Explaining Chest X-ray Pathology Models using Textual Concepts

Vijay Sadashivaiah
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
sadasv2@rpi.edu

Pingkun Yan
Department of Biomedical Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180
yanp2@rpi.edu

James A. Hendler
Department of Computer Science
Rensselaer Polytechnic Institute
Troy, NY 12180
hendler@cs.rpi.edu

Abstract

Deep learning models have revolutionized medical imaging and diagnostics, yet their opaque nature poses challenges for clinical adoption and trust. Amongst approaches to improve model interpretability, concept-based explanations aim to provide concise and human-understandable explanations of any arbitrary classifier. However, such methods usually require a large amount of manually collected data with concept annotation, which is often scarce in the medical domain. In this paper, we propose Conceptual Counterfactual Explanations for Chest X-ray (CoCoX), which leverages the joint embedding space of an existing vision-language model (VLM) to explain black-box classifier outcomes without the need for annotated datasets. Specifically, we utilize textual concepts derived from chest radiography reports and a pre-trained chest radiography-based VLM to explain three common cardiothoracic pathologies. We demonstrate that the explanations generated by our method are semantically meaningful and faithful to underlying pathologies.

1 Introduction

Over the past decade, deep learning (DL) has achieved unprecedented performance in medical imaging-based diagnosis (20). This success is partly attributed to the increasing availability of data and the rapid development of deep neural network (DNN) technologies. Despite these impressive results, DNNs have been criticized for their black-box nature, which has limited the adoption and regulatory approval of medical imaging-based DNNs (19). The combination of lack of explainability (challenging to interpret their inner workings) (27), lack of robustness (being easily fooled using adversarial data) (32), bias (tendency to amplify inequalities that exist in data) (22), and the high stakes nature of clinical applications (21; 31) prevents deployment of black-box models in clinical practice. Explainable artificial intelligence (XAI) methods aim to tackle this by making the decisions and processes of black-box systems understandable and interpretable to humans.

In the medical imaging field, heatmap-style explanations such as saliency maps have been extensively studied to interpret the decisions made by a classifier (26; 25). However, simply highlighting the relevant pixels that influence classifiers' decisions does not answer *why* that region is important(7). Additionally, previous work has shown that these methods are susceptible to adversarial attacks

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

(33; 10), prone to confirmation bias (1; 4) and require domain expert's intervention to *interpret* the explanation leading to subjective bias (16).

On the other hand, concept-based explanation methods provide human-understandable and high-level semantically meaningful explanations (16; 8). In natural images, a *concept* represents any semantically meaningful attribute and is typically represented by words such as "stripes", "smile", etc. To measure the importance of such concepts on the output of a black-box classifier, concept activation vector (CAV) (16) was proposed. A CAV measures the direction of each concept in the embedding space of a pre-trained classifier. Using CAVs, one can then calculate which concepts influenced the output of a particular class. However, in order to train CAVs, one needs human-annotated concept labels for images, which can be expensive and often infeasible. Another line of work aims to explain pre-trained classifiers using counterfactual explanations (9; 29). In simple terms, counterfactual explanation means identifying the feature perturbations required in input images to lead the pre-trained model to a different output prediction. Few methods have explored this line of work in medical imaging (3; 28). A key limitation of this approach is the need for subject matter experts to evaluate and assign meaning to generated counterfactual images(3). Additionally, generating counterfactual images requires training GAN-based methods that can capture latent representations from the input data. Training these models can be expensive (14) and often infeasible due to limited data availability in medical domains (5). While (15) introduced a natural language explanation (NLE) dataset derived from MIMIC-CXR, which involved extracting explanations for different pathologies from radiology reports and suggesting retraining black-box classifiers to align with these NLEs, we propose a different approach. Our method focuses on explaining black-box classifiers without the need for retraining by utilizing existing vision-language models (VLMs).

To tackle this challenge, we propose to develop a conceptual counterfactual explanation (CoCoX) that combines the benefits of concept-based and counterfactual explanations. Specifically, we gathered chest radiology-related natural language concepts from the PadChest dataset (6), which is annotated by radiologists and mapped against the Unified Medical Language System (UMLS) database for validity and augmented additional concepts by querying ChatGPT (23) for pathology-specific visual attributes. We then leverage the joint embedding space of a pre-trained vision-language model (VLM) to derive concept directions and create a *concept bank*. Next, we train simple projection models that allow us to move from the black-box classifier's latent space to VLM's latent space. Finally, we generate *conceptual counterfactuals* by learning perturbations in the direction of each concept in our concept bank. In essence, we perturb the projected image embedding such that the black-box model's prediction changes to the target class, similar to a previous method CounTEX (17).

Given an input image, an actual prediction by a black-box model, the target prediction, and a concept bank, our approach answers the question of *how much* should each concept be perturbed to change the output from current prediction to the target prediction? To achieve this, we learn perturbation weights for each concept in our concept bank, which serves as an importance score for each concept in changing the target classifier output. We present CoCoX's overall model architecture in Figure 1. The main contributions of this paper are summarized as follows: 1) we propose a conceptual counterfactual explanation method for chest radiography image classifiers that integrate the benefits of concept-based explanations and counterfactual explanations, and 2) we conduct in-depth experimental analysis to demonstrate that our proposed approach provides clinically relevant conceptual explanation for several chest x-ray pathologies.

2 Method

Our proposed method can be divided into three stages: 1) Construction of concept bank, 2) Learning Projector functions, and 3) Learning conceptual perturbations. We discuss each of these steps in more detail below.

2.1 Constructing concept bank

We construct our concept bank (C) using natural language concepts derived from PadChest dataset (6). First, for each textual concept x_T, we generate a pair of text prompts composed of a *neutral* sequence (t_n) and a *stimuli* sequence (t_s). The neutral sequence is fixed for all concepts with a generic phrase "`An image of chest xray with No Finding`", following the zero-shot prompt strategy from the original CLIP paper (24). Stimuli sequence is generated using the phrase "`An image of`

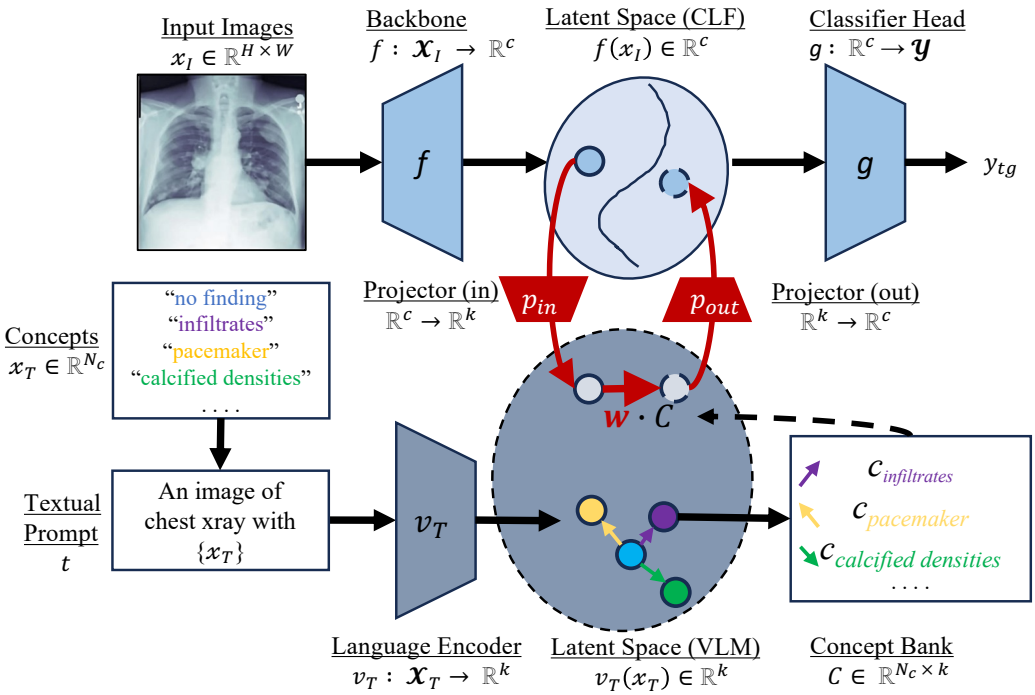

Figure 1: The overall architecture of CoCoX. We create the Concept Bank (C) by encoding natural language concepts with a chest radiography pre-trained language encoder of a VLM. We then learn Projector(in) and Projector(out) functions p_{in} and p_{out} using MLPs to transfer feature representation between CLF latent space and VLM latent space. Additionally, we learn perturbation weights for each concept in our concept bank, which serves as an importance score in changing the target classifier output. Learnable parameters are highlighted in red arrows.

`chest xray with {`x_T`}`", for each concept x_T. We then use the pair of prompts $[t_n, t_s]$ to derive the concept direction. First, the pair is tokenized and encoded using the pre-trained language encoder (v_T), yielding the embedded prompt pair $[v_T(t_n), v_T(t_s)]$. Then, the direction of textual concept $c \in C$ can be obtained by taking the difference between these two embedded prompt pairs and normalizing it to a unit vector, i.e.,

$$c = \frac{v_T(t_s) - v_T(t_n)}{||v_T(t_s) - v_T(t_n)||_2}. \tag{1}$$

We compute a textual concept $c \in \mathbb{R}^k$ for each stimulus to obtain our final concept bank $C = \{c_i | i = 1, \ldots, N_c\}$, where k is the embedding dimension of VLP and N_c is the number of concepts.

2.2 Learning projection functions

Since our goal is to operate on textual concepts encoded in VLM latent space, we need to learn the projections p_{in} and p_{out}, that can move image embedding from classifier (CLF) latent space to VLM latent space and vice versa. First, to train the p_{in} function, we project the input embedding $f(x_I)$ into VLM latent space resulting in the projected embedding $p_{in}(f(x_I))$. Next, we embed the same image using the vision encoder v_I of VLM to obtain $v_I(x_I)$. Ideally, we want these two embeddings to be close to each other, and hence, the p_{in} is optimized to minimize this distance. Similarly, for p_{out}, we first project the VLM vision encoder's embedding ($v_I(x_I)$) to classifier latent space using p_{out}, resulting in the projected embedding $p_{out}(v_I(x_I))$. We also encode the same image using the pre-trained classifier, which results in embedding $f(x_I)$. Again, p_{out} is trained to minimize the distance between these two embeddings. Finally, we want to account for the round-trip error that minimizes the distance between embedding $p_{out}(p_{in}(f(x_I)))$ and $f(x_I)$. Taking all these three into

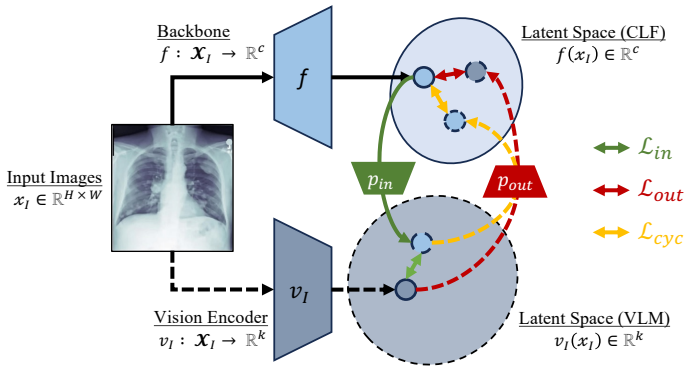

Figure 2: Training paradigm for projection functions. Each $\leftrightarrow$ indicates a distance minimization problem. Total loss is a sum of all loss functions, $\mathcal{L}_{total} = \mathcal{L}_{in} + \mathcal{L}_{out} + \mathcal{L}_{cyc}$.

account, we get the following loss function,

$$\begin{aligned}\mathcal{L}_{total} &= \mathcal{L}_{in} + \mathcal{L}_{out} + \mathcal{L}_{cyc} \\ &= ||p_{in}(f(x_I)) - v_I(x_I)||^2 + ||p_{out}(v_I(x_I)) - f(x_I)||^2 \\ &\quad + ||p_{out}(p_{in}(f(x_I))) - f(x_I)||^2\end{aligned} \tag{2}$$

We pre-train $\mathcal{L}_{in}$ and $\mathcal{L}_{out}$ independently and fine-tune $\mathcal{L}_{total}$ for a few epochs. See Figure 2 for a depiction of this training paradigm.

2.3 Learning conceptual perturbations

Once the concept bank is established and the projection functions are trained, we need to learn concept-wise perturbation weight parameter $\boldsymbol{w}$, such that the gap between perturbed prediction y_p and target prediction y_t is minimized. Since we apply this perturbation in VLM latent space, we first project our input image x_I to VLM latent space using p_{in}, apply learnable perturbation, project back to classifier latent space using p_{out}, and predict the output of perturbed embedding. We define this output below

$$y_p = g(p_{out}(p_{in}(f(x_I)) + \boldsymbol{w} \cdot C)), \tag{3}$$

where g is the classifier head of our pre-trained classifier. In order to optimize $\boldsymbol{w}$, we introduce the following constraints. First, the perturbed embedding should predict the target class y_t. We propose using cross-entropy loss between the perturbed and target classes, $\mathcal{L}_{CE} = CE((y_p, y_t))$. Next, we introduce sparseness through L1 regularization $||\boldsymbol{w}||_1$, and ensure that the perturbed embedding does not deviate from image embedding using L2 regularization $||\boldsymbol{w}||_2$. Our final loss $\mathcal{L}_{final}$ is a weighted sum of these three loss functions

$$\mathcal{L}_{final} = \mathcal{L}_{CE} + \alpha||\boldsymbol{w}||_1 + \beta||\boldsymbol{w}||_2, \tag{4}$$

where α and β are weighting parameters.

3 Experiments

In this section, we describe experimental details, including datasets, implementation, and metrics. We also present the quantitative and qualitative results, comparing the explanations generated by CoCoX against clinical literature and an existing baseline.

3.1 Datasets

Imaging datasets: We used two large-scale chest x-ray image datasets in our experiments. First, we used the CheXpert dataset (12) to train our pathology backbone classifiers. Following the previous work (3), we trained for three different pathologies, Cardiomegaly, Atelectasis, and Pleural Effusion. Next, to train our projector models, we used the MIMIC-CXR dataset (13).

Concept bank: To create our concept bank C, we started with the radiological findings annotated in the PadChest dataset (6). This yielded 174 natural language concepts. Next, we queried ChatGPT (23) with prompt "`What visual attributes are seen in` $pathology$ `chest xray?`" for each of the pathologies and augmented our concept library yielding a total of 192 concepts[1]. It is important to note the ease of augmenting additional concepts to the library. CoCoX allows the addition/removal of textual concepts, unlike other concept-based explanation methods that need annotated images. We then proceed with Eq. 1 to create the concept bank.

3.2 Implementation details

Backbone classifiers and VLM: We implemented three different architectures for our backbone image classifiers. Specifically, we chose DenseNet121, ResNet34, and VLM+Linear for our experiments. For VLM+Linear based architecture, we just added a linear layer on top of our VLM vision encoder (ViT-B/32). Unlike the other two architectures, the VLM-based one does not require separate projector models since are already operating in VLM latent space. Each model was trained for pathologies using the CheXpert dataset for a maximum of 50 epochs with early stopping. For our VLM, we used CheXzero model (30).

Projectors: We used a simple 2-layer multi-layer perceptron (MLP) to implement p_{in} and p_{out} functions. Both these models comprised of (512, 512) hidden units and were trained with the MIMIC-CXR dataset for a maximum of 50 epochs with early stopping to minimize the loss function in Equation 2. A batch size of 64 was used for all the experiments.

Conceptual perturbation: The weight parameter $\boldsymbol{w}$ is optimized to minimize the loss $\mathcal{L}_{final}$ (see Eq. 4). Parameters α and β are set to 0.1 for all experiments following (17). We used stochastic gradient descent (SGD) with momentum and a learning rate of 10^{-2} for a maximum of 100 steps. We stopped training early if the prediction of input image changes to the desired target class. This weight parameter serves as an importance score for each concept in changing the output of the black box classifier to the target output.

3.3 Results and Evaluation

For each of the three pathologies, Cardiomegaly, Pleural Effusion, and Atelectasis, we optimize the perturbations to change a "No Finding" image to corresponding pathology finding. To verify that the proposed approach works as intended, we conducted an experiment where we *included* the pathology name as one of the concepts in our concept bank. After perturbing 100 random example images from each pathology to the target class using our VLM + Linear model, we found that the pathology name was **top@1** concept for 99% of Cardiomegaly, 100% of Pleural Effusion and 98% of Atelectasis cases. The corresponding pathology name is removed from the concept bank for the next set of experiments.

We present random images for each pathology and corresponding top 5 conceptual counterfactuals generated by our model for VLM + Linear and DenseNet121 model in Figure 3. For each pathology, the emerging top concepts represent the main features of each disease. For instance, the top concepts in converting a "No Finding" image to "Cardiomegaly" are `increased cardiac diameter`, `increased cardio-thorasic ratio`, correspond to prominent features in Cardiomegaly(2). For Pleural Effusion, our method picks up relevant features such as `increased opacity in lower lungs`, `fluid accumulation` in the top 5 concepts(18). In essence, these concepts had the most impact when *added* to input images to change from "No Finding" to a pathology finding. Similarly, it is also relevant to analyze conceptual counterfactuals for changing pathology input to "No Finding" output. We present this result in Figure S1 for all pathologies. Additionally, we observe alignment between top concepts chosen by DenseNet121 and VLM+Linear model, demonstrating the efficacy of our approach on various model architectures.

In order to validate that the generated conceptual counterfactuals are clinically relevant, we compare our method against CheXplaining in Style (3). The authors used conditional StyleGAN (14) based architecture to generate counterfactual images for three chest x-ray pathologies. In their evaluation, the radiologists first listed the primary features and possible secondary findings they rely on while diagnosing each disease. Then, the authors perturbed ten random "No Finding" samples to generate

[1]See Supplementary Section 2 for an example query and response

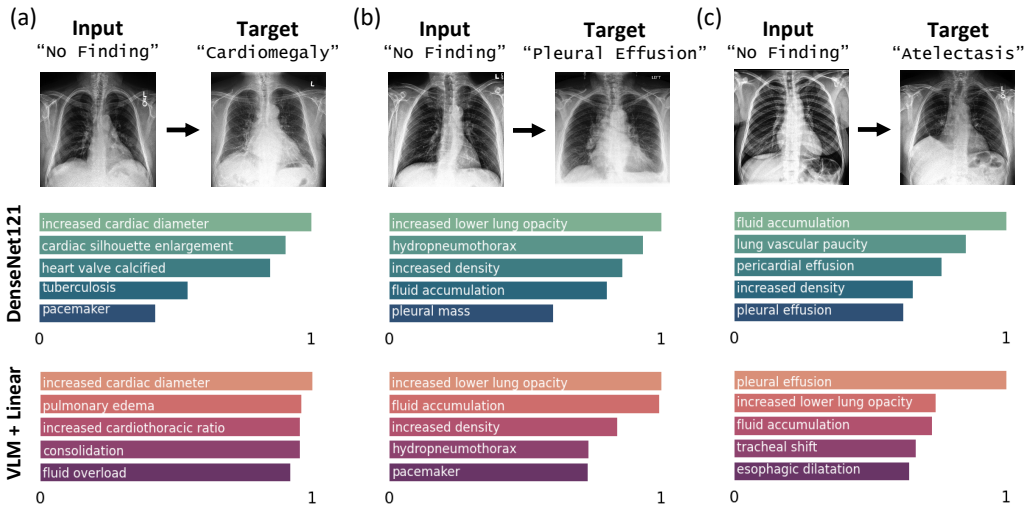

Figure 3: **Conceptual counterfactual generated by CoCoX.** The input image into the model for each pathology is shown under the heading Input ("No Finding"). The target image presented under Target ("Pathology") is a representative image labeled positive for the corresponding pathology. The concept importance scores for the top 5 concepts are visualized with blue-green for the DenseNet-121 model and orange-purple for VLM + Linear. These concepts had the most impact when added to input images to change from "No Finding" to a pathology finding.

Table 1: Comparison of recall (R) at k concepts to radiologists' evaluation.

Pathology	**Finding**	**DenseNet121**		**VLM + Linear**		**ResNet34**	
		R@5	R@10	R@5	R@10	R@5	R@10
Cardiomegaly	*Primary(1)*	0.97	**1.00**	**1.00**	**1.00**	0.98	**1.00**
	Secondary(4)	0.35	0.48	0.50	0.50	0.32	**0.52**
Pleural Effusion	*Primary(2)*	0.46	0.78	0.50	**1.00**	0.43	0.81
	Secondary(3)	0.17	0.29	**0.33**	0.33	0.19	0.25
Atelectasis	*Primary(2)*	0.55	0.69	0.50	**0.85**	0.48	0.73
	Secondary(3)	0.26	0.61	0.33	**0.66**	0.21	0.59

counterfactual images, which radiologists evaluated and annotated for the presence or absence of primary and secondary features. However, the authors did not present the fraction (out of ten) of counterfactuals that contained these features. In our experiment, we first verified if the features relied on by radiologists were already in our concept bank. If a feature was absent, we appended that to our concept bank. Next, we calculate the overlap between features identified by radiologists and concepts identified by our approach. We use the recall@k measure typically used in recommender systems evaluation. Recall@k measures the fraction of ground-truth concepts (i.e., radiologists' list) present in top@k predictions made by our model. We present mean recall@k results for all three models in Table 1 using 1000 images, where we separate the scores by primary and secondary findings[2]. We also list the number of primary and secondary features annotated by radiologists within the parentheses for each pathology. Table 1 shows that our model can recall the primary findings for all three pathologies with high recall scores. Recall@k measure penalizes more when the ground truth list is extensive, which is the case for secondary features. Hence, we observe a lower recall@5 values for secondary features. Additionally, concepts such as "Older patients" in the secondary list for Cardiomegaly and Atelectasis are non-visual attributes and, hence, challenging to generate conceptual counterfactuals. We bold the highest score in each row.

[2]see Supplementary Table S1 for the list of primary and secondary findings

Finally, we demonstrate that our model generates conceptual counterfactuals faster than (3). Our model using VLM+Linear generated explanations for 99% of input images (n=600 images) in under 30 seconds, whereas CheXplaining in Style takes 5 minutes to create explanations with an average of 94% coverage.

4 Conclusions

In order to improve the explainability of black-box medical-image classifiers, we explored the chest x-ray domain and investigated conceptual counterfactual explanations that combine concept-based and counterfactual explanations. Our approach highlights important concepts that contribute to changing an image from "No Finding" to a pathology finding and vice-versa. We created a concept bank using radiological findings from the PadChest (6) dataset and augmented additional concepts by querying ChatGPT (23). Next, to generate conceptual counterfactual explanations, we manipulate the latent embedding of input image with perturbations, such that the classifier's output changes to target class. We found that the top concepts generated by our method are relevant to the underlying pathology and align with radiologists' evaluation. While we have qualitatively demonstrated clinical relevance by aligning with radiologists' assessments, we plan to explore standardized metrics like expert consensus scores, and counterfactual faithfulness in future iterations.

With the proliferation of vision-language models (30; 11), our method can easily be extended to studying black-box classifiers in other medical imaging domains. This extension would involve the careful curation of domain-specific concepts tailored to the target medical imaging modality. For instance, in MRI or CT scans, concepts related to soft tissue abnormalities or bone structure could be incorporated into the concept bank. Additionally, training the CoCoX modules on these curated concepts allows us to generate useful conceptual counterfactual explanations for various medical conditions beyond chest X-rays. We leave this for future work, along with exploring ways to improve concept bank creation.

References

[1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. Advances in neural information processing systems **31** (2018)

[2] Amin, H., Siddiqui, W.J.: Cardiomegaly. In: StatPearls [internet]. StatPearls Publishing (2021), `https://www.ncbi.nlm.nih.gov/books/NBK542296/`

[3] 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)

[4] Bornstein, A.M.: Is artificial intelligence permanently inscrutable. Nautilus **40** (2016)

[5] Bowles, C., Chen, L., Guerrero, R., Bentley, P., Gunn, R., Hammers, A., Dickie, D.A., Hernández, M.V., Wardlaw, J., Rueckert, D.: Gan augmentation: Augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)

[6] Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)

[7] Ghassemi, M., Oakden-Rayner, L., Beam, A.L.: The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health **3**(11), e745–e750 (2021)

[8] Ghorbani, A., Wexler, J., Zou, J.Y., Kim, B.: Towards automatic concept-based explanations. Advances in neural information processing systems **32** (2019)

[9] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: International Conference on Machine Learning. pp. 2376–2384. PMLR (2019)

[10] Gu, J., Tresp, V.: Saliency methods for explaining adversarial attacks. arXiv preprint arXiv:1908.08413 (2019)

[11] Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

[12] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)

[13] Johnson, A.E., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. Scientific data **6**(1), 317 (2019)

[14] Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)

[15] Kayser, M., Emde, C., Camburu, O.M., Parsons, G., Papiez, B., Lukasiewicz, T.: Explaining chest x-ray pathologies in natural language. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 701–713. Springer (2022)

[16] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning. pp. 2668–2677. PMLR (2018)

[17] Kim, S., Oh, J., Lee, S., Yu, S., Do, J., Taghavi, T.: Grounding counterfactual explanation of image classifiers to textual concept space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10942–10950 (2023)

[18] Krishna, R., Antoine, M.H., Rudrappa, M.: Pleural effusion. In: StatPearls [internet]. StatPearls Publishing (2023), `https://www.ncbi.nlm.nih.gov/books/NBK448189/`

[19] Langlotz, C.P., Allen, B., Erickson, B.J., Kalpathy-Cramer, J., Bigelow, K., Cook, T.S., Flanders, A.E., Lungren, M.P., Mendelson, D.S., Rudie, J.D., et al.: A roadmap for foundational research on artificial intelligence in medical imaging: from the 2018 nih/rsna/acr/the academy workshop. Radiology **291**(3), 781–791 (2019)

[20] Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical image analysis **42**, 60–88 (2017)

[21] Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., Jha, N.K.: Systematic poisoning attacks on and defenses for machine learning in healthcare. IEEE journal of biomedical and health informatics **19**(6), 1893–1905 (2014)

[22] Obermeyer, Z., Powers, B., Vogeli, C., Mullainathan, S.: Dissecting racial bias in an algorithm used to manage the health of populations. Science **366**(6464), 447–453 (2019)

[23] OpenAI: Chatgpt 3.5, `https://chat.openai.com` [Accessed: February 2024]

[24] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (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] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)

[27] Singh, A., Sengupta, S., Lakshminarayanan, V.: Explainable deep learning models in medical image analysis. Journal of imaging **6**(6), 52 (2020)

[28] Singla, S., Eslami, M., Pollack, B., Wallace, S., Batmanghelich, K.: Explaining the black-box smoothly—A counterfactual approach. Medical Image Analysis **84**, 102721 (Feb 2023)

[29] Stepin, I., Alonso, J.M., Catala, A., Pereira-Fariña, M.: A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. IEEE Access **9**, 11974–12001 (2021)

[30] Tiu, E., Talius, E., Patel, P., Langlotz, C.P., Ng, A.Y., Rajpurkar, P.: Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. Nature Biomedical Engineering **6**(12), 1399–1406 (2022)

[31] Vayena, E., Blasimme, A., Cohen, I.G.: Machine learning in medicine: addressing ethical challenges. PLoS medicine **15**(11), e1002689 (2018)

[32] Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M.K., Yan, P.: Overlooked Trustworthiness of Saliency Maps. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 451–461. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2022)

[33] Zhang, J., Chao, H., Dasegowda, G., Wang, G., Kalra, M.K., Yan, P.: Revisiting the trustworthiness of saliency methods in radiology ai. Radiology: Artificial Intelligence **6**(1), e220221 (2023)