https://doi.org/10.7840/kics.2024.49.3.393

Two-Fold Differentially Private Mechanism for Big Data Analysis

Assem Utaliyeva◆, Yoon-Ho Choi°

ABSTRACT

Differential privacy (DP) has emerged as a gold standard for privacy preservation in many applications, particularly in light of recent advancements in machine learning for big data for cloud services and the growing threat of privacy attacks. However, the addition of random noise to data for privacy preservation often results in decreased data quality and utility. To address this challenge, we propose a novel twofold differentially private data generation method that leverages the power of denoising autoencoders to preserve higher data quality and utility. Our approach combines traditional additive differential privacy with a novel reductive differential privacy approach that uses a denoising autoencoder to restore the original distribution of the data, increasing the data utility in machine learning tasks. We also experimentally show the effectiveness of the proposed method by experimental evaluation.

Key Words : Differential Privacy, ML for Big Data, Privacy Preservation

Ⅰ. Introduction

Nowadays, machine learning (ML) is primarily involved in big data and cloud-based applications. With such an increase in the ML demand, concerns about data privacy being used to derive valuable insights and train ML models are also increasing. Both ML models and cloud-based systems are vulnerable to various privacy attacks that can exploit sensitive information from the data itself. This is particularly true for applications that involve sensitive or personal information, such as healthcare, finance, or government.

For instance, privacy attacks that target ML models include membership inference attacks (MIA)[1], model inversion[2,3], and attribute inference attacks (AIA)[4,5], where the adversary tries to infer sensitive information about the training data by exploiting the model itself. In the case of privacy attacks for cloud-based systems, it includes side-channel attacks[6], and man-in-the-middle (MITM) attacks[7], where the adversary can obtain or eavesdrop on training data in the cloud-based system.

The threat model in Fig. 1 shows the practical scenario for privacy attacks in cloud-based systems. Assume the client sends structured data in table format to the cloud server to train the ML model in a cloud computing environment. Data privacy could be exploited by an MITM attack during data transmission or by MIA and AIA attacks that target the fitted ML model on the cloud server side.

To address such concerns, we consider differential privacy (DP)[8], which is currently recognized as a golden standard for privacy preservation. DP is a mathematical concept of privacy that guarantees single-user or record indistinguishability by the addition of randomly generated noise to the data. To the best of our knowledge, differentially private data generation methods[9-11] face large privacy and utility trade-off problem, due to the large performance degradation of such noisy data. To preserve the high usefulness of the data, most methods sacrifice privacy

※ This work was supported by a 2-Year Research Grant of Pusan National University

◆ First Author : Pusan National University, School of Computer Science and Engineering, assemakx@pusan.ac.kr, 학생회원

° Corresponding Author : Pusan National University, School of Computer Science and Engineering, yhchoi@pusan.ac.kr, 종신회원

논문번호 : 202309-075-A-RN , Received September 8, 2023; Revised November 27, 2023; Accepted November 29, 2023

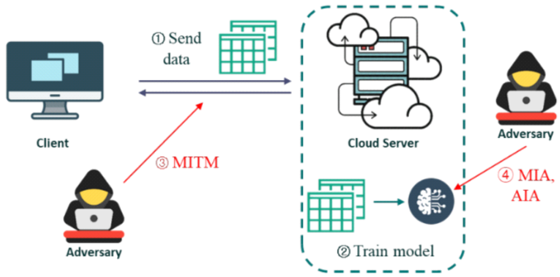


Fig. 1. Threat model

guarantees or vice versa.

In this paper, we propose a novel two-fold DP method of random noise addition inspired by the concept of *no operation* to guarantee high privacy and high data utility simultaneously. Here, the term no operation refers to the no-op machine language instruction that does nothing. In other words, the random noise added to the data is not supposed to change the utility of the data. The proposed method combines the regular concept of DP i.e., additive with the novel concept of reductive DP, where the former adds random noise, and the latter subtracts random noise. However, it is worth noting that reductive DP is not capable of removing the noise completely due to the random nature of the noise. Thus, it does not result in the cancellation of differentially private noise. The key concept of the proposed method is to decrease the degradation of the data utility while preserving privacy, specifically mitigating the MITM and attribute inference attack in the ML models for cloud-based systems.

The rest of the paper is organized as follows. First, we briefly introduce details of DP and mechanisms in section II. Next, we introduce the proposed method and provide a mathematical evaluation in section III. In section IV, we experimentally evaluate the proposed method. Finally, we conclude this paper in section V.

Ⅱ. Preliminaries

2.1 Differential Privacy

DP is the mathematical framework for preserving the privacy of individuals in a dataset while enabling useful analysis. The basic idea behind DP is to add various types of random noise calibrated to hide a single record, such that its presence or absence can cause at most exp(ϵ) + δ change, where ϵ is the privacy parameter and δ is a relaxation parameter.

1) Laplace Mechanism[8]: Since DP is not an algorithm but a notion of privacy, there are various techniques to ensure DP is called differentially private mechanisms. The Laplace mechanism is a representative output perturbation mechanism. It adds random noise drawn from the Laplace distribution.

$$M_{Lap}(D, f(\cdot), \varepsilon) = f(D) + Lap\left(\frac{\Delta f}{\varepsilon}\right) \quad (1)$$

where *f*(*D*) is the original numerical value, *Lap* is the probability density function of Laplace random distribution, ε is the privacy parameter, and *Δf* is the sensitivity, that quantifies the maximum change that can occur in the absence of a single user in the dataset.

2) Composition Property[8]: DP has several properties that enable the building of complex mechanisms and applications. Sequential composition implies that if *F1(x)* satisfies ε1-DP and *F2(x)* satisfies ε2-DP, then the combined mechanism *G(F1(x), F2(x))*, which sequentially releases results, satisfies (ε1 + ε2)-DP.

2.2 Denoising Autoencoders

Denoising autoencoders (DAE) are a type of neural network that is trained to remove the noise from the input data. Similar to traditional autoencoders, DAE consists of encoder and decoder networks, that produce compressed representation and reconstruct it back. However, to denoise the input DAE encourages the network to capture the most salient features while ignoring the noise and tries to minimize the difference between the reconstructed output (denoised) and the clean input. DAEs are commonly used in denoising unstructured data such as images[12,13] and signals [14] but can also be adapted to denoise structured data[15].

Ⅲ. Proposed Method

In this section, we introduce the concept of the proposed method. As shown in Figure 2, the proposed

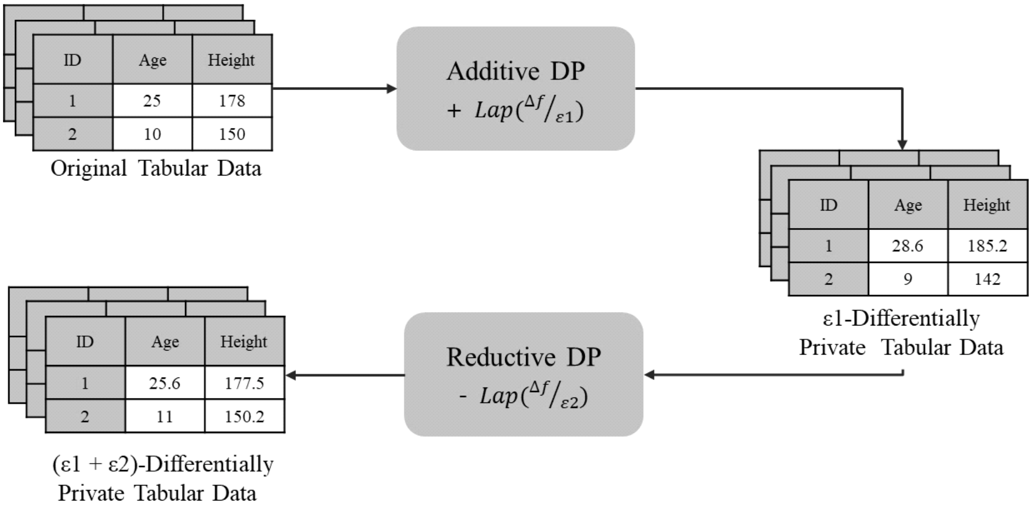

ID
Age
Height
1
25
178
2
10
150
Original Tabular Data
Additive DP
$+ Lap(^{\Delta f}/_{\varepsilon 1})$
ID
Age
Height
1
28.6
185.2
2
9
142
ε1-Differentially Private Tabular Data
Reductive DP
$- Lap(^{\Delta f}/_{\varepsilon 2})$
ID
Age
Height
1
25.6
177.5
2
11
150.2
(ε1 + ε2)-Differentially Private Tabular Data

Fig. 2. Operational overview of the proposed method

method is twofold. Thus, it consists of (1) Additive DP and (2) Reductive DP steps.

3.1 Additive DP

The first step of the proposed method is the well-known additive application of random noise drawn from the Laplace Mechanism given the privacy budget ε1 and sensitivity *Δf*. As a result, we obtain an ε1-differentially private version of the data. However, at this stage, such a straightforward addition of noise provides high privacy guarantees while reducing the usefulness and utility of that data. Thus, it corrupts the distribution of data and negatively impacts the performance of the ML model, which will be trained on ε1-differentially private data.

3.2 Reductive DP

The second step of the proposed method is a novel reductive DP concept implemented using a DAE that tries to remove the ε2-differentially private noise distribution from the ε1-differentially private generated in the previous step. Here, we assume that ε1 = ε2 to efficiently estimate the amount of random noise, while the sensitivity *Δf* remains the same.

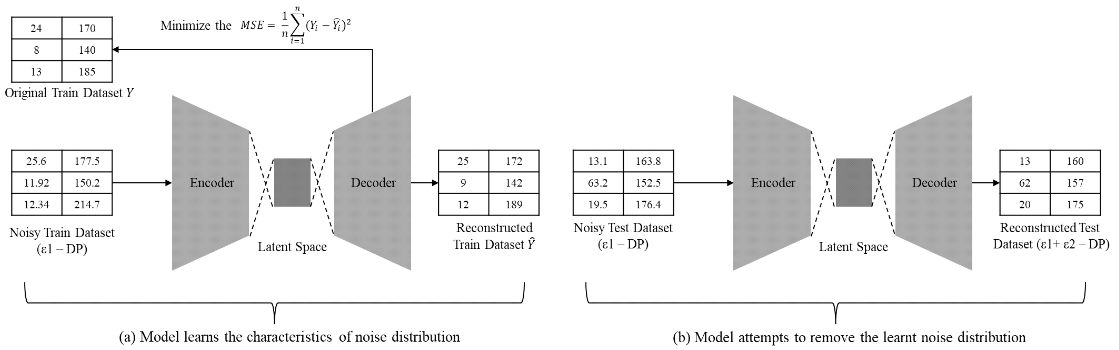
The key concept behind Reductive DP is an attempt to preserve the distribution of differentially private data to be as close as possible to the distribution of original data while offering strong privacy guarantees. It is important to note, that the nature of the noise is random, and DAE cannot completely remove the random noise that is added in the additive DP step. Instead, DAE tries to learn the noise distribution (ε2-DP) and generates denoised version of the ε1-differentially private data. As a result, the generated data can be referred as (ε1 + ε2)-differentially private data according to the sequential composition property.

Figure 3 illustrates the high-level architecture of the proposed Reductive DP step implemented with the DAE.

During the training phase, as shown in Figure 3 (a) the network receives two inputs: the differentially private (noisy) dataset and the original dataset defined as *Y*. In the forward pass, the noisy dataset is inputted into the encoder, which compresses it into a lower-dimensional representation (latent space). Subsequently, the decoder attempts to reconstruct the original data from this compressed representation.

The process is optimized using a loss function defined as the mean squared error (MSE) between the reconstructed $\hat{Y}$ and original data representations *Y*, aiming to minimize this loss. The MSE is defined as follows:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y_i})^2 \quad (2)$$


24 170
8 140
13 185
Original Train Dataset Y
Minimize the $MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
25.6 177.5
11.92 150.2
12.34 214.7
Noisy Train Dataset (ε1 – DP)
Encoder
Latent Space
Decoder
25 172
9 142
12 189
Reconstructed Train Dataset Ŷ
13.1 163.8
63.2 152.5
19.5 176.4
Noisy Test Dataset (ε1 – DP)
Encoder
Latent Space
Decoder
13 160
62 157
20 175
Reconstructed Test Dataset (ε1+ ε2 – DP)

(a) Model learns the characteristics of noise distribution

(b) Model attempts to remove the learnt noise distribution

Fig. 3. Overview of the DAE model as Reductive DP step

where the n is the number of samples, Y_i is the original value of the i^{th} sample, and $\hat{Y}_i$ is the reconstructed value of the i^{th} sample.

By consistently training on such pairs of noisy and original data, the DAE is also able to learn the characteristics of the noise distribution. Since the training dataset includes noise derived from the Laplace distribution, the model learns to identify and mitigate this specific type of noise without explicitly modeling it using parameter values.

Figure 3 (b) illustrates the test phase when the differentially private test data is denoised using the trained DAE model by removing the previously learned noise distribution. However, it is impossible to learn and predict the exact values of the noise distribution, as it is randomly generated. Therefore, we assume that the reconstructed dataset will have an approximately similar distribution to the original data, with the Laplace noise substantially reduced denoted as (ε1 + ε2)-differentially private data.

3.3 Analysis

To prove that the proposed method satisfies DP, we show that both components satisfy ε1 and ε2-DP separately. Specifically, equation 3 is the mathematical proof of the regular additive DP concept[8], where the random noise drawn from the Laplace distribution is added to the original values, and equation 4 is our proof of the novel reductive DP concept, where the noise, similarly drawn from the Laplace distribution, is subtracted. Furthermore, we proceed to show that, in its entirety, the proposed method conforms to the principles of DP.

$$
\begin{aligned}
&\frac{\Pr(M_{Lap}(x,f,\varepsilon)=z)}{\Pr(M_{Lap}(y,f,\varepsilon)=z)} = \frac{\Pr\left(f(x)+Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=z\right)}{\Pr\left(f(y)+Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=z\right)}\\
&= \frac{\Pr\left(Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=z-f(x)\right)}{\Pr\left(Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=z-f(y)\right)}\\
&= \frac{\frac{1}{2b}\exp(\frac{-|z-f(x)|}{b})}{\frac{1}{2b}\exp(\frac{-|z-f(y)|}{b})}\\
&= \exp\left(\frac{|z-f(y)|-|z-f(x)|}{b}\right)\\
&\le \exp\left(\frac{|f(y)-f(x)|}{b}\right) \le \exp\left(\frac{\Delta f}{b}\right)\\
&\le \exp(\varepsilon)
\end{aligned}
\tag{3}
$$

In the equation 4, we argue that the subtraction of random noise drawn from the Laplace distribution also satisfies DP, since $-|-(z-f(x))|$ and $-|-(z-f(y))|$ equal to the $-|z-f(x)|$ and $-|z-f(y)|$ respectively due to the absolute value principle.

$$
\begin{aligned}
&\frac{\Pr(M_{Lap}(x,f,\varepsilon)=z)}{\Pr(M_{Lap}(y,f,\varepsilon)=z)} = \frac{\Pr\left(f(x)-Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=z\right)}{\Pr\left(f(y)-Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=z\right)}\\
&= \frac{\Pr\left(Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=-(z-f(x))\right)}{\Pr\left(Lap\left(0,\frac{\Delta f}{\varepsilon}\right)=-(z-f(y))\right)}\\
&= \frac{\frac{1}{2b}\exp(\frac{-|-(z-f(x))|}{b})}{\frac{1}{2b}\exp(\frac{-|-(z-f(y))|}{b})}\\
&= \exp\left(\frac{|z-f(y)|-|z-f(x)|}{b}\right)\\
&\le \exp\left(\frac{|f(y)-f(x)|}{b}\right) \le \exp\left(\frac{\Delta f}{b}\right)\\
&\le \exp(\varepsilon)
\end{aligned}
\tag{4}
$$

Additionally, in accordance with the composition property introduced in Section II, the sequential application of DP to the data satisfies (ε1 + ε2)-DP. The additive DP phase complies with ε1-DP. The subsequent reductive DP phase aims to remove the learned noise distribution, which is close to ε1, but denoted as ε2. Since both phases are applied consecutively, they satisfy the sequential composition property of DP.

Ⅳ. Experiments

To evaluate the proposed method, we compared the performance of the ML model in a cloud server trained with data generated by the two-fold DP method with the performance of the ML model trained with original data.

4.1 Experimental Environment

We evaluate the proposed method by running several experiments on the environment with the following features: - Windows 10, AMD Ryzen 5 3600 6-Core Processor,16 Gb RAM, Python-3.8, and Jupyter Notebook. As an input dataset, we used the Bank Loan Prediction dataset [16] which is a real-life tabular dataset mainly with numeric features. Since the dataset is quite imbalanced, we apply the synthetic oversampling method SMOTE[17] to balance out the classes.

4.2 Implementation

As an Additive DP mechanism, we implemented the function that injects noise element-wisely to the original data, leveraging the *numpy()* library to generate Laplace-distributed random noise, parametrized by the ε and sensitivity *Δf*. Here, the *Δf* is set to 1 in all cases for evaluation convenience.

As a Reductive DP mechanism, we implemented the DAE, whose architecture consists of an encoder and a decoder, each with multiple hidden layers of 16 nodes activated by the ReLU function. Also, batch normalization layers are incorporated to enhance model stability and training speed into both encoder and decoder modules. The encoder compresses the multi-dimensional input data into a one-dimensional latent space, and the decoder subsequently reconstructs it, trying to minimize the loss. The loss function is configured as the mean squared error between the original data and the reconstructed data, and the objective function is to minimize the loss.

During the training phase, the model processes both noisy (differentially private) and original clean data. In an ideal scenario, the loss would reach zero, indicating that the reconstructed data is perfectly identical to the original. However, given the randomness of the added noise, a perfect reconstruction is unattainable. Despite this, the network is capable of approximating and mitigating the noise distribution. This is evidenced by the achieved training mean squared error of 15.209 and the validation mean squared error of 11.222 between the reconstructed and original data.

Figure 4 represents the change of the mean squared error during the training process for different epochs. From this figure, it is evident that the mean squared error decreases significantly around the 1000th epoch.

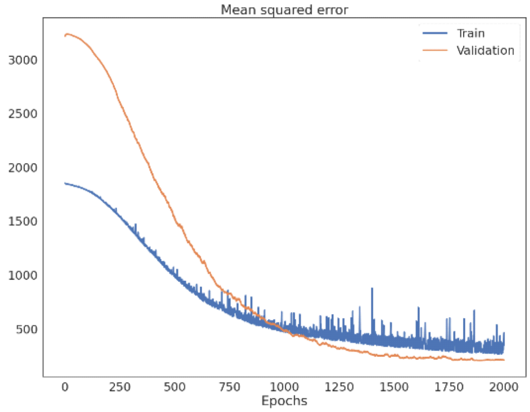


Fig. 4. MSE of the DAE

4.3 Performance Evaluation

To evaluate the effectiveness of the proposed two-fold method, we compare the performance of multiple ML models trained on original data, differentially private data generated by the addition of straightforward Laplace noise, and data generated by the proposed method.

We trained and evaluated 3 different ML models with original data. Namely, the random forest (RF) model with 50 estimators and maximum depth of 5 with the baseline accuracy of 98.4%, the artificial neural network (ANN) with the baseline accuracy of 94.9%, and the support vector machine (SVM) with the baseline accuracy of 95.3%.

We also compare the performance of the ML models trained on differentially private data generated by the straightforward Laplace mechanism(standalone additive noise) and by the proposed two-fold method for privacy budgets ε equal to 1, and 0.1. The definition of DP implies that the smaller the privacy budget, the less information leakage is allowed, and the larger the noise perturbs the data. Consequently, a decrease in ε value inevitably leads to accuracy degradation.

The accuracy of the standalone Laplace mechanism under ε = 1 was 88.1% for the random forest model, 82.5% for the artificial neural network, and 83.8% for the support vector machine. The accuracy of the proposed two-fold method under ε = 1 was 92.8% for the random forest model, 89.4% for the artificial neural network, and 89.7% for the support vector machine.

Similarly, the performance of the Laplace mechanism under ε = 0.1 was 79.0%, 75.1%, and 78.9% for the random forest model, artificial neural network, and support vector machine, respectively. The accuracy of the proposed methods was 84.5%, 87.2%, and 90.1% for the random forest model, artificial neural network, and support vector machine, respectively.

For the smallest privacy parameter ε = 0.01, the accuracy of models on differentially private data generated by the standalone Laplace mechanism was 64.1% for the random forest model, 59.2% for the artificial neural network, and 62.7% for the support vector machine. The performance of the data generated by the proposed method was 72.5% for the random forest model, 74.2% for the artificial neural network, and 70% for the support vector machine.

Table 1. Accuracy of the ML models

	RF	ANN	SVM
Baseline	98.4%	94.9%	95.3%
Laplace (ε = 0.01)	64.1%	59.2%	62.7%
Laplace (ε = 0.1)	79.0%	75.1%	78.9%
Laplace (ε = 1)	88.1%	82.5%	83.8%
Proposed (ε = 0.01)	**72.5%**	**74.2%**	**70.0%**
Proposed (ε = 0.1)	**84.5%**	**87.2%**	**90.1%**
Proposed (ε = 1)	**92.8%**	**89.4%**	**89.7%**

Table 1 summarizes the accuracy of the three distinct ML models under the straightforward Laplace mechanism and proposed two-fold DP mechanisms. As we can observe from the table, overall, the proposed method's performance outperforms the Laplace mechanism's performance across all ML models and ε values, highlighting its efficiency in comparison. The proposed method demonstrates the average performance improvement of approximately 10.23%, 9.6%, and 5.83% for the ε values 0.01, 0.1 and 1. This method effectively mitigates privacy and utility trade-offs, delivering superior performance even with smaller ε values.

Ⅴ. Conclusion

In this paper, we introduce a novel approach for generating differentially private tabular data, with the aim of preserving both accuracy and utility. Our method uniquely combines traditional additive differential privacy mechanisms with a new reductive differential privacy strategy. This latter approach employs a denoising autoencoder to approximate the original data distribution. Experimental results demonstrate that the data generated through our method achieves high levels of accuracy. The proposed method demonstrates the average performance improvement of approximately 10.23%, 9.6%, and 5.83% for the ε values 0.01, 0.1, and 1.

Overall, our proposed methodology represents a promising avenue for future research and development in the realm of privacy-preserving data analytics. Despite these advances, much work remains to be done to enhance the quality and efficiency of differentially private data generation techniques.

References

[1] R. Shokri, M. Stronati, C. Song, and V.

Shmatikov, “Membership inference attacks against machine learning models,” in *Proc. IEEE Symp. Secur. and Privacy*, pp. 3-18, 2017. (https://doi.org/10.1109/SP.2017.41).

[2] M. Fredrikson, S. Jha, and T. Ristenpart, “Model inversion attacks that exploit confidence information and basic countermeasures,” in *Proc. 22nd ACM SIGSAC Conf. Comput. and Commun. Secur.,* pp. 1322-1333, Oct. 2015. (https://doi.org/10.1145/2810103.2813677)

[3] D. Yoo, J. S. Kim, J. Seo, Y. Kang, S.-J. Hwang, and Y.-H. Choi, “A study case for optimizing model inversion attack for prediction systems,” in *Proc. KSIS (KCC 2023)*, pp. 1222-1224, Jun. 2023.

[4] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, “Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing,” in *23rd USENIX Security 14*, pp. 17-32, 2014. (ISBN 978-1-931971-15-7)

[5] B. Z. H. Zhao, A. Agrawal, C. Coburn, H. J. Asghar, R. Bhaskar, M. A. Kaafar, D. Webb, and P. Dickinson, “On the (in) feasibility of attribute inference attacks on machine learning models,” in *2021 IEEE EuroS&P*, pp. 232-251, 2021. (https://doi.org/10.1109/EuroSP51992.2021.00025)

[6] A. K. Khan and H. J. Mahanta, “Side channel attacks and their mitigation techniques,” in *2014 First Int. Conf. ACES*, pp. 1-4, 2014. (https://doi.org/10.1109/ACES.2014.6807983)

[7] G. Nath Nayak and S. Ghosh Samaddar, “Different flavours of man in-the-middle attack, consequences and feasible solutions,” in *2010 3rd Int. Conf. Comput. Sci. and Inf. Technol.*, vol. 5, pp. 491-495, 2010. (https://doi.org/10.1109/ICCSIT.2010.5563900)

[8] C. Dwork, A. Roth, et al., “The algorithmic foundations of differential privacy,” *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3-4, pp. 211-407, 2014. (https://doi.org/10.1561/0400000042)

[9] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, “Differentially private generative adversarial network,” *arXiv preprint arXiv:1802.06739*, 2018.

[10] J. Jordon, J. Yoon, and M. Van Der Schaar, “Pate-gan: Generating synthetic data with differential privacy guarantees,” in *Int. Conf. Learning Representations,* 2018. (https://openreview.net/forum?id=S1zk9iRqF7)

[11] H. Ping, J. Stoyanovich, and B. Howe, “Datasynthesizer: Privacypreserving synthetic datasets,” in *Proc. 29th Int. Conf. Scientific and Statistical Database Manag.,* pp. 1-5, Jun. 2017. (https://doi.org/10.1145/3085504.3091117)

[12] P. Vincent, H. Larochelle, Y. Bengio, and P. A. Manzagol, “Extracting and composing robust features with denoising autoencoders,” in *Proc. 25th Int. Conf. Mach. Learn.,* pp. 1096-1103, Jul. 2008. (https://doi.org/10.1145/1390156.1390294)

[13] A. Ashfahani, M. Pratama, E. Lughofer, and Y. S. Ong, “DEVDAN: Deep evolving denoising autoencoder,” *Neurocomputing,* vol. 390, pp. 297-314, 2020. (https://doi.org/10.1016/j.neucom.2019.07.106)

[14] P. Xiong, H. Wang, M. Liu, and X. Liu, “Denoising autoencoder for eletrocardiogram signal enhancement,” *J. Medical Imaging and Health Informatics,* vol. 5, no. 8, pp. 1804-1810, 2015. (https://doi.org/10.1166/jmihi.2015.1649)

[15] T. Sattarov, D. Herurkar, and J. Hees, “Explaining anomalies using denoising autoencoders for financial tabular data,” *arXiv preprint arXiv:2209.10658*, 2022.

[16] A. Datta, “Personal loan modeling,” Accessed on Mar. 15, 2022. (https://www.kaggle.com/datasets/teertha/personal-loan-modeling.)

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artificial Intelligence Res.*, vol. 16, pp. 321-357, 2002.

Utaliyeva Assem

Feb. 2020 : B.S. degree, Pusan National University

Feb. 2022 : M.S. degree, Pusan National University

Mar. 2022~Current : Ph.D. candidate, Pusan National University

<Research Interests> Differential Privacy, Privacy-Preserving Deep Learning, Security

[ORCID:0000-0002-4271-3944]

Yoon-Ho Choi

2004 : M.S. degree, Seoul National University

2008 : Ph.D. degree, Seoul National University

2009 : Postdoctoral Scholar at Pennsylvania State University, University Park, PA, USA.

2010~2012 : Senior Engineer at Samsung Electronics.

2012~2014 : Assistant Professor in the Department of Convergence Security at Kyonggi University, Suwon, South Korea.

2014~Present : Professor in the School of Computer Science and Engineering at Pusan National University, Busan, South Korea.

<Research Interests> Privacy-Preserving Deep Learning, Adversarial Examples, Anomaly Detection, Deep Packet Inspection for Intrusion detection, IoT Security

[ORCID:0000-0002-3556-5082]