mohammed et al. “Differentially private data release for data mining”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011, pp. 493–501.

[32] Payman Mohassel and Yupeng Zhang. “Secureml: A system for scalable privacy-preserving machine learning”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 19–38.

[33] Benjamin IP Rubinstein et al. “Learning in a large function space: Privacy-preserving mechanisms for SVM learning”. In: *arXiv preprint arXiv:0911.5708* (2009).

[34] Ahmed Ali Mohammed Al-Saffar, Hai Tao, and Mohammed Ahmed Talab. “Review of deep convolution neural network in image classification”. In: *2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*. IEEE. 2017, pp. 26–31.

[35] Sagar Sharma and Keke Chen. “Image disguising for privacy-preserving deep learning”. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. 2018, pp. 2291–2293.

[36] Reza Shokri and Vitaly Shmatikov. “Privacy-preserving deep learning”. In: *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*. 2015, pp. 1310–1321.

[37] Reza Shokri et al. “Membership inference attacks against machine learning models”. In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 3–18.

[38] Krishna Mohan Pd Shrivastva, MA Rizvi, and Shailendra Singh. “Big data privacy based on differential privacy a hope for big data”. In: *2014 International Conference on Computational Intelligence and Communication Networks*. IEEE. 2014, pp. 776–781.

[39] Latanya Sweeney. “k-anonymity: A model for protecting privacy”. In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[40] Antony Selvadoss Thanamani. “Comparison and analysis of anonymization techniques for preserving privacy in big data”. In: *Adv. Comput. Sci. Technol* 10.2 (2017), pp. 247–253.

[41] David Van Riper, Tracy Kugler, and Steven Ruggles. “Disclosure Avoidance in the Census Bureau’s 2010 Demonstration Data Product”. In: *International Conference on Privacy in Statistical Databases*. Springer. 2020, pp. 353–368.

[42] Qian Wang et al. “Privacy-preserving collaborative model learning: The case of word vector training”. In: *IEEE Transactions on Knowledge and Data Engineering* 30.12 (2018), pp. 2381–2393.

[43] Rong Wang et al. “Privacy-preserving high-dimensional data publishing for classification”. In: *Computers & Security* 93 (2020), p. 101785.

[44] Stanley L Warner. “Randomized response: A survey technique for eliminating evasive answer bias”. In: *Journal of the American Statistical Association* 60.309 (1965), pp. 63–69.

[45] Liyang Xie et al. “Differentially private generative adversarial network”. In: *arXiv preprint arXiv:1802.06739* (2018).

[46] Xin-She Yang et al. *Information analysis of high-dimensional data and applications*. 2015.

[47] Jinao Yu et al. “GAN-Based Differential Private Image Privacy Protection Framework for the Internet of Multimedia Things”. In: *Sensors* 21.1 (2021), p. 58.

[48] Hessam Zakerzadeh, Charu C Aggarwal, and Ken Barker. “Privacy-preserving big data publishing”. In: *Proceedings of the 27th international conference on scientific and statistical database management*. 2015, pp. 1–11.

[49] Hessam Zakerzadeh, Charu C Aggarwal, and Ken Barker. “Towards breaking the curse of dimensionality for high-dimensional privacy”. In: *Proceedings of the 2014 SIAM International Conference on Data Mining*. SIAM. 2014, pp. 731–739.

[50] Jinxue Zhang et al. “Privacy-preserving social media data outsourcing”. In: *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE. 2018, pp. 1106–1114.

[51] Lingchen Zhao et al. “Privacy-preserving collaborative deep learning with unreliable participants”. In: *IEEE Transactions on Information Forensics and Security* 15 (2019), pp. 1486–1500.