Latent Diffusion Models for Privacy-preserving Medical Case-based Explanations

Filipe Campos[1,2,3,*], Liliana Petrychenko[3], Luís F. Teixeira[1] and Wilson Silva[1,2,3]

[1]INESC TEC, Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
[2]AI Technology for Life, Department of Information and Computing Sciences, Department of Biology, Utrecht University, Utrecht, Netherlands
[3]Department of Radiology, The Netherlands Cancer Institute, Amsterdam, Netherlands

Abstract

Deep-learning techniques can improve the efficiency of medical diagnosis while challenging human experts' accuracy. However, the rationale behind these classifier's decisions is largely opaque, which is dangerous in sensitive applications such as healthcare. Case-based explanations explain the decision process behind these mechanisms by exemplifying similar cases using previous studies from other patients. Yet, these may contain personally identifiable information, which makes them impossible to share without violating patients' privacy rights. Previous works have used GANs to generate anonymous case-based explanations, which had limited visual quality. We solve this issue by employing a latent diffusion model in a three-step procedure: generating a catalogue of synthetic images, removing the images that closely resemble existing patients, and using this anonymous catalogue during an explanation retrieval process. We evaluate the proposed method on the MIMIC-CXR-JPG dataset and achieve explanations that simultaneously have high visual quality, are anonymous, and retain their explanatory value.

Keywords

Privacy-preserving machine learning, medical imaging, case-based explainability, latent-diffusion models

1. Introduction

Various medical imaging techniques such as X-ray or MRI are important to detect and diagnose multiple medical conditions. These imaging modalities provide clinicians insights into patients' health, facilitating accurate and timely diagnoses. In recent years, Deep Learning techniques have shown the capability to rival human performance in diverse diagnosis tasks while being more efficient. However, this leap brings a large challenge – the inherent lack of interpretability in Deep Learning models. While these models exhibit remarkable diagnostic capabilities, their decision-making processes remain largely opaque, posing a barrier to understanding the rationale behind their predictions. This lack of interpretability raises legitimate concerns, especially in critical medical decisions where transparency is necessary for establishing trust.

EXPLIMED - First Workshop on Explainable Artificial Intelligence for the medical domain - 19-20 October 2024, Santiago de Compostela, Spain
*Corresponding author.
filipe.p.campos@inesctec.pt (F. Campos); l.petrychenko@nki.nl (L. Petrychenko); luisft@fe.up.pt (L. F. Teixeira); w.j.dossantossilva@uu.nl (W. Silva)
0009-0006-4753-8846 (F. Campos); 0009-0008-5522-9562 (L. Petrychenko); 0000-0002-4050-7880 (L. F. Teixeira); 0000-0002-4080-9328 (W. Silva)

One way to better understand deep-learning reasoning is through case-based explanations, which explain by example, closely mimicking the rationale of human professionals. Using these mechanisms, a doctor would have access to a prediction and a set of image explanations for each machine-made diagnosis, allowing the professional to compare the current situation with previous historical examples. Ideally, these explanations should be shareable between medical professionals, doctors, and patients or even between hospitals, which is particularly important to build a diverse catalogue of explanations. Yet, this is not possible since medical images contain personally identifiable data, which may allow for the re-identification of patients; therefore, they are protected under strict regulations such as the GDPR [1] and the HIPAA [2]. These issues can be overcome by anonymizing said explanations, therefore protecting patients' privacy while maintaining the utility of the images.

The generation of anonymous case-based explanations using GAN architectures has been previously explored [3], yet the perceptual utility of these explanations is limited by their reduced image quality. We propose to solve this issue by synthesizing and anonymizing images using Latent Diffusion Models [4], which have emerged as a leading approach for image generation tasks in recent years. The proposed methodology builds upon an anonymization method proposed by Packhäuser *et al.* [5] and follows three key processes: generating a synthetic dataset using a latent diffusion model, removing images from the synthetic dataset that closely resemble patients in the training data and retrieving explanations from the newly created anonymous synthetic dataset. The key contributions of this work are the following:

- Generate a synthetic dataset based on the MIMIC-CXR-JPG [6] dataset using a latent diffusion model and anonymize it using a post-model approach.
- Retrieve anonymous case-based explanations with high visual quality and utility.
- Evaluate the proposed solution quantitatively and with an experienced radiologist's aid.

2. Related Work

This work combines three distinct research areas: anonymization techniques, diffusion models and explainability. Although some works connect the bridge between two of these topics, the combination of all three is still largely unexplored.

Diffusion Models Recently, diffusion models such as DDPM [7] have emerged as an alternative to GANs to generate high-quality images while avoiding common difficulties of GANs, such as mode collapse. The diffusion method consists of two distinct processes. During the forward process, Gaussian noise is iteratively added to a training image. In contrast, in the reverse process, a Deep-learning model, typically a U-Net [8], starting from pure noise, predicts what noise was added at each step and iteratively removes it, generating a new image.

Since the diffusion process is computationally expensive to scale to higher resolutions, instead of performing the diffusion process on the image space, Latent Diffusion Models (LDM) [4] perform it on a low-dimensionality latent space. Instead of generating images, during sampling, the diffusion process will generate new latent vectors, which can be decoded back into the image space using a variational autoencoder (VAE) [9]. To better adapt to different medical imaging domains Medfusion [10] modifies the traditional LDM architecture by modifying the

different number of channels in the VAE [9] used to encode and decode latent vectors. While the original LDM architecture employed 4 channels, the Müller-Franzes *et al.* found that, for medical images, using 8 channels led to fewer visual artefacts.

Anonymization Visual anonymization methods can generally be divided into classical and machine-learning-based. Within the classical methods, two of the most common anonymization techniques include blurring the image or K-Same [11], which overlays K images, therefore obtaining K-Anonymity. Yet, these methods require aggressive image modifications, making them difficult to analyse visually. Recently, there has been a focus on generating synthetic images using machine learning methods. These methods can be divided into two sub-groups: differential-privacy methods [12, 13], which provide strong privacy guarantees at the cost of having a reduced image quality, struggling to be scaled beyond 32×32 pixel images due to their computational complexity. The other methods present a more ad-hoc approach; instead of providing strong statistical guarantees, they typically employ identity verification networks [5, 3], which empirically promote privacy.

Case-based Explanations Case-based explanations are a *post hoc* mechanism which justifies a model's decisions by retrieving cases or examples from the training data. This method is analogue to human reasoning, making them easy to interpret. While certain image-retrieval scenarios employ similarity metrics such as the Euclidean distance or the Structural similarity index measure (SSIM) [14], which analyses an image as a whole, for case-based explanations, the focus is set on small localized features which would be overlooked by said measures. One way of retrieving relevant images is by comparing the features obtained by a task-specific Deep-Learning classifier [15]. Additionally, Montenegro *et al.* [3] proposes a GAN architecture which generates anonymous case-based explanations by employing a privacy loss function that increases the distance between an identity embedding calculated for the generated image and the identity embeddings of the remaining training data images. The main shortcoming of this work is the limited image quality of the synthetic explanations caused by said privacy loss function.

3. Method

Our methodology, summarized in Figure 1, follows the anonymization protocol defined by Packhäuser *et al.* [5]. Initially, we train a latent diffusion model and synthesize a dataset. Afterwards, a retrieval model for each image within this dataset will be used to identify which images in the training set closely resemble the synthetic image. Then, we employ an identity verification network to compare the anonymous image with its closest match and determine whether it likely belongs to the same patient. This information allows us to remove non-anonymous images and, consequently, create an anonymous catalogue from which we can retrieve case-based explanations.

3.1. Image Generation using Latent Diffusion Models

We train a latent diffusion model which uses the Medfusion [10] architecture for image generation. Using it, we sample synthetic images to build a synthetic dataset that will be anonymized.

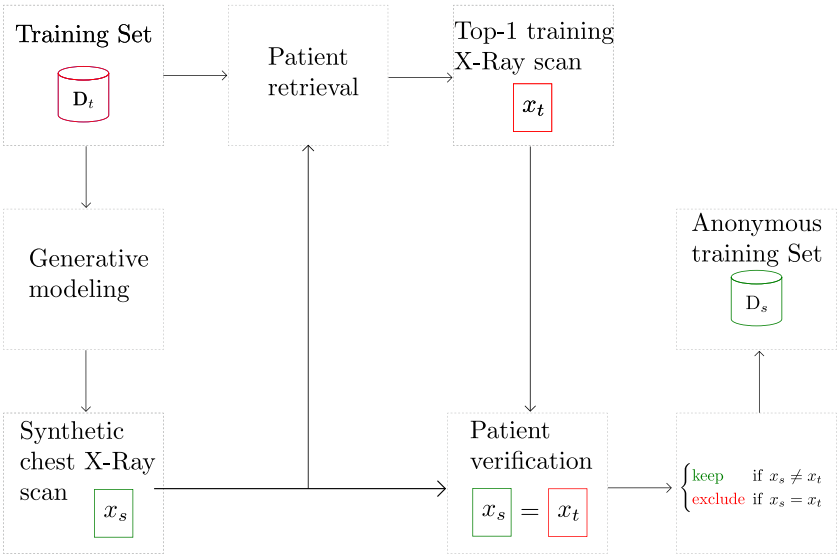

Figure 1: Anonymization Pipeline. A generative model generates synthetic samples based on the real training data. From these samples, we remove those closely resembling patients in the training set based on a patient retrieval and a patient verification model. This anonymous dataset serves as our knowledge base from which to retrieve explanations.

3.2. Retrieval and Verification Network

We must identify if each synthetic image closely resembles any real patient in the training set. To do so, first, a patient retrieval network consists of a Siamese Neural Network (SNN) [16] based on the ResNet-50 [17] architecture, which takes as input two images and, using a contrastive loss function, approximates embeddings so that the embeddings from the same patients are grouped together.

The verification model follows the same architecture, but instead of using a contrastive loss, this network compares both input images and classifies whether or not they belong to the same patient. For both models, the training data consists of positive training pairs obtained based on the patient identity information of a dataset and randomized negative pairs.

3.3. Anonymization Pipeline

Considering the synthetic dataset, the verification model, and the retrieval network, we have all the components required to obtain an anonymous dataset. The anonymization procedure we employ can be broken down into three separate steps:

1. **Compute training identity embeddings**: We obtain an identity embedding for each training set image using the identity retrieval network. These embeddings are stored in a KD-tree [18], used for more efficient queries using a nearest-neighbour strategy later.
2. **Search nearest training image**: For each synthetic sample, we retrieve the top-1 most similar training sample. In this step, we compute the identity embedding of each image

and perform a lookup of the previously mentioned KD-tree. Each real-synthetic image pair is stored in a list used in the next step.

3. **Remove non-anonymous image pairs**: For each real-synthetic image pair, we use the identity verification network to obtain a prediction of whether or not both images belong to the same patient. If this likelihood exceeds a predefined threshold, the synthetic image is deemed non-anonymous and removed from the synthetic dataset.

3.4. Case-based Explanation Retrieval

We aim to retrieve explanations similar in task-related features and not merely structurally similar. We employ a DenseNet121 [19] classifier and use the feature vector present in the last layer, which contains 1024 features, to compare images. The intuition behind this method is that the classifier will represent the most important diagnosis characteristics in its feature space.

During inference, an image is passed through the classifier, and we use the features to obtain the most similar features from a catalogue of synthetic images. The catalogue size will vary depending on the application. In a real-case scenario, we want to have as many images as possible to capture the most variability and obtain images that are as relevant as possible.

4. Experiments

Data We perform experiments on the MIMIC-CXR-JPG dataset [6], consisting of 377,110 chest X-ray images in the JPG format from 65,379 unique patients. We chose to predict cardiomegaly as our classification task since it is a common diagnosis with 66,799 samples. It makes up 29.32% of the available samples and is easily identified visually, making it ideal to be explained by case-based explanations. From the available images, we select the ones from the posteroanterior (PA) view (96,161) since it is the view commonly used by radiology to diagnose Cardiomegaly since the anteroposterior (AP) view tends to magnify the heart [20] hindering the diagnosis. We use the recommended data splits provided alongside the dataset.

Image Generation We generate MIMIC-CXR-JPG images using the Medfusion [10] architecture, the latent embeddings are encoded using a VAE, which accepts $3 \times 512 \times 512$ pixel images and has 8 embedding channels. The diffusion model uses DDIM [21], has label embedding and time embeddings of size 1024 and employs a scaled linear noise scheduler with $T = 1000$, which varies from $\mathrm{B}_0 = 0.002, \mathrm{B}_T = 0.02$. Using a U-Net [8] architecture, we train the model for a maximum of 1001 epochs with early stopping with a patience of 30 epochs. We generate 1000 synthetic images for both datasets, with a balanced split between positive and negative cases.

Retrieval and Verification For the retrieval phase, similarly to Packhäuser *et al.* [5], we train our models for 30 epochs in the first phase, during which the model backbone is frozen, and 50 epochs during the second phase where all parameters are unfrozen. We use a learning rate of 0.158489 with weight decay of $1e^{-5}$ and batch size 32. Our verification model was trained using the Adam optimizer [22] with $3 \times 256 \times 256$ images and a learning rate of 0.0001. We used a batch size of 32 and limited training to a maximum of 100 epochs with an early stopping criteria with 5 epoch patient. During the anonymization step, we considered images non-anonymous if

the verification model prediction of a synthetic image and a real image belonging to the same patient was above a threshold of 0.5.

Explanation Retrieval For explanation retrieval, we employ a DenseNet121 [19] network trained using the Adam optimizer with a learning rate of $1e^{-3}$ up to 10 epochs with an early stopping criteria and patience $p = 3$. To search for similar images, we use the last feature vector of the model, located before the dense layer and sized 1×1024.

5. Results and discussion

Table 1 shows a quantitative image quality evaluation of the synthetic images generated by the Medfusion architecture. We report precision (P), recall (R) and the Fréchet inception distance (FID) [23], which compares the distribution of real and generated images based on the deepest layer of an Inception V3 network. To evaluate the image utility for classification tasks, we compare a classifier trained on synthetic data with a classifier trained on real data in Table 2. The classifier trained on synthetic data still achieves acceptable performance even though it has been trained on a comparatively small dataset. This indicates that the signal required to diagnose Cardiomegaly is still present in the new images.

Table 1
Quantitative evaluation of the images generated using the Medfusion model.

FID	P	R
62.25	68.00	22.17

Table 2
Classification performance reported for both the MIMIX-CXR-JPG and a synthetic dataset.

Dataset	F1	ACC	Image Count
MIMIC-CXR-JPG	91.67	87.33	96,161
Synthetic	75.97	69.00	1,000

Both the verification and retrieval models showcase the ability to re-identify patients, as demonstrated by the metrics highlighted in Tables 3 and 4, respectively. For the verification model, we report the Area under the Receiver operating characteristic (AUC), precision (P), recall (R) and F1-score. For the retrieval task, we track the R-precision, a common information retrieval metric which is calculated as shown in Equation 1 using the number of relevant images retrieved (r) within the first R images, where R is the total number of relevant images existing in the dataset. Finally, the $mAP@R$, shown in Equation 2, is the mean of the average precision at R ($AP@R$) for each query.

$$\text{R-Precision} = \frac{r}{R} \tag{1}$$

$$mAP@R = \frac{1}{Q} \sum_{i=1}^{Q} AP_i@R \tag{2}$$

Table 3
Quantitative evaluation of patient verification model.

AUC	ACC	F1	P	R
93.72	95.32	90.92	90.77	91.07

Table 4
Quantitative evaluation of patient retrieval model.

mAP@R	P@1	R_prec
79.32	94.27	80.87

From the original 1,000 synthetic images, we use the anonymization procedure proposed in Section 3.3 to remove all synthetic images similar to their closest real counterpart based on the predictions made by the verification system. Using this process, 161 images were removed. In Figure 2, we show image pairs at different prediction ranges, and, as expected, the higher-value pairs are more similar than those within the lower ranges.

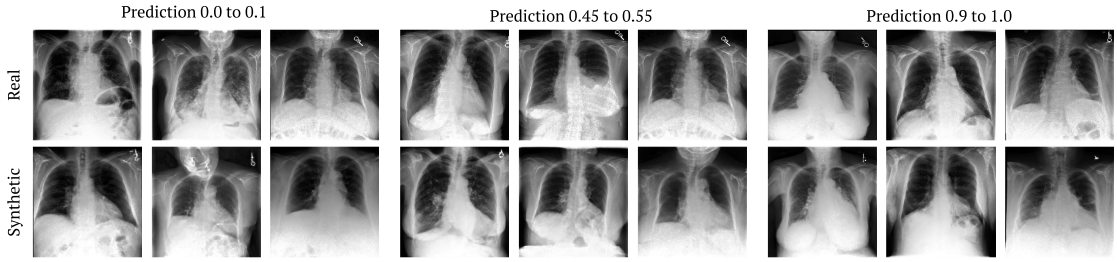

Figure 2: Different image pairs of anonymous images and their closest real counterpart in the training set, sorted by the predicted likelihood of belonging to the same patient. The images on the left are less likely to contain identifiable patient data, while images on the right are highly likely.

Similar to previous studies [15], the explanation ranking is evaluated with the help of a radiologist. For this purpose, we perform two different experiments. For the first experiment, we used 5 test cases, each containing a test image we aim to diagnose and 10 synthetic catalogue images from which we will retrieve explanations. The second experiment differs by using a catalogue of 5 synthetic images and 5 real images to compare the utility of both image types. All the images were randomly sampled from the test set while ensuring a balanced class distribution. The catalogue is limited to 10 images due to the time-consuming nature of evaluating the images.

To evaluate ranking performance, we use a metric commonly used in ranking tasks, the normalized Discounted Cumulative Gain (nDCG_p) [24] metric (Equation 3) where p is the number of retrieved images. The relevance values for each image (rel_i) vary from 5.5, the most similar image, to 1.0, the least similar example, with a 0.5 step between each relevance value. And IDCG_p is the ideal DCG_p value.

$$\text{nDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p} \tag{3}$$

$$\text{DCG}_p = \sum_{i}^{p} \frac{2^{\text{rel}_i} - 1}{\log_2(i+1)} \tag{4}$$

The results of the ranking evaluation are shown in Figure 3. We can notice that our CNN-based method, on average, outperforms its SSIM counterpart. An example of the retrieval test we used to evaluate the solution is shown in Figure 4. Interestingly, the CNN-based retrieval mechanism obtained the same Top-3 images as the expert-based choices, although in a different order.

Additionally, a radiologist accessed each image belonging to the test cases. In Table 5, we see that the generative model can conditionally generate images with a specific diagnosis while maintaining a similar class agreement compared to real images. We can also note that synthetic and real images are similarly relevant, supporting the fact that the generated images can be used as alternatives to real ones.

Finally, this method still presents some limitations. There is no upper bound on the number

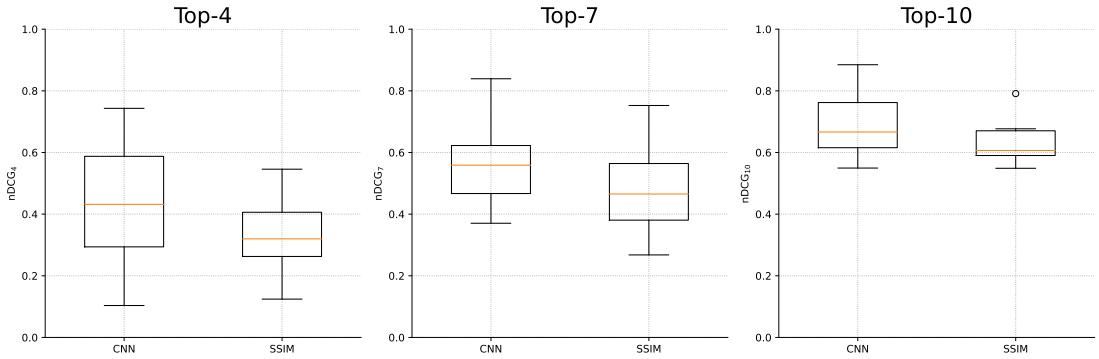

Figure 3: Boxplots regarding nDCG for two retrieval approaches, CNN and SSIM for the Top-4, Top-7 and Top-10 retrieved images.

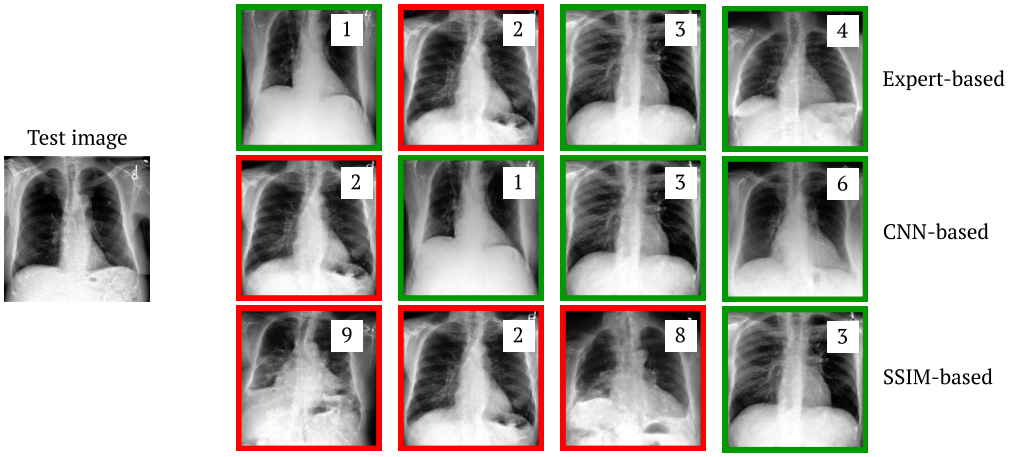

Figure 4: Example test image and corresponding Top-4 catalogue images retrieved by each corresponding method. The color outline around each retrieved image indicates whether the test and catalogue images share the same class (green) or not (red). The ranking annotated on the top-right corner of each image indicates the ground truth based on expert rating.

of synthetic images that will be removed, leading to a waste of computing power on generating images that will effectively be deleted. In future work, the anonymization mechanism could be integrated into the generative model to combat this issue. Another key issue is the empirical nature of this anonymization procedure, making it unproven. Unfortunately, the current best method to provide anonymization, with strong guarantees, is through Differential Privacy, which struggles to scale beyond low-resolution images, making it ineffective in generating explanations which are interpretable by humans.

Table 5
Label agreement between expert evaluation and the ground-truth label. A classification is correct if the image label and the radiologist agree on the diagnosis (Cardiomegaly or No Cardiomegaly). Cases where it is impossible to diagnose a sample confidently are deemed non-conclusive. We also report the average relevance of each type of image for the test cases with mixed image types.

Type	Correct	Incorrect	Non-conclusive	Average Relevance
Synthetic	70.67%	20.00%	9.33%	16.00 ± 4.30
Real	76.00%	20.00%	4.00%	16.50 ± 4.30

6. Conclusion

We proposed a method to generate visually anonymous case-based explanations that retain their utility and realism by levering latent diffusion models. This solution demonstrates that the generated images are empirically unlikely to be traced back to the original patients, allowing for the visual explanation of classifier decisions without exposing patient data. This is a step towards integrating trustworthy and interpretable classifiers in the medical domain while preserving patient privacy.

Acknowledgments

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 the project CAGING, with reference 2022.10486.PTDC (DOI 10.54499/2022.10486.PTDC).

References

[1] Council of European Union, Council regulation (EU) no 679/2016, Online at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:02016R0679-20160504, 2016.

[2] U.S. Department of Health and Human Services, The Health Insurance Portability and Accountability Act of 1996 (HIPAA), Online at http://www.hhs.gov/hipaa/, 1996.

[3] H. Montenegro, W. Silva, J. S. Cardoso, Privacy-preserving generative adversarial network for case-based explainability in medical image analysis, IEEE Access 9 (2021) 148037–148047. doi:10.1109/ACCESS.2021.3124844.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10684–10695.

[5] K. Packhäuser, S. Gündel, N. Münster, C. Syben, V. Christlein, A. Maier, Deep learning-based patient re-identification is able to exploit the biometric nature of medical chest x-ray data, Scientific Reports 12 (2022) 14851. doi:10.1038/s41598-022-19045-3.

[6] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs, 2019. doi:10.48550/arXiv.1901.07042. arXiv:1901.07042.

[7] A. Nichol, P. Dhariwal, Improved Denoising Diffusion Probabilistic Models, 2021. doi:10.48550/arXiv.2102.09672. arXiv:2102.09672.
[8] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, CoRR (2015). arXiv:1505.04597.
[9] D. P. Kingma, M. Welling, Auto-encoding variational bayes, 2022. arXiv:1312.6114.
[10] G. Müller-Franzes, J. M. Niehues, F. Khader, S. T. Arasteh, C. Haarburger, C. Kuhl, T. Wang, T. Han, T. Nolte, S. Nebelung, J. N. Kather, D. Truhn, A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis, Scientific Reports 13 (2023) 12098. doi:10.1038/s41598-023-39278-0.
[11] R. Gross, E. Airoldi, B. Malin, L. Sweeney, Integrating utility into face de-identification, in: G. Danezis, D. Martin (Eds.), Privacy Enhancing Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 227–242.
[12] T. Dockhorn, T. Cao, A. Vahdat, K. Kreis, Differentially Private Diffusion Models, Transactions on Machine Learning Research (2023). URL: https://openreview.net/forum?id=ZPpQk7FJXF.
[13] Z. Chu, J. He, D. Peng, X. Zhang, N. Zhu, Differentially private denoise diffusion probability models, IEEE Access 11 (2023) 108033–108040. doi:10.1109/ACCESS.2023.3315592.
[14] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13 (2004) 600–612.
[15] W. Silva, A. Poellinger, J. S. Cardoso, M. Reyes, Interpretability-guided content-based medical image retrieval, in: A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, L. Joskowicz (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2020, Springer International Publishing, Cham, 2020, pp. 305–314.
[16] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, R. Shah, Signature verification using a "siamese" time delay neural network, in: Proceedings of the 6th International Conference on Neural Information Processing Systems, NIPS'93, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993, p. 737–744.
[17] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). arXiv:1512.03385.
[18] J. L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18 (1975) 509–517. doi:10.1145/361002.361007.
[19] G. Huang, Z. Liu, L. Van Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 2261–2269. doi:10.1109/CVPR.2017.243.
[20] E. Puddy, C. Hill, Interpretation of the chest radiograph, Continuing Education in Anaesthesia Critical Care & Pain 7 (2007) 71–75. doi:10.1093/bjaceaccp/mkm014.
[21] J. Song, C. Meng, S. Ermon, Denoising diffusion implicit models, CoRR abs/2010.02502 (2020). arXiv:2010.02502.
[22] D. P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2017. arXiv:1412.6980.
[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018. arXiv:1706.08500.
[24] K. Fernandes, J. S. Cardoso, Hypothesis transfer learning based on structural model similarity, Neural Computing and Applications 31 (2019) 3417–3430.