Pacific Symposium on Biocomputing 25:563-574(2020)

AnomiGAN: Generative Adversarial Networks for Anonymizing Private Medical Data

Ho Bae [1], Dahuin Jung [2], Hyun-Soo Choi [2], and Sungroh Yoon [1, 2, 3, 4, 5, *]

[1] *Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul 08826, Korea*
[2] *Electrical and Computer Engineering, Seoul National University, Seoul 08826, Korea*
[3] *Biological Sciences, Seoul National University, Seoul 08826, Korea*
[4] *ASRI and INMC, Seoul National University, Seoul 08826, Korea*
[5] *Institute of Engineering Research, Seoul National University, Seoul 08826, Korea*
[*] *E-mail: sryoon@snu.ac.kr*

Typical personal medical data contains sensitive information about individuals. Storing or sharing the personal medical data is thus often risky. For example, a short DNA sequence can provide information that can identify not only an individual, but also his or her relatives. Nonetheless, most countries and researchers agree on the necessity of collecting personal medical data. This stems from the fact that medical data, including genomic data, are an indispensable resource for further research and development regarding disease prevention and treatment. To prevent personal medical data from being misused, techniques to reliably preserve sensitive information should be developed for real world applications. In this paper, we propose a framework called anonymized generative adversarial networks (AnomiGAN), to preserve the privacy of personal medical data, while also maintaining high prediction performance. We compared our method to state-of-the-art techniques and observed that our method preserves the same level of privacy as differential privacy (DP) and provides better prediction results. We also observed that there is a trade-off between privacy and prediction results that depends on the degree of preservation of the original data. Here, we provide a mathematical overview of our proposed model and demonstrate its validation using UCI machine learning repository datasets in order to highlight its utility in practice. The code is available at https://github.com/hobae/AnomiGAN/

Keywords: Deep neural networks, generative adversarial networks, anonymization, differential privacy

1. Introduction

To restrain the use of medical data for illegal practices, the right to privacy has been introduced and is being adaptively amended. The right to privacy of medical data should be enforced because medical data contains static sensitive information of all individuals including genetic information; therefore, a leak of such irreversible information could be very dangerous. For example, Homer et al.[1] and Zerhouni et al.[2] proposed a statistical-based attacks to GWAS demonstrating the possibility of reviling the presence of an individual in a group. The genetic markers (short DNA sequences) of an individual constitutes a very sensitive piece of information regarding their identity. Patterns of genetic markers can easily be used to identify

individuals and their relatives. If proper security of genetic information is not achieved, there could be a risk of genetic discrimination such as denial of insurance or blackmail (e.g., planting fake evidence at crime scenes).[3] To protect the risk of illegal access to genetic information, the Global Initiative on Asthma (GINA) was launched in 1995 in the United States. Nonetheless, as GINA has not been implemented in other countries, their citizens are still at risk of the issues related to the bias based on leaked genetic information.

The advent of next-generation sequencing technology has led to the progress of DNA sequencing at an unprecedented rate, thereby enabling significant scientific achievements.[4] Using information gathered from the Human Genome Project, international efforts have been made to identify the hereditary components of the diseases, which will allow their earlier detection and more effective treatment strategies.[5] Thus, data sharing among medical institutions is essential for the development of novel treatments for rare genetic diseases and seamless progress in genomic research largely depends on the ability to share data among different institutions.[6] Patient portals and telehealth programs have recently gained popularity among patients allowing them to interact with their healthcare service using online tools.[7] Although these online health services provide convenience by allowing patients to order prescriptions remotely, they also require patients to transmit their private data over the Internet. Most health services follow the guidelines of the Accountability Act of 1996 (HIPPA[a]) to protect patient records, but these guidelines may not be upheld when data are shared with a third party.

Development of deep learning (DL) algorithms has transformed the solution of data-driven problems for various applications, including problems associated with the use of large amounts of patient data for health prediction services.[8] Since patient data are private, several studies have been conducted to resolve privacy issues for DL based applications. The two main approaches involved are: 1) encryption and 2) statistics-based anonymization. Most encryption techniques based on DL methods[9–11] exploits homomorphic properties that enables the computation of encrypted data via simple operations such as summation and multiplication. DL approaches based on homomorphic encryption allow the reliable sharing of private data, while providing accurate results, but a single query can takes hundreds of seconds to be processed.[12] In addition, the nature of homomorphic encryption allows limited compatibility with artificial intelligence techniques such as neural networks.[13] Differential privacy (DP)[14] is a state-of-the-art method that guarantees strong privacy for statistics-based approaches.[1,15–18] In addition, DP has been widely used for deep learning and has recently been applied to medical data. For example, DP generative adversarial networks (GANs) framework has been applied to blood pressure data to protect patient privacy.[19] However, DP based approaches have a significant trade-off between privacy and performance of prediction accuracy.

In this paper, we propose a method based on GANs that preserves a level of privacy similar to that provided by DP while achieving a better prediction performance. Our framework is a generic method that exploits any target predictive classifier to preserve the original prediction

[a]The HIPAA states that, by definition linked to an identifiable person, should not be disclosed or made accessible to third parties, in particular, employers, insurance companies, educational institutions, or government agencies, except as required by law or with the separate express consent of the person concerned.

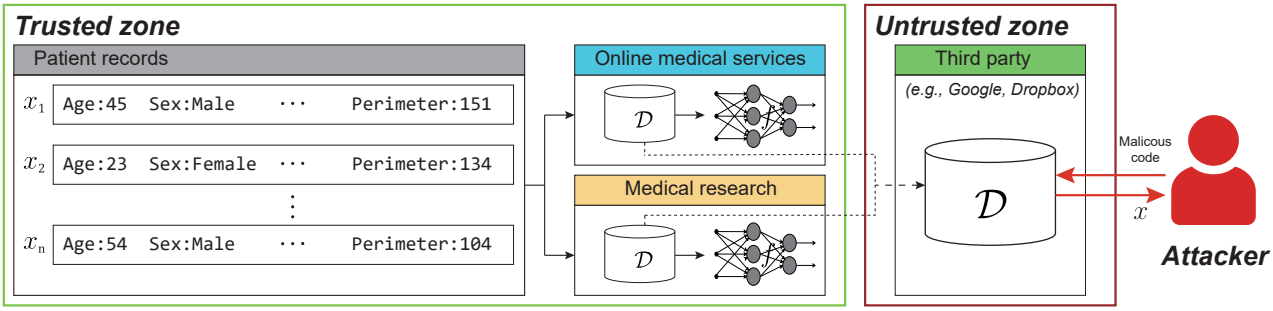

Fig. 1: A trusted zone and an untrusted zones; Patient's medical data are transferred to the online medical service that, in turn, provides diagnostic results to the user. If a user gives consent for data sharing, her or his data may be propagated to third parties (e.g., Google, Dropbox, and Amazon).

result. We explored here, whether a generative model can be constructed to produce meaningful synthetic while also preserving the original predictions and protecting private data. We evaluated the proposed method using target classifiers for four diseases (breast cancer, chronic kidney disease, heart disease, and prostate cancer), and found that its performance was similar to that of the original classifiers. Finally, we compared our method to state-of-the-art privacy techniques and provide a mathematical overview of the privacy parameters.

2. Background

2.1. *Problem Statement*

Fig. 1 shows a scenario in which there is a trusted zone and an untrusted zone. In the trusted zone, patient data are not under any threat from an adversary because service groups in this zone will follow the US health insurance portability and accountability provisions of the 1996 US HIPPA act for the protection of patients records. Groups in the trusted zone, there are groups (online medical services, and medical research services) that follow the above mentioned guidelines[20] may propagate them to a third party if a patient gives consents for data sharing. Service parties that are in the trusted zone may use external storage such as Google, Dropbox, and Amazon. Once given permission, these third parties may no longer uphold the deidentification guideline[20] when interacting with the trusted groups.

The scenario described in the present study, service providers use supervised machine learning classifiers to make predictions based on personal medical data such that a machine learning-based classifier attempts to find a function f that classifies medical data points associated with genetic variants such as tumor data. In this context, AnomiGAN can be used to anonymize personal medical data. Our model can operate even under a strong and realistic black-box scenario in which the classification score revealed from the service providers is in binary rather than a continuous classification score.

2.2. *Adversarial Goal and Capabilities*

It is important to define the capability of an adversary to assess the flaws and the risk to privacy. In our case, the adversary's goal is to compromise an individual's private medical

data. An adversary can be present in online health services, and in third parties that work with medical institutions. An adversary often makes an effort to estimate a posterior probability distribution of query results with the resources available to them including computation power, time, and bandwidth.[3] The probability of success can be quantified by many trials with the adversary's choice of input. This probability distribution can then be used a metric of privacy violation.

2.3. *Differential Privacy*

DP is a state-of-the-art technique for preserving privacy model[21] that guarantees the protection of query results from privacy losses to an adversary. Several methods[22,23] have been proposed based on the following reasoning. DP promises that the probability of harm can be minimized by adding noise to the output of the query as follows:

$$M(\mathcal{D}) = \mathcal{F}(\mathcal{D}) + \tau \tag{1}$$

where $M : \mathcal{D} \rightarrow \mathbb{R}$ is a random function that adds a noise τ to the query output, $\mathcal{D}$ is the target database, and $\mathcal{F}$ is the deterministic original-query response.

Definition 1. (δ-DP). A random algorithm M with domain $\mathbb{N}^{|D|}$ is δ-DP if for all $D, \hat{D} \in \mathbb{N}^{|D|}$ such that $|D - \hat{D}| \leq 1$:

$$\Pr[M(\mathcal{D}) \in S] \leq \exp(\delta)\Pr[M(\mathcal{D}') \in S] \tag{2}$$

where $\mathcal{D}$, and $\hat{\mathcal{D}}$ are the target databases with one element different in $\hat{\mathcal{D}}$; $S \subseteq \text{Range}(M)$ is a subset of $\mathbb{R}$; $M(\mathcal{D})$ and $M(\mathcal{D}')$ are the absolute value of the privacy loss that are bounded by δ with probability of at least $1 - \delta$.[24]

By definition 1, the adversary has no information to gain if an algorithm satisfies δ close to 0. This means that an algorithm with a value of δ that is close to 0 does not reveal significant information on any particular tuple in the input. Privacy is, thus, preserved.

2.4. *Generative Adversarial Networks*

GANs[25] are designed to complement other generative models by introducing a new concept of adversarial learning between a generator and a discriminator instead of maximizing a likelihood. The generator produces real-like samples by transformation function mapping of a prior distribution from the latent space into the data space. The discriminator acts as an adversary to distinguish whether samples produced by the generator derive from the real data distribution. Although theoretically the optimal state of GANs is guaranteed, a fatal limitation of GAN is that its learning is unstable. Therefore, several studies focused on stabilizing GAN's learning by regularization such as weight clipping[26] and gradient penalty.[27,28] As a result, GANs have achieved astonishing results in synthetic image generation,[29,30] and the application of GANs has extended to various fields of studies. For example, GANs have recently been employed in steganalysis[31] and steganography.[32]

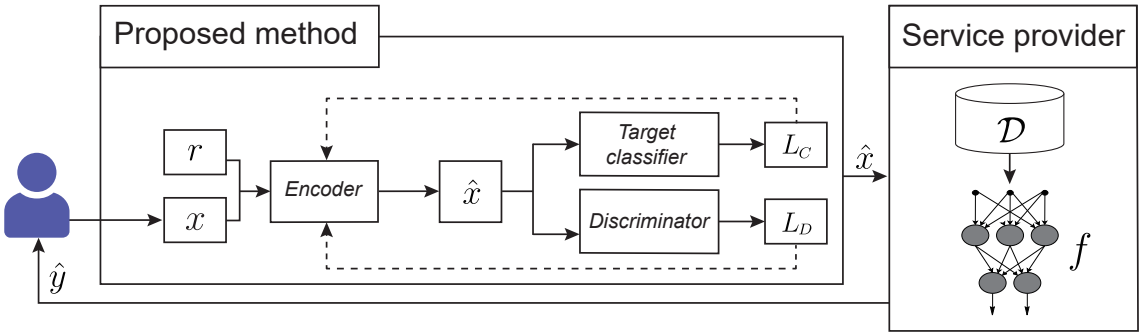

Fig. 2: Architecture of the model presented in this study. The dotted line represents gradients that are fed into the encoder.

3. Methods

Our method involves an encoder, a discriminator, and a target classifier that act as an additional pre-trained discriminator. The encoder generates synthetic data with the aim of mimicking the input data, and the target classifier gives a score to each data item. The discriminator then outputs a confidence score of whether that piece of data is synthetic or original. Starting with random noise, the encoder learns to generate synthetic data such that the prediction result of a synthetic data from a target model is identical to the original data. The optimization of the objective function is equivalent to finding a Nash equilibrium of a min-max game between the generator and the two cooperative relations of the discriminator and the target classifier.

3.1. *Notations*

We will use the following notations: x is the input data; r is the random matrix with same length of the input x; $\hat{x}$ is the anonymized output corresponding to x and r; $\mathbf{M}$ is a trained model; y is the output score given by the trained model given input x, $\mathbf{M}(x) \rightarrow y$; $\hat{y}$ is an output score given by the trained model given input $\hat{x}$, $\mathbf{M}(\hat{x}) \rightarrow \hat{y}$; $\mathcal{A}$ is a probabilistic polynomial-time adversary that queries input to an oracle model; and δ is a privacy parameter that controls privacy levels.

3.2. *Anonymization using GANs*

The architecture of this model is illustrated in Fig. 2. The encoder takes an input x and outputs $\hat{x}$, which is given to both the discriminator and the target classifier. The discriminator outputs the probability L_D that $x = \hat{x}$. The target classifier outputs scores for x and $\hat{x}$ to minimize scores between them. The learning objective of the encoder is to optimize the discriminator's probability to 1/2 while maximizing the prediction score of the target classifier L_C.

The encoder accepts messages of length n as input and r is the n length of random matrix. The input $(r \times x \bmod n)$ is then fed into a neural network. As illustrated in Fig. 3, the first layer of this encoder network consists of n input feature size of filters; the second layer consists of 64 filters; The third layer consists of 32 filters; the fourth layer consists of 16 filters; the fifth layer consists of 8 filters. Additional layers are added in the reverse order of the number

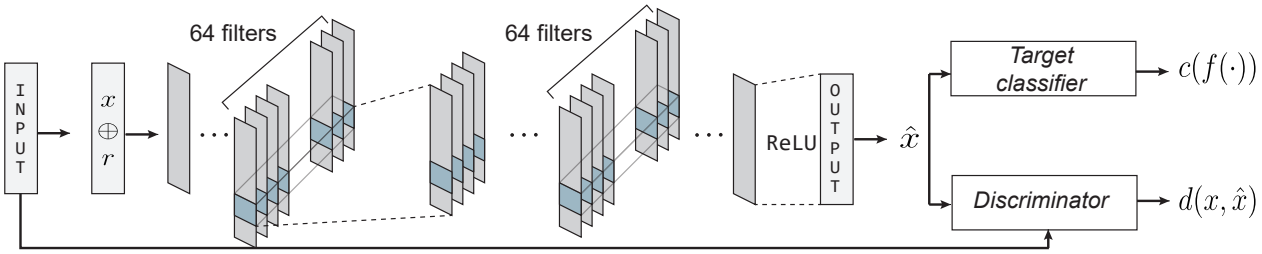

Fig. 3: Model training. The encoder accepts x and r as input and that are fed into the neural network. The discriminator takes an original input and output of the encoder to output probabilities from the last fully connected layer. The target classifier takes an input $\hat{x}$ and outputs the prediction score.

of filters. All layers are constructed with kernel size of 3, strides of 1, and same padding. Batch normalization[33] is used at each layer and tanh[34] is used as the activation function at each layer, except for the final layer where ReLU[35] is used for the activation function. The discriminator takes the output the encoder as an input to determine whether the output is real or generated. A sigmoid activation function is used to output probabilities from the logits.

The discriminator takes an output of the encoder as an input to determine whether the output is real or generated. The first layer of this discriminator network consists of n input feature size of filters; the second layer consists of 10 filters; the third layer consists of 20 filters; the fourth layer consists of 30 filters; the fifth layer consists of n input feature size of filters. The kernel size of each layer is 3 with stride 1. Tanh[34] is used at each layer as the activation, except for the final layer where a sigmoid activation function is used to output probabilities from the logits. The target classifier is a fixed pre-trained model and exploited in the GANs training model. To define the learning objective, let θ_E, θ_D, and θ_C denote parameters of the encoder, discriminator, and target classifier. Let $E(x; r, \theta_E, \delta)$ be the output on x, $C(\hat{x}; \theta_C)$ be the output on $\hat{x}$, and $D(x, E(\theta_E, x, r, \delta), \theta_C; \theta_D)$ be the output on x and $\hat{x}$. Let λ_e and λ_d denote the weight parameters of the encoder, the discriminators to maximize the prediction performance. Let L_E, L_D, L_C denote the loss of encoder, discriminator, and target classifier. The encoder then has the following objective functions:

$$\begin{aligned} \mathrm{L}_E(x, r, \delta; \theta_E) &= \lambda_e \cdot d(x, E(x, r, \delta; \theta_E)) + \lambda_d \cdot (\mathrm{L}_D + \mathrm{L}_C) \\ &= \lambda_e \cdot (d(x, \hat{x}) + \delta) + \lambda_d \cdot (\mathrm{L}_D + \mathrm{L}_C) \end{aligned} \tag{3}$$

where $d(x, \hat{x})$ is the Euclidean distance between synthetic and original data and δ controls the privacy level. δ is updated at each learning epochs. A discriminator has the sigmoid cross entropy loss of:

$$\mathrm{L}_D(\theta_D, \theta_C, x, \hat{x}; \theta_E) = -\, y \cdot \log(D(\theta_D, p)) - (1 - y) \cdot \log(1 - D(\theta_D, p)), \tag{4}$$

where $y = 0$ if $p = \hat{x}$ and $y = 1$ if $p = x$, where p is the score of given input x and $\hat{x}$. A target classifier has a loss of:

$$\mathrm{L}_C(x, \hat{x}; \theta_C) = ||C(f(x)) - C(f(\hat{x}))||_2, \tag{5}$$

where $C(f)$ is a cost function of a pre-defined classifier.

3.3. *Security Principle of Anonymized GANs*

In this section, we show that the AnomiGAN has a scheme that is indistinguishable from real data for an $\mathcal{A}$. For each training steps of the encoder, a multiplicative perturbation by random orthogonal matrices are computed for the inner product matrix. Formally, we have constructed as input entries of the $k \times m$ medical record $x \in X^{k \times m}$, and random matrix $r \in \mathcal{R}^{k \times m}$ is chosen from Gaussian distribution with mean zero variance σ_r^2. Now, assume that the $\mathcal{A}$ has a generated random matrix $\hat{r}$ according to a probability density function. Then the $\mathcal{A}$ need to estimate x given $\hat{x} \leftarrow \mathbf{M}$. A simple intuition of the indistinguishable scheme is that the $\mathcal{A}$ is allowed to choose multiple data from the synthesized data. Then, the $\mathcal{A}$ has the estimation of:

$$\hat{x}_i = \frac{1}{k\sigma_r^2} \sum_t \epsilon_{i,t} x_t, \tag{6}$$

where $\epsilon_{i,j}$ is the i,j-th entry of $\hat{r}^T r$ such that $\epsilon_{i,j} = \sum_t \hat{r}_{t,i} r_{t,j} \forall i,j$. From [Lemma 5.6],[36] it is proven that $\epsilon_{i,j}$ is approximately Gaussian, $\mathrm{E}[\epsilon_{i,j}] = 0$, $Var[\epsilon_{i,j}] = k\sigma_r^4, \forall i,j, i \neq j$. Thus, the expectation of $\mathrm{E}[\hat{x}_i]$ is $\mathrm{E}[\hat{x}_i] = \mathrm{E}[\frac{1}{k\sigma_r^2} \sum_t \epsilon_{i,t} x_t] = 0$, and the variance of $\hat{x}_i$ is $\frac{1}{k} \sum_t x_t^2$. The random matrix r is replaced for each iteration epoch. The variances of each layer are stored during the learning process. Among the stored variances, the randomly selected variance is added to the corresponding layer in inference time to ensure that the encoder does not produce the same output from the same input. Intuitively, AnomiGAN is a probabilistic model; thus $\mathbf{M}$ appears completely random to an $\mathcal{A}$ who observes a medical record $\hat{x}$.

Note that the discussion below is based on the assumption that r is given to the $\mathcal{A}$. However, in our scenario, r is owned by the data owner and an $\mathcal{A}$ has no access to the r, which makes the process even more complex than in the below settings.

Theorem 1. *If* $\mathbf{M}$ *is a probabilistic model and* r *is a random matrix from Gaussian distribution, then* $\mathbf{M}$ *has a scheme that is indistinguishable from real data to an* $\mathcal{A}$*.*

Proof. The rationale for the proof is that if $\mathbf{M}$ is a probabilistic model and a r is the random matrix of each entry that is independently chosen from a Gaussian distribution with mean zero variance σ_r^2; then the resulting scheme is identical to the random projection scheme.[36] Let $\mathcal{A}$ constructs a distinguisher for $\mathbf{M}$. The distinguisher is given an input r, and the goal is to determine whether r is a truly random or r is generated by $\mathbf{M}$. The distinguisher has two observations. If input r is truly random, then the distinguisher has a success probability of $\frac{1}{2}$. If input r is generated by $\mathbf{M}$, then the distinguisher has a success probability of

$$\begin{aligned} \Pr[\mathbf{M} \text{ of success}] &\leq \frac{1}{2} + \frac{1}{k\sigma_r^4} \\ &\leq \frac{1}{2} + \frac{1}{2^l k\sigma_r^4} \end{aligned} \tag{7}$$

where l is the layer number of the encoder. Thus, $\mathcal{A}$ has the success probability as defined in Equation 7. □

Table 1: Performance results of the model upon adding variance to the layer.

Layer	1	2	3	4	5	6	7	8	9	10
Breast Cancer										
Correlation coefficient	0.783	0.793	0.799	0.795	0.829	0.802	0.810	0.780	0.788	0.803
Accuracy (%)	93.18	91.81	93.18	95.45	94.09	95.45	95.73	95.45	95.45	95.45
AUPR	0.991	0.985	0.997	0.995	0.925	0.965	0.981	0.987	0.995	0.991
Chronic Kidney Disease										
Correlation coefficient	0.727	0.766	0.770	0.745	0.775	0.756	0.775	0.760	0.785	0.767
Accuracy (%)	92.00	90.00	92.00	90.00	92.00	94.00	94.00	90.00	90.00	92.00
AUPR (%)	0.828	0.883	0.822	0.871	0.856	0.898	0.880	0.881	0.915	0.913
Heart Disease										
Correlation coefficient	0.856	0.835	0.865	0.845	0.841	0.858	0.854	0.856	0.827	0.851
Accuracy (%)	83.33	80.00	80.00	83.33	86.67	83.33	86.67	86.67	83.33	86.67
AUPR	0.836	0.922	0.853	0.918	0.963	0.927	0.922	0.955	0.924	0.906
Prostate Cancer										
Correlation coefficient	0.379	0.482	0.423	0.419	0.427	0.4340	0.456	0.440	0.479	0.479
Accuracy (%)	69.99	69.99	69.99	69.99	69.99	69.99	69.99	69.99	69.99	69.99
AUPR (%)	0.859	0.804	0.816	0.768	0.797	0.802	0.780	0.810	0.778	0.755

4. Results

4.1. *Datasets*

We simulated our approach using the Wisconsin breast cancer, chronic kidney disease, heart disease, and prostate cancer datasets from the UCI machine learning repository.[37,38] The Wisconsin breast cancer, chronic kidney disease, heart disease and prostate cancer datasets consist of 30, 24, 13, and 8 features, respectively. We carried out five-fold cross-validation with the datasets that were randomly partitioned to training and validation sets of 90% and 10%, respectively.

4.2. *Target Classifiers*

Many services are incorporate disease classifiers using machine learning techniques. For our experiments, we selected breast cancer, chronic kidney disease, heart disease and prostate cancer models from the kaggle competitions as the target classifiers. The classifiers were used as a black-box access to our target classifier in our method. We selected these classifiers for two reasons: a) both classifiers achieve high accuracy in disease detection in their testing datasets, and b) these classifiers are open source implementations, which allows them to be easily accessed as our target classifiers.

4.3. *Model Training*

For the training model, we used Adam[39] optimizer for multi-class loss function with a learning rate of 0.001, a beta rate of 0.5, the epoch of 50000, and mini-batch size of 10. The objective function $\mathcal{L}_E$ was minimized as described in Eq (3). Most of these parameters and the networks structure were experimentally determined to achieved optimal performance. The discriminator achieves the optimal loss after 3000 epochs, whereas the encoder required 5000 epochs to generate synthesize data similar to original sample.

4.4. *Evaluation Process*

We exploited DP, in particular, the Laplacian mechanism,[40] to compare the anonymization performance against the corresponding accuracy and area under the precision recall (AUPR). For the evaluation metric, the accuracy and the AUPR were used to measure performance be-

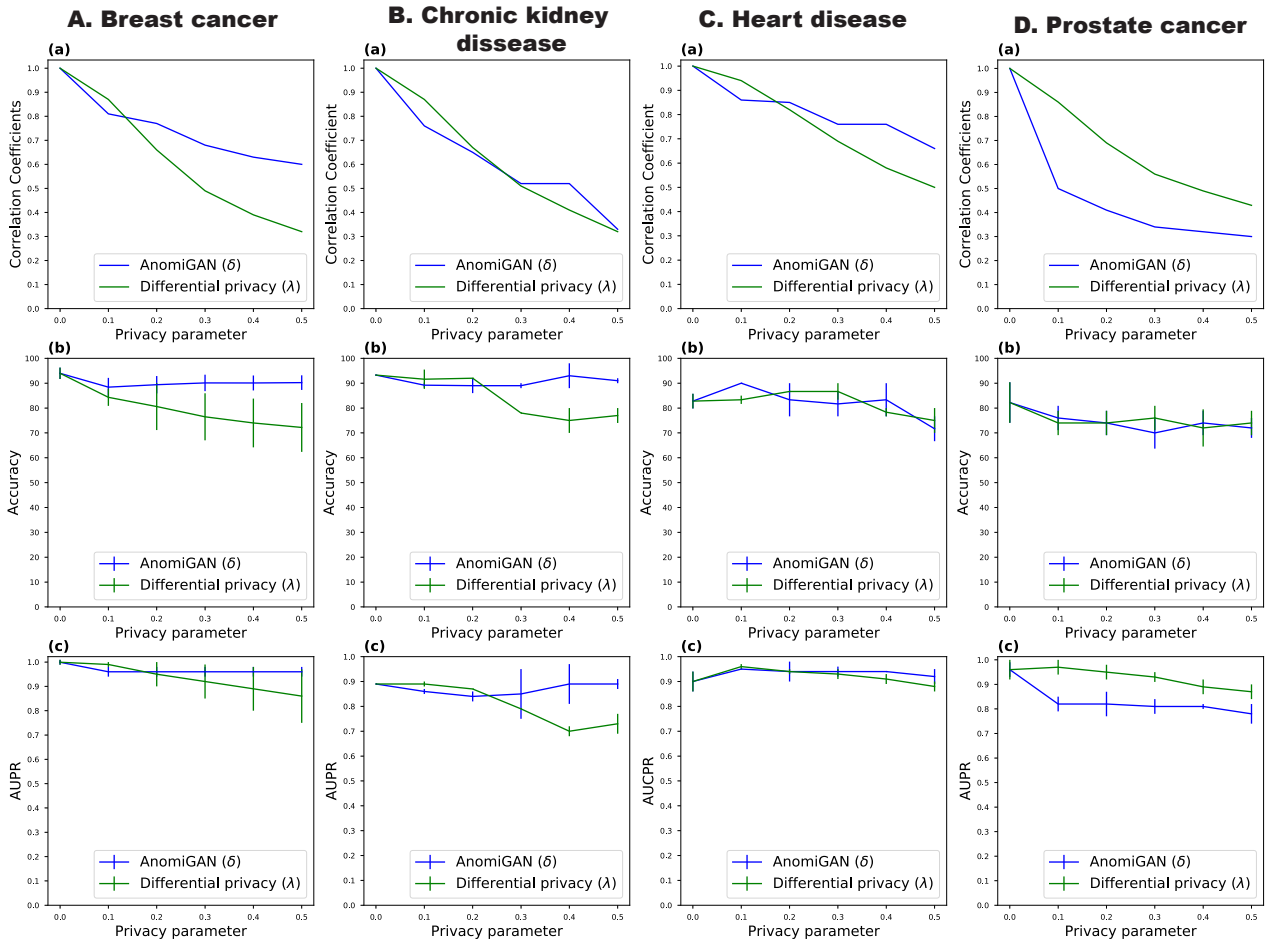

Fig. 4: Anonymization performance using breast cancer, chronic kidney, heart disease, and prostate cancer datasets: A fixed test dataset was selected from the UCI machine learning repository. Correlation coefficient, accuracy, and AUPR were measured by changing 0.1 of the privacy parameter for the fixed test data, δ.

tween original samples and anonymized samples according to the model's parameter changes. The correlation coefficient was used to measure the linear relationship between the original samples and anonymized samples by changing the privacy parameters. We generated the anonymized data according to privacy parameter δ and λ_e by randomly selecting 1,000 cases, and obtained the average prediction of accuracy, AUPR, and correlation coefficient against the corresponding original data. In the next step, we fixed data and generated anonymized data to validate the probabilistic behavior of our model. A variance of each encoder layers was added to the corresponding encoder layers in the inference time. The process was repeated 1,000 times with the fixed test data. We measured the mean of the correlation coefficient, AUPR, and the accuracy for each of the 10 encoder layers as shown in Table 4. The results indicate that adding variance to each of layers influences the correlation coefficient with limited effects on accuracy.

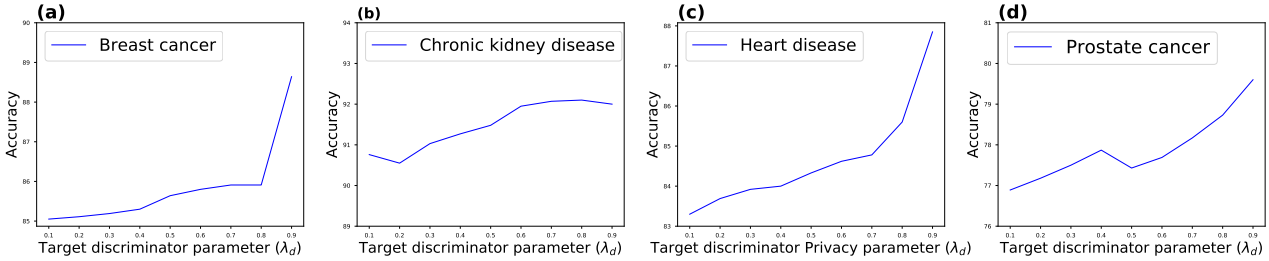

Fig. 5: Comparison of the target discriminator parameter λ_d. Accuracy is measured by changing 0.1 of the λ_d for the fixed privacy parameter δ.

4.5. *Comparison to Differential Privacy*

DP achieves plausible privacy by adding Laplacian noise $\mathbf{Lap}(\lambda) = \mathbf{Lap}(S/\delta)$ to a statistics.[40] The parameter $\lambda = 0.1$ has a minimal effect on privacy and the risk of privacy increases as the parameter λ increases. The amount of noise presents a trade-off between accuracy and privacy. Note that the standard DP of unbounded noise version of Laplacian[41] was applied for the experiments. The experiments were conducted by increasing the parameter λ by 0.1. Note that x-axis (0) in all Fig. 4 represents the prediction scores of original data.

Fig. 4.B shows an experiment for our proposed algorithm and DP algorithm using a fixed chronic kidney disease tests. Both DP and our methodology showed similar performance in the correlation coefficient, but our method showed a better performance in terms of accuracy and AUPR. Fig. 4.A, and 4.C show experimental results for our proposed algorithm and the DP algorithm with respect to the breast cancer and heart disease datasets. Both DP and our proposed method showed similar performance in terms of the correlation coefficient, accuracy, and AUPR. In the case of prostate cancer dataset (Fig. 4.D), our approach shows a better performance in terms of the correlation coefficient and accuracy.

4.6. *Performance Comparison*

We evaluated the performance of our proposed method based on four classifiers (breast cancer, chronic kidney disease, heart disease, and prostate cancer) to measure the prediction performance of the additional discriminator (target classifier). The experiments were conducted by changing 0.1 of the λ_d with the fixed privacy parameter of δ. Because the additional discriminator relies on the original classifier, the performance of the prediction accuracy should increase as λ_d increases. As shown in Fig. 5, the prediction accuracy was increased by a minimum of 2% to a maximum of 6% depending on the datasets. Note that the value of the correlation coefficient value remained constant as the privacy parameter δ was fixed.

5. Discussion

Here, we have introduced a novel approach for anonymizing private data while preserving the original prediction accuracy. We showed that under a certain level of privacy parameters, our approach preserves privacy while maintaining a better performance of accuracy and AUPR compared to the DP. Moreover, we provide a mathematical overview showing that our model is secure against an efficient adversary demonstrate the estimated behavior of the model, and the

performance of our model compared to that of the state-of-the-art privacy preserving method. One of our primary motivations for this study was that many companies are providing new services based on both traditional machine learning and deep neural networks, and we believe this will extend to online medical services. Potential risks regarding the security of medical information (including genomic data) are higher compared to the current risks to private information security, as demonstrated by Facebook's recent privacy scandal.[42] In addition, it is difficult to notice a privacy breach even when there are privacy policies in place. For example, when a patient consents to the use of medical diagnostic techniques, the propagation of that information to a third party cannot guarantee that the same privacy policies will be adhered to by them. Finally, machine learning as a service (MLaaS) is mostly provided by Google, Microsoft, or Amazon owing to hardware constraints, and it is even more challenging to maintain user data privacy when using such services.

Exploiting traditional security in the deep learning requires encryption and decryption phases, which make its use impractical in the real world due to a vast amount of computation complexity. As a result, other privacy preserving techniques such as DP will be exploited in deep learning. Towards this objective, we developed a new approach of privacy-preserving method based on deep learning. Our method is not limited to the medical data. Our framework can be extended in many various ways to the concept of exploiting a target classifier as a discriminator. Unlike a statistics-based approach, our method does not require a background population to achieve good prediction results. AnomiGAN also provides the ability to share data while minimizing privacy risks. We believe that online medical services using the deep neural networks technology will soon be available in our daily lives, and it will no longer be possible to overlook issues regarding the privacy of medical data. We believe that our methodology will encourage the anonymization of personal medical data. As part of future studies, we plan to extend our model to genomic data. The continuous investigation of privacy in medical data will benefit human health and enable the development of various diagnostic tools for early disease detection.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (Ministry of Science and ICT) [2014M3C9A3063541, 2018R1A2B3001628], the Brain Korea 21 Plus Project, and Strategic Initiative for Microbiomes in Agriculture and Food, Ministry of Agriculture, Food and Rural Affairs, Republic of Korea (as part of the multi-ministerial Genome Technology to Business Translation Program, grant number 918013-4).

References

1. N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson and D. W. Craig, *PLoS genetics* **4**, p. e1000167 (2008).
2. E. A. Zerhouni and E. G. Nabel, *Science* **322**, 44 (2008).
3. I. Wagner and D. Eckhoff, *ACM Computing Surveys (CSUR)* **51**, p. 57 (2018).
4. S. C. Schuster, *Nature methods* **5**, p. 16 (2007).

5. F. S. Collins and M. K. Mansoura, *Cancer: Interdisciplinary International Journal of the American Cancer Society* **91**, 221 (2001).
6. B. Oprisanu and E. De Cristofaro, *bioRxiv* , p. 262295 (2018).
7. B. H. Crotty and W. V. Slack, *Israel journal of health policy research* **5**, p. 22 (2016).
8. B. Shickel, P. Tighe, A. Bihorac and P. Rashidi, *arXiv preprint arXiv:1706.03446* (2017).
9. A. Sanyal, M. J. Kusner, A. Gascón and V. Kanade, *arXiv preprint arXiv:1806.03461* (2018).
10. J. Kim, H. Ha, B.-G. Chun, S. Yoon and S. K. Cha, 743 (2016).
11. E. Hesamifard, H. Takabi and M. Ghasemi, *arXiv preprint arXiv:1711.05189* (2017).
12. R. Gilad-Bachrach, N. Dowlin, K. Laine, K. Lauter, M. Naehrig and J. Wernsing, 201 (2016).
13. H. Bae, J. Jang, D. Jung, H. Jang, H. Ha and S. Yoon, *arXiv preprint arXiv:1807.11655* (2018).
14. C. Dwork, Differential privacy, in *Encyclopedia of Cryptography and Security*, (Springer, 2011) pp. 338–340.
15. Y. Erlich and A. Narayanan, *Nature Reviews Genetics* **15**, p. 409 (2014).
16. X. Zhou, B. Peng, Y. F. Li, Y. Chen, H. Tang and X. Wang, 607 (2011).
17. S. Sankararaman, G. Obozinski, M. I. Jordan and E. Halperin, *Nature genetics* **41**, p. 965 (2009).
18. S. Simmons and B. Berger, **2015**, p. 41 (2015).
19. B. K. Beaulieu-Jones, Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd and C. S. Greene, *Circulation: Cardiovascular Quality and Outcomes* **12**, p. e005122 (2019).
20. M. Berhane Russom, Concepts of privacy at the intersection of technology and law, PhD thesis2012.
21. C. Dwork, 1 (2008).
22. M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar and L. Zhang, 308 (2016).
23. R. Shokri and V. Shmatikov, 1310 (2015).
24. C. Dwork, A. Roth *et al.*, *Foundations and Trends® in Theoretical Computer Science* **9**, 211 (2014).
25. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, 2672 (2014).
26. M. Arjovsky, S. Chintala and L. Bottou, **70**, 214 (06–11 Aug 2017).
27. I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, 5767 (2017).
28. L. Mescheder, A. Geiger and S. Nowozin, **80**, 3481 (10–15 Jul 2018).
29. T. Miyato, T. Kataoka, M. Koyama and Y. Yoshida (2018).
30. T. Karras, S. Laine and T. Aila, *arXiv preprint arXiv:1812.04948* (2018).
31. D. Jung, H. Bae, H.-S. Choi and S. Yoon, *arXiv preprint arXiv:1902.11113* (2019).
32. S. Baluja, 2069 (2017).
33. S. Ioffe and C. Szegedy, *arXiv preprint arXiv:1502.03167* (2015).
34. Y. A. LeCun, L. Bottou, G. B. Orr and K.-R. Müller, 9 (2012).
35. V. Nair and G. E. Hinton, 807 (2010).
36. K. Liu, H. Kargupta and J. Ryan, *IEEE Transactions on knowledge and Data Engineering* **18**, 92 (2005).
37. C. Blake, *http://www. ics. uci. edu/˜ mlearn/MLRepository. html* (1998).
38. L. J. Rubini and P. Eswaran, *International Journal of Modern Engineering Research (IJMER)* **5**, 49 (2015).
39. D. Kingma and J. Ba, *arXiv preprint arXiv:1412.6980* (2014).
40. C. Dwork and R. Pottenger, *Journal of the American Medical Informatics Association* **20**, 102 (2013).
41. S. Simmons, B. Berger and C. Sahinalp, **24**, 403 (2019).
42. C. Analytica, *https://en.wikipedia.org/wiki/FacebookCambridge_Analytica_data_scandal* (2018).