DiffExplainer: Unveiling Black Box Models Via Counterfactual Generation

Yingying Fang[1], Shuang Wu[2], Zihao Jin[3], Shiyi Wang[4], Caiwen Xu[4], Simon Walsh[1†], and Guang Yang[1,4,5,6†]

[1] National Heart and Lung Institute, Imperial College London, London, UK
[2] Black Sesame Technologies, Singapore, Singapore
[3] Department of Metabolism, Digestion and Reproduction, Imperial College London, London, UK
[4] Bioengineering Department and Imperial-X, Imperial College London, London, UK
[5] Cardiovascular Research Centre, Royal Brompton Hospital, London, UK
[6] School of Biomedical Engineering & Imaging Sciences, King's College London, London, UK
{s.walsh;g.yang}@imperial.ac.uk

Abstract. In the field of medical imaging, particularly in tasks related to early disease detection and prognosis, understanding the reasoning behind AI model predictions is imperative for assessing their reliability. Conventional explanation methods encounter challenges in identifying decisive features in medical image classifications, especially when discriminative features are subtle or not immediately evident. To address this limitation, we propose an agent model capable of generating counterfactual images that prompt different decisions when plugged into a black box model. By employing this agent model, we can uncover influential image patterns that impact the black model's final predictions. Through our methodology, we efficiently identify features that influence decisions of the deep black box. We validated our approach in the rigorous domain of medical prognosis tasks, showcasing its efficacy and potential to enhance the reliability of deep learning models in medical image classification compared to existing interpretation methods. The code will be publicly available at https://github.com/ayanglab/DiffExplainer.

Keywords: Counterfactual explanation · Diffusion model · Teacher-Student learning · Explainable AI

1 Introduction

In this age where deep learning-based models are increasingly predominant, it is of paramount importance for users to comprehend the rationale behind the decisions made by these models, especially when these models are engaged in high stake applications such as AI-based diagnosis and prognosis. Over the past decade, Explainable AI (XAI) methods for medical image analysis have been

---

[†] S. Walsh and G. Yang are co-senior authors.

developed, aiming to reveal semantically meaningful insights [10,17], which is essential for providing guidance to medical practitioners in making informed decisions or discovering novel biomarkers.

One major class of XAI methods falls under the paradigm of attribution maps. Various methods for computing attribution maps include backpropagation [28,31], activation map [25,14], and perturbation-based[32,21,18], with the common goal of highlighting the relative importance of each pixel to the model prediction. While attribution maps have been relatively useful in identifying salient regions for classification tasks on natural images, their utility has been suboptimal for tasks where different classes share many common features, a situation often seen in image-based prognosis tasks. In these cases, patients may have similar baseline images but different outcomes, and the granularity of traditional attribution maps can hardly pinpoint the human-interpretable differences that lead to different predictions and, therefore, fails to provide the exact reasoning behind their decisions. To address this issue, Rudin et al. proposed an interpretable model that calculates the similarity of query images to predefined disease-outcome-related prototypes and integrates these similarities into the final decision [22,2]. This approach enhances interpretability and reliability substantially by grounding the model's decisions in domain knowledge. However, it requires both extensive prior knowledge and fine annotations, which are not feasible when such knowledge is lacking, and this method may miss predictive features not identified by clinicians. Hence, leveraging model-agnostic explanation methods to detect fine-grained contributive features from deep classifiers offer a promising solution to maintain both interpretability and performance.

Counterfactual explanation, initiated in [30], became an alternative line of the traditional model-agnostic interpretation methods. Essentially, this approach generates a variant of the original input, *i.e.* the counterfactual, which alters the model's prediction, serving to pinpoint the crucial features that are responsible for this change. Notably, the means for generating the counterfactual examples are greatly facilitated by the parallel developments of the Generative Adversarial Networks (GANs) [7,8]. There has been an emergence of GAN-based counterfactual explanation methods in medical classification tasks across various modalities, including X-Ray [1,15,26,24,23], Magnetic Resonance Imaging (MRI) [29,6] and ultrasound images [20] and histopathology images [12,24].

While making significant progress in generating realistic counterfactual images, existing counterfactual generation models still face three major limitations that hinder their widespread adoption for explaining various classifiers. Firstly, training GAN models is notoriously difficult, often requiring meticulous tuning of hyperparameters to ensure stable training [13,4,6]. Secondly, explaining black-box decisions requires the model to have the ability to reconstruct the original query image without any manipulation applied. However, GAN-based models often struggle to accurately reconstruct original images with rich textures, resulting in ineffective counterfactuals where only certain features are altered from the original images with more complex contents [24] and impeding their adoption in more complex modalities such as histopathology and CT imaging. Thirdly,

most counterfactual generation models require training with inputs from the black-box classifiers they aim to explain [1]. This requirement significantly increases training costs and complicates the application of these methods.

In this work, we develop a novel counterfactual method to address the aforementioned limitations, leveraging diffusion-based generation [9,27] which has surpassed GANs for many generative tasks [5,16]. We aim to achieve a more stable training procedure, higher-quality, controllable counterfactual image generation performance and faster development of explanation models . Our main contributions can be summarized as: (1). The proposed **DiffExplainer**, combines teacher-student learning and Diffusion Autoencoders [19] for generating counterfactual images to decode a given black box model. (2). Compared to traditional attribution maps, DiffExplainer can accurately identify fine-grained features underlying a black box's decision for any given query image. Even for instances with indeterminate classification results - typically a significant challenge for existing methods - our method can provide robust and coherent explanations. (3). Compared to the existing counterfactual models, we are also the first to develop counterfactual methods specifically for CT modalities and enable the manipulation of lossless reconstructed images. Moreover, DiffExplainer is trained independently of the classifier, allowing for its direct adoption with different classifiers using the same modalities. This feature may facilitate its wider adoption in various tasks in the future.

2   Method

An overview of our DiffExplainer framework is presented in Fig. 1 (A), which showcases elucidating the operations of a black box pretrained classifier for lung disease prognosis. The input of this classifier, which we hereafter refer to as the teacher model, consists of four slices extracted from a CT scan of a patient. Our DiffExplainer, crafted to discern correlated features for model's decisions, comprises three sequentially applied steps with three sub-models. First, an encoder model maps each slice to a latent representation, which are then combined into a latent feature. Here, the encoder is trained as part of a Diffusion autoencoder model, so that the latent feature therefore contains sufficient information for reconstruction. The feature vector is then sent into an agent model, which performs knowledge distillation and arrives at a final feature representation that is aligned with that of the black box teacher model, hence predicting similar classification scores. This shallow student model is composed of two-layer linear model and will provide the manipulation direction on the latent feature to increase/decrease the score in a target class. In the second stage, we generate counterfactuals from the manipulated features and reconstruct the original images from the original feature from the diffusion model. Lastly, we can generate the difference heatmap from the counterfactual images and the reconstructed images as well as their scores in the black box. The difference maps will give clear indications of the image regions that affect the predictions of the black box.
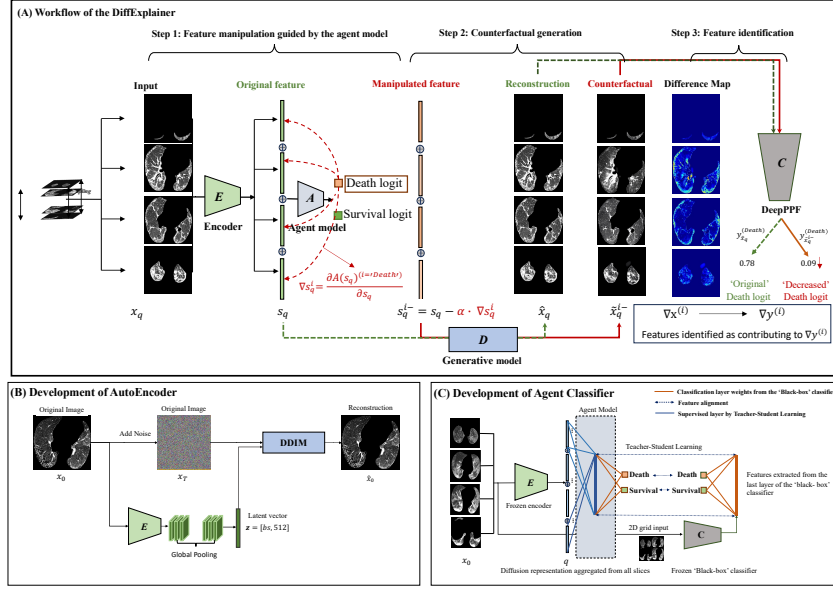
Fig. 1: Framework of DiffExplainer. (A) The workflow of using DiffExplainer to perform feature manipulation and counterfactual generation to understand the impact of different regions in affecting the teacher model's predictions. The two key components are: (B) Diffusion autoencoder consisting of an encoder and a DDIM generative decoder [27]; (C) Knowledge distillation for aligning the latent feature from the diffusion autoencoder to that of the given black box.

2.1   Diffusion Autoencoder

A diffusion autoencoder model consists of an encoder that learns a latent representation and a generative decoder that learns to reconstruct the original CT slices. Both the encoder and decoder play crucial roles in the proposed Diff-Explainer method and are pretrained jointly, as illustrated in Fig. 1 (B). The encoder $\mathbf{E}$ compresses the input image $\mathbf{x}_0$ into a semantically meaningful low-dimensional representation $\mathbf{z} \in \mathbb{R}^{512}$. This feature space can be manipulated through a subsequently introduced agent model. The decoder then generates the reconstructed slice $\hat{\mathbf{x}}_0$ from $\mathbf{z}$ and the noise injected input $\mathbf{x}_T$ (T iterations of Gaussian noise injection). The reconstruction is achieved by iteratively performing noise removal with a conditional diffusion model $\mathbf{D}_\theta$:

$$\hat{\mathbf{x}}_{t-1} = \sqrt{\alpha_{t-1}} \left(\frac{\hat{\mathbf{x}}_t - \sqrt{1-\alpha_t}\mathbf{D}_\theta\left(\hat{\mathbf{x}}_t, t, \mathbf{z}\right)}{\sqrt{\alpha_t}} \right)$$
$$+ \sqrt{1-\alpha_{t-1}}\mathbf{D}_\theta\left(\hat{\mathbf{x}}_t, t, \mathbf{z}\right), \text{ with } t = T, T-1, ..., 1 \quad (1)$$

where α_t characterizes the variance schedule (for noise injection) [9], $\hat{\mathbf{x}}_T = \mathbf{x}_T = \sqrt{\alpha_T}\boldsymbol{x}_0 + \sqrt{1-\alpha_T}\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The encoder $\mathbf{E}$ and decoder $\mathbf{D}$ are trained

concurrently as the way proposed in [19], optimizing for the noise reconstruction loss with respect to θ and ϕ: $\mathcal{L}_{(\phi,\theta)} = \mathbb{E} \left\| \boldsymbol{\epsilon} - \mathbf{D}_\theta \left(\mathbf{x}_t, t, \mathbf{E}_\phi(\mathbf{x}_0) \right) \right\|_1$.

2.2 Teacher-Student Learning for Agent Classifier

The agent classifier is designed to provide decisions aligned with the pretrained teacher model, with additional interpretability. To achieve this goal, it is implemented as a two-layer linear architecture, taking the latent space of the previously introduced autoencoder as input, as illustrated in Fig. 1 (C).

The key role of this two-layer agent classifier is to provide controllable manipulation directions for the input features, consistent with the classification scores of the black box teacher model. To align the classification ability of the two models, we transfer knowledge from the pretrained 'black-box' classifier to the agent model by teacher-student learning techniques. Specifically, we engage a linear layer to map the encoder extracted features to align with that of the teacher model's features, supervised by a L1 loss. The weight of the classification layer in the agent model is fixed to match that of the teacher model.

2.3 Counterfactual Generation

After training the diffusion autoencoder and the agent model, counterfactual images can be generated from a query image. These images are manipulated to potentially modify the prediction scores of the teacher model, providing insights into the reasoning behind the teacher model's decisions. As illustrated in Fig. 1 (A), we denote a query image as I_q, the teacher model as $\mathbf{C}$, and its classification score as y, the features extracted from the diffusion model as q, the agent model as $\mathbf{A}$, and its corresponding prediction score as $y' = \mathbf{A}(q)$.

The feature q can be readily manipulated to respectively increase or decrease its score in the agent model as:

$$\begin{cases} q_{i+} = q + \alpha \frac{dy'_i}{dq}, I'_{q_{i+}} = \mathbf{D}(q_{i+}), y_{i+} = \mathbf{C}(I'_{q_{i+}}) \\ q_{i-} = q - \alpha \frac{dy'_i}{dq}, I'_{q_{i-}} = \mathbf{D}(q_{i-}), y_{i-} = \mathbf{C}(I'_{q_{i-}}) \end{cases} \tag{2}$$

where $\alpha > 0$ is the stepsize for controlling the manipulation scale, and i is the target class for score manipulation. q_{i+} (resp. q_{i-}) denotes manipulated features with increased (resp. decreased) agent prediction score y'_{i+} (resp. y'_{i-}). The corresponding counterfactual images are denoted as $I'_{q_{i+/-}}$, while the reconstructed original image is $I'_q = \mathbf{D}(q)$.

2.4 Evaluation criteria for counterfactual generation

We define the following criteria as essential conditions that counterfactual images need to meet in order to effectively explain a black box model: (1) (Reconstruction ability) The generated counterfactual images should retain most of the features present in the query sample ($I_{q_{i\pm}} \approx I_q$). When there is no manipulation, the generated counterfactual should have minimal deviations from its

query image and similar classification scores. (2) (Observable) The generated counterfactual images should change certain features, that can be detected and discerned by the human eye, while also preserving the overall context of the query image unchanged. (3) (Manipulation) The classification score of the generated counterfactual for the black box should change in the intended manipulation direction, *i.e.* $y_{i\pm}$ should change in the same direction as $y'_{i\pm}$ changes.

3   Experiments

3.1   Dataset

We trained a DiffExplainer for counterfactual generations for a pretrained classification model that predicts the one-year mortality of patients with Fibrotic Lung Disease (FLD). Specifically, we utilized two FLD datasets. The first cohort is sourced from OSIC , with 27 patients deceased within one year and 704 patients surviving for one year. The second cohort is an in-house dataset collected from hospitals in Australia, with 43 patients deceased within one year and 458 patients surviving for one year. In our experiments, we used only the first dataset for training, and switched to the the second in-house dataset for evaluation.

3.2   Experimental setting

The black box classification model employs DenseNet121 as backbone, and correspondingly the two-layer Agent model consists of 1024 neurons in the first layer (aligned to the feature dimension of the black box) and two neurons in the second layer for the two-classification task. For training the black box and agent model, we partitioned this dataset into 5 folds for training and selected the models with the highest sum of specificity and sensitivity. Both models are trained with a batch size of 16 for 100 epochs, with learning rates of 1e-6 and $1e^{-4}$, respectively. The autoencoder, with an input size of 256×256, was trained on slices from the OSIC dataset using eight V100 GPUs with a learning rate of $1e^{-4}$ and a batch size of 64 for 100 epochs.

3.3   Validity of the Agent Model

We first evaluate the consistency of the agent model in producing similar predictions as its black box teacher model. To proceed, we designate the predictions provided by the black box as the 'ground truth'. We calculate the accuracy, sensitivity, specificity, and F1 score for the agent model's results with respect to this 'ground truth' on the test dataset, to validate if the agent model is indeed aligned with its teacher model. Furthermore, we report the Kullback–Leibler (KL) divergence for the two models' predicted probability. We also conducted ablation experiments to study how different training objectives affects knowledge distillation and feature alignment, including L2 and L1 loss on feature space and KL

---

https://www.osicild.org/

divergence on the model's output. The alignment performance under different constraints is presented in Table 1. It can be seen that these different training objectives are all effective in ensuring feature alignment.

Table 1: Alignment of agent model to teacher model.

Method	AUC	Accuracy	Sensitivity	Specificity	F1	KL
MSE	0.99	0.95	0.91	0.97	0.92	0.18
L1	0.99	0.96	0.91	0.98	0.93	0.17
KL	0.99	0.96	0.91	0.98	0.93	0.17

3.4   Performance of the counterfactual generation

The Diffusion autoencoder achieves the best reconstruction quality while GAN-based models fails to achieve acceptable reconstruction. The reconstruction from our Diffusion autoencoder achieved an average PSNR of 35.28, SSIM of 0.99, and a KL divergence for classification, demonstrating outstanding reconstruction capability. On the other hand, although capable of synthesizing realistic-looking CT images, the conditional GAN and StyleGAN-based autoencoder models in our experiments failed to accurately reconstruct the same textures such as airway and vessel from the encoded features of the query image (this is also observed in the histology reconstruction task in [24]).Since GAN based autoencoders failed the reconstruction criterion, we did not continue to generate counterfactual examples.

DiffExplainer consistently pinpoints observable features upon which the decision is based. We select the cases which are indeterminate to the 'black box' (similar logits for each class) and manipulate it to achieve a confident score for each class. As presented in Fig. 2, we generated the counterfactual examples which can be confidently classified as the 1-year mortality and 1-year survival. The difference between the counterfactuals and the original images are presented as difference heatmaps, where red indicates the increased textures and blue indicate the decreased textures. We can attribute the indeterminacy of these cases to the lack of features which are leveraged by the 'black-box' classifier.

DiffExplainer allows for fine-grained control over the counterfactual generation, enabling smooth transition from one classification result to another. In Fig. 2, we illustrate the controllability of the counterfactual generation method by incrementally increasing the manipulation weight. The resulting generated images were validated to exhibit a gradual augmentation in the decisions of the black box classifier and manifestation of features.
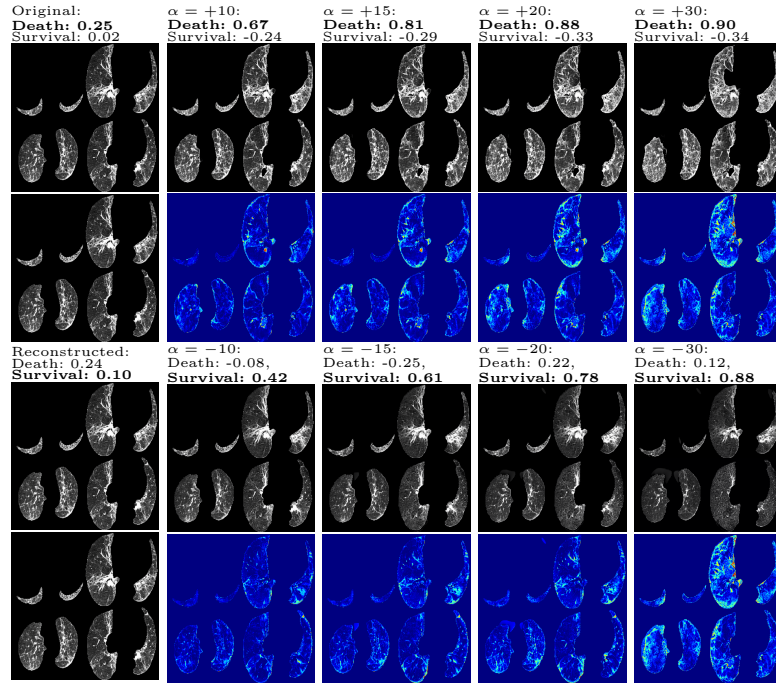
Fig. 2: Counterfactual generation for 'hard' cases that the pretrained classifier failed to assign confident predictions. Row 1: Counterfactuals with increasing 'Death score'; Row 2: Counterfactuals with increasing 'Survival score'.

3.5   Comparison to other XAI methods

We compare the difference heatmaps obtained by DiffExplainer approach to the attribution maps from existing widely-employed XAI methods. We observe that DiffExplainer can more accurately pinpoint theregions within the query image influencing model decisions, with finer granularity, whereas other methods can only localize to a coarse area (Grad-CAM methods) or to regions that are incomprehensible (Saliency maps). For an indeterminate case, DiffExplainer is capable of generating the missing features that can lead to confident decisions for either class, thus explaining why the model arrived at its indeterminate decision, a capability not matched by existing methods.

4   Conclusion

In this paper, we developed a counterfactual generation approach named Diff-Explainer for explaining the decisions of a given black box classifier. Compared to existing attribution methods, our approach can more accurately locate the features influencing model decisions within the query image, has finer granularity, and also allows enhanced controllability. We validate our approach on the
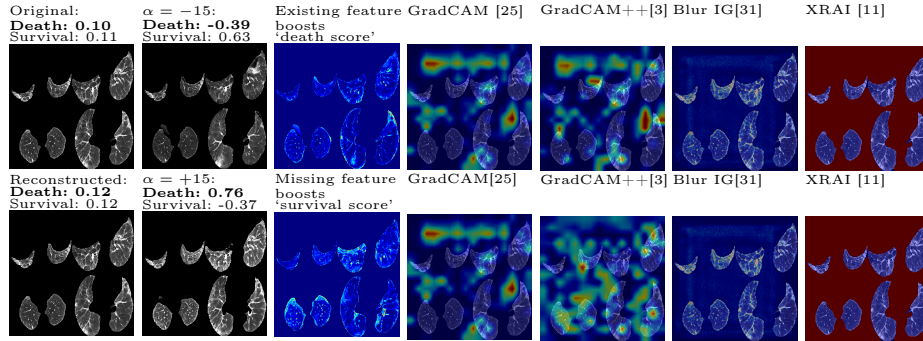
Fig. 3: Comparison to other XAI methods. Blue and red areas indicate the existing features and missing features that contributes to the change in prediction.

challenging modality of CT scans, which posed significant difficulties for existing methods. Our DiffExplainer distinguishes itself by providing high-quality reconstructions, and the capacity to identify semantically meaningful features which impact the classifier's predictions. This capability enables users to uncover the reasons behind each decision made by a given black box model. Combining DiffExplainer with highly accurate black-box models presents an opportunity to discover novel biomarkers.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Atad, M., Dmytrenko, V., Li, Y., Zhang, X., Keicher, M., Kirschke, J., Wiestler, B., Khakzar, A., Navab, N.: Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan. arXiv preprint arXiv:2207.07553 (2022)
2. Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., Rudin, C.: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence **3**(12), 1061–1070 (2021)
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Gradcam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)

4. Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: a simple autoencoder approach to counterfactual generation for chest x-rays. In: Medical Imaging with Deep Learning. pp. 74–104. PMLR (2021)

5. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

6. Fontanella, A., Antoniou, A., Li, W., Wardlaw, J., Mair, G., Trucco, E., Storkey, A.: Acat: Adversarial counterfactual attention for classification and detection in medical imaging. arXiv preprint arXiv:2303.15421 (2023)

7. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)

8. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

10. Jin, W., Li, X., Fatehi, M., Hamarneh, G.: Guidelines and evaluation of clinical explainable ai in medical image analysis. Medical Image Analysis **84**, 102684 (2023)

11. Kapishnikov, A., Bolukbasi, T., Viégas, F., Terry, M.: Xrai: Better attributions through regions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4948–4957 (2019)

12. Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T.: Training generative adversarial networks with limited data. In: Proc. NeurIPS (2020)

13. Kodali, N., Abernethy, J., Hays, J., Kira, Z.: On convergence and stability of gans. arXiv preprint arXiv:1705.07215 (2017)

14. Li, M., Fang, Y., Tang, Z., Onuorah, C., Xia, J., Del Ser, J., Walsh, S., Yang, G.: Explainable covid-19 infections identification and delineation using calibrated pseudo labels. IEEE Transactions on Emerging Topics in Computational Intelligence **7**(1), 26–35 (2022)

15. Mertes, S., Huber, T., Weitz, K., Heimerl, A., André, E.: Ganterfactual—counterfactual explanations for medical non-experts using generative adversarial learning. Frontiers in artificial intelligence **5**, 825565 (2022)

16. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nebelung, S., Kather, J.N., et al.: Diffusion probabilistic models beat gans on medical images. arXiv preprint arXiv:2212.07501 (2022)

17. Patrício, C., Neves, J.C., Teixeira, L.F.: Explainable deep learning methods in medical image classification: A survey. ACM Computing Surveys **56**(4), 1–41 (2023)

18. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)

19. Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10619–10629 (2022)

20. Reynaud, H., Vlontzos, A., Dombrowski, M., Gilligan Lee, C., Beqiri, A., Leeson, P., Kainz, B.: D'artagnan: Counterfactual video generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 599–609. Springer (2022)

21. Ribeiro, M.T., Singh, S., Guestrin, C.: "why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. pp. 1135–1144 (2016)
22. Rudin, C.: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence **1**(5), 206–215 (2019)
23. Sankaranarayanan, S., Hartvigsen, T., Oakden-Rayner, L., Ghassemi, M., Isola, P.: Real world relevance of generative counterfactual explanations. In: Workshop on Trustworthy and Socially Responsible Machine Learning, NeurIPS 2022 (2022)
24. Schutte, K., Moindrot, O., Hérent, P., Schiratti, J.B., Jégou, S.: Using stylegan for visual interpretability of deep learning models on medical images. arXiv preprint arXiv:2101.07563 (2021)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
26. Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly—a counterfactual approach. Medical Image Analysis **84**, 102721 (2023)
27. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
28. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International conference on machine learning. pp. 3319–3328. PMLR (2017)
29. Tanyel, T., Ayvaz, S., Keserci, B.: Beyond known reality: Exploiting counterfactual explanations for medical research. arXiv preprint arXiv:2307.02131 (2023)
30. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
31. Xu, S., Venugopalan, S., Sundararajan, M.: Attribution in scale and space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9680–9689 (2020)
32. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13. pp. 818–833. Springer (2014)