Evaluating Privacy on Synthetic Images Obtained using Deep Generative Models

Helena Montenegro [1,2]
maria.h.sampaio@inesctec.pt
Pedro Neto [1,2]
pedro.d.carneiro@inesctec.pt
Cristiano Patrício [2,3]
cristiano.p.patricio@inesctec.pt
Isabel Rio-Torto [2,4]
isabel.riotorto@inesctec.pt
Tiago Gonçalves [1,2]
tiago.f.goncalves@inesctec.pt
Luís F. Teixeira [1,2]
luisft@fe.up.pt

[1] Faculdade de Engenharia da Universidade do Porto
Porto, Portugal
[2] INESC TEC
Porto, Portugal
[3] Universidade da Beira Interior
Covilhã, Portugal
[4] Faculdade de Ciências da Universidade do Porto
Porto, Portugal

Abstract

The generation of synthetic data is often used as a data augmentation technique for training deep learning models. In this work, we investigate whether synthetic medical datasets obtained through generative adversarial networks contain identifiable characteristics of the training data, threatening patient privacy. We propose various methods to classify a set of images as having been used or not used in the training of the model that originated a set of synthetic images. The empirical results support the hypothesis that synthetic data compromises the privacy of patients in the training data and, thus, should be subjected to the same regulations as real data when used in real-world clinical applications.

1 Introduction

Deep learning models have achieved outstanding results in medical image analysis. Nevertheless, these models rely on heavy amounts of data which is often difficult to obtain in clinical settings. As such, the generation of synthetic data has been explored as a data augmentation technique to aid the training of deep learning models. Since deep generative models model the probability distribution of the data, the synthetic images generated using such models may contain identifiable characteristics of patients contained in the training data. Verifying whether synthetic data leaks the identity of patients is of the utmost importance to determine whether this data can be safely used and shared in various real-world applications without compromising patient privacy.

In this work, we verify whether synthetic images can be used to identify patients from the real data through the binary classification task suggested in the GANs task of the medical track of the ImageCLEF Challenge 2023 [1]. Given a set of synthetic images and a set of real images, the task aims to predict which real images were used in the training of the generative adversarial network (GAN) used to obtain the synthetic images. We propose and compare various strategies to classify the images, based on their similarity to the synthetic data, using outlier detection techniques, and making comparisons between their patches [4].

2 Methods

The following subsections describe the methods developed to classify the real images as having been "used" or "not used" in the training of the GAN used to obtain the synthetic images.

2.1 Similarity-based Methods

The similarity-based methods calculate the similarity between real and synthetic images, which is then used to classify the real images. As similarity metrics, we use the Structural Similarity Index Measure (SSIM) [5], and the Euclidean distance between the images' latent representations obtained using a ResNet network [3] pre-trained on ImageNet [2] and the autoencoders described in the next section. The methods used to classify the real images based on their similarity to the synthetic ones are:

- **Threshold**: If the similarity between the real image and any of the generated images is higher than the maximum similarity calculated between real images, then the real image is classified as "used". Otherwise, it is classified as "not used".
- **Retrieval**: For each synthetic image, we retrieve its most similar real image. All retrieved images are classified as "used", while images that were not retrieved and, consequently, are not the most similar to any of the synthetic images, are classified as "not used".
- **Ranking**: We rank the real image according to its similarity to each generated image. Then, we calculate the average ranking of the real image. If the average ranking is higher than the threshold of average ranking calculated only within the real data, the image is classified as "used", otherwise, it is classified as "not used".

2.2 Autoencoder-based Methods

As autoencoder-based methods, we develop the following autoencoders using convolutional neural networks:

- **Autoencoder for Outlier Detection**: is trained only on synthetic images, modelling their probability distribution. On inference, it is applied to the real data to detect outliers, verifying which of the real samples do not follow the probability distribution of the synthetic data. To do so, we measure the reconstruction error of the autoencoder when applied to the real image, such that if this error is higher than the average error obtained on synthetic images plus two times its standard deviation, the sample is considered an outlier. Outliers are classified as "not used", while the real images whose reconstruction error is small and that are, therefore, more similar to the synthetic data, are classified as "used".
- **Autoencoder for Similarity-Based Methods**: is trained on both synthetic and real images and is exclusively used to obtain latent representations for the similarity-based methods.
- **Two Decoder Autoencoder**: contains one encoder and two decoders, one trained on real images and the other on synthetic images, as depicted in Figure 1. Since the encoder is trained simultaneously on real and synthetic data, it is capable of obtaining meaningful latent representations for the similarity-based methods. We also use this model for outlier detection by applying the decoder trained on synthetic images to the real images and measuring its reconstruction error.

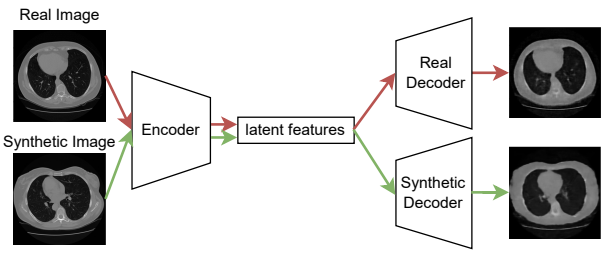

Figure 1: Overview of Two Decoder Autoencoder.

2.3 Patch-based Methods

The patch-based methods operate on patches extracted from the images. We implement the following approaches:

- **Matching Patches**: compares two patches and predicts whether they belong to the same image. To do so, it trains a feature extractor model to obtain latent representations patches, by minimising the Euclidean distance between patches of the same image, while maximising the distance between patches of different images, as depicted in Figure 2. On inference, it compares a patch of a real image with patches of all the synthetic images. If there is a synthetic patch whose distance to the real patch is lower than the maximum distance between patches of the same image, the image is classified as "used". Otherwise, it is classified as "not used".
- **Replacing Patches**: replaces a patch from the real image with a patch from a synthetic image and applies the Autoencoder for Similarity-Based Methods, measuring its reconstruction error. If there is a modified image whose reconstruction error is smaller than the average error on the original data, then its original real image is considered to be similar to the synthetic image from where the patch was extracted and is, therefore, classified as "used".

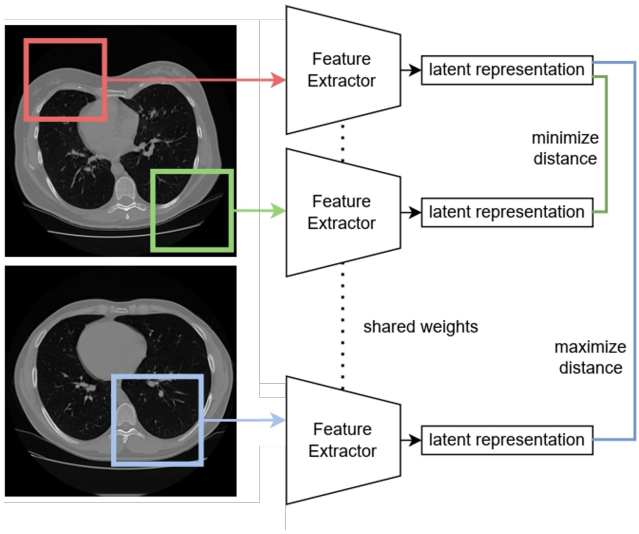

Figure 2: Overview of patch-based method: Matching Patches.

3 Results

The methods were applied to the datasets provided in the GANs task of the medical track of the ImageCLEF Challenge 2023 [1]. The dataset contains axial slices of 3D computed tomography images taken from a dataset of around 8,000 lung tuberculosis patients. Two versions of the dataset are available: a development set with 500 synthetic images and 160 real images, and a test set with 10,000 synthetic images and 200 real images. In both datasets, the proportion of used and not used images in the real data is balanced. Table 1 exposes the accuracy and F1-score obtained by the proposed methods on the development and test datasets.

The results of the methods on the development set differ substantially from the results on the test set. On the development set, the method that achieved the best results was ranking using the distance between latent representations obtained using the basic autoencoder as a similarity metric. In the test set, the highest results were achieved by the threshold method using SSIM as a similarity metric.

Overall, the similarity-based methods outperformed the remaining methods, achieving the best results in both sets. Nevertheless, the results of the outlier detection methods are comparable to the results of some of the similarity-based methods, with the basic autoencoder achieving the second-highest accuracy on the test set. The patch-based methods obtained the worst results, with the method that matches patches being incapable of identifying used images in both sets.

The results support the hypothesis that synthetic data exposes the identity of patients, as it was possible to identify the real images used during the GAN's training with high accuracy and F1-score in both sets.

Method	Accuracy (Dev)	F1-Score (Dev)	Accuracy (Test)	F1-Score (Test)
Similarity-Based Methods				
Threshold (SSIM)	0.675	0.559	**0.810**	**0.802**
Retrieval (SSIM)	0.613	0.687	0.590	0.707
Ranking (SSIM)	0.650	0.650	0.685	0.731
Ranking (ResNet)	0.731	0.711	0.460	0.448
Autoencoder and Similarity-Based Methods				
Threshold (AE)	0.644	0.642	-	-
Retrieval (AE)	0.600	0.674	-	-
Ranking (AE)	**0.850**	**0.846**	0.635	0.621
Threshold (2D AE)	0.606	0.577	-	-
Retrieval (2D AE)	0.550	0.633	-	-
Ranking (2D AE)	0.575	0.575	-	-
Autoencoder-Based Outlier Detection Methods				
Basic AE	0.650	0.582	0.720	0.654
Two Decoder AE	0.613	0.570	-	-
Patch-Based Methods				
Matching Patches	0.525	0.612	0.500	0.514
Replacing Patches	0.644	0.596	0.615	0.594

Table 1: Results on development (Dev) and test sets. AE stands for autoencoder. 2D AE refers to the two decoder autoencoder. The similarity metrics used with each similarity-based method are shown in parentheses.

4 Conclusions

We proposed various methods to classify a set of real images as having been used or not used in the training of the GAN that originated a set of synthetic images. The proposed similarity-based methods were capable of achieving high accuracy and F1-score in this task, confirming the hypothesis that synthetic medical data compromises the privacy of patients.

Future work considers the application of the proposed methods on medical datasets with more variability in the images, to verify whether the similarity-based methods outperform the remaining methods even when there are more pronounced differences between the images. Future work also considers the further development of the proposed models.

To conclude, this paper serves to raise awareness about the privacy risks of using and sharing synthetic images in real-world clinical contexts.

Acknowledgements

This work is financed by National Funds through the FCT - Fundação para a Ciência e a Tecnologia, I.P. (Portuguese Foundation for Science and Technology) within project CAGING, with reference 2022.10486.PTDC, and within PhD grants 2020.06434.BD, 2020.07034.BD, 2021.06872.BD, 2022.11566.BD, 2022.14516.BD.

References

[1] A.-G. Andrei et al. Overview of ImageCLEFmedical GANs 2023 Task – Identifying Training Data "Fingerprints" in Synthetic Biomedical Images Generated by GANs for Medical Image Security. In *CLEF2023 Working Notes*, pages 1305–1315, 2023.

[2] J. Deng et al. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[3] K. He et al. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] H. Montenegro et al. Evaluating Privacy on Synthetic Images Generated using GANs: Contributions of the VCMI Team to ImageCLEFmedical GANs 2023. In *CLEF2023 Working Notes*, CEUR Workshop Proceedings, pages 1596–1610, 2023.

[5] Z. Wang et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612, 2004.