Deep Learning for Medical Imaging

From Diagnosis Prediction to its Explanation

by

Sumedha Singla

Masters in Computing, University of Utah, 2015

Submitted to the Graduate Faculty of

the Department of Computer Science in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF COMPUTING AND INFORMATION

This dissertation was presented

by

Sumedha Singla

It was defended on

July 12th 2022

and approved by

Dr. Adriana Kovashka, Assistant Professor, Department of Computer Science, School of Computing and Information

Dr. Xiaowei Jia, Assistant Professor, Department of Computer Science, School of Computing and Information

Dr. Sofia Triantafillou, Assistant Professor, Department of Mathematics and Applied Mathematics, University of Crete

Dr. Milos Hauskrecht, Professor, Department of Computer Science, School of Computing and Information (Dissertation Co-Chair)

Dr. Kayhan Batmanghelich, Assistant Professor, Department of Biomedical Informatics, University of Pittsburgh School of Medicine (Dissertation Chair)

Deep Learning for Medical Imaging
From Diagnosis Prediction to its Explanation

Sumedha Singla, PhD

University of Pittsburgh, 2022

Deep neural networks (DNN) have achieved unprecedented performance in computer-vision tasks almost ubiquitously in business, technology, and science. While substantial efforts are made to engineer highly accurate architectures and provide usable model explanations, most state-of-the-art approaches are first designed for natural vision and then translated to the medical domain. This dissertation seeks to address this gap by proposing novel architectures that integrate the domain-specific constraints of medical imaging into the DNN model and explanation design.

Prior work on DNN design commonly performs lossy data manipulation to make volumetric data compatible with 2D or low-resolution 3D architectures. To this end, we proposed a novel DNN architecture that transforms volumetric medical imaging data of any resolution into a robust representation that is highly predictive of disease. For DNN model explanation, current explanation methods primarily focus on highlighting the essential regions (where) for the classification decisions. The location information alone is insufficient for applications in medical imaging. We designed counterfactual explanations to visually demonstrate how adding or removing image-features changes the DNN decision to be positive or negative for a diagnosis.

Further, we reinforced the explanations by quantifying the causal relationship between neurons in DNN and relevant clinical concepts. These clinical concepts are derived from radiology reports and are corroborated by the clinicians to be useful in identifying the underlying diagnosis. In the medical domain, multiple conditions may have a similar visual appearance, and it's common to have images with conditions that are novel for the pre-trained DNN. DNN should refrain from making over-confident predictions on such data and mark them for a second reading. Our final work proposed a novel strategy to make any off-the-shelf DNN classifier adhere to this clinical requirement.

Table of Contents

List of Tables

List of Figures

Preface

First and foremost, I would like to thank my advisor Dr. Kayhan Batmanghelich. My research has been guided by his leadership, persistence, and encouragement and I thank him for all the time and energy that he has invested in me. Thank you for your trust in my capacities and your guidance. No matter how busy his schedule was, he was the most available person to me, just a message away. Dr. Kayhan has always given me the best of his mentorship, and I will be forever grateful to have benefited from his abundance of experience and knowledge. I would also like to thank his postdocs and my close collaborators, Brian Pollock, Junxiang Chen and Mingming Gong. I had a great experience in learning and working with them.

The final part of this dissertation is the outcome of collaboration with Dr. Sofia Triantafillou. Sofia has been an awesome person to work with. I really enjoyed working and collaborating with her during the last few years, especially discussing research problems through our long meetings. Thank you so much for all your time. I am also grateful to Forough Arabshahi who has been a great industrial collaborator.

I could not thank enough Dr. Stephen Wallace for his time and patience. As my clinical collaborator he has given me his precious time, have patiently answers all my questions, and have helped me develop research from a clinical prospective. A special thanks to Dr. Frank Sciurba for his interesting discussions and his passion for small things. I will always remember our meetings and your joyful nature.

I am grateful to Dr. Motahhare Eslami who have rescued me at my time of need by providing her expertise in human computer interaction. Her smiling face and words of encouragement have made our interactions so much more than academic collaboration. Thank you so much for making my research human-relevant.

I am also thankful to former and current members of Batman Lab who have been a great source of support during my PhD years. I am grateful to all of them, particularly Yanwu Xu, Rohit Jena, Ke Yu, Li Sun, Matthew Ragoza, Maxwell Reynolds, Nihal Murali, Sead Nikšić, Shantanu Ghosh and Payman Yadollahpour.

I am also thankful to the staff members in DBMI and Computer Science department who have been always extremely helpful during the last 6 years. Special thanks goes to Keena Walker and Genine Bartolotta for all their help.

Finally and most importantly, I would like to thank my family, Vasu Gangrade and my beloved Vanya. You both have been the pillars of my strength. Your support has given me courage and determination over this long journey. With all my heart and soul, thanks to my mother, father and sister Suruchi. You three have been there for me all this time, despite the distance. Thank you mom for sharing your experiences and making sure that I keep making progress. I undoubtedly could not have achieved this without them.

1.0 Overview

Decision-making pipelines are inclining towards using deep neural networks (DNN) for high-stake applications especially medical diagnostics [57, 74, 53, 82]. DNNs are sophisticated and complex machine learning (ML) models trained using large datasets and computational resources [98, 116]. Many advancements have been proposed to use different sources of medical data such as imaging, electronic health records [168, 206, 158], patient discharge summaries [88], diagnosistic tests, and others to build deep learning (DL) systems for diagnostic applications. The primary focus of this thesis is on using medical imaging information for diagnosis prediction. Further, in medical image analysis, DL models are used for solving different computer vision-related problems such as classification [288], detection [286], segmentation [77], and registration [55]. Among them, we primarily focus on classification methods that classify an input image or a series of images with diagnostic labels of some predefined diseases.

Over the last few years, convolutional neural networks (CNNs) have become the de-facto backbone of numerous models using medical-imaging data for diagnostic classification [53, 208, 281, 170]. CNNs superior predictive performance compared to simpler counterparts makes them a lucrative option for real-world deployment. But this improvement comes at the cost of decreased model interpretability [176]. They are essentially *Black box* predictive models that often predict the correct answer for the wrong reasons [64, 188] and are over-confident on out-of-domain data [85, 65]. These challenges remain a primary reason for lack of trust and a barrier to broader acceptance of such algorithms in practice.

The clinical deployment of DNN models is contingent on satisfying three essential requirements. First, DNN model design should integrate domain-specific constraints of medical imaging into its architectural design while providing sufficiently high predictive performance [240, 110, 251]. Second, the decision-making process of these models should be explainable to clinicians to obtain their trust in the model [73, 67, 111]. Third, the model should communicate the uncertainty in its prediction and raise a flag when it doesn't have sufficient information to make a confident prediction [64, 92]. This dissertation proposes models to account for subtleties of medical imaging and add support for these clinical needs. At the same time, our foundation is a general approach and applies to different model and dataset domains. Next, we will discuss each requirement in more detail.

1.1 DNN model design

Researchers have proposed numerous DNN models that use medical imaging data for autonomous diagnostics [153]. In this thesis, we primarily focus on lung imaging and associated diseases such as Chronic obstructive pulmonary disease (COPD), pleural effusion, and others. However, our proposed models apply to a wide range of heterogeneous disorders. Having an objective way to characterize local patterns of disease is essential in diagnosis, risk prediction, and sub-typing [56, 89, 182, 233]. Previously, researchers have proposed intensity and texture-based feature descriptors to represent the visual appearance of a disease. However, most image features are generic and not necessarily optimized for a given diagnosis [29, 248, 289]. Recent advances in deep learning (DL) have enabled researchers to use raw images directly for predicting clinical outcomes without specifying radiological features [74]. These clinical outcomes may include diagnosis identification, symptom score, and mortality. The classical DL methods, which operate on entire volumes or slices, are challenging to interpret and require resizing the input images to a fixed dimension [74]. My first project proposes an attention-based method that aggregates local image features to a subject-level representation for predicting disease severity. Our proposed method operates on a set of image patches; hence it can accommodate variable-length input volumes without image resizing.

1.2 DNN model explanation

Explainability is essential for auditing DNNs [281]; identifying various failure modes [52, 193] or hidden biases in the data [39] or the model [135]; evaluating the model's fairness [51];

and obtaining new insights from large-scale studies [218]. There are two general approaches to model explanation: (1) developing *interpretable* models and (2) *posthoc explaining* a pre-trained model. Interpretable predictive models are constrained to make their reasoning processes more understandable to humans, making them much easier to troubleshoot and use in practice [220]. However, such models may impose simplifications to ensure interpretability and thus achieve a lower predictive accuracy [9, 263]. Other times, they have a complicated design, making them difficult to train [25]. Another popular line of work builds attention-based interpretability into the DNN. These methods use attention weights to highlight parts of an input that the network focuses on while making its decision [284, 63].

Post-hoc explanation aims to improve human understanding of a pre-trained model [51, 84, 120, 175, 297]. Hence, the performance of the model is not compromised. Post-hoc explanation comprises several broad approaches, such as example-based explanations [123, 121]; approximating DNN models with simpler models [212, 213]; understanding feature attribution [255, 151] or importance [231, 125, 10, 61]. These methods provide a local (image-level) or a global (target label-level) explanation. They explain by highlighting the critical regions (*where*) for the classification decisions. However, the location information alone is insufficient for applications in medical imaging. My second project focuses on providing posthoc model explanations in the form of counterfactual images. Counterfactual images show *what* image features are important for the classification decision and *how* to modify the critical features to change the classifier's decision.

Recently, researchers have focused on providing explanations that resemble a domain expert's decision-making process. Very often, such explanations are expressed using human-understandable concepts or terminology [291, 81, 34]. Existing approaches for concept-based explanation depend on explicit concept annotations. The concept annotations are provided either as a representative set of images [121] or as semantic segmentation [12]. Such annotations are expensive to acquire, especially in the medical domain. Furthermore, these methods measure correlations between concept perturbations and classification predictions to quantify the concept's relevance [121, 12, 304]. However, NN may not use the information from learned concepts to arrive at its decision. My third project focuses on using counterfactual images to quantify the causal effect of a concept on the classification decision.

1.3 DNN model uncertainty quantification

Most DNN models are deterministic functions that provide point estimates of parameters and predictions. In the absence of probabilistic distribution, it is essential to convey the uncertainty in the DNN model's decision to the end-user [188, 85]. For example, consider a DNN model trained on adult face images, predicting whether the person is young or old. Such a model should refrain from making an overconfident decision on ambiguous images from middle-aged people or out-of-distribution (OOD) images of children and animals [179, 267]. Uncertainty in prediction may arise from noisy data or data in high class-overlap regions, leading to *aleatoric uncertainty* [2]. The other kind of uncertainty is related to the model, *epistemic uncertainty*, that arises from the model's limited information on unseen data or due to a mismatch between training and testing data distributions [104]. Communicating predictive uncertainty can help the end-user understand the predictions better, anticipate when uncertainty is irreducible, prioritize gathering more data to reduce model uncertainty, and decide when to discard the model prediction and rely on expert knowledge [69].

Much of the prior work focused on deriving uncertainty measurements from a pre-trained DNN output [92, 85, 142, 148], feature representations [144, 139], or gradients [100]. Such methods use a threshold-based scoring function to identify OOD samples. The scoring function is derived from softmax confidence scores [92], scaled logit [85, 144], energy-based scores [148, 277], or gradient-based scores [100]. These methods help in identifying OOD samples but did not address the over-confidence problem of DNN, which made identifying OOD non-trivial in the first place [91, 188]. My final project focuses on mitigating the over-confidence issue in a pre-trained classifier by efficiently capturing both epistemic or aleatoric uncertainty.

1.4 Explanation Framework

This dissertation propose an **explanation framework** to explain the DNN classification model's decision. The explanation framework has two primary models. The first is an *inter-*

pretable model that uses a carefully designed attention mechanism to provide interpretability while achieving high predictive performance. The second is a *progressive counterfactual explainer* (PCE) that provides a posthoc explanation for a pre-trained classifier. A counterfactual image is a perturbation of the input image with an opposite classification decision compared to an input image. It shows what imaging features are present in salient locations and how changing such features modify the classification decision. The generative explainer is constrained to create natural-looking images as explanations that resemble medical-imaging data, thus ensuring the clinical usability of our explanations. My work presents a thorough human-grounded experiment with diagnostic radiology residents to compare different styles of explanations (no explanation, saliency map, cycleGAN explanation, and our counterfactual explanation) by evaluating different aspects of explanations. The results show that the counterfactual explanations from my proposed method, were the only explanations that significantly improved the users' understanding of the classifier's decision compared to the no-explanation baseline.

Further, in my next project, I extended the explanation framework to support two applications. The first application focuses on enriching the explanation with conceptual information. Specifically, I integrated counterfactual explanations with tools from Causal Inference literature [106] to quantify the causal relationship between the building units of a DNN, neurons, and clinically relevant concepts [273]. The weak annotations from radiology reports were used to derive concept annotations. The second application focuses on fixing an over-confident pre-trained classifier. The counterfactual images derived from PCE were used to fine-tune the classifier. The empirical results show that fine-tuning helps in smoothing the decision boundary and helps in preventing the classifier from being over-confident on samples near the decision boundary. Further, the discriminator of the GAN-generator was used to provide a density score to identify OOD samples.

1.5 Dissertation structure

Chapter 2 provides a detail literature review on different deep learning models in medical imaging. It also provides a thorough background on different paradigms of deep learning model explanation.

Chapter 3 proposes an attention-based DNN model aggregating local image features from volumetric medical images into a compact latent representation. This representation is then used to predict multiple patient-relevant outcomes such as symptom scores and disease severity. The model provides interpretability by learning an attention weight for each anatomical feature that reflects its contribution to the final prediction decision. We evaluated our proposed model in a large clinical study of over 10K participants with chronic obstructive pulmonary disease (COPD). Our results show that our model independently predicted spirometric obstruction, emphysema severity, exacerbation risk, and mortality from CT imaging alone.

Chapter 4 proposes a Progressive Counterfactual Explainer (PCE) to explain the decision of a pre-trained image classifier. The explainer generates a progressive set of perturbations to a query image, such that the classification decision changes from its original class to its negation. We used counterfactual explanations derived from our framework to audit a classifier. We conducted experiments on a natural image dataset of face images and a medical chest x-ray (CXR) dataset. To quantitatively evaluate our explanations, we proposed new metrics that consider the clinical definition of a target disease while comparing counterfactual changes between normal and abnormal populations, as identified by the classifier.

We conducted a human-grounded experiment with diagnostic radiology residents to compare different styles of explanations (no explanation, saliency map, cycleGAN explanation, and our counterfactual explanation) by evaluating different aspects of explanations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users. Our results show that our counterfactual explanation was the only explanation method that significantly improved the users' understanding of the classifier's decision compared to the no-explanation baseline. Our metrics established a benchmark for evaluating model explanation methods in medical images. Our explanations revealed that the classifier relied on clinically relevant radiographic features for its diagnostic decisions, thus making its decision-making process transparent to the end-user.

Chapter 5 shows an application of PCE to provide concept-based explanations. In this project, we aim to quantify causal associations between the hidden units of the DNN and human-understandable concepts [121, 273]. We take advantage of radiology reports accompanying the chest X-ray images to define concepts. First, we solve sparse linear logistic regression to identify hidden units that are positively correlated with the presence of a concept. Next, we viewed these concept units as a mediator in the treatment-mediator-outcome framework [106] from mediation analysis. Using PCE to define counterfactual interventions, we measure the in-direct causal effect of a concept on the network's prediction. Finally, we present our findings as a low-depth decision tree over causally relevant concepts, providing the global explanation for the model in the form of clinically relevant decision rules.

Chapter 6 demonstrates an application of PCE in improving the uncertainty quantification of an existing pre-trained DNN. Ideally, the DNN model's output should reflect its confidence in its decision. This project proposes fine-tuning an existing pre-trained classifier on counterfactually augmented data (CAD) generated using PCE to improve its uncertainty estimates. Further, the GAN-PCE discriminator helps identify and reject far-OOD samples. In our experiments, we out-performed state-of-the-art methods for uncertainty quantification on multiple datasets with varying difficulty levels. Chapter 7 summarizes this thesis and suggests future extensions.

All chapters in this dissertation address unique DNN challenges motivated by specific clinical requirements. We investigated and explored efficient DNN architectural, explanation and training paradigms while keeping our end-users "clinicians" in focus. At the same time, the methods developed in this research have broad applicability and have been used by many researchers in different domains [293]. We have released open-source implementations of all these methods.

1.6 Contributions

The most notable contributions of this dissertation are the development of:

1. An interpretable, attention-based DNN architecture that processes an entire 3D volumetric image without any resizing and predicts multiple disease outcomes with high predictive accuracy (summarized in Chapter 3).
2. A new posthoc explainability method that provides visual counterfactual explanations. These explanations not only highlight the important regions but also shows how the image features should be transformed to flip the classification decision (summarized in Chapter 4).
3. A concept-based explanation method that explains the classification decision in terms of clinically relevant concepts. This method uses the explanations from the technique described in Chapter 2 to quantify the causal effect of a concept on the network's prediction (summarized in Chapter 5).
4. A methodology to fine-tune the existing DNN on counterfactually augmented data to improve its estimates for aleatoric uncertainty. Further, using the discriminator of the GAN-counterfactual explainer as a selection function to identify and reject samples with high epistemic uncertainty (summarized in Chapter 6).

1.7 List of publications

Material presented in this dissertation has been published in peer-reviewed conferences and journal papers:

1. Singla, S.; Gong, M.; Ravanbakhsh, S.; Sciurba, F.; Poczos, B. and Batmanghelich, K.N.;,"Subject2Vec: Generative-Discriminative Approach from a Set of Image Patches to a Vector," MICCAI, Part I, pp. 502-510, September 2018. [240]
2. Singla, S.; Gong, M.; Riley, C.; Sciurba, F. and Batmanghelich, K.N.;, "Improving clinical disease subtyping and future events prediction through a chest CT-based deep learning

approach," Medical physics, 48, no 3, pp. 1168-1181, March 2021. [241]

3. Singla, S.; Pollack, B.; Chen, J. and Batmanghelich, K.N.;, "Explanation by Progressive Exaggeration," International Conference on Learning Representations, September 2019. [242]

4. Singla, S.; Pollack, B.; Wallace, S. and Batmanghelich, K.N.;, "Explaining the Black-box Smoothly-A Counterfactual Approach," arXiv e-prints, pp.arXiv-2101, Jan 2021. [243]

5. Singla, S.; Wallace, S.; Triantafillou, S. and Batmanghelich, K.N.;, "Using Causal Analysis for Conceptual Deep Learning Explanation," MICCAI, Vol. 12903, pp. 519-528, January 2021. [244]

6. Singla, S.; Murali, N; Arabshahi, F; S.; Triantafillou, S. and Batmanghelich, K.N.;, "Augmentation by Counterfactual Explanation - Fixing an Overconfident Classifier," under review WACV, 2022.

2.0 Literature Review

2.1 Deep learning for medical imaging

With the expanding development of deep learning (DL) techniques, utilizing advanced deep neural networks (DNNs) for medical image analysis has become an active field of research. DNNs have shown superior performance over clinicians in many tasks, primarily due to the availability of large training datasets and increased computational power. Applications of DL in medical image analysis involve different computer vision-related problems such as classification [288], detection [286], segmentation [77], and registration [55]. Among them, we primarily focus on classification methods that classify an input image or a series of images with diagnostic labels of some predefined diseases [74, 200]. Traditional computer algorithms for image classification use feature extractors and statistical models that translate human intuition into handcrafted features [47, 8, 105]. These features were then used in a supervised setting to train specialized image classifiers. In contrast, DNN models follow a data-driven approach and learn to optimally represent the data for the given classification task with minimum human intervention. The resulting models are complex functions with millions of parameters but are much more accurate, efficient, generalizable, and easier to scale.

The commonly used image modalities for diagnostic analysis in clinic include projection imaging such as X-ray imaging and computed tomography (CT). As a working example, we focus on DNN models that are developed for chest imaging. Chest CT imaging comprises a continuous sequence of 2D slices that vary in depth and resolution with changes in patient and scanner settings. Many DL architectural designs have been applied to applications in CT analysis to solve specific clinical tasks such as nodule detection [269], fibrosis [31], emphysema [103], COPD [258], and cancer diagnosis [252]. The most common setting is to sub-sample 2D slices from volumetric images and concatenate, join, or crop them in different ways to create a 2D image [252, 74, 247, 6]. The primary motivation behind this rescaling is to make the input images compatible with the classical DNN architectures, which were originally designed for natural images [90, 239]. Extensive pre-processing pipelines are proposed which include sub-sampling spatially aligned CT volumes into three slices in either axial, sagittal or coronal directions, to accommodate for the RGB input [258]. Distortion of CT imaging may lead to undesired artefacts and information loss, leading to sub-optimal performance.

Further, researchers have explored variants of recurrent neural networks (RNN) [296] to process consecutive slices from sub-sampled 3D volumes [103, 59]. These methods include using long short-term memory (LSTM) networks capable of learning dependencies between a sequence of images [50]. Such algorithms can take multiple slices as input and provide better global features for downstream use-cases, such as classification. More recently, efforts are being made to integrate various designs, such as 3D CNN with RNN [290]; multiple resolution CNNs with two, two-and-a-half, and three-dimensional architectures [15, 202, 32]; and 3D multi-scale capsule networks [4]. These methods aim to better capture information from 3D imaging at different spatial resolutions with minimum information loss. Their primary motivation is high discriminative performance, while little attention is paid to models' interpretability.

Although deep learning models have achieved great success in medical image analysis, minimum interpretability is still the main bottleneck in the clinical deployment of these methods [222]. The key benefit of DNN is that it identifies essential features without human intervention. However, this makes the model opaque as the end-user has no intuition on how the decision was being made. The legal ramifications of black-box functionality could have severe consequences; hence, healthcare professionals may decline to work with such systems.

2.2 Interpretable deep learning

The overarching goal of any deep learning method for medical imaging is to aid clinicians in their workflow by increasing their efficiency by removing redundancies [145]. This requires a partnership between the clinical experts and the AI system, which in turn requires the clinical experts' trust. Interpretability, or the ability of a DNN model to explain its outcomes and assist clinicians in rationalizing the model prediction, is critical to establishing trust [261]. Interpretable DL models aim to incorporate interpretability during the design process of the DNN and, thus, alter the network structure to encourage interpretability. They learn to provide both prediction, and explanation are gaining the interest of the medical research community.For example, DNN have been designed to perform case-based reasoning [25], to incorporate logical structures [282], to incorporate hard attention to do classification [54], and to learn a disentangled latent space [28]. One of the early methods modified the CNN architecture to extract prototypical examples [25]. In another attempt, Song *et al.*proposed a student-teacher network, where one network is optimized for superior interpretability while the other network is trained to achieve high discriminative performance [41]. Some methods provide interpretability by performing multi-modality learning by integrating radiology reports [300] or electronic health record data [101, 136]. This data provides additional information for assisting clinical decision-making.

Furthermore, many variants of the attention-based model have been proposed that learn an attention mechanism to highlight the most relevant part of the input for the prediction decision [228]. For instant, Choi *et al.*proposed a multi-level attention model on time series data for detecting influential past visits, and clinical variables while predicting diagnosis. Another example is interpretable R-CNN [282], which is an object detection-based DNN that provides a classification score and a bounding box on the region of interest. Recently, a concept whitening approach was proposed that learns a DNN where the latent space of each layer is aligned with a known set of concepts [28]. Creating an interpretable model is much more complicated than a black-box model, as it involves solving a complex optimization problem while satisfying the interpretability constraints. Nevertheless, the benefits of having an explanation built into the model have far better deployment prospects than a highly accurate but opaque model.

2.3 Post-hoc deep learning model explanation

Post-hoc explanation methods provide explanations for the predictions after the DL model has been trained. Such methods can provide local or global explanations. Local explanations provide explanation for individual data point. It identifies attributes or features in a particular image that are important for the DNN model's prediction. On the other hand, global explanations aim at providing an overall summarization of the model behaviour for a particular class. Post-hoc explanation methods can be model-specific *i.e.*, they are applicable to only certain types of models and require access to model-specific information. On the other hand, they can be model-agnostic methods, that is they are applicable to any DNN model in general.

2.3.1 Feature attribution-based explanation

Feature attribution methods provides an explanation as a saliency map that reflects the importance of each input component (*e.g.*, pixel) to the classification decision. Saliency-based methods are the most common form of post hoc explanations for neural networks. Gradient-based methods for obtaining saliency maps is mostly DNN-specific and provides local explanation [236, 255, 151]. Some earlier work in this direction [237, 249, 10] focuses on computing the gradient of the target class with respect to input image and considers the image regions with large gradients as most informative. Building on this work, the class activation map (CAM) [303] and its generalized version Grad-CAM [231] uses the gradients of the target class, flowing into the final convolutional layer to produce a saliency map. The Layer-Wise Relevance Propagation (LRP) [10] method back-propagates a class specific error signal through the DNN and considered its product with each convolutional layer's activation to derive the saliency map. DeepLift [236] is a version of LRP method that back-propagates the contribution back to every feature of the input. The above gradient-based methods are not model-agnostic and require access to intermediate layers. Recently, [3] have showed that some saliency methods are independent both of the model and of the data generating process. The saliency maps are also prone to adversarial attacks as shown by [71] and [125].

In another line of work, perturbation-based methods provide interpretation by showing what minimal changes are required in input image to induce a desirable classification output. Some methods employed image manipulation via the removal of image patches [302, 297], masking with constant values [42, 204] or the occlusion of image regions [302] to change the classification score. Recently, the use of influence function, as proposed by [129] are applied as a form of data perturbation to modify a classifier's response. The authors in [61] proposed the use of optimal perturbation, defined as removing the smallest possible image region that results in the maximum drop in classification score. In another approach, [24] proposed a generative process to find and fill the image regions that correspond to the largest change in the decision output of a classifier. To switch the decision of a classifier, [79] suggested generating counterfactuals by replacing the image regions with patches from images with a different class label. All of the aforementioned works perform pixel- or patch-level manipulation to input image, which may not result in natural-looking images. Especially for medical images, such perturbations may introduce anatomically implausible features or textures.

Another interesting approach is to use game theory to compute the Shapley value of each pixel as its marginal contribution to the final prediction decision [151, 254]. The idea of Shapely values is that all features cooperate to produce model prediction. This is a local interpretation method that can be either model agnostic or model specific depending on the formulation. Classical SHAP method required repeated predictions from the model, as it exhaustively try all possible configuration of the features. This is computationally expensive and hence, multiple approximations are been proposed [26].

Saliency map-based methods are frequently applied to the medical imaging studies, *e.g.*, chest x-rays [208], skin imaging [294], brain MRI [53] and retinopathy [226]. Saliency maps lack a clear interpretation and provide incomplete explanation especially when different diagnoses affect the same regions of the anatomy. Although objects in natural images have a distinct appearance and are easier to identify and isolate by humans, the visual variations in different diagnoses, in medical images, are very subtle and require expert observation. Thus, very similar explanations are given for multiple diagnosis, and often none of them are useful explanations [219].

2.3.2 Counterfactual explanation

Recently, researchers have explored generative models that provides explanation by modifying existing examples [79] or generating new examples [224, 114]. A popular direction is to generate counterfactual explanations. Counterfactual explanations are a type of contrastive [46] explanation that are generated by perturbing the real data such that the classifier's prediction is flipped. Similar to our method, generative models like GANs and variational autoencoders (VAE) are used to compute interventions that generate realistic counterfactual explanations [224, 114, 147, 157, 270, 199, 5]. Much of this work is limited to simpler image datasets like MNIST, celebA [147, 157, 270] or simulated data[199]. An extension of these methods on large datasets will actually show their scalability and generalizability strengths. This work is yet to be explored by the community in general and provides a great venue for future exploration.

For more complex natural images, previous studies [24, 5] focused on finding and in-filling salient regions, in order to generate counterfactual images. In contrast, at inference time, our explanation model doesn't require any re-training for generating explanations for a new image. In another line of work [278, 79] provide counterfactual explanations that explains both the predicted and the counter class. Recently [185, 44] used a cycle-GAN [305] to perform image-to-image translation between normal and abnormal images. While images generated by such independently trained GANs may look realistic, these generative models are not explicitly coupled to the classifier that they are aiming to explain. Hence, cycle-GAN may end up learning features that do not reflect the true behaviour of the classifier. In contrast, our model uses the classifier's predicted probabilities and gradients during the training of the GAN-generator, and hence the generated images are tied to the classifier.

Since the inception of our work, various extensions to our counterfactual generation process have been proposed. These include adding support for creating diverse and multiple counterfactual explanations [214, 70], enhancing compatibility with smaller datasets [117] and inducing a bijective transformation through normalizing flow [49].

2.3.3 Concept-based explanation

Concept learning was used in traditional machine learning to identify and classify samples based on a list of concepts. A concept, in this case is a feature that is discriminative and whose presence is highly associated with the presence of a class label. To summarize, a concept is semantically meaningful attribute that is visually coherent across images and is important for the prediction of a given class *i.e.*, its presence is a necessary condition for the classification decision to be true for a given class [72]. Another benefit of concept-based explanation is, usually concepts are high level attributes that are mentioned in human-friendly manner. Recent studies have thus focused on bringing such concept-based explainability to DNNs.

Concept-based explanation methods aim to recover concept information from the intermediate DNN activations and then relate them to the classification decision and the data. The essential first step towards deriving concept-based explanations is defining concepts. Some methods used human-labelled supervised data to mark the salient concepts [122, 304], while others used purely unsupervised approaches such as clustering of the DNN activations [72]. TCAV [122] learns a concept classifier by training a linear classification model on the activations of an arbitrary intermediate layer, using the ground truth labels for each concept. Gohorbani *et al.*extended TCAV to used self-supervised labels obtained from automatically super-pixel segmentation followed by k-means clustering. Zhou *et al.* [304] decomposes the prediction of one image into multiple human-interpretable conceptual components. Concept activation vectors (CAV) are used in medical imaging analysis for solving particular tasks such as retina disease diagnosis [259], skin lesion classification [150], breast tumor detection [215], cardiac MRI classification [34], tumor segmentation in liver CT [37] and radiomics [291].

In another line of work, researchers explore training both a classification model and a concept classifier to obtain an inherently explainable model [17]. Similar to this, concept bottleneck method [130] first learns to predict the concepts, then uses only those predicted concepts to make a final prediction [221]. Such approach are popular in medical domain, with applications in lung nodule malignancy classification [234]. Goyal *et al.*measures the causal effect of concepts by using a conditional VAE model [78]. Measuring the causal effect is essential, as presence of concept information in the latent space of the DNN doesn't necessarily means the network is using that information to make its decision. To provide causal explanation, Harradon *et al.*build a bayesian causal model using these extracted concepts as variables in order to explain image classification [87].

2.4 DNN model uncertainty quantification

To facilitate the real-world deployment of a DNN model, it is essentially important to understand what a DNN model does not know. State-of-the-art classification models are mostly DNN such as DenseNet, ResNet and more. These models are deterministic, in the sense they only provide point estimates for the posterior. The gold standard for UQ is *Bayesian Neural Network* (BNN) [186]. BNN are an alternative to DNN, as they provide a distribution over the model parameters which helps in quantifying model uncertainty. However, computing this information comes at an extra computational cost while also increasing the inference complexity [37, 151]. Moreover, training a BNN is often intractable, and they arguably result in sub-optimal accuracy as compared to deterministic approaches. This is perhaps due to the difficulty in tuning their hyper-parameters [280].

Alternatively, approaches such as *Deep Ensembles* [133] and *MC Dropout* [65] has been introduced as an approximation of BNN that are compatible with the deterministic DNN architecture with minimal changes at the inference time. Deep ensembles require training multiple copies of the DNN with either random initialization of the weights or the training data or both. In MC-dropout, weights are randomly dropped at training as well as inference time. Deep ensembles, however, require multiple DNN models to be trained using different initialization seeds, making them computationally expensive to train. MC Dropout is computationally less expensive, but cannot be used for UQ in pre-trained models that are trained without dropout.

The recent interest in single forward pass UQ techniques [267, 266] has led to less expensive alternatives for MC Dropout. However, they require a DNN to be trained from scratch using specific constraints or loss functions, and hence cannot be used to fix pre-trained DNNs with poor uncertainty estimates. Further, [118] proposed a novel method that learns the observation noise parameter, which enables it to model both epistemic and aleatoric uncertainty in a single forward pass. Model uncertainty or epistemic uncertainty [64], measure the uncertainty in estimating the DNN model parameters given the training data. Epistemic uncertainty measures how well the model learns the data. It is reducible as the size of the training data increases. Data uncertainty, or aleatoric uncertainty [64], is irreducible uncertainty that arises from the natural complexity of the data, such as class overlap or label noise. Data uncertainty is also considered as a 'known-unknown' *i.e.*, the DNN model understands (knows) the data distribution and can confidently predict whether a given input is difficult to classify ı.e an unknown [159]. However, epistemic uncertainty may also arise when there is a mis-match between the training and testing data distribution. This is 'unknown-unknown' as the model is unfamiliar with the test data and hence, cannot confidently make predictions.

2.4.1 Uncertainty quantification in pre-trained DNN models

Much of the prior work focused on deriving uncertainty measurements from a pre-trained DNN output [92, 85, 142, 148], feature representations [144, 139] or gradients [100]. Such methods use a threshold-based scoring function to identify OOD samples. A baseline method for OOD detection was introduced by Hendrycks *et al.*. They showed that simple statistics derived from softmax distributions provide an effective way to identify out of distribution (OOD) data [92]. Guo*et al.*extended this work by demonstrating that a single-parameter variant of Platt scaling, also known as temperature scaling is an effective method to obtain calibrated probabilities, which in turn helps in better OOD detection [85]. Very recently, researchers have proposed energy-based scores for OOD detection [148, 277]. The energy score helps in mitigating a critical problem with softmax confidence that assigns arbitrarily high values for OOD examples. Further, several works are been proposed that attempts to improve the OOD uncertainty quantification by using ODIN score [142] and its variant [96]. Specifically, ODIN proposed adding small perturbations to the input and gradually increasing the softmax score of any input by reinforcing the model's belief in the predicted label.

Further, proposed to use Mahalanobis distance-based confidence score to identify and reject OOD samples [139].

In another attempt, Huang *et al.*proposed to use GradNorm, a simple and for detecting OOD inputs by utilizing information extracted from the gradient space. Gradient norm uses the vector norm of gradients, backpropogated from the KL divergence between the softmax output and uniform probability distribution. All these methods help in identifying OOD samples but did not address the over-confidence problem of DNN, that made identifying OOD non-trivial in the first place [91, 188]. Our work focuses on mitigating the over-confidence issue by fine-tuning a pre-trained classifier on counterfactually augmented data (CAD). Further, we used the discriminator of the GAN-generator to provide a density score to identify OOD samples.

2.4.2 DNN designs for improved uncertainty estimation

Designing generalized DNN that provides robust uncertainty estimates has gained significant research attention. The Bayesian neural networks are the gold standard for reliable uncertainty quantification [186]. Multiple approximate Bayesian approaches have been proposed to achieve tractable inference and to reduce computational complexity [80, 16, 127, 65]. Popular non-Bayesian methods include deep ensembles [133] and their variant [97, 66]. However, most of these methods are computationally expensive and requires multiple passes during inference. An alternative approach is to modify DNN training [256, 299, 271], loss function [181], architecture [253, 146, 68] or end-layers [267, 96] to support improved uncertainty estimates in a single forward-pass. Further, methods such as DUQ [267] and DDU [179] proposed modifications to enable the separation between aleatoric and epistemic uncertainty. Unlike these methods, our approach improves the uncertainty estimates of any existing pre-trained classifier, without changing its architecture or training procedure. We used the discriminative head of the fine-tuned classifier to capture aleatoric uncertainty and the density estimation from the GAN-generator to capture epistemic uncertainty.

2.4.3 Uncertainty estimation using GAN

A popular technique to fix an over-confident classifier is to regularize the model with an auxiliary OOD data which is either realistic [93, 174, 198, 27, 142] or is generated using GAN [210, 138, 160, 285, 232]. Such regularization helps the classifier to assign lower confidence to anomalous samples, which usually lies in the low-density regions. On of the earlier methods proposed outlier exposure (OE) that leverages diverse, realistic datasets for exposing the model training to OOD distribution [93]. Chen *et al.*showed that randomly selecting outlier samples for training may yield uninformative samples. They proposed an adversarial training with informative outlier mining (ATOM) technique to selectively collect auxiliary outlier data for estimating a tight decision boundary between ID and OOD data, which leads to robust OOD detection performance [27].

Another line of researchers investigate deep generative model based approaches for OOD detection. Such methods use generative modeling to detect OOD samples by setting a threshold on the likelihood. An application of generative model such as GAN in OOD detection is the use of entropy loss in the construction of an OD detector for generalized zero-shot action recognition [160]. They learn an OOD detector using real and GAN-generated features from seen and unseen categories, respectively. In another attempt, Ren *et al.*propose the use of a likelihood-ratio test by taking the ratio between the likelihood obtained from the model and from a background model which is trained on random perturbations of input data [210]. Further, [232] proposed to offset the bias of the generative models by a factor that measures the input complexity, such as the length of lossless compression of the image. Further, [210, 232, 223] obtain high OOD detection performances with Glow, VAE and Pixel-CNN generative models.

Defining the scope of OOD a-priori is generally hard and can potentially cause a selection bias in the learning. Alternative approaches resort to estimating in-distribution density [250]. Our work fixed the scope of GAN-generation to CAD [242]. Rather than merging the classifier and the GAN training, we train the GAN in a post-hoc manner to explain the decision of an existing classifier. This strategy defines OOD in the context of pre-trained classifier's decision boundary. Previously, training with CAD have shown to improved generalization performance on OOD samples [119]. However, much of this work is limited to Natural Language Processing, and requires human intervention while curating CAD [118]. In contrast, we train a GAN-based counterfactual explainer [243, 134] to derive CAD.

2.4.4 Data augmentation for improving uncertainty estimation

There is a rich literature on data augmentation (DA) for improving the classification performance of DNNs [45, 40, 301, 235]. However, most of the classical DA literature is task agnostic and focused on improving accuracy. While GAN-based DA is popular, they are mainly used to generate samples that are consistent with the underlying distribution without taking the DNN into account. In contrast, our GAN-based augmentation network is closely coupled with the pre-trained DNN, and generates samples in ambiguous regions of the distribution to enhance the uncertainty characteristics of the pre-trained model. We take inspiration from recent works [242, 243, 134] on counterfactual explanations which focus on explaining a DNN. However, they do not explore whether the generated samples can improve a downstream task. Additionally, there is research showing that models trained on counterfactually augmented data have improved generalization performance on out-of-domain samples [119]. However, much of this work is limited to Natural Language Processing, and our work differs in terms of both the application and the architecture we use for our proposed method.

3.0 Improving Clinical Disease Sub-typing and Future Events Prediction through a Chest CT based Deep Learning Approach

3.1 Introduction

Chronic obstructive pulmonary disease (COPD) is characterized by persistent respiratory symptoms and irreversible airflow obstruction [274]. The measurement of spirometric obstruction, while traditionally used to define COPD severity, is not sufficient to explain the many critical dimensions required to characterize and manage COPD [38]. Airflow obstruction can result from varying combinations of emphysematous parenchymal destruction [194], chronic airway remodelling [83], and other poorly characterized imaging patterns, including fibrotic changes common in smokers [279]. Hence, clinicians must adopt a comprehensive approach while assessing patients with COPD, including identifying risk factors, standardized assessment of symptoms and comorbidities, estimating exacerbation risk [246], and prognostication of survival. Other established tools for assessing COPD symptoms are the modified Medical Research Council (mMRC) dyspnea scale and prognostication of survival using the body mass index, obstruction, dyspnea and exercise capacity (BODE) index [21, 162]. Though radiography has not been historically utilized in routine diagnosis or management of COPD [196], the growing use of CT imaging for pulmonary nodule assessment and cancer screening [257, 197], provides a novel opportunity to leverage imaging data to investigate patients with COPD.

Despite much interest in using CT imaging in subtyping COPD [153], stratification of patients as obstructed or non-obstructed is currently based on spirometric pulmonary function testing findings according to the Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines [274]. Much of the clinical workflows rely heavily on qualitative visual assessment for characterizing COPD. Visual assessment includes identifying image features highlighting air trapping in small airways [164], characterizing local patterns for emphysema [89, 182], bronchial wall thickening, or endobronchial mucus [124], and calculating the percentage of low attenuation area (LAA) [192], blood vessel volume[56], or airway counts[47]. Also, various intensity and texture-based feature descriptors are proposed to characterize the visual appearance of COPD [29, 248, 289]. But most of these image features are generic and are not necessarily optimized for characterizing COPD. Furthermore, some of these methods rely on manual segmentation methods and are thus both labour-intensive and prone to operator error[153, 154, 173, 182].

While visual CT analysis remains the mainstay of clinical imaging interpretation, there has been growing research interest in quantitative image analysis techniques to quantify abnormalities on CT and characterize disease subtypes [103]. Recent advances in deep learning (DL) enable researchers to go directly from raw images to clinical outcomes without specifying radiological features [74]. However, most of the existing work concentrate on some aspect of COPD disease like only spirometry or only emphysema or COPD sub-typing [258]. There is room for improvement to bring the prediction of multiple patient-centred outcomes to quantify COPD. Further, much impact can be made by predicting patients' future exacerbation or survival, thus providing helpful input to construct personalized treatment plans.

This paper proposes a novel DL model that takes an entire 3D volumetric image as input and provides a holistic view of a patient's health in terms of multiple COPD outcomes. Our novel DL model followed a data-driven approach and directly analyzed raw HRCT data without manually segmenting or specifying radiological features. Previous, DL approaches [74] processed slices (three orthogonal slices) of CT images and hence may not be able to characterize the volumetric impact of the disease. In contrast, our proposed method views each subject as a *set* of image patches from the lung region. It can analyze the entire 3D CT scan and requires no image distortion due to resizing or cropping. Previously, [29, 227] also viewed CT images as a set and extracted handcrafted image features from each input element. In contrast, the *discriminative* part of our model uses a deep learning approach and directly extracts features from the volumetric patches. Further, we use an attention mechanism [287] to adaptively weigh local features and build the subject level representation, which is predictive of the disease severity. Our model is inspired by the Deep Set [295]. We extend it by adapting *generative* regularization, which prevents the redundancy of the hidden features. Furthermore, the *attention mechanism* provides interpretability by quantifying the relevance of a region to the disease.

We predict multiple patient-relevant outcomes such as symptom scores, emphysema severity and pattern, exacerbation risk, and mortality. When compared to other DL model [74], our method improved the prediction of important clinical variables, such as COPD disease severity and exacerbation risk. Furthermore, it can distinguish between centrilobular and paraseptal emphysema and quantify the future risk of exacerbation based on the current CT image. Estimating these clinically relevant features using only CT images has a potential application both to clinical care and research.

3.2 Method

We represent each subject as a set (bag) of volumetric image patches extracted from the lung region $\mathcal{X}_i = \{x_{ij}\}_{j=1}^{N_i}$, where N_i is the number of patches for subject i, which varies with subject.The model learned to extract informative regional features from these patches x_{ij}, and then adaptively weight these features to form a fix-length representation for each patient. This patient-representation is then used to predict disease severity (y_i). The general idea of our approach is shown in Figure.1.

The method consists of three networks that are trained jointly: (1) a *discriminative* network, that aggregates the local information from patches in the set $\mathcal{X}_i$ to predict the disease severity y_i, (2) an *attention* mechanism, that helps discriminative network to selectively focus on patch-features by assigning weights to the patches in $\mathcal{X}_i$, and (3) a *generative* network, that regularizes the discriminative network to avoid redundant representation of patches in the latent space. The model is trained end to end, by minimizing the below objective function:

$$\min_{\omega,\theta_e,\theta_d,\theta_a} \sum_i \mathcal{L}_d\left(y_i, \hat{y}_i(\mathcal{X}_i); \theta_e, \omega\right) + \lambda_1 \mathcal{L}_g\left(\mathcal{X}_i, \hat{\mathcal{X}}_i; \theta_e, \theta_d\right) + \lambda_2 \mathcal{R}\left(\mathcal{X}_i; \theta_e, \theta_a\right), \qquad (1)$$

where $\mathcal{L}_d(\cdot,\cdot)$ and $\mathcal{L}_g(\cdot,\cdot)$ are the discriminative and generative loss functions respectively and $\mathcal{R}(\cdot)$ is a regularization over the attention. The θ_e, θ_d, θ_a and ω are the parameters of each term. λ_1, λ_2 controls the balance between the terms. The sum is over number of subjects. Next, we discuss each term in more detail.

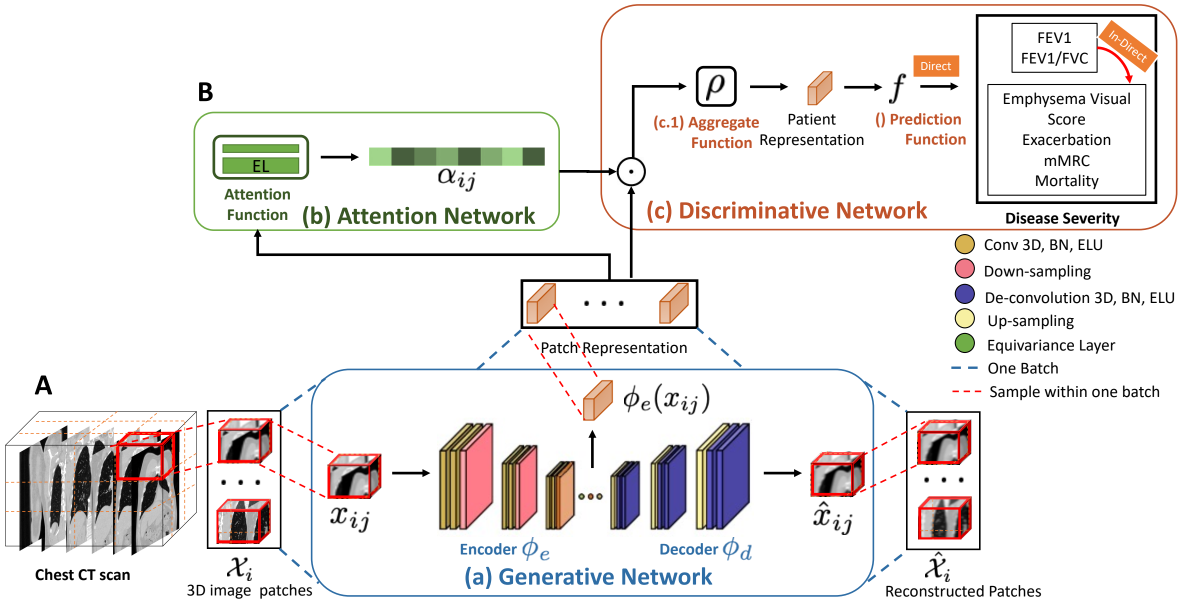


Figure 1: The schematic of our model. **A.** The input to our model is a 3D CT scan of the lung. The lung is divided into a set of equally sized, overlapping 3D image patches. (a) The **Generative Network** is a convolutional auto-encoder (CAE). The encoder function projects the raw image patch to a latent space and the decoder function reconstructs the image patch from the extracted latent features. **B.** The **Attention Network** provides interpretability by weighting the patches based on their importance in predicting the disease severity. **C.** The **Discriminative Network** (c.1) aggregates the local patch-level information information, based on their attention weights, to create a patient-level representation, and (c.2) uses it to predict disease severity.

3.2.1 Generative Network

The Generative Network is a convolutional auto-encoder (CAE)[163]. CAE consists of an encoder $\phi_e(\cdot)$, that extracts local image features from each patch $\big(i.e., \phi_e(x_{ij};\theta_e) \in \mathbb{R}^d\big)$. These features are a summarization of the information in the raw image patch (or region) in a low dimensional "feature space". To regularize the feature extraction process, CAE have a decoder $\phi_d(\cdot)$. The decoder recovers the input patch back from the low dimensional feature space as $\hat{x}_{ij} = \phi_d(\phi_e(x_{ij}; \theta_e); \theta_d)$. In the absence of the decoder function, the feature extractor ϕ_e will be forced to retain only information that is sufficient for the underlying task of predicting y. If y is low dimensional as compared to d, ϕ_e learns a highly redundant latent space representation for each patch. To prevent this information loss, we regularize the auto-encoder using a distance loss defined as, $\mathcal{L}_g(\mathcal{X}_i, \hat{\mathcal{X}}_i; \theta_e, \theta_d) = \frac{1}{|\mathcal{X}_i|} \sum_{x_{ij} \in \mathcal{X}_i} ||x_{ij} - \hat{x}_{ij}||_2$.

3.2.2 Attention Network

The goal of our proposed model is twofold: first to provide a prediction of the disease severity and secondly, to provide a qualitative assessment of our prediction. Here, it is reasonable to assume that different regions in the lung contribute differently to the disease severity. We model this contribution by adaptively weighting the patches. The weight indicates the importance of a patch in predicting the overall disease severity of the lung. This idea is similar to attention mechanism in Computer Vision [287] and Natural Language Processing [152] communities.

The goal of the attention network is to learn a weight for each of the input image patches, such that the weight indicates the importance of a patch in predicting the overall disease severity of the lung. We used another neural network to learn these weights for i^{th} subject as $(\boldsymbol{\alpha_i} = \{\alpha_{i1}, \cdots, \alpha_{i,N_i}\})$ where $\boldsymbol{\alpha_i} = A\left(\phi_e\left(\mathcal{X}_i; \theta_e\right); \theta_a\right)$. We formulated the attention network $A(\cdot)$ as a feed-forward network, consisting of multiple equivariant layers (EL)[295]. Assuming $\mathbf{H}_i \in \mathbb{R}^{N_i, d}$ where k^{th} row is $\phi(x_{ik}; \theta_e) \in \mathbb{R}^d$, an equivariant layer is defined as

$$[\mathbf{H}_i]_k = \mathbf{W}\left([\mathbf{H}_i]_k - \max(\mathbf{H}_i, 1)\right) + \mathbf{b}, \tag{2}$$

where $[\mathbf{H}_i]_k$ denotes k^{th} row of $\mathbf{H}_i$ and $\max(\mathbf{H}_i, 1)$ is the max over rows. $\mathbf{W} \in \mathbb{R}^{L \times d}$, $\boldsymbol{b} \in \mathbb{R}^L$ are the parameters of the EL. Such formulation ensures that the weight of any patch depends not only on the corresponding patch feature but also on the features of all the other patches in a patient. Next, we pass the output of the EL layers to a softmax function, to obtain a distribution of weights over the patches. This ensures that the weights $(\boldsymbol{\alpha_i})$ are non-negative numbers that sums to one.

To enable interpretability, the weight vector, $\boldsymbol{\alpha_i}$, should follow a sparse distribution. Increased sparsity pushes some weights terms, α_{ij}, to zero, and hence, it increases the interpretation by focusing on only the patches relevant for the prediction task. In our formulation, the weights α_{ij}, have non-negative values that sum to 1 *i.e.*, ($||\boldsymbol{\alpha_i}|| = \sum_j \alpha_{ij} = 1$). Hence, its derivative is zero, and using an ℓ_1 norm over the weight vector will not result in a sparse solution. To ensure high sparsity, we use a log-sum function as a regularizer. Minimizing $\sum_j \log \alpha_{ij}$ is equivalent of maximizing KL-divergence from the uniform distribution. The uniform distribution assigns the same weight to all the patches within one subject, *i.e.*, $\max_{\boldsymbol{\alpha_i}} \mathrm{KL}([\frac{1}{N_i}, \cdots, \frac{1}{N_i}], \boldsymbol{\alpha_i}) = \max_{\boldsymbol{\alpha_i}} \sum_j \frac{1}{N_i} \log \frac{1}{N_i} - \sum_j \frac{1}{N_i} \log \alpha_{ij} \equiv \min_{\alpha_i} \sum_j \log \alpha_{ij}$. We defined the regularization term as, $\mathcal{R}\left(\mathcal{X}_i; \theta_e, \theta_a\right) = \sum_{j=1}^{N_i} \log(\alpha_{ij} + \epsilon)$ and add it to the loss function in Equation. 1.

3.2.3 Discriminative Network

The discriminative network predicts the disease severity as

$$\hat{y}_i(\mathcal{X}_i) = f\left(\rho\left(\phi_e\left(\mathcal{X}_i, \theta_e\right)\right), \omega\right). \tag{3}$$

The discriminative network takes the patch-level features (*i.e.*, $\phi_e(x_{ij}; \theta_e)$) extracted by the encoder as input. It transforms the patch-level features using composition of two functions: (1) The aggregate function $\rho(\cdot)$. It is a permutation invariant function that aggregates the patch-level features to form a fixed length patient-representation. (2) A prediction function $f(\cdot; \omega)$, parameterized by ω. It takes the patient representation extracted by $\rho(\cdot)$ as input, and estimates the disease severity. Finally, $\mathcal{L}_d\left(y_i, \hat{y}_i(\mathcal{X}_i); \theta_e, \omega\right)$ is a regression or classification loss function between predicted and true value.

Conceptually, the aggregate function makes the prediction of disease severity less sensitive to the precise location within an image. It does so by aggregating the information from the local patches. One possible formulation of aggregate function is an average function, defined as $\rho(\cdot) = \frac{1}{N_i} \sum_{j=1}^{N_i} \phi_e(x_{ij})$. It considers all the feature values and hence, spread out the volume of the latent space evenly. The average function assumes an equal contribution of all the local patches towards final disease severity. However, COPD disease is often attributed to the diffuse air-sacks obstruction spread unevenly throughout the lung. To incorporate the disease's diffused effect, we adaptively weight the patch-level features to create the patient-representation as, $\rho(\cdot) = \sum_{j=1}^{N_i} \alpha_{ij}\phi_e(x_{ij})$. An attention network, described in Section 3.2.2, learns the weights (α_{ij}).

3.2.4 Architecture Details

The architecture of the encoder function consists of stacked convolutional layers which down-sampled the patches while doubling the number of channels. The decoder function consists of transposed convolutional layer (or deconvolutional layer) which up-sample the features while cutting the number of channels to half. Each convolutional layer employs batch-normalization for regularization, followed by an exponential linear unit (ELU) [33] for non-linearity. The attention network has 2 equivalence layers with sigmoid activation function, followed by a softmax layer. The model is trained using Adam optimizer [126] with hyper-parameters $\beta_1 = 0$ and $\beta_2 = 0.999$ and a fixed learning rate of 0.001. The dimension of the feature vector is 128. The trade-off hyper-parameters are $\lambda_1 = 10$ and $\lambda_2 = 1$. The experiments are performed on two NVIDIA p100 GPUs, each with 16GB GPU memory. The source code is available at https://github.com/batmanlab/Subject2Vec.

3.3 Experiments and Results

3.3.1 Study cohort

We evaluated our method on a dataset from the COPDGene study; an NIH funded multi-center clinical trial focused on the genetic epidemiology of COPD [209]. COPDGene includes 10,300 baseline participants, all of which were either current or former smokers. Each participant performed spirometry and had a high resolution inspiratory and expiratory CT scan, using a standardized protocol [209]. The acquired CT scan images were assessed by trained experts to provide a visual quantification of the centrilobular and paraseptal emphysema severity. Survival information was collected using the Social Security Death Index (SSDI) search and the COPDGene longitudinal follow-up (LFU) program.

3.3.2 Experimental setup

In our analysis, we used full-inspiration CT images, which were re-sampled to isotropic 1 mm^3. We worked on the fixed range of intensity values between -1024 HU and 240 HU, as suggested by Bhalla et al. [7]. We represented each subject as a set of equally sized 3-dimensional patches. To extract these patches, we first segmented the chest using Chest Imaging Platform (CIP) [225], open-source software for quantitative CT imaging assessment. Next, we extracted 3D overlapping patches from parenchyma region of the chest. The number of patches in a subject (N_i) may vary between subjects. A large patch size or a high overlap between the patches increases the N_i for a subject. All the patches of a subject must be processed in the same batch, as they are required to learn the patient-representation, which is then used to predict the disease severity. The available GPU memory restricts the maximum number of patches that can be processed in a single batch. We experimented with different values and finally used a patch-size of 32×32×32 with a 40% overlap and an upper limit of 1000 patches per batch in our experiments. The average Ni for this setting is 700 patches per subject.

We presented an analysis of the performance of our model for predicting patient-centered outcomes related to COPD. We trained two versions; 1) **Direct**: the model was trained to predict forced expiratory volume in 1 second (FEV1) and the FEV1/forced vital capacity (FVC) ratio, along with a clinical outcome of interest to represent disease severity. We separately trained one such model for each of the target outcomes. 2) **Indirect**: the model was trained only once, to predict FEV1 and FEV1/FVC as disease severity. The patient-representations from such model were then used in a separate regression analysis to predict other clinical outcomes of interest. The idea is to learn generalized patient-representations by training the model for one clinical variable (spirometry) and testing on another clinical output (emphysema score) which the models haven't seen previously. If two clinical variables are correlated, we should be able to capture much variance. Ofcourse, training directly for the clinical variable, as in direct version, will achieve better results. For all results, we reported average test performance in five-fold cross-validation. We compared the performance of our method against

1. Baseline: The low attenuation area (LAA) features. **LAA-950** is defined as the total percentage of both lungs with attenuation values less than -950 Hounsfield units on inspiratory images. LAA-950 signifies radiographic emphysema [192].
2. The **non-parametric** method proposed by Schabdac et al. [227]. In this method, hand-crafted image features were extracted for each patient, and non-parametric density estimation was performed to assign a characteristic vector to each patient.
3. The classical **k-means** algorithm applied to image features extracted from local lung regions [227]. A similar approach was suggested by Ash *et al.* [7].
4. The previous state-of-the-art method based on **CNN** also, applied to the COPDGene [74].

We perform three experiments: (1) *Predicting COPD outcomes:* we compare the performance of our method against the sate-of-art for different prediction tasks, (2) *Generative regularizer (λ_1):* we study the effect of the generative regularizer (*i.e.*, λ_1) in terms of prediction accuracy and information preserved in latent space, (3) *Visualization:* we visualize the interpretation of the model on the subject and population level.

3.3.3 Predicting COPD outcomes

We evaluated our proposed model over multiple COPD outcomes. These outcomes are summarized in Table 1. Next, we discuss each COPD outcome in more details and summarize our results.

3.3.3.1 Spirometry Measures

As part of the pulmonary function test, following spirometry values were evaluated for all the participants in COPDGene: forced expiratory volume in 1 second (FEV1) and the FEV1/forced vital capacity (FVC) ratio. All spirometric values were expressed as percentage of predicted values. Participants were classified as obstructed or non-obstructed under the

Table 1: Summarization of the clinical outcomes considered in the experiments and their numerical type and values.

Clinical Outcomes	**Type**	**Values**	**Description**
Spirometry Measures - Section 3.3.3.1			
FEV1	Continuous		Percentage predicted forced expiratory volume in 1 sec.
FEV1/FVC	Continuous		FEV1 ratio with forced vital capacity (FVC)
COPD	Binary	0 or 1	True if FEV1/FVC $>$ 0.7
GOLD stages	Categorical	0 - 4	GOLD stages 0 (non-obstructed) through 4 (severely obstructed).
Visual Emphysema Score - Section 3.3.3.2			
Centrilobular Emphysema (CLE)	Categorical	0 - 5	CLE emphysema severity score: none (0) to advanced destruction (6).
Paraseptal Emphysema	Categorical	0 - 2	Three severity scores: none, mild and substantial.
Acute Exacerbation - Section 3.3.3.3			
Historic Exacerbation	Binary	0 or 1	True if patient have experienced exacerbation in the last 1 year.
Future Exacerbation	Binary	0 or 1	True if patient reported an exacerbation by the 5th year followup.
Others - Section 3.3.3.4			
mMRC Dyspnea Scale	Categorical	0 - 4	Dyspnea with strenuous exertion (0) to dyspnea in daily activities (4)
Mortality	Binary	0 or 1	Vital status

2019 Global Initiative for Chronic Obstructive Lung Disease (GOLD) guidelines using a fixed FEV1/FVC ratio of 0.7 [274]. We defined the disease severity as the GOLD stages of 0 (non-obstructed) through 4 (very severely obstructed). Following the GOLD guidelines, in our experiments, we first train the model to predicted FEV1 and FEV1/FVC ratio, and then use these values to diagnose and stage COPD.

Table 2: Results for predicting spirometry measurements and using them to diagnose and stage COPD.

Method	FEV1	FEV1/FVC	COPD Diagnosis			GOLD	
	R-Square	R-Square	AUC ROC	AUC PR	Recall	% Accuracy	% Accuracy *one-off*
Ours (direct)	**0.67±0.03**	**0.74±0.01**	0.82	**0.72**	**0.80**	**65.44**	**89.14**
CNN[74]	0.53	-	**0.86**	-	-	51.10	74.90
Non-Parametric [227]	0.58±0.03	0.70±0.02	0.79	0.70	**0.80**	58.85	84.15
K-Mean	0.56±0.01	0.68±0.02	0.77	0.68	**0.81**	57.27	82.28
LAA-950	0.45±0.02	0.60±0.01	0.75	0.64	0.70	55.75	75.69

Results: Our model attained an r^2 of 0.67 ± 0.03 for the FEV1 and 0.74 ± 0.01 for the FEV1/FVC ratio, which is significantly better than previously reported approaches (see Table 2, Figure. 2). Next, we used the model-predicted FEV1/FVC ratio to diagnose COPD which achieved an AUC-ROC of 0.82. For the GOLD stage severity classification, our model achieved 65.4% and 89.1% exact and one-off accuracy's, respectively. Figure. 2 shows the confusion matrix for the COPD-GOLD stage classification.

3.3.3.2 Visual Emphysema Score

In the COPDGene cohort, radiographic centrilobular (CLE) and paraseptal emphysema were scored on inspiratory scans by a trained research analysts using the Fleischner Society classification system. Detailed methods for emphysema visual quantification are provided by Lynch *et al.* [154]. They grade the severity of CLE parenchymal emphysema on a scale of zero to five using labels: none, trace, mild, moderate, confluent, and advanced destructive emphysema. While paraseptal emphysema was scored using three labels: none, mild and substantial.

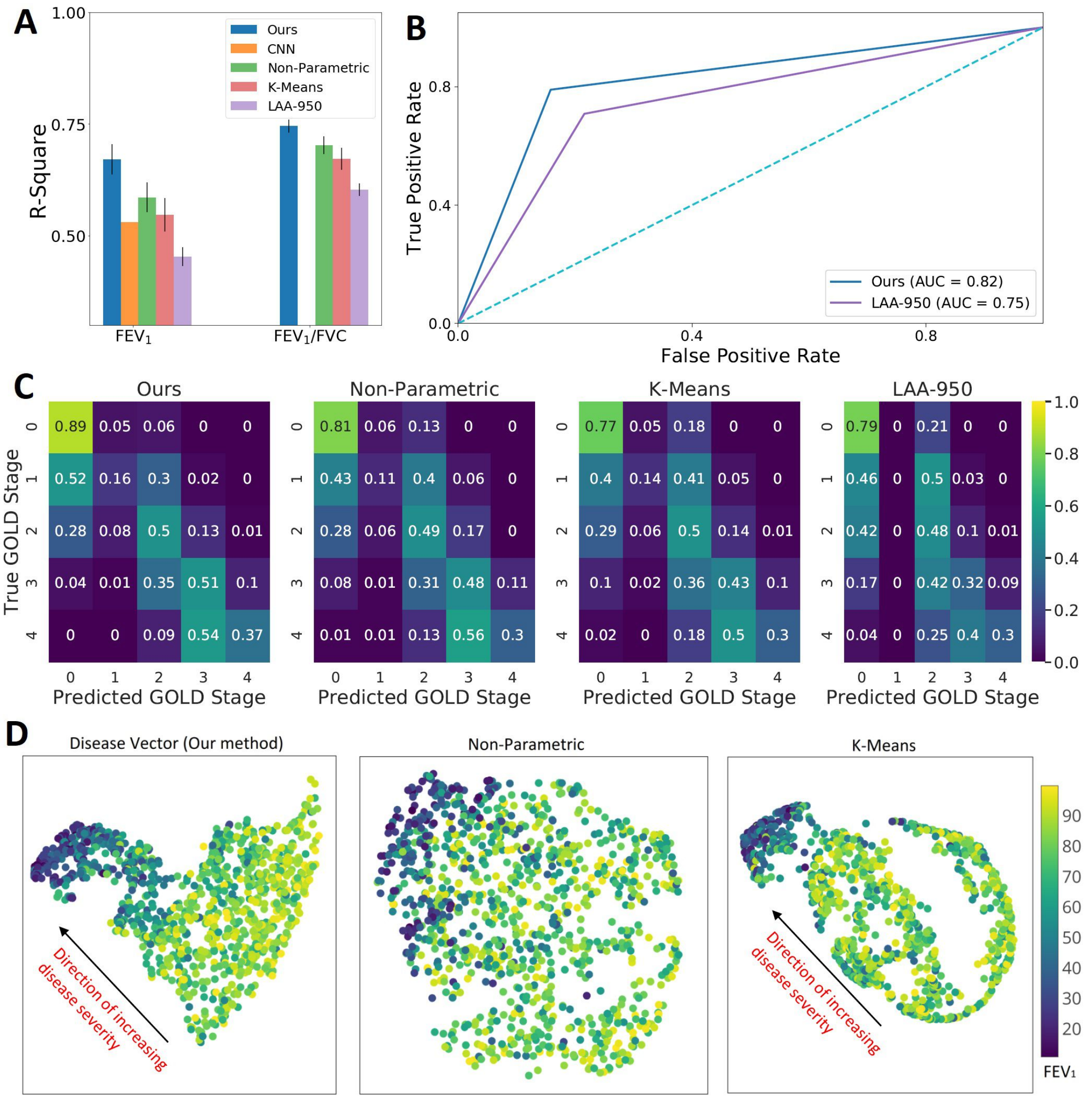


Figure 2: **A.** Bar graph comparing the r-square, coefficient of determination, for regression analysis of FEV1 and FEV1/FVC. **B.** Receiver Operating Characteristic (ROC) curve for prediction of COPD. Higher AUC-ROC suggests better classification. **C.** Confusion matrix plot for staging subjects using the GOLD stage. Following the GOLD guidelines [274], we used the model predicted FEV1 and FEV1/FVC ratio to diagnose and stage COPD. **D.** Visualizing the population by projecting the patient-level representations to 2D space using a dimensionality reduction method called UMAP [165]. Each dot represents one subject colored by percentage predicted FEV1. The relative position of a subject can be used to monitor the progression. We use two dimensions for the sake of visualization; it is straightforward to use a higher dimension and improve patient characterization. Figure is best viewed in color.

Results: Our model can identify subjects with different degrees of visual emphysema severity. The model correctly identified CLE visual emphysema score in 40.6% of the subjects in the COPDGene cohort and was within ± one score 74.8% of the time. Figure. 3 compares the confusion matrices of our method and LAA-950 features. In staging Paraseptal emphysema, the proposed model has an exact and on-off accuracy of 52.8% and 82.99% respectively. Results are summarized in Table 3, and the confusion matrix for Paraseptal emphysema prediction is shown in Figure. 3. Application of the Hosmer-Lemeshow [141] test did not suggest evidence of poor calibration (p-value 0.079).

Table 3: Results classifying subjects based on their emphysema visual score.

Method	CLE		Para-septal	
	% Acc.	% Acc. *one-off*	% Acc.	% Acc. *one-off*
Ours (direct)	**40.61**	**74.68**	**52.82**	82.99
Ours (in-direct)	36.30	61.33	46.87	75.97
Spirometry (FEV1)	33.52	63.96	44.64	72.77
LAA-950	31.89	77.74	33.32	**87.64**

3.3.3.3 Acute Exacerbations

In the COPDGene study, the exacerbations of COPD were self-reported and were quantified by the subject recall on questionnaires. A participant recorded a positive experience of an acute exacerbation if, in the last year, they had experienced at least one episode of increased dyspnea, cough or sputum production, resulting in admission to the hospital or changing of their treatment plan. Approximately 20% of the subjects reported experiencing at least one exacerbation before enrolling in the study. We used the HRCT acquired at the baseline visit to predict both historical and future exacerbations. The future exacerbation prediction used exacerbations reported by the longitudinal follow-up participants at the subsequent 5-year follow-up visit.

Results: Our model achieved an AUC-ROC of 0.70 in identifying the subjects who reported experiencing at least one exacerbation before enrolling in the study. We compared our performance against the intensity-based LAA feature in Figure. 4 (see Table 4).

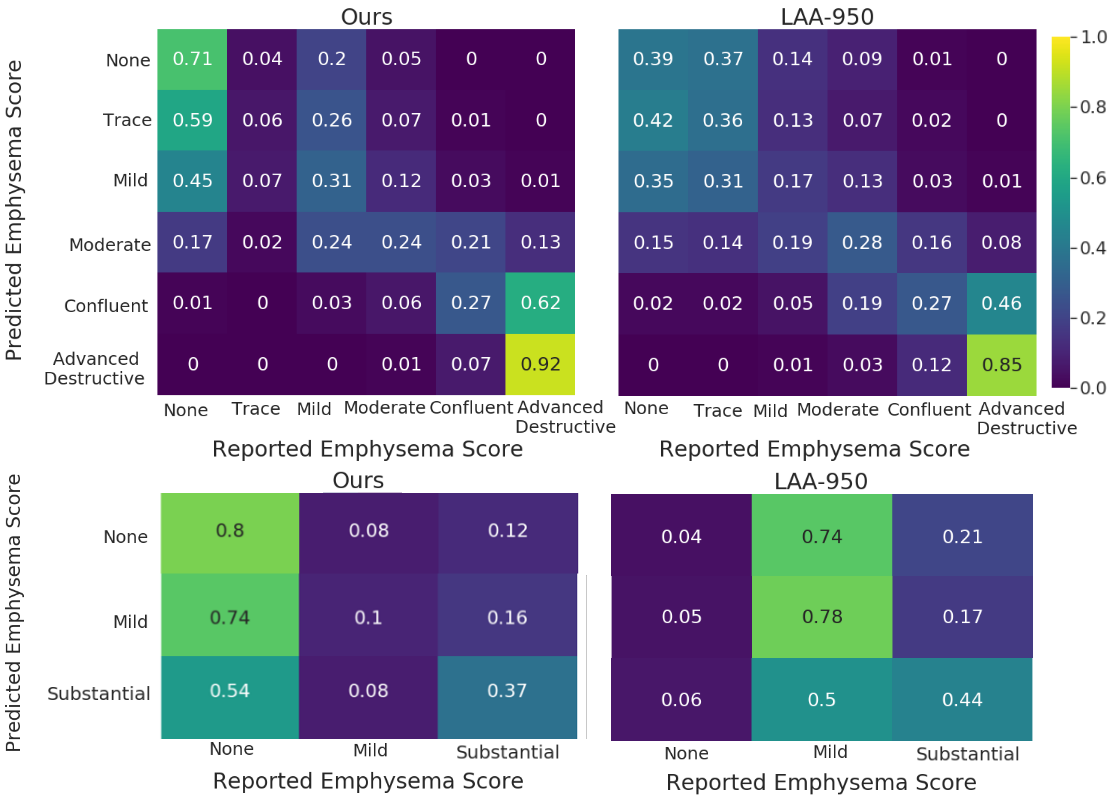


Figure 3: Comparing our method against traditionally used CT quantification measures (LAA-950). We stratify the population based on centrilobular and paraseptal emphysema severity score. Ours (direct) model is trained to predict spirometry measures and emphysema visual score together in a single loss function. The emphysema visual score is predicted in ordinal multi-class classification analysis. **A.** Confusion matrix plot for grouping the COPDGene population-based on **centrilobular emphysema** and **B. paraseptal emphysema**. Our proposed method performed better than LAA features and created a more significant separation between little and substantial emphysema. Figure is best viewed in color.

We also evaluated the performance of our model in identifying the population who reported subsequent exacerbations at the time of the 5-year follow-up. Our model achieved an AUC-ROC of 0.68. Our experiments show that the previous exacerbation history, together with imaging features from our method performs better (AUC-ROC 0.73), in predicting future exacerbation events than using exacerbation history alone (AUC-ROC 0.67). A quantitative comparison between different methods is shown in Table 4. Figure. 4 shows the ROC curve and the PR curve for binary classification. The p-value of the null hypothesis, using the Hosmer-Lemeshow test, is 0.08, suggesting no evidence of poor calibration.

Table 4: Results for identifying subjects with exacerbation risk and dyspnea.

Method	**Exacerbation History (EH)**			
	ROC-AUC	PR-AUC	Recall	% Accuracy
Ours (direct)	0.68±0.02	**0.38±0.03**	0.27±0.14	**76.93**
Ours (in-direct)	**0.73±0.01**	**0.43±0.03**	**0.59±0.03**	74.75
CNN [74]	0.643	-	0.18	60.40
LAA-950	0.65±0.01	0.35±0.02	0.43±0.02	73.78
	Future Exacerbation in longitudinal followup			
	ROC-AUC	PR-AUC	Recall	% Accuracy
Ours (direct)	0.65±0.01	0.32±0.02	0.43±0.01	68.30
Ours (in-direct)	0.70±0.02	0.35±0.02	**0.57±0.02**	73.87
LAA-950	0.64±0.01	0.31±0.02	0.43±0.04	73.80
EH	0.67±0.02	0.37±0.02	0.47±0.04	80.60
Ours (in-direct) + EH	**0.73±0.01**	**0.42±0.02**	0.47±0.04	**80.83**
	mMRC Dyspnea Score			
	% Accuracy		% Accuracy *one-off*	
Ours (direct)	**46.40**		67.04	
Ours (in-direct)	38.94		59.86	
Spirometry (FEV1)	42.63		**69.07**	
LAA-950	41.52		63.45	

3.3.3.4 mMRC Dyspnea Scale

Subjects completed the modified Medical Research Council (mMRC) dyspnea scale during their baseline visit. The scale ranges from zero (dyspnea only with strenuous exertion) to four (dyspnea with daily activities)

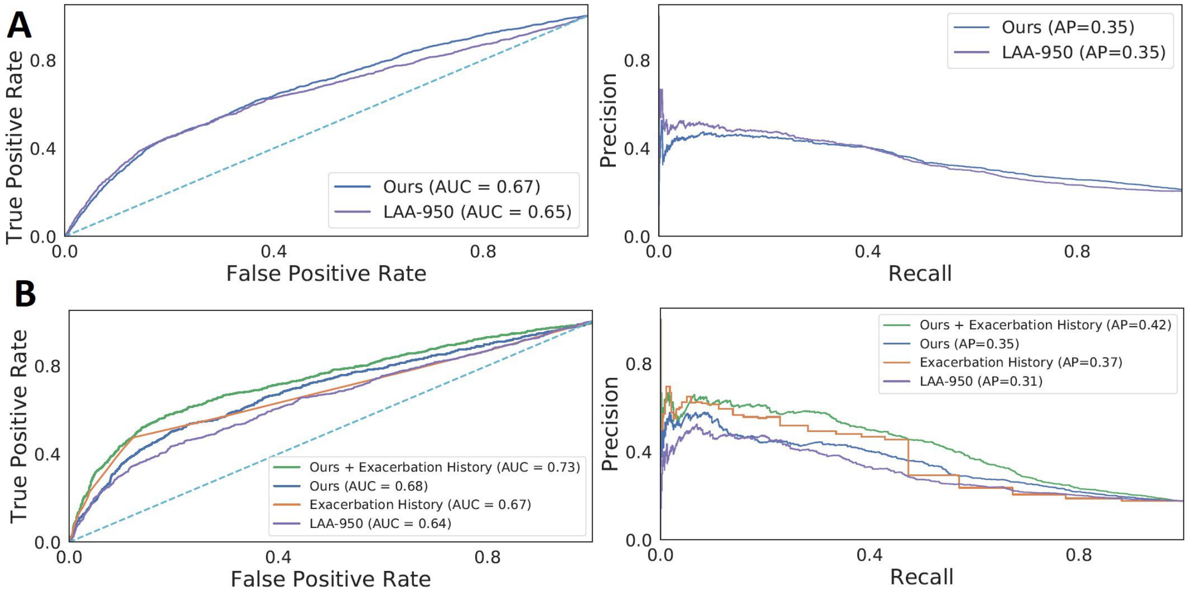


Figure 4: Receiver Operating Characteristic (ROC) curve and Precision-Recall (PR) curves. Identifying subjects with **A. exacerbation history** and **B. future exacerbation** as given in longitudinal follow up. The ROC curve shows how the true positive vs. false positive relationship changes as we vary the threshold of the positive class. In the top row, the positive class represents those subjects in COPD Cohort who reported experiencing at least one exacerbation before enrolling in the study. In the bottom row, the positive class represents those subjects who reported experiencing at least one exacerbation at the 5-year longitudinal follow up. Higher AUC-ROC number indicates better classification performance. Higher average precision (AP) in the PR curve means the better ability of the model in identifying subjects in a positive class. The plot shows that combining the history of past exacerbation with deep learning features from our model improves the prediction of future exacerbation. Figure is best viewed in color.

Results: Our proposed model was successful in classifying subjects in the COPDGene cohort based on their mMRC dyspnea scale with an accuracy of 43.5% and was within one score, 64.3% of the time (Table 4). Dyspnea scale is used to guide therapeutic strategies in patients with COPD [205, 203].

3.3.3.5 Mortality

We used the vital status and censoring time information provided in the mortality dataset to perform survival analysis. In the COPDGene cohort, the mean time between phase 1 data and the censoring time is approximately five years. Nearly 13% of subjects were reported deceased either in the SSDI search or in the COPDGene LFU. We used Cox proportional hazards (PH) model [143] to predict survival utilizing the probability of death predicted by patient-representation against age, gender, smoking status and center of enrollment as fixed covariates. Next, we used Kaplan-Meier plots stratified by quantile of predicted probabilities of death to visualize the results. Kaplan-Meier plot shows the probability of survival plotted against time.

We tested the PH assumption by performing a correlation between each of the covariates and their corresponding set of scaled Schoenfeld residuals with time [229]. A non-significant p-value for this test supported the PH assumption. In another test, we checked the global statistical significance of the Cox model. The test validated the null hypothesis that the variables have no association with survival. If the test failed to reject the null hypothesis, this would suggest that removing the variables from the model will not substantially harm the fit of that model. This global test is performed using three alternative tests: the likelihood-ratio test, the Wald test, and the score log-rank statistic. The survival analysis was performed using the lifelines library in Python[43] and the survival package in R[260]. We also compared the performance of our survival model against the uni-variate Cox regression model using intensity features (LAA-950) and the BODE index. The multidimensional BODE index has been shown to predict survival in cohort studies of COPD [162]. For the Cox PH model, we reported the results in terms of concordance, similar to AUC-ROC in binary classification.

Results: Our proposed method achieved a concordance of 0.61 in Cox regression[143]

Table 5: Results of Cox Proportional-Hazard (PH) model for survival analysis. The probability of death, learned from binary classification of mortality, is used as covariate in Cox regression.

Method	Hazard Ratio[b]	Quantile p-value[c]	Concordance[d]	Global statistical significance[e] Max p-value (LR, Wald, log Rank)	PH-Assumption (Global p-value)[f]
Ours (direct)	1.04 [CI: 0.09, 1.87]	<2e-16	0.590	p=<2e-16	0.514
Ours (in-direct)	1.54 [CI: 1.09, 2.17]	<2e-16	0.615	p=<2e-16	0.598
CNN [74]	2.69 [CI: 1.19, 6.05]	0.017	0.72	-	-
Spirometry (FEV1)	1.20 [CI: 0.94, 1.54]	6.91e-07	0.525	p=4e-06	-
BODE Index [21][a]	1.68 [CI: 1.21, 2.31]	<2e-16	0.568	p=<2e-16	0.462
LAA-950	1.13 [CI: 0.93, 1.37]	6.35e-07	0.537	p=4e-06	0.391

PH: proportional hazards; BODE = Body-mass index, airflow Obstruction, Dyspnea and Exercise index; CI = Confidence interval;
All the models have age, gender, smoking pack-years, and center of enrollment as covariates.

[a] BODE index is the clinical index used to predict the mortality rate from COPD [162].

[b] The Hazard ratio is the exponential coefficient ($\exp(\beta)$) of the covariate. A covariate is positively associated with the event probability when the hazard ratio is above one and, thus, is negatively associated with the length of survival. We also report 95% confidence intervals for the hazard ratio.

[c] A significant p-value with > 1 hazard ratio indicates a strong relationship between the covariate and increased risk of death.

[d] The concordance shows the fraction of pairs, where the observations with higher survival time have a higher probability of survival predicted by the model. It is analog to the area under the ROC curve in classification analysis.

[e] The Global statistical significance of the model is tested using three alternative tests namely the likelihood-ratio (LR) test, the Wald test, and the score log-rank statistics. $p < 0.001$ indicates that the model fits significantly better than the null hypothesis. The null hypothesis states that all the betas (β) are 0.

[f] We used scaled Schoenfeld residuals to check the proportional hazards assumption. A non-significant p-value shows no evidence of violation of PH assumption by survival model.

analysis compared to 0.56 for the BODE index and 0.53 for LAA-950 features (Table 5). In testing the proportional hazard (PH) assumption of our model using scaled Schoenfeld residues, we achieved a p-value > 0.3 for all the covariates and a global p-value of 0.59 for the model. A significant p-value for this test provided no evidence for the violation of the PH assumption made by the Cox model.

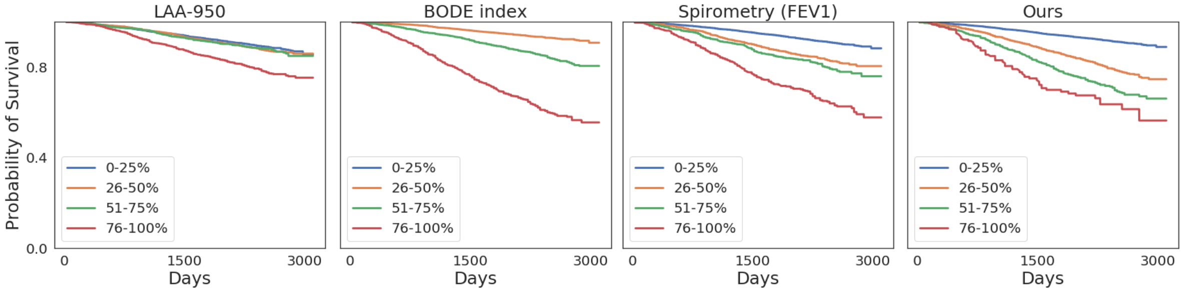


Figure 5: Kaplan Meier plot for visualizing the results of survival analysis. The plot is obtained by performing Cox regression analysis stratified on the quantile of predicted probability of mortality in binary classification. A good Kaplan Meier plot has large separations between the groups. BODE index is the Body-mass index, airflow Obstruction, Dyspnea and Exercise index which is highly correlated with mortality [162]. Our model performed better than the conventional emphysema quantification, the BODE index, and spirometry measures for mortality assessment.

Next, we tested the global statistical significance of the Cox model using three alternative tests: the likelihood-ratio test, the Wald test, and the score log-rank statistic. We achieved a p-value of < 0.001 in all three tests. Hence, we can reject the null hypothesis that all the coefficients are 0, with high confidence. Figure. 5 shows the Kaplan Meier (KM) plots to visualize the subjects grouped by quantile of predicted probability of 5-year survival. The KM plot for our method has a large separation between different quantile groups. Thus, our model can divide the population into distinct groups based on their survival risk.

3.3.4 Generative regularizer

Hyper-parameter λ_1 in our overall loss function in Eq. 1 balances between the discriminative and the generative setting. $\lambda_1 = 0$ represents a fully discriminative setting, in which the decoder of the auto-encoder is not trainer. Hence, there is no reconstruction of the input patch back from the low dimensional embedding learned by the encoder. In the absence of the decoder function, the feature extractor (encoder) is forced to retain only information that is sufficient for the underlying discriminative task. The Figure. 6(a) reports the spectral behaviors of the latent features ($i.e.$, $\phi_e(\mathcal{X}_i)$) for varying λ_1. For fully discriminative setting with $\lambda_1 = 0$, we observe a highly redundant latent space, with almost similar patch-level features $\phi_e(x_{ij})$, with attention weights α_{ij} converging to $\frac{1}{|\mathcal{X}_i|}$.

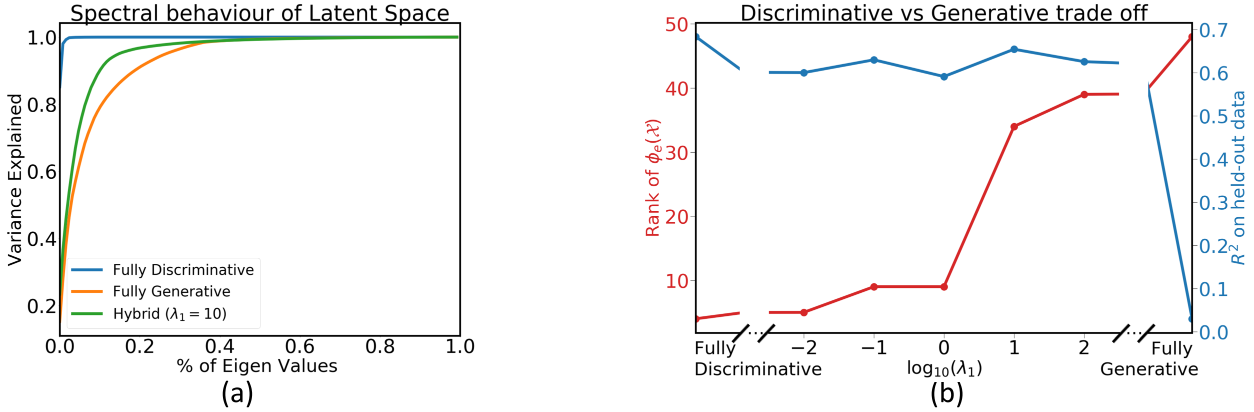


Figure 6: Evaluating generative regularizer (a) Spectral properties of patch-level features for different values of λ_1. (b) The trade-off between rank of latent space (red, y-axis on left) and predictive power (blue, y-axis on right) for different values of λ_1. Left represents fully discriminative and right represents fully generative models.

As $\lambda_1 \to \infty$, the network mostly focuses on the generative task of reconstructing the patch back from patch-level features. In such fully generative setting, the encoder features are not optimized for the downstream prediction task. Hence, though the patch-level features are much diverse, as seen in Figure. 6(a), there is a significant drop in R^2 for predicting FEV1, as shown in the Figure. 6(b).

We demonstrate the effect of regularization through Figure. 6(b). It shows the trade-off between effective rank of the latent feature and R^2 for predicting FEV1. Although, the R^2 drops a little, the rank, which represents the diversity of the latent features, improves drastically. The gap between accuracy's of $\lambda_1 = 0$ and $\lambda_1 > 0$ is the price we pay for the interpretability.

3.3.5 Visualization

Figure. 7(b) visualizes the attention weights on a subject. The dark area on the left lung (tope-row), which is severely damaged, received high attention, while same regions in a different lung (bottom-row) have minimum attention.

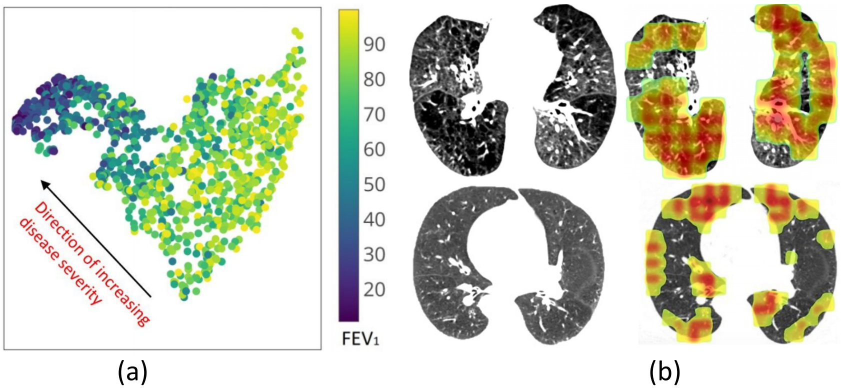


Figure 7: An axial view of the attention map on a subject. Red color indicate higher relevance to the disease severity.

3.4 Discussion and Conclusion

Our proposed Deep Learning-based method demonstrates the ability to predict multiple aspects of COPD disease pattern, severity, and future events. It does so by extracting the most relevant information from volumetric HRCT images of the subject. Unlike previous Deep Learning methods that process a collection of 2D slices, our method works on the entire 3D inspiratory scan of the subject. Deep Learning enables us to go beyond standard radiographic features such as LAA and construct data-driven radiological features that are optimal for a specific task. Our results show that large cohorts such as COPDGene enable DL methods to learn meaningful patterns and converge to reliable predictions. Another advantage of our method lies in its generalizability and flexibility to incorporate different aspects of COPD. Using the same DL model and architecture, we were not only able to predict spirometric obstruction but were also successful in predicting all-cause mortality and future exacerbations, quantifying emphysema burden and disease pattern, and evaluating symptom scores.

In the direct approach, our model achieved high predictive strength by explicitly training to predict a target outcome. Our cross-validation experiments showed that the model was well-calibrated and achieved consistent performance over all folds. While in the in-direct approach, the model was trained only once, to predict respiratory measurements, this model performed well in predicting COPD outcomes including, acute exacerbations, and mortality.

Our predictions of spirometry measurements outperformed previously reported methods, including the previous DL method. Our method has a potential translational impact if it is utilized as a clinical screening tool, e.g., when obtained during routine cancer screening, to identify subjects with a high likelihood of COPD for further assessment. Our visualization of the COPDGene population colored by the FEV1 value shows subjects with high FEV1 clustered together and a progression of disease severity from low to high (**Figure. 2(d)**). This population-level analysis may be helpful in prospectively identifying unique clinical subgroups or in quantifying disease severity across research cohorts.

This is the first study to use DL-based method to predicted various clinical outcomes associated with COPD like spirometric obstruction, emphysema severity, current and future exacerbation risk and mortality, using CT imaging alone. The results of our study conclude that DL-based method can provide a holistic view of disease severity and progression from a single set of CT images. Further work toward developing interpretable DL models is essential for the development of standardized CT-based assessment of COPD.

High-resolution CT evaluation by a deep learning algorithm might provide low-cost, reproducible, near-instantaneous classification of fibrotic lung disease with human-level accuracy. These methods could be of benefit to centres at which thoracic imaging expertise is scarce, as well as for stratification of patients in clinical trials.

4.0 Progressive Counterfactual Explainer

4.1 Introduction

With the explosive adoption of deep learning for real-world applications, explanation and model interpretability have received substantial attention from the research community [51, 84, 120, 175]. Primarily, DL is used for Computer-Aided Diagnosis [95] and other tasks in the medical imaging domain [207, 215]. However, for real-world deployment [276], the decision-making process of these models should be explainable to humans to obtain their trust in the model [67, 111]. Explainability is essential for auditing the model [281], identifying failure modes [52, 193] or hidden biases in the data or the model [135], and obtaining new insights from large-scale studies [218]. For example, consider evaluating a computer-aided diagnosis of Alzheimer's disease from medical images. The physician should be able to assess whether or not the model pays attention to age-related or disease-related variations in an image to trust the system.

With the advancement of DL methods for medical imaging analysis, deep neural networks (DNNs) have achieved near-radiologist performance in multiple image classification tasks [230, 208]. However, DNNs are criticized for their "black-box" nature, *i.e.*, they fail to provide a simple explanation as to why a given input image produces a corresponding output [261]. To address this concern, multiple model explanation techniques have been proposed that aim to explain the decision-making process of DNNs [231, 35]. The most common form of explanation in medical imaging is a class-specific heatmap overlaid on the input image. It highlights the most relevant regions (*where*) for the classification decisions [10, 151, 231, 236, 237, 249, 255]. The location information alone is insufficient for applications in medical imaging. Different diagnoses may affect the same anatomical regions, resulting in similar explanations for multiple diagnosis, resulting in inconclusive explanations. A thorough explanation should explain *what* imaging features are present in those important locations, and *how* changing such features modifies the classification decision.

Although not always clear, there are subtle differences between interpretability and *expla-*

nation [264]. While the former mainly focuses on building or approximating models that are locally or globally interpretable [212], the latter aims at explaining a predictor a-posteriori. The explanation approach does not compromise the prediction performance. However, a rigorous definition for what is a good explanation is elusive. Some researchers focused on providing feature importance (*e.g.*, in the form of a heatmap [231]) that influence the outcome of the predictor. In some applications (*e.g.*, diagnosis with medical images) the causal changes are spread out across a large number of features (*i.e.*, large portions of the image are impacted by a disease). Therefore, a heatmap may not be informative or useful, as almost all image features are highlighted. Furthermore, those methods do not explain *why* a predictor returns an outcome. Others have introduced local occlusion or perturbations to the input [302, 61] by assessing which manipulations have the largest impact on the predictors. There is also recent interest in generating counterfactual inputs that would change the black box classification decision with respect to the query inputs [79, 147]. Local perturbations of a query are not guaranteed to generate realistic or plausible inputs, which diminishes the usefulness of the explanation, especially for end users (*e.g.*, physicians). We argue that the explanation should depend not only on the predictor function but also on the data. Therefore, it is reasonable to train a model that learns from data as well as the black-box classifier (*e.g.*, [24, 42, 61]).

To address these gaps, we propose a novel explanation method to provide a counterfactual explanation. A counterfactual explanation is a perturbation of the input image such that the classification decision is flipped [114, 147, 157, 224]. By comparing, the input image and its corresponding counterfactual image, the end-users can visualize the difference in important image features that leads to a change in classification decision. Our proposed method falls into the local explanation paradigm. Our approach is model agnostic and only requires access to the predictor values and its gradient with respect to the input. Given a query input to a black-box, we aim at explaining the outcome by providing *plausible* and *progressive* variations to the query that can result in a change to the output. The plausibility property ensures that perturbation is natural-looking. A user can employ our method as a "tuning knob" to progressively transform inputs, traverse the decision boundary from one side to the other, and gain understanding about how the predictor makes a decision.

We adopted a conditional Generative Adversarial Network (cGAN) as our explanation framework. However, using cGAN is challenging, as GANs with an encoder may ignore small or uncommon details during image generation [14]. This is particularly important in our application, as the missing information includes foreign objects such as a pacemaker that influence human users' perception. To address this issue, we stipulate when the input image has reconstructed the shape of the anatomy and that foreign objects are preserved. We achieve this by incorporating semantic segmentation and object detection into our loss function.

We introduce three principles for an explanation function that can be used beyond our application of interest. We evaluate our method on a set of benchmarks as well as real medical imaging data. Our experiments show that the counterfactually generated samples are realistic-looking and in the real medical application, satisfy the external evaluation. We also show that the method can be used to detect bias in training of the predictor. Our contributions are summarized as follows:

1. We developed a progressive counterfactual explainer (PCE) that generates visual explanations for a black-box classifier. PCE explains the decision for a query image by gradually changing the image such that the classification decision is flipped.
2. Our method accounts for subtleties of medical imaging by preserving the anatomical shape and foreign objects such as support devices across generated images. The specialized reconstruction loss is proposed to incorporate context from semantic segmentation and foreign object detection networks.
3. We evaluated our method extensively on both natural and medical datasets.
4. We proposed quantitative metrics based on clinical definition of two diseases (cardiomegaly and PE). We are one of the first methods to use such metrics for quantifying DNN model explanation. Specifically, we used these metrics to quantify statistical differences between counterfactual and query images.
5. We are one of the first methods to conduct a thorough human-grounded study to evaluate different counterfactual explanations for medical imaging task. Specifically, we collected and compared feedback from diagnostic radiology residents, on different aspects of ex-

planations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users.

6. On the face images dataset, we show that our method successfully detects confounding bias in the classifier.

4.2 Method

Consider a *black box* classifier that maps an input space $\mathcal{X}$ (*e.g.*, images) to an output space $\mathcal{Y}$ (*e.g.*, labels). In this paper, we consider binary classification problems where $\mathcal{Y} = \{0, 1\}$. However, the proposed method is general and can be used for multi-class or multi-label settings. We use $f(\mathbf{x}) = \mathbb{P}(y|\mathbf{x}) \in [0, 1]$ to denote the posterior probability of the classifier. We assume that f is a differentiable function and we have access to its value as well as its gradient with respect to the input $\nabla_{\mathbf{x}} f(\mathbf{x})$.

We view the (visual) explanation of the black-box as a generative process that produces a plausible and realistic perturbation of the query image such that the classification decision is changed to a desired value $\mathbf{c}$. By gradually changing the desired output $\mathbf{c}$ in range $[0, 1]$, we can traverse the prediction space while visualizing the gradual exaggeration of the target class in generated images. We conceptualize this traversal from one side of the decision boundary to the other as walking across a data manifold, $\mathcal{M}_x$. Directly manipulating the high dimensional image space is very challenging. Hence, we assume that there is a low-dimensional embedding space ($\mathcal{M}_z$) that encodes the walk. An encoder, $E : \mathcal{M}_x \rightarrow \mathcal{M}_z$, maps an input, $\mathbf{x}$, from the data manifold, $\mathcal{M}_x$, to the embedding space. The desired output $\mathbf{c}$ represents the step size of the walk. We gradually increase or decrease $\mathbf{c}$ to generate a new image to represent each step of this walk. A generator, $G : \mathcal{M}_z \rightarrow \mathcal{M}_y$, takes both the embedding coordinate and the desired output $\mathbf{c}$ (current step-size) and maps it back to the data manifold (see Figure. 8).

The PCE is denoted as $\mathcal{I}_f(\cdot, \cdot)$. Formally, $\mathcal{I}_f(\mathbf{x}, \mathbf{c}) : (\mathcal{X}, \mathbb{R}) \rightarrow \mathcal{X}$ is a function that takes two arguments: a query image $\mathbf{x}$ and the desired posterior probability $\mathbf{c}$ for some target class y. This function generates a perturbed image $\mathbf{x}_c$ such that $f(\mathbf{x}_c)[y] \approx \mathbf{c}$. This formulation allows us to view $\mathbf{c}$ as a "knob" that gradually perturb the input image to achieve visually perceptible differences in $\mathbf{x}$ while crossing the decision boundary given by function f. PCE function $\mathcal{I}_f$ should satisfy the following properties:

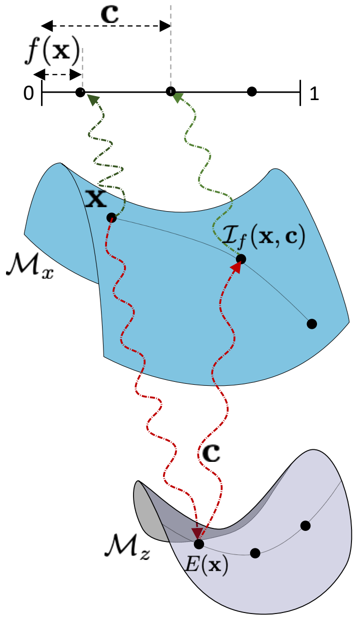


Figure 8: The schematic of the method. f is the black-box function producing the posterior probability $f(\mathbf{x})$. $\mathbf{c}$ is the desired probability. $\mathcal{I}_f(\mathbf{x},\mathbf{c})$ is an explainer function for f, which creates a perturbation of $\mathbf{x}$ that produce a classifier's prediction of $\mathbf{c}$. The $E(\cdot)$ is an encoder that maps the data manifold $\mathcal{M}_x$ to the embedding manifold $\mathcal{M}_z$. Explanation function is generator that conditionally maps embedding back to the data manifold.

A.) **Data consistency**: The perturbed image, $\mathbf{x_c}$ should resemble data instance from input space $\mathcal{X}$ *i.e.*, if input space comprises chest x-rays, $\mathbf{x_c}$ should look like a chest x-ray with minimum artifacts or blurring.

B.) **Classification model consistency**: The perturbed image, $\mathbf{x_c}$ should produce the desired output from the classifier f, *i.e.*, $f(\mathcal{I}_f(\mathbf{x},\mathbf{c})) \approx \mathbf{c}$.

C.) **Context-aware self-consistency**: To be self-consistent, the PCE should satisfy three criteria (1) Reconstructing the input image by setting $\mathbf{c} = f(\mathbf{x})$ should return the input image, *i.e.*, $\mathcal{I}_f(\mathbf{x}, f(\mathbf{x})) = \mathbf{x}$. (2) Applying a reverse perturbation on the explanation image $\mathbf{x_c}$ should recover $\mathbf{x}$, *i.e.*, $\mathcal{I}_f(\mathbf{x_c}, f(\mathbf{x})) = \mathbf{x}$. (3) Achieving the aforementioned reconstructions while preserving anatomical shape and foreign objects (*e.g.*, pacemaker) in the input image.

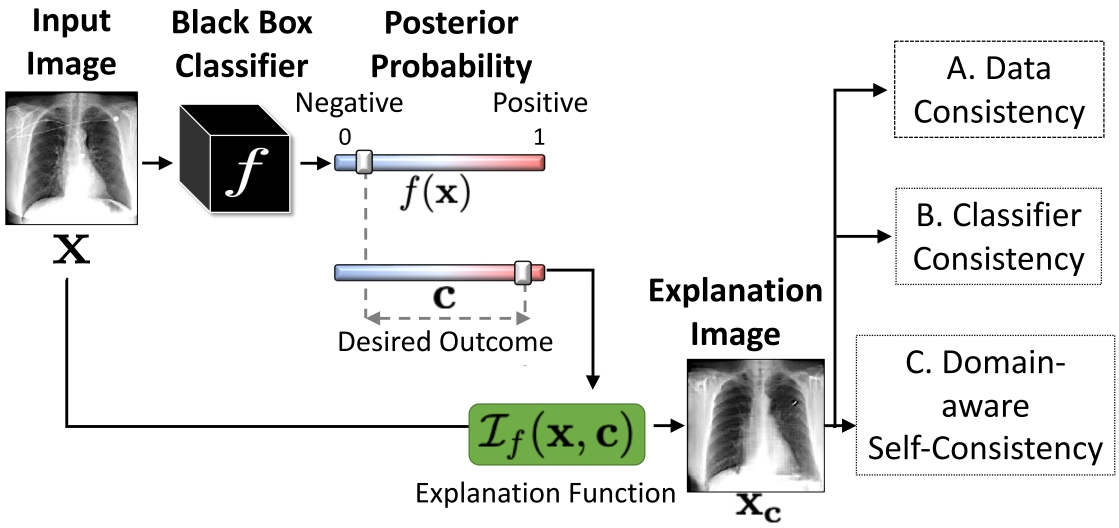


Figure 9: PCE function $\mathcal{I}_f(\mathbf{x}, \mathbf{c})$ for classifier f. Given an input image $\mathbf{x}$, we generates a perturbation of the input, $\mathbf{x_c}$ as explanation, such that the posterior probability, f, changes from its original value, $f(\mathbf{x})$, to a desired value $\mathbf{c}$ while satisfying the three consistency constraints.

We designed PCE as a novel deep learning (DL) framework, which is trained end to end to satisfy the three properties. It minimizing the following objective function:

$$\min_{E,G} \max_{D} \lambda_{cGAN} \mathcal{L}_{\text{cGAN}}(D, G) + \lambda_f \mathcal{L}_f(D, G) + \lambda_{rec} \mathcal{L}_{\text{rec}}(E, G) \tag{4}$$

where $\mathcal{L}_{\text{cGAN}}$ is a conditional GAN-based loss function that enforces data-consistency, $\mathcal{L}_f$ enforces classification model consistency through a KullbackLeibler (KL) divergence loss and $\mathcal{L}_{\text{rec}}$ is a reconstruction loss that enforces self-consistency. The loss function is defined over three networks, an image encoder $E(\cdot)$, a conditional GAN generator $G(\cdot)$ and a discriminator $D(\cdot)$. $\lambda_{cGAN}, \lambda_f$ and λ_{rec} controls the balance between the terms. In the following sections, we will discuss each property and the associated loss term in detail.

4.2.1 Data consistency

We formulated PCE, $\mathcal{I}_f(\mathbf{x}, \mathbf{c})$, as an image encoder $E(\cdot)$ followed by a conditional GAN (cGAN) [172], with $\mathbf{c}$ as the condition. The encoder enables transformation of a given image, while the GAN framework allows to generate realistic looking transformations as explanation image. The Generative Adversarial Network (GANs) [75] are implicitly models, that learn the underlying data distribution $p_{\text{data}}(\mathbf{x})$ by setting up a min-max game between generative (G) and discriminative (D) networks. The $G(\cdot)$ network learns to transform samples drawn from a canonical distribution such that $D(\cdot)$ network fails to distinguish the generated data from the real data. GANs optimizes the following objective function:

$$\mathcal{L}_{\text{GAN}}(D, G) = \mathbb{E}_{\mathbf{x}, c\sim P(\mathbf{x})}\big[\log\big(D(\mathbf{x})\big)\big] + \mathbb{E}_{\mathbf{z}\sim P_{\mathbf{z}}}\big[\log\big(1 - D(G(\mathbf{z}))\big)\big],$$

where $\mathbf{z}$ and $P_{\mathbf{z}}$ are the noise distribution and the corresponding canonical distribution. There has been significant progress toward improving GANs stability as well as sample quality [18, 116]. The advantage of GANs is that they produce realistic-looking samples without an explicit likelihood assumption about the underlying probability distribution. This property is appealing for our application.

Furthermore, we use a Conditional GAN (cGAN) that allows the incorporation of a context as a condition to the GAN [169, 172]. We use the desired classification outcome as our condition $\mathbf{c} \in [0, 1]$. The cGAN optimizes the following loss function:

$$\mathcal{L}_{\text{cGAN}}(D, G) = \mathbb{E}_{\mathbf{x}, c\sim P(\mathbf{x}, c)}\big[\log\big(D(\mathbf{x}, \mathbf{c})\big)\big] + \mathbb{E}_{\mathbf{z}\sim P_{\mathbf{z}}, c\sim P_c}\big[\log\big(1 - D(G(\mathbf{z}, \mathbf{c}), \mathbf{c})\big)\big], \quad (5)$$

where $\mathbf{c}$ denotes a condition. In our formulation, $\mathbf{z}$ is the latent representation of the input image $\mathbf{x}$, learned by the encoder $E(\cdot)$. Finally, the PCE is defined as,

$$\mathcal{I}_f(\mathbf{x}, \mathbf{c}) = G(E(\mathbf{x}), \mathbf{c}). \quad (6)$$

Our architecture is based on Projection GAN [172], a modification of cGAN. An advantage of the Projection GAN is that it scales well with the number of classes allowing to use very small bin size while discretizing $\mathbf{c}$. The Projection GAN imposes the following structure on the discriminator loss function:

$$\mathcal{L}_{\text{cGAN}}(D, \hat{G})(\mathbf{x}, \mathbf{c}) = \log \frac{p_{\text{data}}(\mathbf{c}|\mathbf{x})}{q(\mathbf{c}|\mathbf{x})} + \log \frac{p_{\text{data}}(\mathbf{x})}{q(\mathbf{x})} := r(\mathbf{c}|\mathbf{x}) + \psi(\boldsymbol{\phi}(\hat{G}(\mathbf{z}))), \quad (7)$$

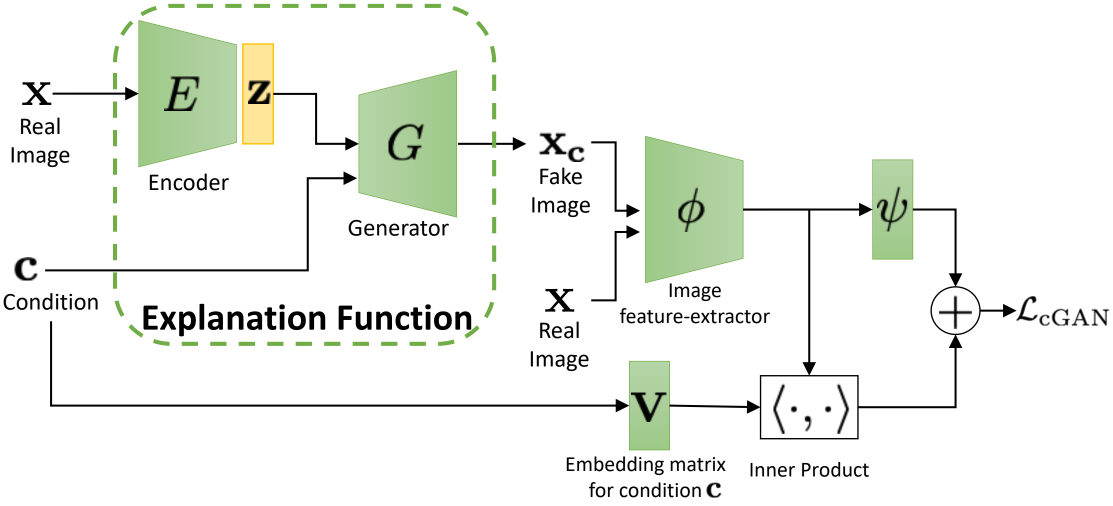


Figure 10: Progressive counterfactual explainer (PCE) as a conditional-GAN with an encoder.

For the discriminator in cGAN, we adapted the loss function from Projection GAN [172] based on our application, as shown in Figure. 10. We can view $\mathbf{c}$ as a one-hot vector over N classes. The loss function of projection cGAN has two terms. The first term is the distribution ratio between marginals *i.e.*, the real data distribution $p_{\text{data}}(\mathbf{x})$ and the learned distribution of the generated data $q(\mathbf{x})$. The second term is the distribution ratio between conditionals. It evaluates the correspondence between the generated image and the condition. This formulation allows us to skip calculating q as we are only interested in the ratio. The overall loss function is as follows,

$$\begin{aligned}\mathcal{L}_{\text{cGAN}}(D, \hat{G})(\mathbf{x}, \mathbf{c}) &= \log \frac{p_{\text{data}}(\mathbf{x})}{q(\mathbf{x})} + \log \frac{p_{\text{data}}(\mathbf{c}|\mathbf{x})}{q(\mathbf{c}|\mathbf{x})} \\ &:= r(\mathbf{x}) + r(\mathbf{c}|\mathbf{x})\end{aligned} \tag{8}$$

where $\mathcal{L}_{\text{cGAN}}(D, \hat{G})$ indicates the loss function in Eq. 5 when $\hat{G}$ is fixed. Further, $r(\mathbf{x})$ is the discriminator logit that evaluates the visual quality of the generated image. It is the discriminator's attempt to separate real images from the fakes images created by the generator. The second term evaluates the correspondence between the generated image $\mathbf{x_c}$ and the condition $\mathbf{c}$.

To represent the condition, the discriminator learns an embedding matrix $\mathbf{V}$ with N rows, where N is the number of conditions. The condition is encoded as an N-dimensional one-hot vector which is multiplied by the embedding-matrix to extract the condition-embedding. When $\mathbf{c} = n$, the conditional embedding is given as the n-th row of the embedding-matrix ($\mathbf{v}_n$). The projection is computed as the dot product of the condition-embedding and the features extracted from the fake image,

$$\mathcal{L}_{\text{cGAN}}(D, \hat{G})(\mathbf{x}, \mathbf{c}) := r(\mathbf{x}) + \mathbf{v}_n^T \phi(\mathbf{x}), \tag{9}$$

where, n is the current class for the conditional generation and ϕ is the feature extractor.

In our use-case, the condition $\mathbf{c}$ is the desired posterior probability from the classification function f. $\mathbf{c}$ is a continuous variable with values in range $[0, 1]$. Projection-cGAN requires the condition to be a discrete variable, to be mapped to the embedding matrix $\mathbf{V}$. Hence, we discretize the range $[0, 1]$ into N bins, where each bin is one condition. One can view change from $f(\mathbf{x})$ to $\mathbf{c}$ as changing the bin index from the current value $C(f(\mathbf{x}))$ to $C(\mathbf{c})$ where $C(\cdot)$ returns the bin index.

4.2.2 Classification model consistency

Ideally, cGAN should generate a series of smoothly transformed images as we change condition $\mathbf{c}$ in range $[0, 1]$. These images, when processed by the classifier f should also smoothly change the classification prediction between $[0, 1]$. To enforce this, rather than considering bin-index $C(\mathbf{c})$ as a scalar, we consider it as an ordinal-categorical variable, *i.e.*, $C(\mathbf{c}_1) < C(\mathbf{c}_2)$ when $\mathbf{c}_1 < \mathbf{c}_2$. Specifically, rather than checking one condition that desired bin-index is equal to some value n, $C(\mathbf{c}) = n$, we check $n-1$ conditions that desired bin-index is greater than all bin-index which are less than n *i.e.*, $C(\mathbf{c}) > i \forall i \in [1, n)$ [62].

We adapted Eq. 9 to account for a categorical variable as the condition, by modifying the second term to support ordinal multi-class regression. The modified loss function is as follows:

$$r(\mathbf{c} = n|\mathbf{x}) := \sum_{i<n} \mathbf{v}_i^T \boldsymbol{\phi}(\mathbf{x}), \tag{10}$$

Along with conditional loss for the discriminator, we need additional regularization for the generator to ensure that the actual classifier's outcome, *i.e.*, $f(\mathbf{x}_\mathbf{c})$, is very similar to the condition $\mathbf{c}$. To ensure this compatibility with f, we further constrain the generator to minimize the KullbackLeibler (KL) divergence that encourages the classifier's score for $\mathbf{x}_\mathbf{c}$ to be similar to $\mathbf{c}$ (*see* Figure. 10(b). Our final condition-aware loss is as follows,

$$\mathcal{L}_f(D, G) := r(\mathbf{c}|\mathbf{x}) + D_{\mathrm{KL}}(f(\mathbf{x}_\mathbf{c})||\mathbf{c}), \tag{11}$$

Here, the first term is a function of both G and D, the second term influences only the G.

4.2.3 Context-aware self consistency

A valid explanation image is a small modification of the input image, and should preserve the inputs' identity *i.e.*, patient-specific information such as the shape of the anatomy. While images generated by a GAN is shown to be realistic looking [116], GAN with an encoder may ignore small or uncommon details in the input image [14]. To preserve these features, we propose a context-aware reconstruction loss (CARL) that exploits extra information from the input domain to refine the reconstruction results. This extra information comes in the form of semantic segmentation and detection of any foreign object present in the input image. The CARL is defined as,

$$\mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}') = \sum_j \frac{S_j(\mathbf{x}) \odot ||\mathbf{x} - \mathbf{x}'||_1}{\sum S_j(\mathbf{x})} + D_{\text{KL}}(O(\mathbf{x})||O(\mathbf{x}')). \tag{12}$$

Here, $S(\cdot)$ is a pre-trained semantic segmentation network that produces a label map for different regions in the input domain. Rather than minimizing a distance such as ℓ_1 over the entire image, we minimize the reconstruction loss for each segmentation label (j). Such a loss heavily penalizes differences in small regions to enforce local consistency.

$O(\mathbf{x})$ is a pre-trained object detector that, given an input image $\mathbf{x}$, outputs a number of bounding boxes called region of interests (ROIs). For each bounding box, it outputs 2-d coordinates in the image where the box is located and an associated probability of presence of an object. Using the input image $\mathbf{x}$, we obtain the ROIs and associated $O(\mathbf{x})$, which is a probability vector, stating probability of finding an object in each ROI. For reconstructed image $\mathbf{x}'$, we reuse the ROIs obtained from image $\mathbf{x}$ and computed the associated probabilities for the reconstructed image as $O(\mathbf{x}')$. Next, we used KL divergence to quantify the difference between probability vectors as $D_{\text{KL}}(O(\mathbf{x})||O(\mathbf{x}'))$, in Eq. 12. Finally, we used the CAR loss to enforce two important properties of the explanation function:

1. If $\mathbf{c} = f(\mathbf{x})$, the self-reconstructed image should resemble the input image.
2. For $\mathbf{c} \neq f(\mathbf{x})$, applying a reverse perturbation on the explanation image $\mathbf{x_c}$ should recover the initial image *i.e.*, $\mathbf{x} \approx \mathcal{I}_f(\mathcal{I}_f(\mathbf{x}, \mathbf{c}), f(\mathbf{x}))$.

We enforce these two properties by the following loss,

$$\mathcal{L}_{\text{rec}}(E, G) = \mathcal{L}_{\text{rec}}(\mathbf{x}, \mathcal{I}_f(\mathbf{x}, f(\mathbf{x}))) + \mathcal{L}_{\text{rec}}(\mathbf{x}, \mathcal{I}_f(\mathcal{I}_f(\mathbf{x}, \mathbf{c}), f(\mathbf{x}))). \tag{13}$$

We minimize this loss only for reconstruction of the input image. Please note, the classifier f and support networks $S(\cdot)$ and $O(\cdot)$ remained fixed during training.

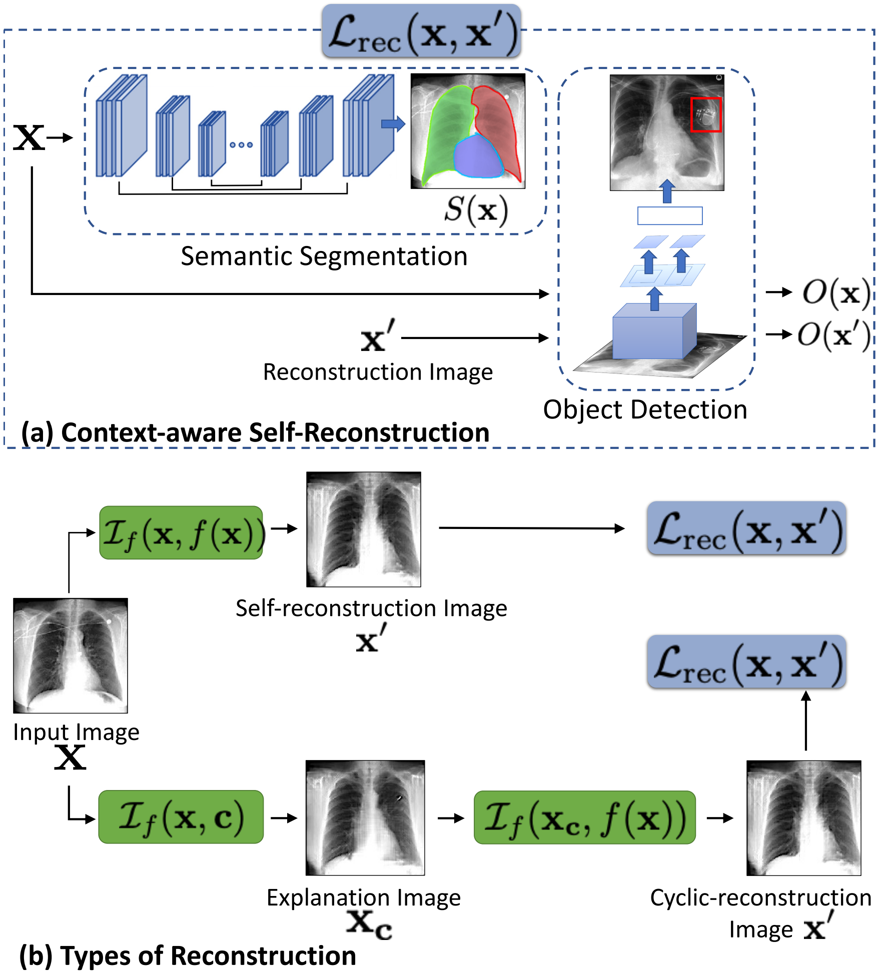

$\mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}')$
$\mathbf{x}$
$S(\mathbf{x})$
Semantic Segmentation
$O(\mathbf{x})$
$O(\mathbf{x}')$
$\mathbf{x}'$
Reconstruction Image
Object Detection
(a) Context-aware Self-Reconstruction
$\mathcal{I}_f(\mathbf{x}, f(\mathbf{x}))$
Self-reconstruction Image $\mathbf{x}'$
$\mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}')$
$\mathcal{L}_{\text{rec}}(\mathbf{x}, \mathbf{x}')$
Input Image $\mathbf{x}$
$\mathcal{I}_f(\mathbf{x}, \mathbf{c})$
$\mathcal{I}_f(\mathbf{x_c}, f(\mathbf{x}))$
Explanation Image $\mathbf{x_c}$
Cyclic-reconstruction Image $\mathbf{x}'$
(b) Types of Reconstruction

Figure 11: (a) A context-aware self-reconstruction loss with pre-trained semantic segmentation $S(\mathbf{x})$ and object detection $O(\mathbf{x})$ networks. (b) The self and cyclic reconstruction should retain maximum information from $\mathbf{x}$. Note, explanation image $\mathbf{x_c}$ may differ from input image, $\mathbf{x}$.

4.3 Experiments and Results

4.3.1 Study cohort and imaging dataset

Our experiments are conducted on the CelebA [149] and MIMIC-CXR [112] datasets. CelebA contains 200K celebrity face images, each with forty attribute labels. MIMIC-CXR is a multi-modal dataset consisting of 473K chest X-ray images and 206K reports from 63K patients. The dataset provides binarized labels over fourteen radio-graphic observations, namely, enlarged cardiomediastinum, cardiomegaly, lung-lesion, lung-opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture, support devices and no-finding. The images are preprocessed using a standard pipeline involving cropping, re-scaling and intensity normalization. We consider a multi-label classifier that takes a frontal view chest x-ray image as input and outputs a posterior probability for the fourteen radio-graphic observations.

4.3.2 Experimental setup

Classification model: CelebA - We considered two independently trained binary classifiers trained on the "smiling" and "age" attributes. The classifiers are deep learning models with a ResNet [90] backbone. For training the classifier, we used the default test and train split as provided by the dataset. The classifiers are very accurate with a test AUC-ROC greater than 0.90. We also experimented with other attributes.

MIMIC-CXR - Following the prior work on diagnosis classification [108], we used DenseNet-121 [98] architecture as the baseline classification model. The model is trained on 198K (∼80%) frontal view CXR from 51K patients and is test on a held-out set of 50K images from 12K non-overlapping patients. We use the Adam optimizer with default β-parameters of $\beta_1 = 0.9$, $\beta_2 = 0.999$ and learning rate $1\times\ 10^{-4}$ which is fixed for the duration of the training. We used a batch size of 16 images and train for 3 epochs, saving checkpoints every 4800 iterations. The classifier have a mean AUC-ROC of 0.75. It is highly discriminative for three diagnosis: cardiomegaly (AUC-ROC = 0.87), pleural effusion (AUC-ROC = 0.95) and edema (AUC-ROC = 0.91). These results are comparable to performance of the published model [108].

Explanation function: Our explanation function is implemented using TensorFlow version 2.0 and is trained on NVIDID P100 GPU. Before training the explanation function, we assume access to the pre-trained classification function, that we aim to explain. We also assume access to pre-trained segmentation and object detection networks, that are used to enforce CARL loss.

The explanation function is a cGAN with an encoder $E(\cdot)$ and generator $G(\cdot)$ network, that follows ResNet [90] architecture. $G(\cdot)$ uses conditional batch normalization (cBN) to incorporate condition information. For discriminator $D(\cdot)$ network, we adapted the architecture from SNGAN [171]. We optimized the adversarial hinge loss for the cGAN training. We set the loss hyper-parameters as $\lambda_1 = 1.0$, $\lambda_2 = 1.0$ and $\lambda_3 = 0.5$. We used the Adam optimizer [126], with default hyper-parameters set to $\alpha = 0.0002, \beta_1 = 0, \beta_2 = 0.9$.

Using PCE, we can derive qualitative explanations for any target class. However, for chest imaging dataset currently we support quantitative metrics only for cardiomegaly and PE. Deriving such metrics for other diagnosis requires understanding of the clinical definition of the disease and is challenging. Previously, researcher have shown qualitative results on other diagnosis [35], but quantitative evaluation is still missing. We train three independent cGANs to explain target classes: cardiomegaly, PE, and edema.

To train PCE, we used 30K images from held-out set, which was not used in training the classifier. We divide $f(\mathbf{x})[y] \in [0, 1]$ into $N = 10$ equally size bins. Here, y is a target class. Each input image is mapped to a bin-index depending on the prediction $f(\mathbf{x})[y]$. We randomly sample images such that each bin has 2500 to 3000 images. After training the cGAN, we audited the classifier by deriving explanations on a held-out set of 20K images. Please refer appendix A.5 for further details.

Semantic segmentation network function: Semantic segmentation network $S(\cdot)$ is a 2D U-Net [216]. It marks the lung and the heart contour in a chest x-ray. The network is trained on 385 chest x-rays and masks from Japanese Society of Radiological Technology (JSRT) [268] and Montgomery [109] datasets.

Object detector: We trained a faster regional CNN [211] network, for detecting FO such as pacemaker and hardware in a chest x-ray. The network learns to detect FO by placing a bounding box over them. To create the training dataset, we extracted 300 x-rays with a positive mention of these objects in the corresponding radiology reports, and collected bounding box annotations to mark the ground truth. We further trained two more detectors to evaluate our explanations. Specifically, we trained detectors for identifying normal and abnormal costophrenic (CP) recess region in the chest x-ray. We associated an abnormal

CP recess with the radiological finding of a blunt CP angle as identified by the positive mention for *"blunting of the costophrenic angle"* in the corresponding radiology report. For the normal-CP recess, we considered images with a positive mention for *"lungs are clear"* in the reports.

We compared our counterfactual explanations with closest existing methods such as xGEM[113] and CycleGAN [185, 44]. For proper comparison, we used the open-source implementation of these models and trained them on the MIMIC-CXR dataset, using the same training-set as our model. Please refer SM-Sec A.8 for more details. We also compared against the saliency-based methods to provide *post-hoc* model explanation. We performed following experiments:

1. *Desiderata of explanation function:* We compared our model with existing methods on the three desiderata of valid explanations and evaluated the following metrics: Fréchet Inception Distance (FID) score to assess visual quality, counterfactual validity (CV) score to quantify compatibility with the classifier, and face verification accuracy and foreign object preservation (FOP) score to evaluate the identity preservation in the explanations.
2. *Comparison with saliency-maps:* We compared the localization ability of our counterfactual explanations against the saliency maps generated by gradient-based methods.
3. *Clinical evaluation:* We used two clinical metrics, namely, cardiothoracic ratio (CTR) and the Score for detecting a normal Costophrenic recess (SCP) to demonstrate the clinical relevance of our explanations.
4. *Bias detection:* We trained two classifiers on biased and unbiased data and examined the performance of our method in identifying the bias.
5. *Evaluating class discrimination:* We trained a multi-label classifier and demonstrate the sensitivity of PCE to the task being explained.
6. *Human evaluation:* We evaluate the strength of PCE in explaining the classifier's decision to the end-user.

4.3.3 Desiderata of explanation function

We evaluated our method on three desiderata of a valid counterfactual [178]. First, *Data consistency:* A counterfactual should be realistic-looking *i.e.*, it should be very similar to the input image. Second, *Classifier consistency:* A counterfactual should flip the classification decision for the input image. Third, *Identify preservation:* A counterfactual should preserve patient-specific details such as FOs.

4.3.3.1 Data consistency

Given an input image, our model generates a series of images $\mathbf{x_c}$ as explanations by gradually changing $\mathbf{c}$ in range $[0, 1]$. Figure. 12 shows the qualitative results on CelebA dataset. We show results for two prediction tasks: smiling or not-smiling and young or old. Bottom two rows of Figure. 12 shows our result on MIMIC-CXR dataset. The left-most image is the input CXR. For **cardiomegaly**, we highlight the heart contour (yellow). Its helps in visualizing enlargement of the cardiac silhouette. For **PE**, we showed the results of an object detector as bounding-box (BB) over the normal (blue) and abnormal (red) CP recess regions. The number on the top-right of the blue-BB is the Score for detecting a normal CP recess (SCP). The number on red-BB is 1-SCP. The CP recess is the potential area to be analyzed for PE [195]. From left to right, the normal CP recess changed into an abnormal CP recess with a high detection score. We observed a gradual increase in posterior probability $f(\mathbf{x_c})$ (bottom label) as we go from left to right.

Quantitatively evaluation: We evaluated the visual quality of our explanations by computing Fréchet Inception Distance (FID) score [94]. FID quantifies the visual similarity between the real images $\mathbf{x}$ and the synthetic counterfactuals $\mathbf{x_c}$ by computing distance between their activation distributions as follow,

$$\text{FID}(\mathbf{x}, \mathbf{x_c}) = ||\mu_{\mathbf{x}} - \mu_{\mathbf{x_c}}||_2^2 + \text{Tr}(\Sigma_{\mathbf{x}} + \Sigma_{\mathbf{x_c}} - 2(\Sigma_{\mathbf{x}}\Sigma_{\mathbf{x_c}})^{\frac{1}{2}}), \tag{14}$$

where μ's and Σ's are mean and covariance of the activation vectors derived from the penultimate layer of a pre-trained Inception v3 network [94]. The pre-trained network is trained

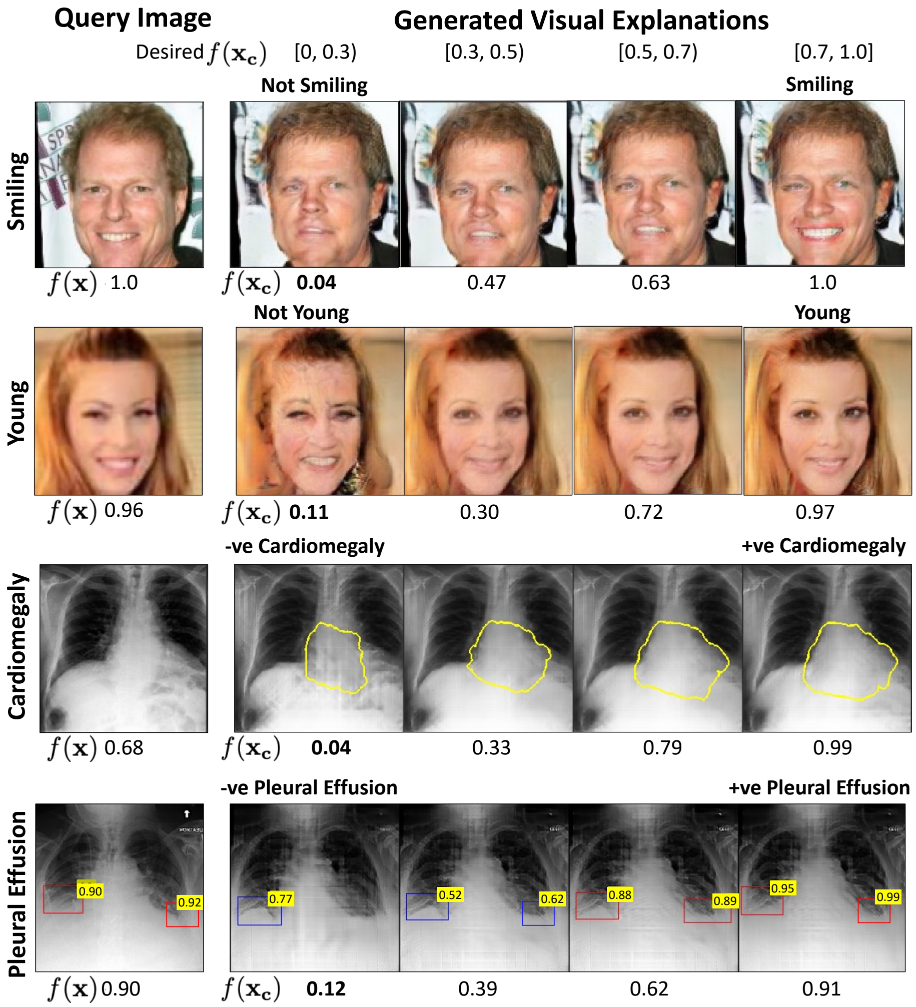


Figure 12: Progressive counterfactual explanations generated for different prediction tasks. The figure shows smiling/not-smiling (first row), young/old face (second row), diagnosis of cardiomegaly (third row) and diagnosis of pleural effusion (last row). The first column shows the query image, followed by the corresponding generated explanations. The bottom label is the output of the classifier f. For Cardiomegaly, we show the segmentation of the heart (yellow edge). For PE, we show the bounding box (BB) for normal (blue) and abnormal (red) costophrenic (CP) recess. Extended results in SM-Figure. 45.

on same dataset as the PCE. We examined real and synthetic (*i.e.*, generated explanations) images on the two extreme of the decision boundary, *i.e.*, a normal group ($f(\mathbf{x}) < 0.2$) and an abnormal group ($f(\mathbf{x}) > 0.8$). In Table. 6, we compared three counterfactual-generating algorithms: ours, xGEM, and cycleGAN, and reported the FID for each group. Our model creates natural-looking counterfactuals compared to xGEM. The cycleGAN model generates the most visually appealing images with the lowest FID score across the classification tasks. However, a visually good image doesn't necessarily means a good counterfactual. It is also equally important to flip the classification decision as explained in next section.

Table 6: The FID score quantifies the visual appearance of the explanations. The counterfactual validity (CV) score is the fraction of explanations that have an opposite prediction compared to the input image. An ideal counterfactual explanation have low FID score and a high CV score.

	Negative ($f(\mathbf{x}), f(\mathbf{x_c}) < 0.2$)			**Positive** ($f(\mathbf{x}), f(\mathbf{x_c}) > 0.8$)		
Target class	Ours	xGEM	CycleGAN	Ours	xGEM	CycleGAN
			Fréchet Inception Distance (FID)			
CelebA:Smiling	56.3	112.9	**30**	46.9	111.0	**37**
CelebA:Young	74.4	170.3	**56**	67.6	115.2	**35**
CXR:Cardiomegaly	166	138	**30**	137	316	**56**
CXR:Pleural Effusion	146	347	**37**	122	355	**35**
CXR:Edema	149	376	**72**	102	274	**77**
			Counterfactual Validity Score			
CXR:Cardiomegaly	**0.91**	**0.91**	0.43			
CXR:Pleural Effusion	**0.97**	**0.97**	0.49			
CXR:Edema	**0.98**	0.66	0.57			

4.3.3.2 Classification model consistency

By definition, a counterfactual image have an opposite classification decision as compared to the query image. A counterfactual provides explanation by showing how image-features should be modified to flip the classification decision. If the decision doesn't flip then the explanation is inconclusive. Counterfactual validity (CV) score is the fraction of counterfactual explanations that successfully flipped the classification decision *i.e.*, if the input image is negative (normal) then the generated explanation is predicted as positive (abnormal) for the target class. We compared different counterfactual-generating algorithms on CV score [178] metric. The last rows of Table. 6, summarizes our result. For all tasks, our model consistently achieved the highest CV score. CycleGAN achieved a low CV score, thus creating explanations that are frequently inconsistent with the classifier.

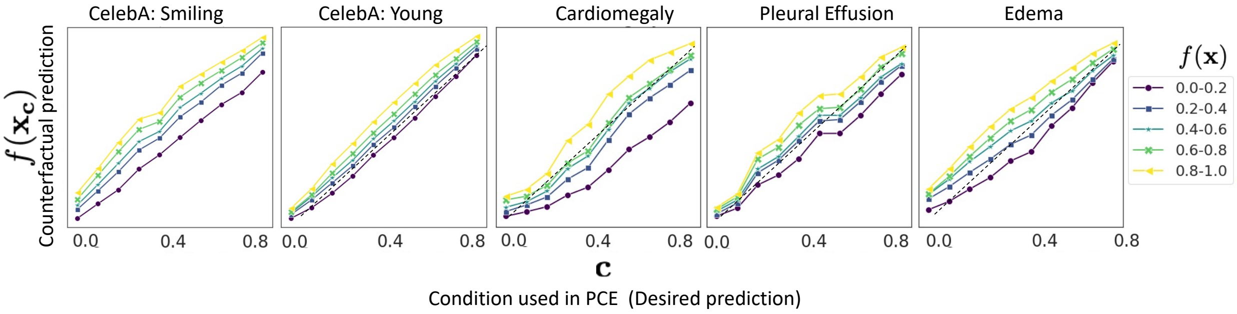


Figure 13: Plot of the desired outcome from the classifier, $\mathbf{c}$, against the actual response of the classifier on generated explanations, $f(\mathbf{x_c})$. The monotonically increasing trend shows a positive correlation between $\mathbf{c}$ and $f(\mathbf{x_c})$, and thus the generated explanations are consistent with the expected condition. Each line represents a set of input images with prediction $f(\mathbf{x})$ in a given range.

Next, we quantify this consistency at every step of the transformation. We divided the prediction range $[0, 1]$ into $N = 10$ equally sized bins. For each bin, we generated an explanation image by choosing an appropriate, $\mathbf{c}$. We further divided the input image space into five groups based on their initial prediction *i.e.*, $f(\mathbf{x})$. In Figure. 13, we represented each group as a line and plotted the average response of the classifier *i.e.*, $f(\mathbf{x_c})$ for explanations in each bin against the expected outcome *i.e.*, $\mathbf{c}$. The positive slope of the line-plot, parallel to $y = x$ line confirms that starting from images with low $f(\mathbf{x})$, our model creates fake images such that $f(\mathbf{x_c})$ is high and vice-versa.

4.3.3.3 Identity preservation

The counterfactual explanations should differ only in semantic features associated with the target class while retaining the identity of the query image. For example, in CelebA, if the classifier is using image features near the lips to decide smiling or not, the other features such as hair and the person in the image should remain the same. Similarly, in CXR, any foreign objects (FOs) such as pacemaker should be preserved. Furthermore, FO provides critical information to identify the patient in an x-ray. The disappearance of FO in explanation images may create confusion that explanation images show a different patient.

For PCE trained on CelebA dataset, we evaluate the identify preservation in counterfactual explanations through face verification. In face verification, we quantify the similarity between the faces in query image and corresponding fake counterfactual explanation image. We used state-of-the-art face recognition model trained on VGGFace2 dataset [19] as feature extractor for both real images and their corresponding fake explanations. For face verification, we calculated the closeness between real and fake image as cosine distance between their feature vectors. The faces were considered as verified *i.e.*, fake explanation have same identity as real image, if the distance is below 0.5. In Table 7, we report face verification accuracy as percentage of the verified query image and fake counterfactual image pairs. We evaluated this metric over a randomly sampled test set of 500 images.

Table 7: Results on face-verification task to demonstrate that the identity of a person is preserved across counterfactual explanations.

Target Class	**Face Verification Accuracy**
CelebA: Smiling	85.3%
CelebA:Young	72.2%

Table 8: The foreign object preservation (FOP) score with and without the context-aware reconstruction loss (CARL). FOP score depends on the performance of FO detector.

Foreign	**FOP score**	
Objects	**Ours with CARL**	**Ours with ℓ_1**
Pacemaker	**0.52**	0.40
Hardware	**0.63**	0.32

For PCE trained on CXR dataset, we quantify the strength of our revised CARL loss in preserving FO in explanation images compared to an image-level ℓ_1 reconstruction loss. We reported results on the FO preservation (FOP) score metric. FOP score is the fraction of real images, with successful detection of FO, in which FO was also detected in the corresponding explanation image $\mathbf{x_c}$. Our model with CARL obtained a higher FOP score, as shown in Table 8. The FO detector network has an accuracy of 80%.

Figure. 14 presents examples of counterfactual explanations generated by our model with and without the CARL. Our results confirm that CARL is an improvement over ℓ_1 reconstruction loss. We further provide a detailed ablation study over different components of our loss in appendix A.12.

4.3.4 Comparison with saliency-maps

Popular existing approaches for model explanation consist of gradient-based methods that provide a qualitative explanation in the form of saliency maps [200, 108]. Saliency maps show the importance of each pixel of an image in the context of classification. Our method is not designed to produce saliency maps as a continuous score for every feature of the input. To compare against such methods, we approximated a saliency map as an absolute difference map between the explanations generated for the two extremes; negative decision with $f(\mathbf{x_c}) < 0.2$ and positive decision with $f(\mathbf{x_c}) > 0.8$. For proper comparison, we considered the absolute values of the saliency maps and normalized them in the range

[0, 1].

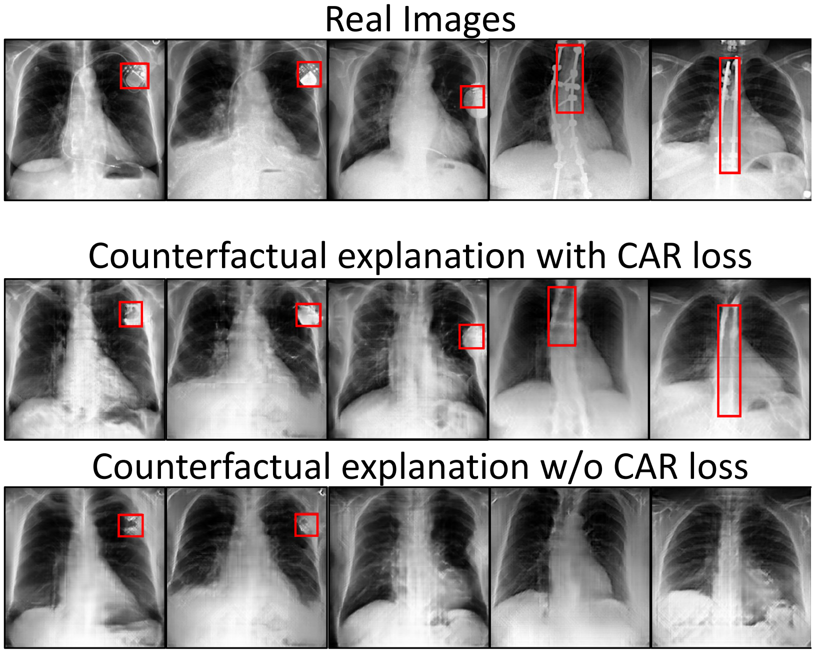

Real Images
Counterfactual explanation with CAR loss
Counterfactual explanation w/o CAR loss

Figure 14: Fidelity of generated images with respect to preserving FO.

Figure. 15 shows the saliency map obtain from our method and its comparison with popular gradient based methods. For CelebA, we compare the explanations derived for the "smiling" classifier. For CXR dataset, we show an example of an input image, where the gradient-based saliency maps highlight almost the same region for two different target tasks. In contrast, our difference map localized disease to specific regions in the chest. Figure. 15.C shows the two extreme explanation images and the corresponding difference map, derived for input images shown in Figure. 15.A.

For quantitative evaluation, we used the *deletion* evaluation metric to compare our difference map with saliency maps produced by different gradient-based methods [204]. The deletion metric quantifies how the probability of the target-class changes as important pixels are removed from an image. A sharp drop in in prediction accuracy, resulting in a low area under the probability curve (AUC) (as a function of the percentage of the salient pixels removed), represents a good explanation. To remove salient pixels from an image, in CXR images, we selectively impaint the removing regions based on its surrounding.

For CXR dataset, in Table 9, we report the mean AUC over a sample of 500 images. The

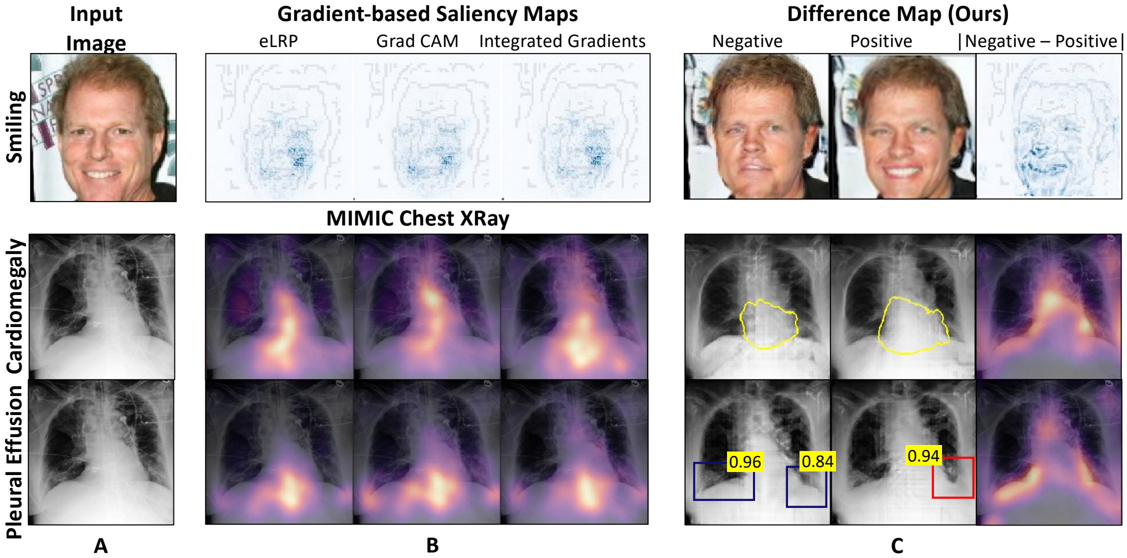


Figure 15: Comparison of our method against different gradient-based methods. A: Input image; B: Saliency maps from existing works; C: Our simulation of saliency map as difference map between the normal and abnormal explanation images. More examples are shown in SM-Figure. 47.

Table 9: Quantity comparison of our method against gradient-based methods. Mean area under the probability curve (AUC), plotted as a function of the fraction of removed pixels. A low AUC shows a sharp drop in prediction accuracy as fraction of removed pixels increases.

Method	**Cardiomegaly**	**Pleural Effusion**	**Edema**
Ours	**0.040±0.04**	**0.023±0.02**	0.083±0.05
eLRP	0.071±0.05	0.033±0.02	0.055±0.03
Grad-CAM	0.045±0.04	0.058±0.05	**0.035±0.02**
Integrated Gradients	0.058±0.06	0.046±0.05	0.077±0.04

images were selected such that the $f(\mathbf{x}) > 0.9$ for the target-disease. Our model achieved the lowest AUC in deletion-by-impainting for cardiomegaly and pleural effusion. In Figure. 16, we show an example of deletion-by-impainting. The results show that the regions modified by our explanation model are important for the classification decision.

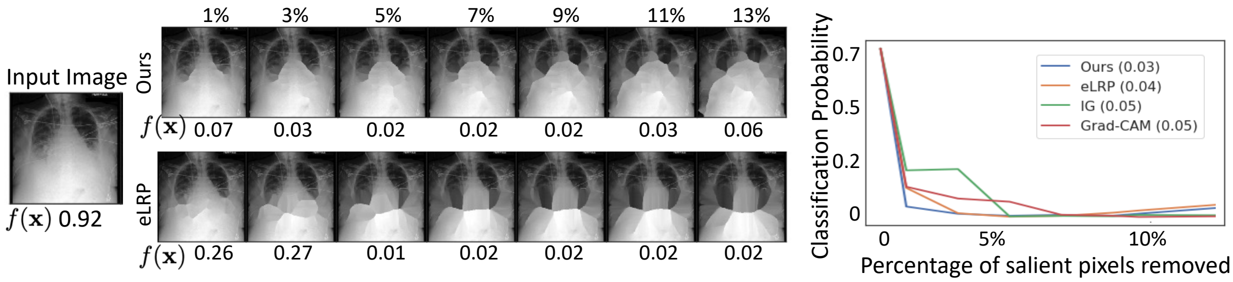


Figure 16: Evaluation using deletion metric. The plot shows the drop in classification probability for pleural effusion as important pixels are removed from the input image. Top label shows the percentage of removed pixels. The bottom label shows the classification prediction.

4.3.5 Clinical evaluation

In this experiment, we demonstrate the clinical relevance of our explanations. First, we translate the clinical definition of two diseases (cardiomegaly and pleural effusion) into quantitative metrics. Next, we used these clinical metrics to quantify the counterfactual changes between normal and abnormal populations, as identified by the given classifier. If the change in classification decision is associated with the corresponding change in clinical-metric, we can conclude that the classifier considers clinically relevant information in its diagnosis prediction. We considered the following two metrics:

4.3.5.1 Cardio Thoracic Ratio (CTR)

The CTR is the ratio of the cardiac diameter to the maximum internal diameter of the thoracic cavity. A CTR ratio greater than 0.5 indicates cardiomegaly [166, 22, 48]. We followed the approach in [23] to calculate CTR from a CXR. In the absence of ground truth lung and heart segmentation on the MIMIC-CXR dataset, we used a segmentation network trained on open-sourced supervised datasets [155, 109]. We calculated heart diameter as the distance between the leftmost and rightmost points from the lung centerline on the heart segmentation. The thoracic diameter is calculated as the horizontal distance between the widest points on the lung mask. Please refer appendix A.6 for details on segmentation network.

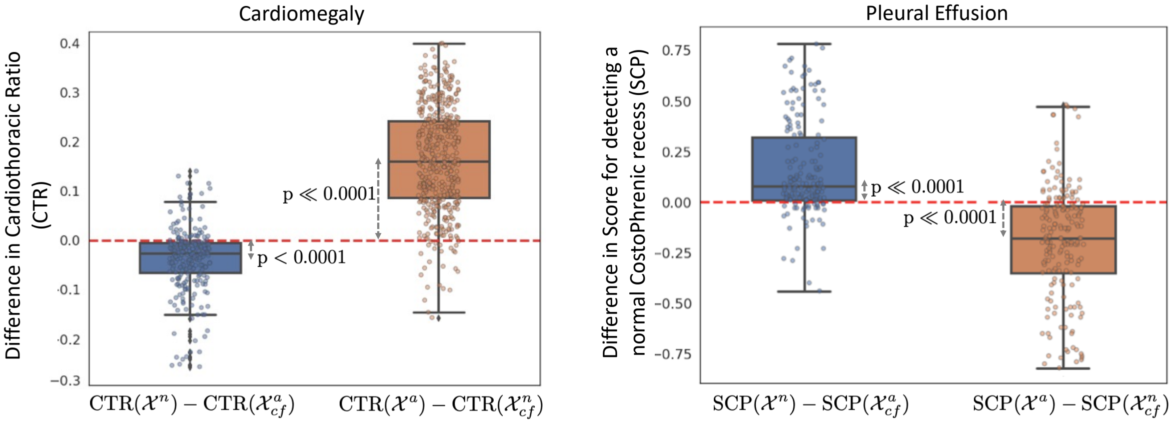


Figure 17: Box plots to show distributions of pairwise differences in clinical-metrics. We consider clinical metrics such as CTR for cardiomegaly and the Score of normal CP recess (SCP) for pleural effusion, before (real) and after (counterfactual) our generative counterfactual creation process. The mean value corresponds to the average causal effect of the clinical-metric on the target disease. The low p-values for the dependent t-test statistics confirms the statistically significant difference in the distributions of metrics for real and counterfactual images. Further numbers are summarized in SM-Table 20.

4.3.5.2 Costophrenic recess

The fluid accumulation in costophrenic (CP) recess may lead to the diaphragm's flattening and the associated blunting of the angle between the chest wall and the diaphragm arc, called costophrenic angle (CPA). The blunt CPA is an indication of pleural effusion [156].

Marking the CPA angle on a CXR requires expert supervision, while annotating the CP region with a bounding box is a much simpler task (*see* SM-Figure. 41). We learned an object detector to identify normal or abnormal CP recess in the CXRs and used the Score for detecting a normal CP recess (SCP) as our evaluation metric. Further details on the training of the object detector are provided in appendix A.7.

We performed a statistical test to quantify the differences in real images and their corresponding counterfactuals based on the above two metrics. We randomly sample two groups of real images (1) a *real-normal* group defined as $\mathcal{X}^n = \{\mathbf{x}; f(\mathbf{x}) < 0.2\}$. It consists of real CXR that are predicted as normal by the classifier f. (2) A *real-abnormal* group defined as $\mathcal{X}^a = \{\mathbf{x}; f(\mathbf{x}) > 0.8\}$. For $\mathcal{X}^n$ we generated a counterfactual group as, $\mathcal{X}^a_{cf} = \{\mathcal{I}_f(\mathbf{x}, \mathbf{c}); \mathbf{x} \in \mathcal{X}^n; \mathbf{c} > 0.8\}$. Similarly for $\mathcal{X}^a$, we derived a counterfactual group as $\mathcal{X}^n_{cf} = \{\mathcal{I}_f(\mathbf{x}, \mathbf{c}); \mathbf{x} \in \mathcal{X}^a; \mathbf{c} < 0.2\}$.

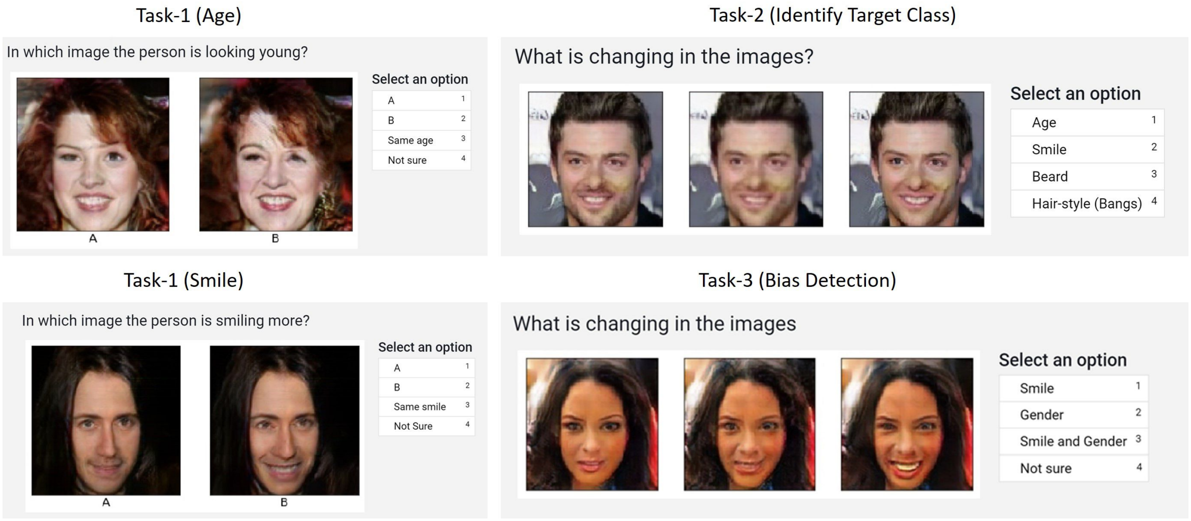
In Figure. 17, we showed the distribution of differences in CTR for cardiomegaly and SCP for PE in a pair-wise comparison between real (normal/abnormal) images and their respective counterfactuals. Patients with cardiomegaly have higher CTR as compared to normal subjects. Hence, one should expect CTR($\mathcal{X}^n$) $<$ CTR($\mathcal{X}^a_{cf}$) and likewise CTR($\mathcal{X}^a$) $>$ CTR($\mathcal{X}^n_{cf}$). Consistent with clinical knowledge, in Figure. 17, we observe a negative mean difference for CTR($\mathcal{X}^n$) $-$ CTR($\mathcal{X}^a_{cf}$) (a p-value of < 0.0001) and a positive mean difference for CTR($\mathcal{X}^a$) $-$ CTR($\mathcal{X}^n_{cf}$) (with a p-value of $\ll 0.0001$). The low p-value in the dependent t-test statistics supports the alternate hypothesis that the difference in the two groups is statistically significant, and this difference is unlikely to be caused by sampling error or by chance.

By design, the object detector assigns a low SCP to any indication of blunting CPA or abnormal CP recess. Hence, SCP($\mathcal{X}^n$) $>$ SCP($\mathcal{X}^a_{cf}$) and likewise SCP($\mathcal{X}^a$) $<$ SCP($\mathcal{X}^n_{cf}$). Consistent with our expectation, we observe a positive mean difference for SCP($\mathcal{X}^n$) $-$ SCP($\mathcal{X}^a_{cf}$) (with a p-value of $\ll 0.0001$) and a negative mean difference for SCP($\mathcal{X}^a$) $-$ SCP($\mathcal{X}^n_{cf}$) (with a p-value of $\ll 0.0001$). A low p-value confirmed the statistically significant difference in SCP for real images and their corresponding counterfactuals. For further details and visual examples of samples in normal and abnormal groups, please refer appendix A.11.

4.3.6 Human Evaluation

4.3.6.1 CelebA dataset

We used Amazon Mechanical Turk (AMT) to conduct human experiments to demonstrate that the progressive exaggeration produced by our model is visually perceivable to humans. We presented AMT workers with three tasks. In the first task, we evaluated if humans can detect the relative order between two explanations produced for a given image. We ask the AMT workers, “Given two images of the same person, in which image is the person younger (or smiling more)?” (*see* Figure 18). We experimented with 200 query images and generated two pairs of explanations for each query image (*i.e.*, 400 hits). The first pair (*easy*) imposed the two images are samples from opposite ends of the explanation spectrum (counterfactuals), while the second pair (*hard*) makes no such assumption.



Task-1 (Age)

In which image the person is looking young?

Task-2 (Identify Target Class)

What is changing in the images?

Task-1 (Smile)

In which image the person is smiling more?

Task-3 (Bias Detection)

What is changing in the images

Figure 18: The interface for the human evaluation done using Amazon Mechanical Turk (AMT). Task-1 evaluated if humans can detect the relative order between two explanations. Task-2 evaluated if humans can identify the target class for which our model has provided the explanations. Task-3 demonstrated that our model can help the user to identify problems like possible bias in the black-box training.

In the second task, we evaluated if humans can identify the target class for which our model has provided the explanations. We ask the AMT workers, "What is changing in the images? (age, smile, hair-style or beard)". We experimented with 100 query images from each of the four attributes (*i.e.*, 400 hits). In the third task, we demonstrate that our model can help the user to identify problems like possible bias in the black-box training. Here, we used the same setting as in the second task but also showed explanations generated for a biased classifier. We ask the AMT workers, "What is changing in the images? (smile or smile and gender)" (*see* Figure 18). We generated explanations for 200 query images each, from a biased-classifier (f_{Biased}) explainer from Section 4.3.7 and an unbiased classifier ($f_{\text{No-biased}}$) explainer (*i.e.*, 400 hits). In all the three tasks, we collected eight votes for each task, evaluated against the ground truth, and used the majority vote for calculating accuracy.

We summarize our results in Table 10. In the first task, the annotators achieved high accuracy for the *easy* pair when there was a significant difference among the two explanation images, as compared to the *hard* pair when the two explanations can have very subtle differences. Overall, the annotators were successful in identifying the relative order between the two explanation images.

In the second task, the annotators were generally successful in correctly identifying the target class. The target class "bangs" proved to be the most difficult to identify, which was expected. The generated images for "bangs" were qualitatively, the most subtle. For the third task, the correct answer was always the target class *i.e.*, "smile". In the case of biased classifier explainer, the annotators selected "Smile and Gender" 12.5% of the times. The gradual progression made by the explainer for a biased classifier was very subtle and was changing large regions of the face as compared to the unbiased explainer. The difference is much more visible when we compare the explanation generated for the same query image for a biased and no-biased classifier, as in Figure 20. But in a realistic scenario, the no-biased classifier would not be available to compare against. Nevertheless, the annotators detected bias at roughly the same level of accuracy as our classifier (Table 12). Future work could improve upon bias detection.

Table 10: Summarizing the results of human evaluation. The κ -statistics measure inter-rater agreement for qualitative classification of items into some mutually exclusive categories. One possible interpretation of κ as given in [272] is < 0.0: Poor, $0.01 - 0.2$: Slight, $0.21 - 0.40$: Fair, $0.41 - 0.60$: Moderate, $0.61 - 0.80$: Substantial and $0.81 - 1.00$: Almost perfect agreement.

Annotation Task	**Overall**		**Sub categories**		
	Accuracy	κ-statistic	Category	Accuracy	κ-statistic
Task-1 (Age)	83.5%	0.41 (Moderate)	Hard	73%	0.31 (Fair)
			Easy	**94%**	0.51 (Moderate)
Task-1 (Smile)	77.5%	0.28 (Fair)	Hard	66%	0.23 (Fair)
			Easy	**89.5%**	0.32 (Fair)
Task-2 (Identify Target Class)	77%	0.35 (Fair)	Age	72%	-
			Smile	**99%**	-
			Bangs	50%	-
			Beard	87%	-
Task-3 (Bias Detection)	93.75%	0.14 (Slight)	f_{Biased}	87.5%	0.09 (Slight)
			$f_{\text{No-biased}}$	**100%**	0.02 (Slight)

4.3.6.2 MIMIC CXR Dataset

We conducted a human-grounded experiment with diagnostic radiology residents to compare different styles of explanations (no explanation, saliency map, cycleGAN explanation, and our counterfactual explanation) by evaluating different aspects of explanations: (1) understandability, (2) classifier's decision justification, (3) visual quality, (d) identity preservation, and (5) overall helpfulness of an explanation to the users.

Our results show that our counterfactual explanation was the only explanation method that significantly improved the users' understanding of the classifier's decision compared to the no-explanation baseline. In addition, our counterfactual explanation had a significantly higher classifier's decision justification than the cycleGAN explanation, indicating that the participants found a good evidence for the classifier's decision more frequently in our counterfactual explanation as compared to cycleGAN explanation.

Further, cycleGan explanation performed better in terms of visual quality and identity preservation. However, at times the cycleGAN explanations were identical to the query image, thus providing inconclusive explanations. Overall the participants found our explanation method the most helpful method in understanding the assessment made by the AI system in comparison to other explanation methods. Below, we describe the design of the study, the data analysis methods, along with the results of the experiment in detail.

Experiment Design: We conducted an online survey experiment with 12 diagnostic radiology residents. Participants first reviewed an instruction script, which described the AI system developed to provide an autonomous diagnosis for CXR findings such as cardiomegaly. The study comprised of the radiologists evaluating six CXR images which were presented in random order to them. For selecting these siz CXR, we first, divided the test-set of the explanation function for cardiomegaly into three groups, positive ($f(\mathbf{x}) \in [0.8, 1.0]$), mild ($f(\mathbf{x}) \in [0.4, 0.6]$) and negative ($f(\mathbf{x}) \in [0.0, 0.2]$). Next, we randomly selected two CXR images from each group. The six CXR images were anonymized as part of the MIMIC-CXR dataset protocol.

For each image, we had the same procedure consisted of a diagnosis tasks, followed by four explanation conditions, and ended by a final evaluation question between the explanation conditions. Further details of the study design are includes in appendix A.1.

Diagnosis: For each CXR image, we first asked a participant to provide their diagnosis for cardiomegaly. This question ensures that the participants carefully consider the imaging features that helped them diagnose. Subsequently, the participants were presented with the classifier's decision and were asked to provide feedback on whether they agreed.

Explanation Conditions: Next, the study provides the classifier's decision with the following explanation conditions:

1. **No explanation (Baseline)**: This condition simply provides the classifier decision without any explanation, and is used as the control condition.
2. **Saliency map**: A heat map overlaid on the query CXR, highlighting essential regions

for the classifier's decision.

3. **CycleGAN explanation**: A video loop over two CXR images, corresponding to the query CXR transformation with a negative and a positive decision for cardiomegaly.
4. **Our counterfactual explanation**: A video showing a series of CXR images gradually changing the classifier's decision from negative to positive.

Please note that after showing the baseline condition, we provided the other explanation conditions in random order to avoid any learning or biasing effects.

Evaluation metrics: Given the classifier's decision and corresponding explanation, we consider the following metrics to compare different explanation conditions:

1. **Understandability**: For each explanation condition, the study included a question to measure whether the end-user understood the classifier's decision, when explanation was provided. The participants were asked to rate agreement with *"I understand how the AI system made the above assessment for Cardiomegaly"*.
2. **Classifier's decision justification**: Human user's may perceive explanations as the reason for the classifier's decision. For the cycleGAN and our counterfactual explanation conditions, we quantify whether the provided explanation were actually related to the classification task by measuring the participants' agreement with *"The changes in the video are related to Cardiomegaly"*.
3. **Visual quality**: The study quantifies the proximity between the explanation images and the query CXR by measuring the participants' agreement with "*Images in the video look like a chest x-ray.*".
4. **Identity preservation**: The study also measures the extent to which participants think the explanation images correspond to the same subject as the query CXR by measuring the participants' agreement with *"Images in the video look like the chest x-ray from a given subject"*.
5. **Helpfulness:** For each CXR image, we asked the participants to select the most helpful explanation condition in understanding the classifier's decision, *"Which explanation helped you the most in understanding the assessment made by the AI system?"*. This evaluation metric directly compares the different explanation conditions.

All metrics, but the helpfulness metric were evaluated for agreement on a 5-point Likert scale, where one means "*strongly disagree*" and five means "*strongly agree*".

Free-form Response: After each question, we also asked the participants a free-form question: "*Please explain your selection in a few words.*" We used answers to these questions to triangulate our findings and complement our quantitative metrics by understanding our participants' thought-processes and reasoning.

Participants. Our participants include 12 diagnostic radiology residents who have completed medical school and have been in the residency program for one or more years. On average, the participants finished the survey in 40 minutes and were paid $100 for their participation in the study.

Data analysis: For each evaluation metric, the study asked the same question to the participants while showing them different explanations. For each question, we gather 72 responses (6 - number of CXR images × 12 - number of participants).

For the understandability and helpfulness metrics, we conducted a one-way ANOVA test to determine if there is a statistically significant difference between the mean metric scores for the four explanation conditions. Specifically, we built a one-way ANOVA with the metric as our dependent variable and analyzed agreement rating as the independent variable. If we found a significant difference in the ANOVA method, we ran Tukey's Honestly Significant Difference (HSD) posthoc test to perform a pair-wise comparison between different explanation conditions.

We measured the classifier's decision justification, visual quality and identity preservation metrics only for the cycleGAN and our counterfactual explanations. We conducted paired t-tests to compare these evaluation metrics between these two explanation conditions. We also qualitatively analyzed the participants' free-form responses to find themes and patterns in their responses.

Results: Fig. 19 shows the mean score for the evaluation metrics of understandability, classifier's decision justification, visual quality, and identity preservation among the different explanation conditions. Below, we report the statistical analysis for these results, followed by analysis of the participants' free-form responses to understand the reasons behind these results.

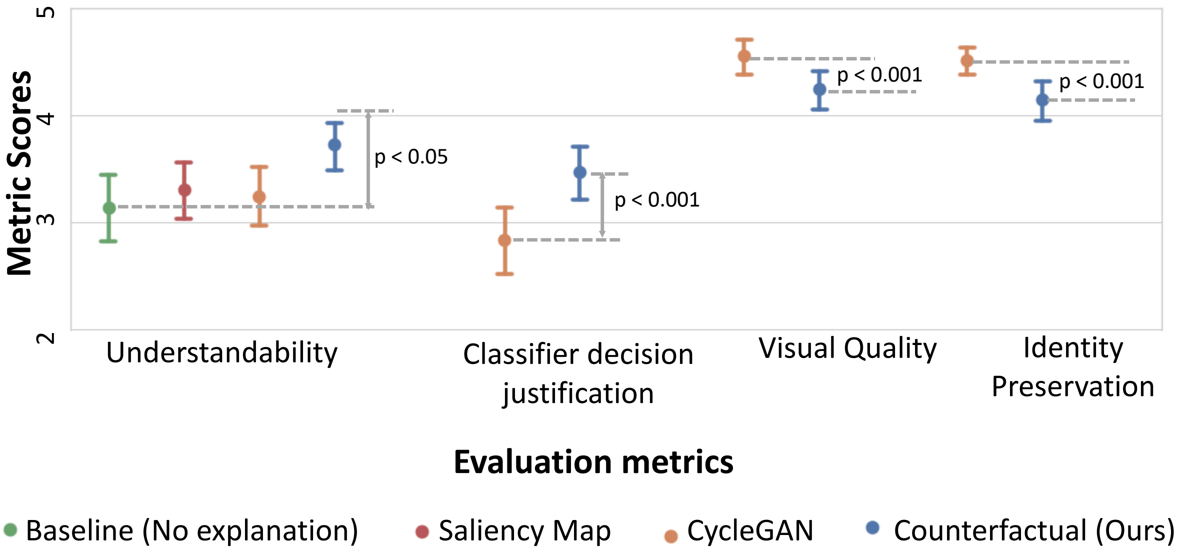


Figure 19: Comparing the different metrics in human evaluation study.

Understandability: The results show that our counterfactual explanation was the most understandable explanation to the participants. A one-way ANOVA revealed that there was a statistically significant difference in the understandability metric between at least two explanation conditions (F(3, 284) = [3.39], p=0.019). The Tukey post-hoc test showed that the understandability metric for our counterfactual explanation was significantly higher than the no-explanation baseline (p = 0.018). However, there was no statistically significant difference in mean scores between other pairs of explanations (refer to Table 11, "Understandability" column). This finding indicates that providing our counterfactual explanations along with the classifier's decision made the algorithm most understandable to our clinical participants, while other explanation conditions, saliency map and cycleGAN failed to achieve significant difference from no-explanation baseline on the understandability metric. Next, we use responses from free-text question to supplement our findings.

For the no-explanation baseline, the primary reason for poor understanding was the absence of explanation (n=30), (*e.g.*, they stated that *"there is no indication as to how the AI made this decision"*). Interestingly, many responses (n=23) either associated their high understanding with the correct classification decision *i.e.*, participants understood the decision as the decision is correct (*"I agree, it is small and normal"*) or they assumed the

AI-system is using similar reasoning as them to arrive at its decision (*"I assume the AI is just measuring the width of the heart compared to the thorax", "Assume the AI measured the CT ratio and diagnosed accordingly."*).

Participants' mostly found saliency maps to be correct but incomplete (n=23), (*"Unclear how assessment can be made without including additional regions"*). Specifically, for cardiomegaly, the saliency maps were highlighting parts of the heart and not its border (*"Not sure how it gauges not looking at the border"*) or thoracic diameter (*"thoracic diameter cannot be assessed using highlighted regions of heat map"*). We observe a similar result in Fig. 15, where the heatmap focuses on the heart but not its border. Further, some participants expressed a concern that they didn't understand how relevant regions were used to derive the decision (*"i understand where it examined but not how that means definite cardiomegaly"*).

For cycleGAN explanation, the primary reason for poor understanding was the minimal perceptible changes between the negative and positive images (n=3), (*"There is no change in the video."*). In contrast, many participant's explicitly reported an improved understanding of the classifier's decision in the presence of our counterfactual explanations (n=33), (*"I think the AI looking at the borders makes sense.", "i can better understand what the AI is picking up on with the progression video"*).

Classifier's decision justification: Our counterfactual explanation (M=3.46; SD=1.12) achieved a positive mean difference of 0.63 on this metric as compared to cycleGAN (M=2.83; SD=1.33), with t(71)=3.55 and $p < 0.001$. This result indicates that the participants found a good evidence for the predicted class (cardiomegaly), much frequently in our counterfactual explanations as compared to cycleGAN.

Most responses (n=25) explicitly mentioned visualizing changes related to cardiomegaly such as an enlarged heart in our explanation video as compared to cycleGAN (n=17). In cycleGAN, many reported that changes in the explanation video was not perceptible (n=23). Further, the participants reported changes in density, windowing level or other attributes which were not related to cardiomegaly (*"Decreasing the density does not impact how I assess for cardiomegaly.", "they could be or just secondary to windowing the radiograph"*). Such responses were observed in both cycleGAN (n=17) and our explanation (n=17). This indicates that the classifier may have associated such secondary information (short-cuts) with cardiomegaly diagnosis. A more in-depth analysis is required to quantify the classifiers' behaviour.

Visual quality and identity preservation: We observe a negative mean difference of 0.31 and 0.37 between our and cycleGAN explanation methods in visual quality and identity preservation metrics, respectively. The mean score for visual quality was higher for cycleGAN (M=4.55; SD=0.71) as compared to our method (M=4.24; SD=0.80) with t(71)=3.49 and $p < 0.001$. Similarly, the mean score for identity preservation was also higher for cycleGAN (M=4.51; SD=0.56) as compared to our method (M=4.14; SD=0.78) with t(71)=3.96 and $p < 0.001$.

Most of the responses (n=69) agreed that the CycleGAN explanation were marked as highly similar to the query CXR image. These results are consistent with our earlier results, that cycleGAN has better visual quality with a lower FID score (*see* Table. 6). However, in some responses, the participants pointed out that the explanation images were almost identical to the query image (*"There's virtually no differences. This is within the spectrum of a repeat chest x-ray for instance."*). An explanation image identical to the query image provides no information about the classifier's decision. Further, similar looking CXR will also result in similar classification decision, and hence will fail to flip the classification decision. As a result, we also observed a lower agreement in the classifier consistency metric and a lower counterfactual validity score in Table. 6 for cycleGAN.

Helpfulness: In our concluding question, *"Which explanation helped you the most in understanding the assessment made by the AI system?"*, 57% of the responses selected our counterfactual explanation as the most helpful method. A one-way ANOVA revealed that there was a statistically significant difference in the helpfulness metric between at least two explanation conditions (F(3, 284) = [21.5], $p < 0.0001$). In pair-wise Tukey's HSD posthoc test, we found that the mean usefulness metric for our counterfactual explanations was significantly different from all the rest explanation conditions($p < 0.0001$). Table 11 ("Helpfulness" column) summarizes these results.

These results indicates that the participant's selected our counterfactual explanations as the most helpful form of explanation for understanding the classifier's decision.

Table 11: Results for one-way ANOVA for understandability metric, followed by Tukey's HSD post-hoc test between different levels of agreement. Note that the mean value for E4 (our counterfactual explanation) is the highest, indicating that our explanations helped users the most in understanding the classifier's decision. *$p < 0.05$; ***$p < 0.0001$.

Understandability			**Helpfulness**		
F(3, 284) = 3.39			F(3, 284) = 21.5		
$p < 0.05$			$p < 0.001$		
Explanation method		p	Explanation method		p
E1 (No explanation)	E2		E1	E2	
M=3.14	E3		M=0.05	E3	
SD=1.39	E4	*	SD=0.23	E4	***
E2 (Saliency Map)	E1		E2	E1	
M=3.31	E3		M=0.18	E3	
SD=1.13	E4		SD=0.39	E4	***
E3 (CycleGAN)	E1		E3	E1	
M=3.24	E2		M=0.16	E2	
SD=1.19	E4		SD=0.37	E4	***
E4 (Our counterfactual	E1	*	E4	E1	***
explanation) **M=3.72**	E2		**M=0.24**	E2	***
SD=0.97	E3		SD=0.42	E3	***

4.3.7 Bias detection

Our model can discover confounding bias in the data used for training the black-box classifier. Confounding bias provides an alternative explanation for an association between the data and the target label. For example, a classifier trained to predict the presence of a disease may make decisions based on hidden attributes like gender, race, or age. In a simulated experiment, we trained two classifiers to identify smiling vs not-smiling images in the CelebA dataset. The first classifier f_{Biased} is trained on a biased dataset, confounded with gender such that all smiling images are of male faces. We train a second classifier $f_{\text{No-biased}}$ on an unbiased dataset, with data uniformly distributed with respect to gender. Note that we evaluate both the classifiers on the same validation set. Additionally, we assume access to a proxy Oracle classifier f_{Gender} that perfectly classifies the confounding attribute *i.e.*, gender.

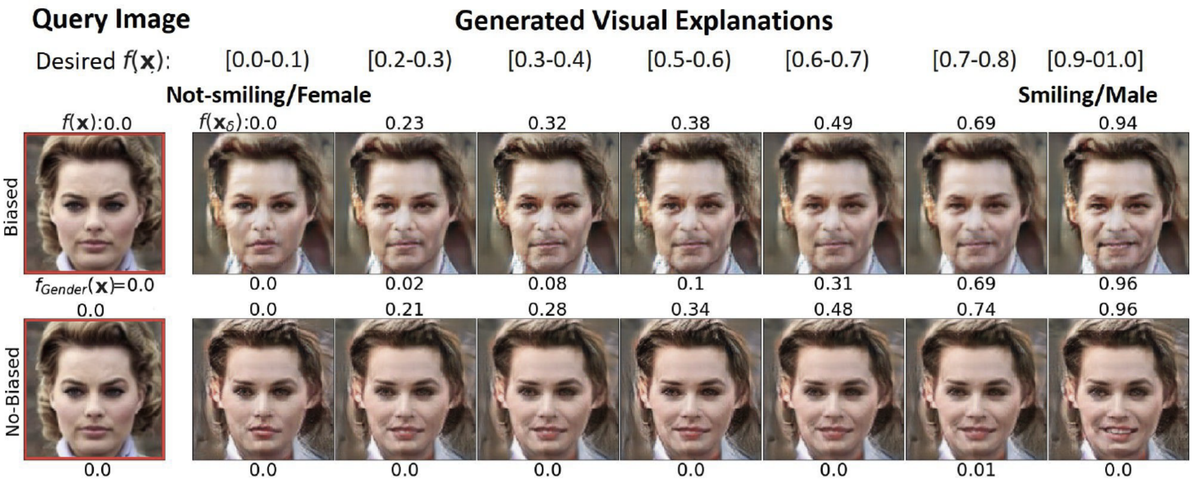


Figure 20: Explanations for two classifiers, both trained to classify "Smiling" attribute on CelebA dataset. For each example, the top row shows results from "Biased" classifier whose data distribution is confounded with "Gender". The bottom row shows explanations from "No-Biased" classifier with uniform data distribution w.r.t gender. The top label indicates output of the classifier and the bottom label is the output of an oracle classifier for the con-founding attribute gender. The visual explanations for the "Biased" classifier changes the gender as it adds smile on the face.

As shown in [36], if the training data for the GAN is biased, then the inference would reflect that bias. In Figure 20, we compare the explanations generated for the two classifiers. The visual explanations for the biased classifier change gender as it increases the amount of smile. We adapted the confounding metric proposed in [113] to summarize our results in Table 12. Given the data $\mathcal{D} = \{(\mathbf{x}_i, y_i, a_i), \mathbf{x}_i \in \mathcal{X}, y_i, a_i \in \mathcal{Y}\}$, we quantify that a classifier is confounded by an attribute a if the generated explanation $x_{\mathbf{c}}$ has a different attribute a, as compared to query image $\mathbf{x}$, when processed through the Oracle classifier f_{a}. The metric is formally defined as $\mathbb{E}_{\mathcal{D}}[1(f_{\mathrm{a}}(\mathbf{x_c}) \neq f_{\mathrm{a}}(\mathbf{x}))]/|\mathcal{D}|$. For a biased classifier, the Oracle function predicted the female class for the majority of the images, while the unbiased classifier is consistent with the true distribution of the validation set for gender. Thus, the fraction of generated explanations that changed the confounding attribute "gender" was found to be high for the biased classifier.

Table 12: Confounding metric for biased detection. For target label "Smiling" and "Not-Smiling", the explanations are generated using condition $\mathbf{c} > 0.9$ and $\mathbf{c} < 0.1$ respectively. The Male and Female values quantifies the fraction of the generated explanations classifier as male or female, respectively by oracle classifier f_{Gender}. The overall value quantifies the fraction of the generated explanations who have different gender as compared to the query image. A small overall value shows least bias.

	Target	**Label**
Black-box classifier	Smiling	Not-Smiling
f_{Biased}	Male: 0.52	Male: 0.18
	Female: 0.48	Female: 0.82
	Overall: **0.12**	Overall: **0.35**
$f_{\text{No-biased}}$	Male: 0.48	Male: 0.47
	Female: 0.52	Female: 0.53
	Overall: 0.07	Overall: 0.08

4.3.8 Evaluating class discrimination

In multi-label settings, multiple labels can be true for a given image. A multi-label setting is common in CXR diagnosis. For example, cardiomegaly and pleural effusion are associated with cardiogenic edema and frequently co-occur in a CXR. Please note that our classification model is also trained in a multi-label setting where the fourteen radiological findings may co-occur in a CXR. In this evaluation, we demonstrate the sensitivity of our generated explanations to the task being explained. Ideally, an explanation model trained to explain a given task should produce explanations consistent with the query image on all the other classes besides the given task. Specifically, if we are training a model to explain "cardiomegaly" then the counterfactual image should flip classification decision only for "cardiomegaly" class and not for any other class.

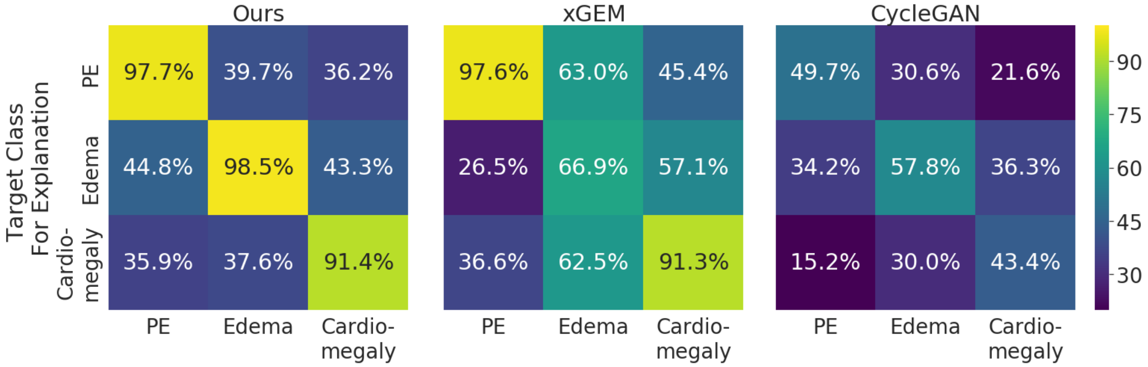


Figure 21: Evaluating class discrimination. Each cell is the fraction of the generated explanations, that have flipped in a class as compared to the query image. The x-axis shows the classes in a multi-label setting, and the y-axis shows the target class for which an explanation is generated. Note: This is not a confusion matrix.

We considered three diagnosis tasks, cardiomegaly, pleural effusion, and edema. For each task, we trained one explanation model. Figure. 21 plots the fraction of the generated explanations, that have flipped in other classes as compared to the query image. In Figure. 21, each column represents one task, and each row is one run of our method to explain a given task. The diagonal values also represent the counterfactual validity (CV) score reported in Table. 6.

4.4 Discussion and Conclusion

We provided a BlackBox a *Progressive Counterfactual Explainer* designed to explain image classification models for medical applications. Our framework explains the decision by gradually transforming the input image to its counterfactual, such that the classifier's prediction is flipped. We have formulated and evaluated our framework on three properties of a valid counterfactual transformation: data consistency, classifier consistency, and self-consistency. Our results showed that our framework adheres to all three properties.

Comparison with xGEM and cycleGAN: Our model satisfy all three essential properties of a valid counterfactual explanation. Our model creates natural-looking explanations that produce a desired outcome from the classification model while retaining maximum patient-specific information. In comparison, both xGEM and cycleGAN failed on at least one essential property. xGEM model fails to create realistic images with a high FID score (> 300). Furthermore, the cycleGAN model fails to flip the classifier's decision with a low CV score ($< 60\%$).

xGEM: The visual quality of images generated by xGEM, is limited by the expressiveness of its generator. xGEM adopted a variational autoencoder (VAE) as the generator. VAE uses a Gaussian likelihood (ℓ_2 reconstruction), an unrealistic assumption for image data, and is known to produce over-smoothed images [99]. In contrast, our model uses an implicit likelihood assumption of GAN [172], resulting in realistic explanation images.

CycleGAN: The cycleGAN model learns two generator networks to transform an input image into a positive or a negative sample for a given target class. However, during training, cycleGAN loss function does not incorporate the external black-box classifier. It primarily follows a data-driven approach to learn all the differences between positive and negative samples. Hence, the cycleGAN model learns to explain the data and not the classification model. As a result, the counterfactual explanations derived from cycleGAN model frequently fails to flip the classification decision, despite their high visual quality, resulting in inconclusive images that are not counterfactual.

Further, we present a thorough comparison between cycleGAN and our explanation in a human evaluation study. The clinical experts' expressed high agreement that explanation images from cycleGAN were of high quality and they resembles the query CXR. But at the same time, users found the explanation images to be too similar to query CXR, and the cycleGAN explanations failed to provide the counterfactual reasoning for the decision. In comparison, our explanation were most helpful in understanding the classification decision.

Though the users reported inconsistencies in the visual appearance, but the overall sentiment looks positive and they selected our method as their preferred explanation method for improved understandability.

Comparison with saliency maps: As compared to saliency maps, counterfactual explanations provide extra information to the end-user to understand the classification decision. Our quantitative experiments show that the region modified by the PCE to create counterfactual image, frequently matches the salient regions highlighted by the saliency-map based explanations models. Also, saliency-map-based explanations may highlight almost the same region for different tasks, resulting in misleading and inconclusive explanations (*see* Figure. 15). In contrast, our counterfactual explanations provide additional information to clarify *how* input features in the important regions could be modified to change the prediction decision. Our difference map localizes disease to specific regions in the chest, and these regions align with the clinical knowledge of the disease. In Figure. 15, our difference map focused on the heart region for cardiomegaly and the CP recess region for PE.

Clinical relevance of the explanations: From a clinical perspective, we demonstrated that the counterfactual changes associated with normal (negative) or abnormal (positive) classification decisions are also associated with corresponding changes in disease-specific metrics such as CTR and SCP. For example, changes associated with an increased posterior probability for cardiomegaly also resulted in an increased CTR. Similarly, for PE, a healthy CP recess with a high SCP score transformed into an abnormal CP recess with blunt CPA, as the posterior probability for PE increases (*see* Figure. 12 and Figure. 17).

To the best of our knowledge, ours is the first attempt to quantify model explanations using clinical metrics. At the same time, our evaluation has certain limitations. Our automatic pipeline to compute CTR and SCP suffers from inaccuracies. This contributed to the large variance in difference plots in Figure. 17. These inaccuracies are due to the sub-optimal performance of the segmentation and object detector networks. In the absence of ground truth annotations for lung and heart segmentation and limited annotations for the CP recess region, these networks have sub-optimal performance. Nevertheless, our goal is not to compute these metrics correctly for each image but to perform a population-level analysis. In our experiments, CTR and SCP successfully captured the difference between normal and abnormal CXR for cardiomegaly and PE, respectively. One may argue using CTR and SCP to perform disease classification. However, models based on these features will also suffer from similar inaccuracies, resulting in poor performance and generalization compared to the DL methods.

Defining clinical metrics for different diseases is a challenging task. For example, consider edema. It may appear as different radiographic concepts (*e.g.*, cephalization, peribronchial cuffing, perihilar batwing appearance, and opacities *etc.*) in different patients [167]. Transforming a healthy CXR to a counterfactual image for edema introduce changes in multiple such concepts. Future research should determine appropriate metrics to quantify and understand these concepts. Manual annotation is one solution for obtaining ground truth to train models that can identify concepts. Efforts should be made to reduce the dependency on manual labelling as it is expensive and not scalable.

Usability of explanations: Counterfactual explanations can help in model auditing and recovering hidden bias in the classifier's training. In our experiments, we visualize counterfactual explanations from a biased classifier and contrast it with a classifier without any data bias. Further, using human evaluation we demonstrate the prospective use-case for counterfactual explanations.

We acknowledge that our GAN-generated counterfactual explanations may have missing details such as small wires. In our extended experiments, we found that the foreign objects such as pacemaker have minimal importance in the classification decision (*see* appendix A.13). We attempted to improve the preservation of such information through our revised context-aware reconstruction loss (CARL). However, even with CARL, the FO preservation score is not perfect. A possible reason for this gap is the limited capacity of the object detector used to calculate the FOP score. Training a highly accurate FO detector is outside the scope of this study.

Further, a resolution of 256×256 for counterfactually generated images is smaller than a standard CXR. Small resolution limits the evaluation for fine details by both the algorithm and the interpreter. Our formulation of cGAN uses conditional-batch normalization (cBN) to encapsulate condition information while generating images. For efficient cBN, the mini-batches should be class-balanced. To accommodate high-resolution images with smaller batch sizes, we must decrease the number of conditions to ensure class-balanced batches. Fewer conditions resulted in a coarse transformation with abrupt changes across explanation images. In our experiments, we selected the smallest bin width, which created a class-balanced batch that fits in GPU memory and resulted in stable cGAN training. However, with the advent of larger-memory GPUs, we intend to apply our methods to higher resolution images in future work; and assess how that impacts interpretation by clinicians.

To conclude, this study uses counterfactual explanations as a way to audit a given black-box classifier and evaluate whether the radio-graphic features used by that classifier have any clinical relevance. In particular, the proposed model did well in explaining the decision for cardiomegaly and pleural effusions and was corroborated by an experienced radiology resident physician. By providing visual explanations for deep learning decisions, radiologists better understand the causes of its decision-making. This is essential to lessen physicians' concerns regarding the "BlackBox" nature by an algorithm and build needed trust for incorporation into everyday clinical workflow. As an increasing amount of artificial intelligence algorithms offer the promise of everyday utility, counterfactually generated images are a promising conduit to building trust among diagnostic radiologists.

By providing counterfactual explanations, our work opens up many ideas for future work. Our framework showed that valid counterfactual can be learned using an adversarial generative process, that is regularized by the classification model. However, counterfactual reasoning is incomplete without a causal structure and explicitly modeling of the interventions. An interesting next step should explore incorporating or discovering plausible causal structures and creating explanations that are grounded with them.

5.0 Concept-based Counterfactual Explanation

5.1 Introduction

Machine Learning, specifically, Deep Learning (DL) methods are increasingly adopted in healthcare applications. Model explainability is essential to build trust in the AI system [73] and to receive clinicians' feedback. Standard explanation methods for image classification delineates regions in the input image that significantly contribute to the model's outcome [231, 151, 212]. However, it is challenging to explain *how* and *why* variations in identified regions are relevant to the model's decision. Ideally, an explanation should resemble the decision-making process of a domain expert. This paper aims to map a DL model's neuron activation patterns to the radiographic features and constructs a simple rule-based model that partially explains the Black-box.

Methods based on feature attribution have been commonly used for explaining DL models for medical imaging [11]. However, an alignment between feature attribution and radiology concepts is difficult to achieve, especially when a single region may correspond to several radiographic concepts. Recently, researchers have focused on providing explanations in the form of human-defined concepts [122, 12, 304]. In medical imaging, such methods have been adopted to derive an explanation for breast mammograms [291], breast histopathology [81] and cardiac MRIs [34]. A major drawback of the current approach is their dependence on explicit concept-annotations, either in the form of a representative set of images [122] or semantic segmentation [12], to learn explanations. Such annotations are expensive to acquire, especially in the medical domain. We use weak annotations from radiology reports to derive concept annotations. Furthermore, these methods measure correlations between concept perturbations and classification predictions to quantify the concept's relevance. However, the neural network may not use the discovered concepts to arrive at its decision. We borrow tools from causal analysis literature to address that drawback [273].

In this work, we used radiographic features mentioned in radiology reports to define concepts. Using a National Language Processing (NLP) pipeline, we extract weak annotations from text and classify them based on their positive or negative mention [107]. Next, we use sparse logistic regression to identify sets of hidden-units correlated with the presence of a concept. To quantify the causal influence of the discovered concept-units on the model's outcome, we view concept-units as a *mediator* in the treatment-mediator-outcome framework [106]. Using measures from mediation analysis, we provide an effective ranking of the concepts based on their causal relevance to the model's outcome. Finally, we construct a low-depth decision tree to express discovered concepts in simple decision rules, providing the global explanation for the model. The rule-based nature of the decision tree resembles many decision-making procedures by clinicians.

5.2 Method

We consider a pre-trained *black-box* classifier $f : \mathbf{x} \rightarrow \mathbf{y}$ that takes an image $\mathbf{x}$ as input and process it using a sequence of hidden layers to produce a final output $\mathbf{y} \in \mathbb{R}^D$. Without loss of generality, we decompose function f as $\Phi_2 \circ \Phi_1(\mathbf{x})$, where $\Phi_1(\mathbf{x}) \in \mathbb{R}^L$ is the output of the initial few layers of the network and Φ_2 denotes the rest of the network. We assume access to a dataset $\mathcal{X} = \{(\mathbf{x}_n, \mathbf{y}_n, \mathbf{c}_n)\}^N$, where $\mathbf{x}_n$ is input image, $\mathbf{y}_n$ is a d-dimensional one-hot encoding of the class labels and $\mathbf{c}_n \in \mathbb{R}^K$ is a k-dimensional concept-label vector. We define concepts as the radiographic observations mentioned in radiology reports to describe and provide reasoning for a diagnosis. We used a NLP pipeline [107] to extract concept annotations. The NLP pipeline follows a rule-based approach to extract and classify observations from the free-text radiology report. The extracted k^{th} concept-label $\mathbf{c}_n[k]$ is either 0 (negative-mention), 1(positive-mention) or -1 (uncertain or missing-mention).

An overview of our method is shown in Fig. 22. Our method consists of three sequential steps:

(1) *Concept associations*: We seek to discover sparse associations between concepts and the hidden-units of $f(\cdot)$. We express k^{th} concept as a sparse vector $\mathbf{v}_k \in \mathbb{R}^L$ that represents a linear direction in the intermediate space $\Phi_1(\cdot)$.

(2) *Causal concept ranking*: Using tools from causal inference, we find an effective

ranking of the concepts based on their relevance to the classification decision. Specifically, we consider each concept as a mediator in the causal path between the input and the outcome. We measure concept relevance as the effect of a counterfactual intervention on the outcome that passes indirectly through the concept-mediator.

(3) *Surrogate explanation function*: We learn an easy-to-interpret function $g(\cdot)$ that mimics function $f(\cdot)$ in its decision. Using $g(\cdot)$, we seek to learn a global explanation for $f(\cdot)$ in terms of the concepts.

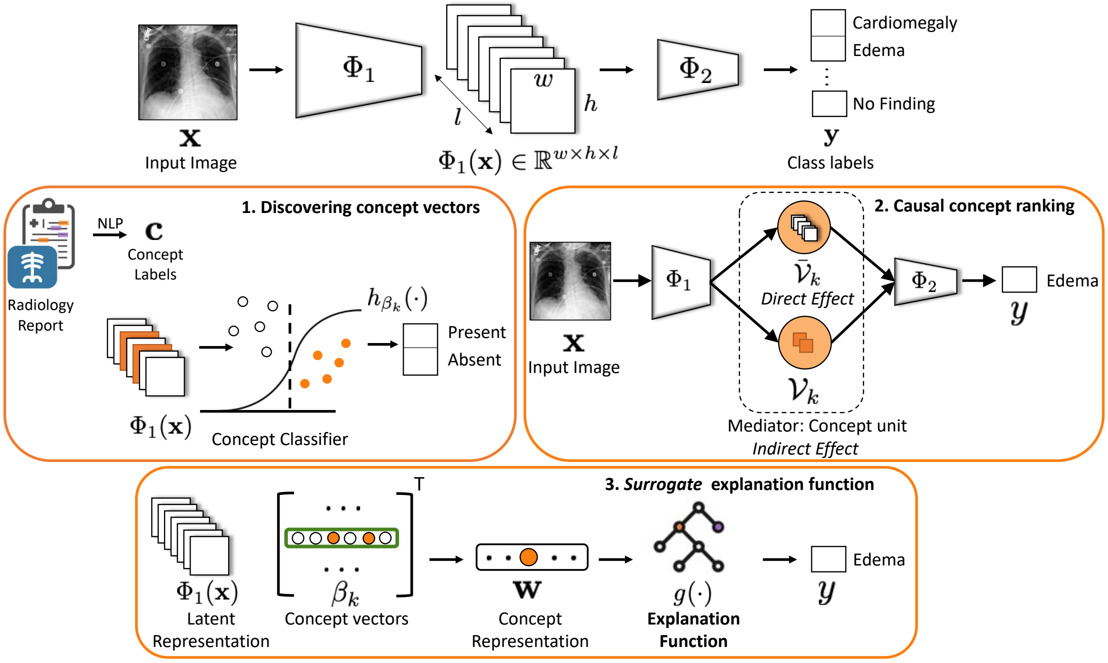


Figure 22: Method overview for concept-based counterfactual explanations. We provide explanation for the black-box function $f(\mathbf{x})$ in-terms of concepts, that are radiographic observations mentioned in radiology reports. 1) The intermediate representation $\Phi_1(\mathbf{x})$ is used to learn a sparse logistic regression $h_{\mathbf{v}_k}(\cdot)$ to classify k^{th} concept. 2) The non-zero coefficients of $\mathbf{v}_k$ represents a set of concept units $\mathcal{V}_k$ that serves as a mediator in the causal path connecting input $\mathbf{x}$ and outcome y. 3) A decision tree function is learned to map concepts to class labels.

5.2.1 Concept associations

We discover concept associations with intermediate representation $\Phi_1(\cdot)$ by learning a binary classifier that maps $\Phi_1(\mathbf{x})$ to the concept-labels [122]. We treat each concept as a separate binary classification problem and extract a representative set of images $\mathcal{X}^k$, in which concept $c_n[k]$ is present and a random negative set. We define concept vector ($\mathbf{v}_k$) as the solution to the logistic regression model $c_n[k] = \sigma(\mathbf{v}_k^T \operatorname{vec}(\Phi_1(\mathbf{x}_n))) + \epsilon$, where $\sigma(\cdot)$ is the sigmoid function. For a convolutional neural network, $\Phi_1(\mathbf{x}) \in \mathbb{R}^{w \times h \times l}$ is the output activation of a convolutional layer with width w, height h and number of channels l. We experimented with two vectorization for Φ_1. In first, we flatten $\Phi_1(\mathbf{x})$ to be a whl-dimensional vector. In second, we applied a spatial aggregation by max-pooling along the width and height to obtain l-dimensional vector. Unlike TCAV [122] that uses linear regression, we used lasso regression to enable sparse feature selection and minimize the following loss function,

$$\min_{\mathbf{v}_k} \sum_{\mathbf{x}_n \in \mathcal{X}_k} \ell(h_{\mathbf{v}_k}(\mathbf{x}), c_n[k]) + \lambda ||\mathbf{v}_k||_1 \tag{15}$$

where $\ell(\cdot,\cdot)$ is the cross entropy loss, $h_{\mathbf{v}_k}(\mathbf{x}) = \sigma(\mathbf{v}_k^T \operatorname{vec}(\Phi_1(\mathbf{x}_n)))$ and λ is the regularization parameter. We performed 10-fold nested-cross validation to find λ with least error. The non-zero elements in the concept vector $\mathbf{v}_k$ forms the set of hidden units ($\mathcal{V}_k$) that are most relevant to the k^{th} concept.

5.2.2 Causal concept ranking

Concept associations identified hidden units that are strongly correlated with a concept. However, the neural network may or may not use the discovered concepts to arrive at its decision. We use tools from causal inference, to quantify what fraction of the outcome is mediated through the discovered concepts.

To enable causal inference, we first define *counterfactual* $\mathbf{x}'$ as a perturbation of the input image $\mathbf{x}$ such that the decision of the classifier is flipped. Following the approach proposed in Chapter 3, we used the Progressive Counterfactual Explainer (PCE), a conditional generative adversarial network (cGAN) to learn the counterfactual perturbation. We conditioned on the output of the classifier, to ensure that cGAN learns a classifier-specific perturbation for the given image $\mathbf{x}$. Next, we used theory from causal mediation analysis to causally relate a concept with the classification outcome. Specifically, we consider concept as a mediator in the causal pathway from the input $\mathbf{x}$ to the outcome $\mathbf{y}$. We specify following effects to quantify the causal effect of the counterfactual perturbation and the role of a mediator in transferring such effect,

1. Average treatment effect (ATE): ATE is the total change in the classification outcome $\mathbf{y}$ as a result of the counterfactual perturbation.
2. Direct effect (DE): DE is the effect of the counterfactual perturbation that comprises of any causal mechanism that *do not* pass through a given mediator. It captures how the perturbation of input image changes classification decision directly, without considering a given concept.
3. Indirect effect (IE): IE is the effect of the counterfactual perturbation which is mediated by a set of mediators. It captures how the perturbation of input image changes classification decision indirectly through a given concept.

Following the potential outcome framework from [217, 273], we define the ATE as the proportional difference between the factual and the counterfactual classification outcome,

$$\mathbf{ATE} = \mathbb{E}\big[\frac{f(\mathbf{x}')}{f(\mathbf{x})} - 1\big]. \tag{16}$$

To enable causal inference through a mediator, we borrow Pearl's definitions of natural direct and indirect effects [201] (*ref* Fig. 23). We consider set of concept-units $\mathcal{V}_k$ as a mediator, representing the k^{th} concept. We decompose the latent representation $\Phi_1(\mathbf{x})$ as concatenation of response of concept-units $\mathcal{V}_k(\mathbf{x})$ and rest of the hidden units $\bar{\mathcal{V}_k}(\mathbf{x})$ *i.e.*, $\Phi_1(\mathbf{x}) = [\mathcal{V}_k(\mathbf{x}), \bar{\mathcal{V}_k}(\mathbf{x})]$. We can re-write classification outcome as $f(\mathbf{x}) = \Phi_2(\Phi_1(\mathbf{x})) = \Phi_2([\mathcal{V}_k(\mathbf{x}), \bar{\mathcal{V}_k}(\mathbf{x})])$. To disentangle the direct effect from the indirect effect, we use the concept of *do*-operation on the unit level of the learnt network. Specifically, we use $do(\mathcal{V}_k(\mathbf{x}))$ to denote that we set the value of the concept-units to the value obtained by using the original image as input. By intervening on the network and setting the value of the concept units, we can compute the direct effect as the proportional difference between the factual and the counterfactual classification outcome, while holding mediator *i.e.*, $\mathcal{V}_k$ fixed to its value before the perturbation,

$$\mathbf{DE} = \mathbb{E}\big[\frac{\Phi_2([do(\mathcal{V}_k(\mathbf{x})), \bar{\mathcal{V}}_k(\mathbf{x}')])}{\Phi_2([\mathcal{V}_k(\mathbf{x}), \bar{\mathcal{V}}_k(\mathbf{x})])} - 1\big]. \tag{17}$$

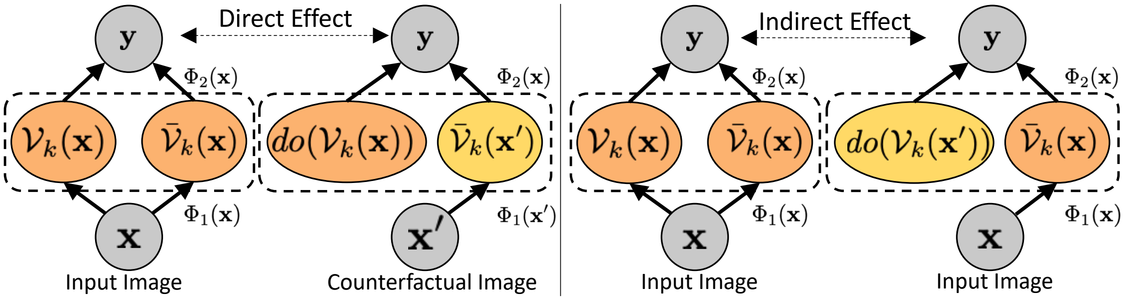

Direct Effect
y
$\Phi_2(\mathbf{x})$
$\mathcal{V}_k(\mathbf{x})$
$\bar{\mathcal{V}}_k(\mathbf{x})$
$do(\mathcal{V}_k(\mathbf{x}))$
$\bar{\mathcal{V}}_k(\mathbf{x}')$
$\Phi_1(\mathbf{x})$
$\Phi_1(\mathbf{x}')$
x
x'
Input Image
Counterfactual Image
Indirect Effect
$do(\mathcal{V}_k(\mathbf{x}'))$

Figure 23: Illustration of direct and indirect effects in causal mediation analysis.

We compute indirect effect as the expected change in the outcome, if we change the mediator from its original value to its value using counterfactual, while holding everything else fixed to its original value,

$$\mathbf{IE} = \mathbb{E}\big[\frac{\Phi_2([do(\mathcal{V}_k(\mathbf{x}')), \bar{\mathcal{V}}_k(\mathbf{x})])}{\Phi_2([\mathcal{V}_k(\mathbf{x}), \bar{\mathcal{V}}_k(\mathbf{x})])} - 1\big]. \tag{18}$$

If the perturbation has no effect on the mediator, then the causal indirect effect will be zero. Finally, we use the indirect effect associated with a concept, as a measure of its relevance to the classification decision.

5.2.3 Surrogate explanation function

We aim to learn a surrogate function $g(\cdot)$, such that it reproduces the outcome of the function $f(\cdot)$ using an interpretable and straightforward function. We formulated $g(\cdot)$ as a decision tree as many clinical decision-making procedures follow a rule-based pattern. We summarize the internal state of the function $f(\cdot)$ using output of k concept regression functions $h_{\mathbf{v}_k}(\cdot)$ as follows,

$$\mathbf{w}_n = [\mathrm{logit}(h_{\mathbf{v}_1}(\mathbf{x}_n)), \mathrm{logit}(h_{\mathbf{v}_2}(\mathbf{x}_n)), \cdots]. \tag{19}$$

Next, we fit a decision tree function, $g(\cdot)$, to mimic the outcome of the function $f(\cdot)$ as,

$$g^* = \arg\min_g \sum_n \mathcal{L}(g(\mathbf{w}_n), f(\mathbf{x}_n)), \tag{20}$$

where $\mathcal{L}$ is the splitting criterion based on minimizing entropy for highest information gain from every split.

5.3 Experiments and Results

5.3.1 Study cohort and imaging dataset

We perform experiments on the MIMIC-CXR [112] dataset, which is a multi-modal dataset consisting of 473K chest X-ray images and 206K reports. The dataset is labeled for 14 radiographic observations, including 12 pathologies. We used state-of-the-art DenseNet-121 [98] architecture for our classification function [107]. DenseNet-121 architecture is composed of four dense blocks. We experimented with three versions of $\Phi_1(\cdot)$ to represent the network until the second, third, and fourth dense block. For concept annotations, we considered radiographic features that are frequently mentioned in radiology reports in the context of labeled pathologies. Next, we used Stanford CheXpert [107] to extract and classify these observations from free-text radiology reports.

5.3.2 Experimental setup

We first evaluated the concept classification performance and visualized concept-units to demonstrate their effectiveness in localizing a concept. Next, we summarized the indirect effects associated with different concepts across different layers of the classifier. We evaluated a proposing ranking of the concepts based on their causal contribution to the classification decision. Finally, we used the top-ranked concepts to learn a surrogate explanation function in the form of a decision tree.

5.3.3 Evaluation of concept classifiers

The intermediate representations from third dense-block consistently outperformed other layers in concept classification. In Fig. 24, we show the testing-ROC-AUC and recall metric for different concept classifiers. All the concept classifiers achieved high recall, demonstrating a low false-negative (type-2) error.

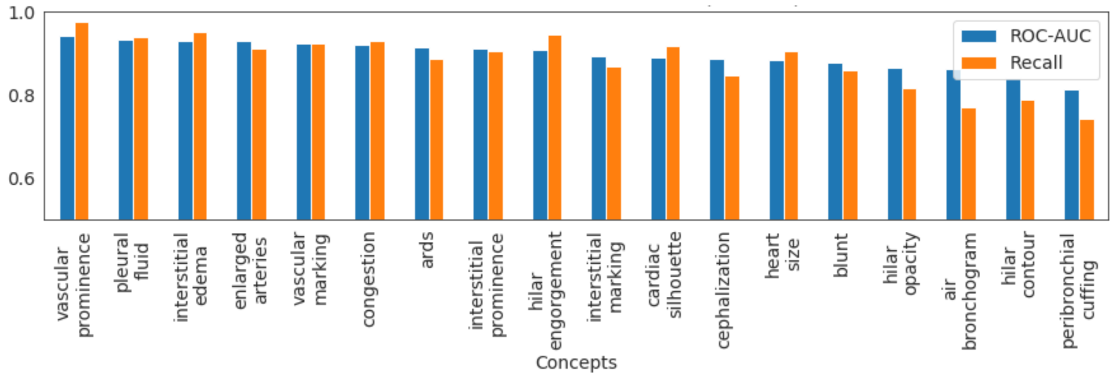


Figure 24: AUC-ROC and recall metric for different concept classifiers.

In Fig. 25, we visualize the activation map of hidden units associated with the concept vector $\mathcal{V}_k$. For each concept, we visualize hidden units that have large logistic regression-coefficient (β_k). To highlight the most activated region for a unit, we threshold activation map by the top 1% quantile of the distribution of the selected units' activations [12]. Consistent with prior work [13], we observed that several hidden units have emerged as concept detectors, even though concept labels were not used while training f. For *cardiac-silhouette*, different hidden units highlight different regions of the heart and its boundary with the lung. For localized concept such as *blunt costophrenic angle*, multiple relevant units were identified that all focused on the lower-lobe regions. Same hidden unit can be relevant for multiple concepts. The top label in Fig. 25. shows the top two important concepts for each hidden unit.

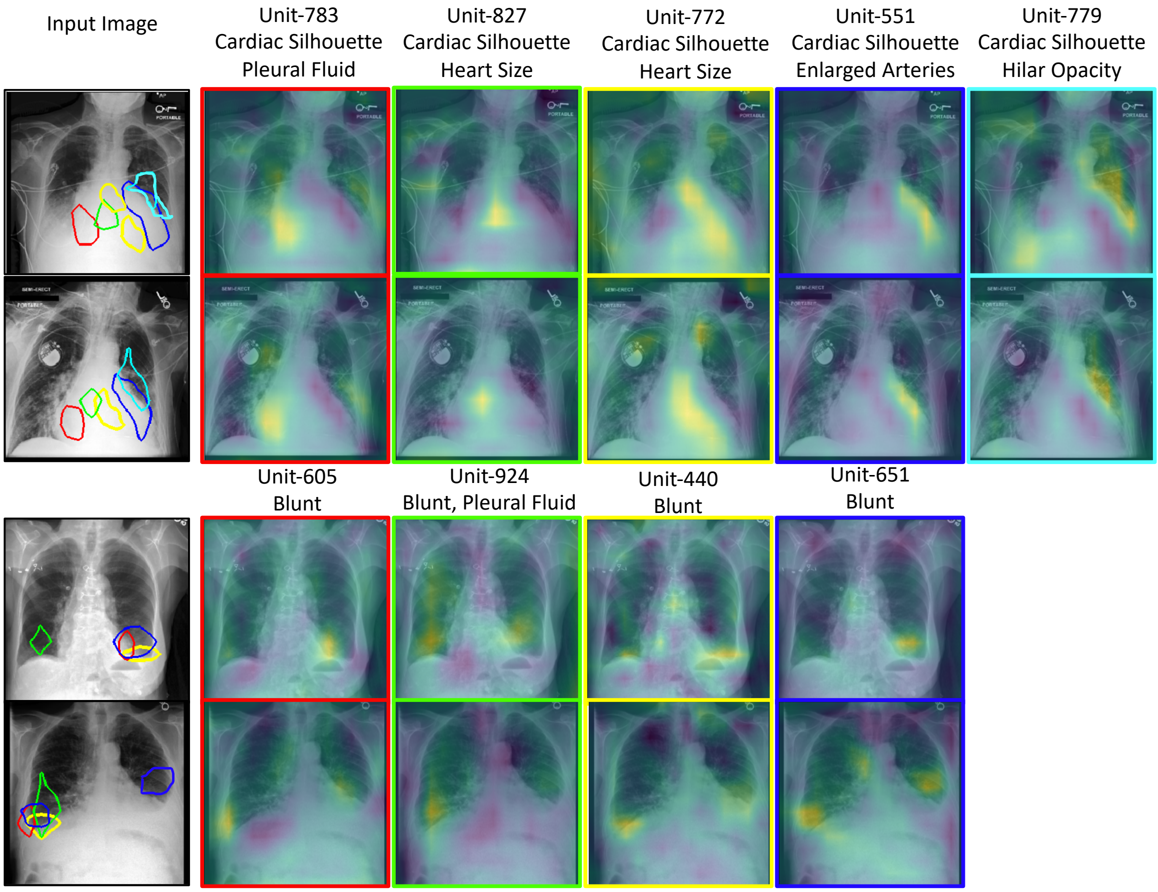


Figure 25: A qualitative demonstration of the activation maps of the hidden units that act as visual concept detectors. Each column represents one hidden unit identified as part of concept vector $\mathcal{V}_k$. Top two rows show k = *cardiac-silhouette* and bottom rows have k =*blunt costophrenic angle.*

5.3.4 Evaluating causal concepts using decision tree as surrogate function

We evaluate the success of the counterfactual intervention by measuring average total effect (ATE). High values for ATE confirms that counterfactual image generated by [242] successfully flips the classification decision. We achieved an ATE of 0.97 for cardiomegaly, 0.89 for pleural effusion and 0.96 for edema. In Fig. 26 (heat-map), we show the distribution of the indirect effect associated with concepts, across different layers. The middle layer demonstrates a large indirect effect across all concepts. This shows that the hidden units in dense-block 3 played a significant role in mediating the effect of counterfactual intervention.

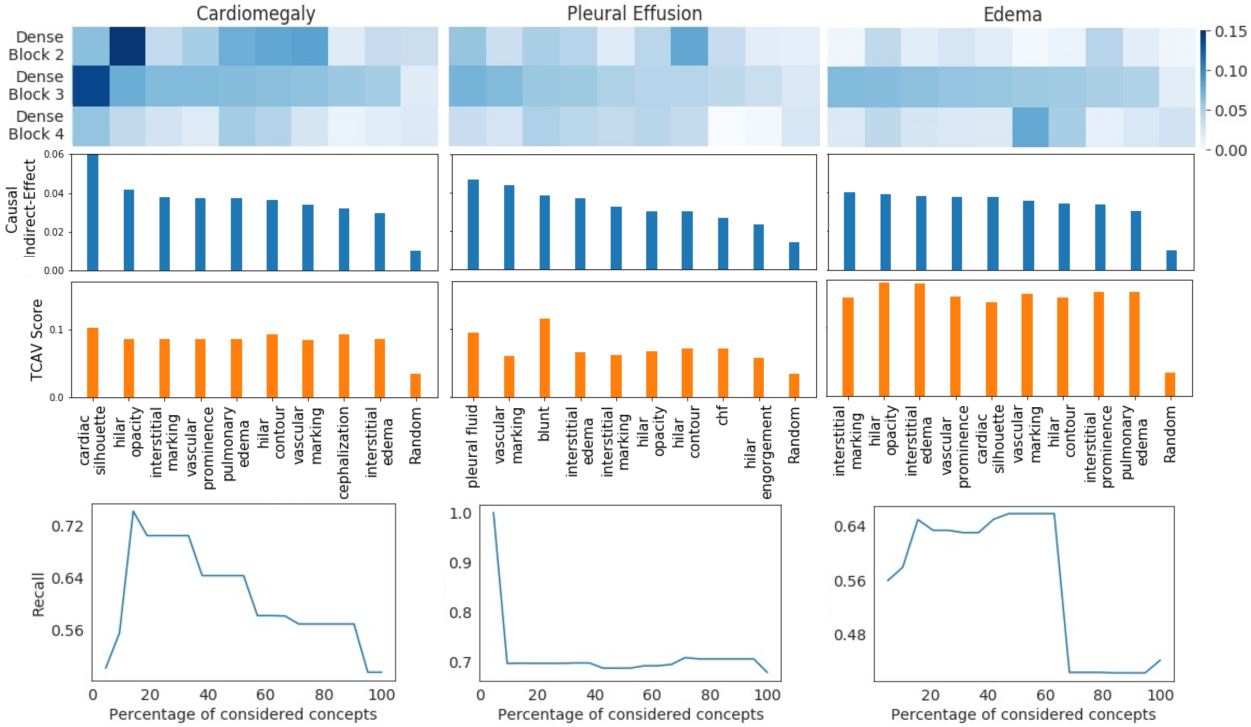


Figure 26: Evaluating concept vectors and their causal effect. Indirect effects of the concepts, calculated over different layers of the DenseNet-121 architecture (heat-map). The derived ranking of the concepts based on their causal relevance to the diagnosis (bar-graph). A comparative ranking based on concept sensitivity score from TCAV [122]. The trend of recall metric for the decision tree function $g(\cdot)$, while training using top x% of top-ranked concepts (trend-plot).

In Fig. 26 (bar-graph), we rank the concepts based on their indirect effect. The top-ranked concepts recovered by our ranking are consistent with the radiographic features that clinicians associates with the examined three diagnoses [115, 167, 184]. Further, we used the concept sensitivity score from TCAV [122] to rank concepts for each diagnosis. The top-10 concepts identified by our in-direct effect and TCAV are the same, while their order is different. The top-3 concepts are also the same, with minor differences in ranking. Both the methods have low importance score for random concept. This confirms that the trend in importance score is unlikely to be caused by chance. For our approach, random concept represents an ablation of the concept-association step. Here, rather than performing lasso regression to identify relevant units, we randomly select units.

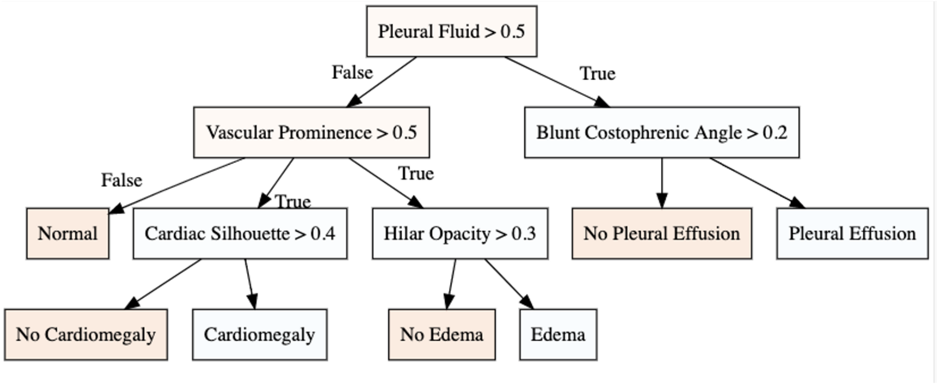


Figure 27: The decision tree for the three diagnosis with best performance on recall metric.

To quantitatively demonstrate the effectiveness of our ranking, we iteratively consider $x\%$ of top-ranked concepts and retrain the explanation function $g(\mathbf{w})$. In Fig. 26 (bottom-plot), we observe the change in recall metric for the classifier $g(\cdot)$ as we consider more concepts. In the beginning, as we add relevant concepts, the true positive rate increases resulting in a high recall. However, as less relevant concepts are considered, the noise in input features increased, resulting in a lower recall. Fig. 27 visualize the decision tree learned for the best performing model.

5.4 Discussion and Conclusion

Model explainability is essential for the creation of trustworthy Machine Learning models in healthcare. An ideal explanation resembles the decision-making process of a domain expert and is expressed using concepts or terminology that is meaningful to the clinicians. To provide such explanation, we grounded our explanation in terms of clinically relevant concepts that are causally influencing the model's decision. We first associate the hidden units of the classifier to clinically relevant concepts. We take advantage of radiology reports accompanying the chest X-ray images to define concepts. We discover sparse associations between concepts and hidden units using a linear sparse logistic regression. To ensure that the identified units truly influence the classifier's outcome, we adopt tools from Causal Inference literature and, more specifically, mediation analysis through counterfactual interventions. Finally, we construct a low-depth decision tree to translate all the discovered concepts into a straightforward decision rule, expressed to the radiologist. We evaluated our approach on a large chest x-ray dataset, where our model produces a global explanation consistent with clinical knowledge. We successfully discovered highly discriminative neurons associated with fine-grains concepts that clinicians uses to explain their decision.

Acknowledgement This work was partially supported by NIH Award Number 1R01HL141813-01, NSF 1839332 Tripod+X, SAP SE, and Pennsylvania's Department of Health. We are grateful for the computational resources provided by Pittsburgh SuperComputing grant number TG-ASC170024.

6.0 Augmentation by Counterfactual Explanation - Fixing an Overconfident Classifier

6.1 Introduction

A highly accurate but overconfident model is ill-suited for decision-making pipelines, especially in critical applications such as healthcare and autonomous driving. The classification outcome should reflect a high uncertainty on ambiguous in-distribution samples that lie close to the decision boundary. The model should also refrain from making overconfident decisions on samples that lie far outside its training distribution, far-out-of-distribution (far-OOD), or on unseen samples from novel classes that lie near its training distribution (near-OOD). This paper proposes a method to fine-tune a given pre-trained classifier to fix its uncertainty characteristics while retaining its predictive performance.

We propose using a Progressive Counterfactual Explainer (PCE) to generate counterfactually augmented data (CAD) for fine-tuning the classifier. The PCE is a form of conditional Generative Adversarial Networks (cGANs) trained to generate samples that visually traverse the separating boundary of the classifier. The discriminator of the PCE serves as a density estimator to identify and reject OOD samples. We perform extensive experiments with detecting far-OOD, near-OOD, and ambiguous samples. Our empirical results show that our model improves the uncertainty of the baseline, and its performance is competitive to other methods that require a significant change or a complete re-training of the baseline model.

Deep neural networks (DNN) are increasingly being used in *decision-making* pipelines for real-world high-stake applications such as medical diagnostics [57] and autonomous driving [60]. For optimal decision making, the DNN should produce accurate predictions as well as quantify uncertainty over its predictions [64, 140]. While substantial efforts are made to engineer highly accurate architectures [98], many existing state-of-the-art DNNs do not capture the uncertainty correctly [65]. This hinders the re-use of openly available pre-trained models for real-world applications. We proposed to fine-tune the given *pre-trained* DNN on counterfactually augmented data, to improve its uncertainty quantification while retaining its original predictive accuracy.

Any classification model is essentially learning a hyperplane to separate samples from different classes. Accuracy only captures the proportion of samples that are on the correct side of the decision boundary. However, it ignores the relative distance of the sample from the decision boundary [131]. Ideally, samples closer to the boundary should have high uncertainty. The actual predicted value from the classifier should reflect this uncertainty via a low confidence score [102]. Conventionally, DNNs are trained on hard-label datasets to minimize a negative log-likelihood (NLL) loss. Such models tend to over-saturate on NLL and end-up learning very sharp decision boundaries [85, 180]. The resulting classifiers extrapolate over-confidently on ambiguous, near boundary samples, and the problem amplifies as we move to OOD regions [64].

We consider two types of uncertainty: *epistemic uncertainty*, caused due to limited data and knowledge of the model, and *aleatoric uncertainty*, caused by inherent noise or ambiguity in the data [128]. We evaluate these uncertainties with respect to three test distributions (*see* Fig 28):

- **Ambiguous in-Distribution (AiD)**: These are the samples within the training distribution that have an inherent ambiguity in their class labels. Such ambiguity represents high aleatoric uncertainty arising from class overlap or noise [245], *e.g.*, an image of a '5' that is similar to a '6'.
- **Near-OOD**: Near-OOD represents a label shift where label space is different between ID and OOD data. It has high epistemic uncertainty arising from the classifier's limited information on unseen data. We use samples from unseen classes of the training distribution as near-OOD.
- **Far-OOD**: Far-OOD represents data distribution that is significantly different from the training distribution. It has high epistemic uncertainty arising from mismatch between different data distributions.

Earlier work focuses on threshold-based detectors that use information from a pre-trained DNN to identify OOD samples [85, 92, 100, 277, 96]. Such methods predominantly focus on far-OOD detection and often do not address the over-confidence problem in DNN.

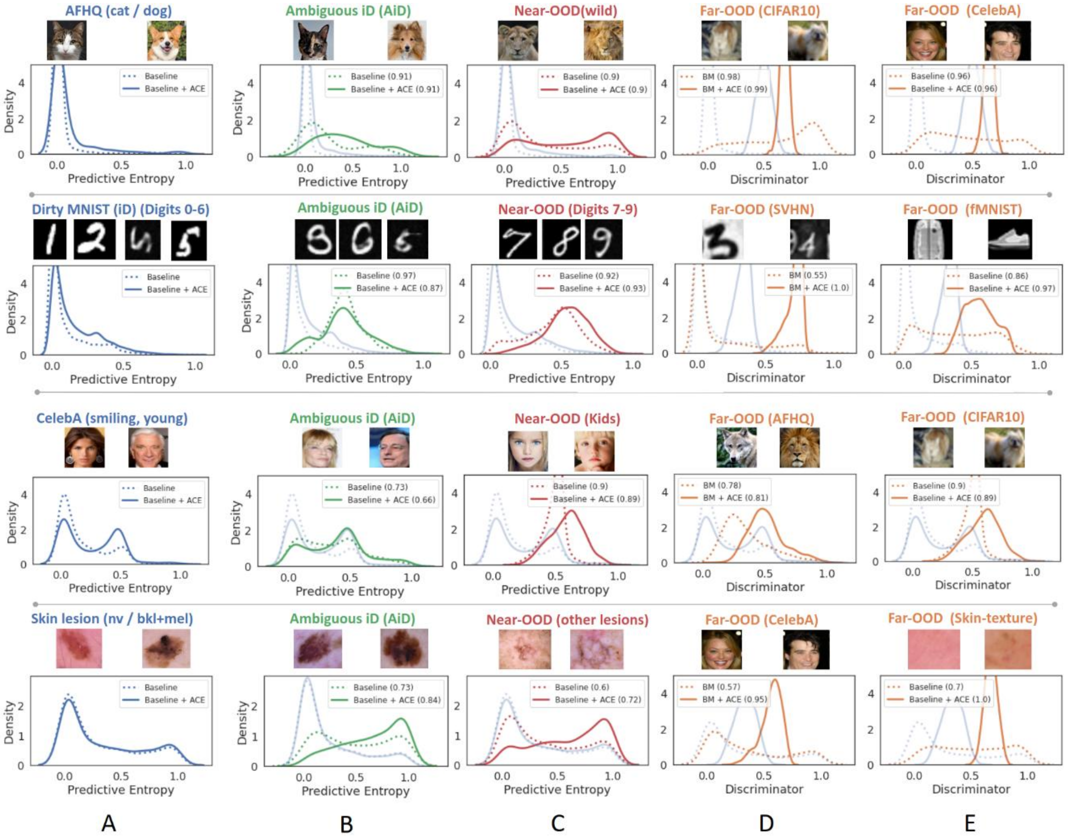


Figure 28: Comparison of the uncertainty estimates from the baseline, before and after fine-tuning with ACE. Each row represents a different dataset. A) Fine-tuning has little effect on the predicted entropy (PE) of **in-distribution (iD)** samples. We use PE to identify **ambiguous iD (AiD)** samples (B) and **near-OOD** samples (C). D-E) We use the discriminator of the PCE to identify **far-OOD** samples (last two columns). The legend shows the AUC-ROC for binary classification over uncertain samples and iD samples. Our method improved the uncertainty estimates across the spectrum.

In another line of research, variants of Bayesian models [186, 65] and ensemble learning [97, 133] were explored to provide reliable uncertainty estimates. Recently, there is a shift towards designing generalizable DNN that provide robust uncertainty estimates in a single forward pass [266, 27, 179]. Such methods usually propose changes to the DNN architecture [253], training procedure [299] or loss functions [181] to encourage separation between ID and OOD data. Popular methods include, training deterministic DNN with a distance-aware feature space [267, 146] and regularizing DNN training with a generative model that simulates OOD data [138]. However, these methods require a DNN model to be trained from scratch and are not compatible with an existing pre-trained DNN. Also, they may use auxiliary data to learn to distinguish OOD inputs [148].

In this work, we introduce an augmentation by counterfactual explanation (ACE) strategy to fine-tune an existing *pre-trained* DNN. Fine-tuning improves the uncertainty estimates without changing the network's architecture or compromising on its predictive performance. ACE uses a progressive counterfactual explainer (PCE) similar to Lang *et al.* [134] and Singla *et al.* [243] to generate counterfactually augmented data (CAD). The discriminator of the PCE is a density estimator and is used in a threshold-based selection function to identify and reject far-OOD samples.

The PCE is a conditional-Generative Adversarial Network (cGAN)-based explanation function that explains the decision of a DNN by gradually perturbing a query image to flip its classification decision. We used PCE to generate augmented samples closer to the decision boundary. We assign soft labels to these generated samples that mimic their distance from the boundary. Fine-tuning on such augmented data helps the DNN to recover from the over-saturation on NLL loss, thus making the decision boundary smoother. Smooth decision boundary facilitates improved uncertainty estimates for AiD and near-OOD samples and also makes the classifier robust to adversarial attacks.

Our contributions are as follows: (1) We propose a novel strategy to fine-tune an existing *pre-trained* DNN to improve its uncertainty estimates and facilitate its deployment in real-world applications. (2) Our approach generates counterfactual augmentations near the decision boundary, allowing the classifier to widen its boundary, to successfully capture the uncertainty over ambiguous-iD and near-OOD samples. (3) Our GAN-based augmenter provides a density estimator (the discriminator) to detect far-OOD samples.

We provide a comprehensive evaluation of our method on specifically defined test datasets to capture different uncertainties. Furthermore, our fine-tuned classifier exhibits better robustness to popular adversarial attacks.

6.2 Method

In this paper, we consider a pre-trained DNN classifier, f_θ, with good prediction accuracy but sub-optimal uncertainty estimates. We assume f_θ is a differentiable function and we have access to its gradient with respect to the input, $\nabla_{\mathbf{x}} f_\theta(\mathbf{x})$, and to its final prediction outcome $f_\theta(\mathbf{x})$. We also assume access to either the training data for f_θ, or an equivalent dataset with competitive prediction accuracy. We further assume that the training dataset for f_θ has hard labels $\{0, 1\}$ for all the classes.

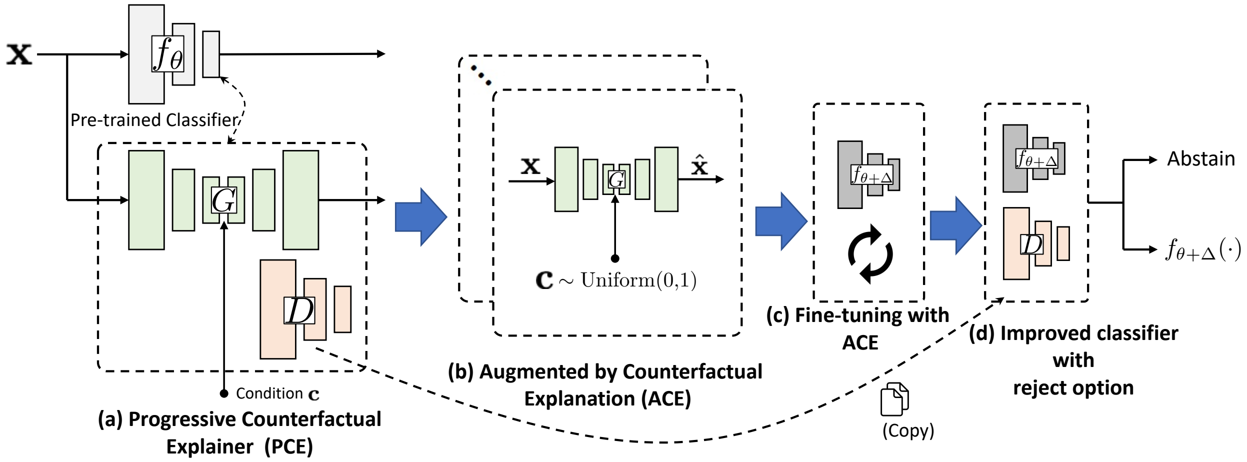

$\mathbf{x}$
f_θ
Pre-trained Classifier
G
D
Condition $\mathbf{c}$
(a) Progressive Counterfactual Explainer (PCE)
$\mathbf{x}$
G
$\hat{\mathbf{x}}$
$\mathbf{c} \sim$ Uniform(0,1)
(b) Augmented by Counterfactual Explanation (ACE)
$f_{\theta+\Delta}$
(c) Fine-tuning with ACE
(Copy)
$f_{\theta+\Delta}$
D
(d) Improved classifier with reject option
Abstain
$f_{\theta+\Delta}(\cdot)$

Figure 29: Overview of the method. (a) Given a *pre-trained* classifier f_θ, we learn a c-GAN based progressive counterfactual explainer (PCE) $G(\mathbf{x}, \mathbf{c})$, while keeping f_θ fixed. (b) The trained PCE creates counterfactually augmented data. (c) A combination of original training data and augmented data is used to fine-tune the classifier, $f_{\theta+\Delta}$. (d) The discriminator from PCE serves as a selection function to detect and reject OOD data.

Our goal is to *fine-tune* f_θ such that the revised model provides better uncertainty estimates, while retaining its original predictive accuracy. More specifically, we aim to improve uncertainty estimates for OOD and ambiguous samples. We use a progressive counterfactual explainer (PCE) to generate counterfactually augmented data. This data is then used to apply a few updates to f_θ, to gradually widen its decision boundary, resulting in improved uncertainty estimates on ambiguous and near-OOD samples. The PCE generates realistic perturbations of a given query image while gradually traversing the decision boundary between the classes, as defined by f_θ[243, 134]. We used the PCE with a conditional-GAN backbone that is trained with respect to f_θ. The discriminator of the cGAN-based PCE models is a density estimator that provides essential information to enhance f_θ far-OOD detection. More specifically, we used the discriminator as a *selection function* to abstain f_θ from making prediction on far-OOD samples. Fig. 29 summarizes our approach.

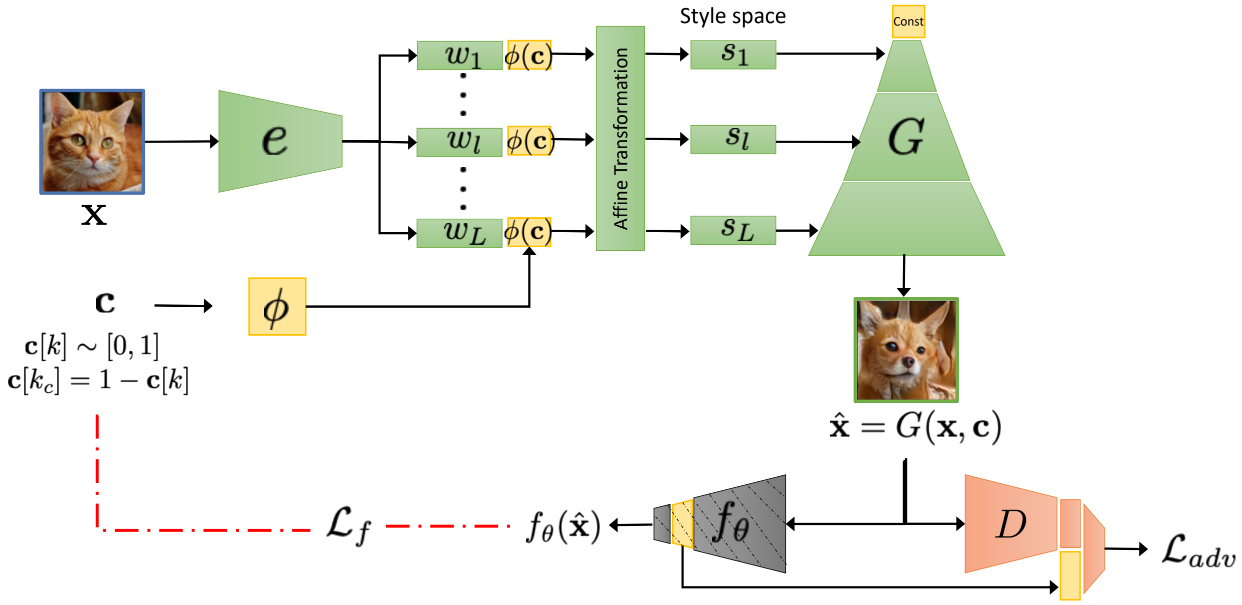


Figure 30: PCE: The encoder-decoder architecture to create counterfactual augmentation for a given query image.

The remaining sections are structured as follow: we first formulate the cGAN-based PCE model in Section 6.2.1. In Section 6.2.2, we describe our novel Augmentation by Counterfactual Explanation (ACE) strategy that uses the trained PCE to generate CDA and fine-tune f_θ. Finally, in Section 6.2.3, we combine the discriminator from the PCE with the fine-tuned f_θ to provide our final classifier.

Notation: The classification function is defined as $f_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^K$, where θ represents model parameters. The training dataset for f_θ is defined as $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathbf{x} \in \mathcal{X}$ represents an input space and $y \in \mathcal{Y} = \{1, 2, \cdots, K\}$ is a label set over K classes. The classifier produces point estimates to approximate the posterior probability $\mathbb{P}(y|\mathbf{x}, \mathcal{D})$.

6.2.1 Progressive Counterfactual Explainer (PCE) v2.0

We designed the PCE network to take a query image ($\mathbf{x} \in \mathbb{R}^d$) and a desired classification outcome ($\mathbf{c} \in \mathbb{R}^K$) as input, and create a perturbation of a query image ($\hat{\mathbf{x}}$) such that $f_\theta(\hat{\mathbf{x}}) \approx \mathbf{c}$. Our formulation, $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{c})$ allows us to use $\mathbf{c}$ to traverse through the decision boundary of f_θ from the original class to a counterfactual class. Following previous work [134, 243], we design the PCE to satisfy the following three properties:

1. **Data consistency:** The perturbed image, $\hat{\mathbf{x}}$ should be realistic and should resemble samples in $\mathcal{X}$.
2. **Classifier consistency:** The perturbed image, $\hat{\mathbf{x}}$ should produce the desired output from the classifier f_θ *i.e.*, $f_\theta(G(\mathbf{x}, \mathbf{c})) \approx \mathbf{c}$.
3. **Self consistency:** Using the original classification decision $f_\theta(\mathbf{x})$ as condition, the PCE should produce a perturbation that is very similar to the query image, *i.e.*, $G(G(\mathbf{x}, \mathbf{c}), f_\theta(\mathbf{x})) = \mathbf{x}$ and $G(\mathbf{x}, f_\theta(\mathbf{x})) = \mathbf{x}$.

Data Consistency: We formulate the PCE as a cGAN that learns the underlying data distribution of the input space $\mathcal{X}$ without an explicit likelihood assumption. The GAN model comprised of two networks – the generator $G(\cdot)$ and the discriminator $D(\cdot)$. The $G(\cdot)$ learns to generate fake data, while the $D(\cdot)$ is trained to distinguish between the real and fake samples. We jointly train G, D to optimize the following logistic adversarial loss [75],

$$\mathcal{L}_{\text{adv}}(D, G) = \mathbb{E}_{\mathbf{x}}[\log D(\mathbf{x}) + \log(1 - D(G(\mathbf{x}, \mathbf{c})))] \tag{21}$$

Our architecture for the cGAN is adapted from StyleGANv2 [1]. We formulate the generator as $G(\mathbf{x}, \mathbf{c}) = g(e(\mathbf{x}), \mathbf{c})$, a composite of two functions, an image encoder $e(\cdot)$ and a conditional decoder $g(\cdot)$ [1]. The encoder function $e : \mathcal{X} \rightarrow \mathcal{W}^+$, learns a mapping from the input space $\mathcal{X}$ to an extended latent space $\mathcal{W}^+$. The $\mathcal{W}^+$ represents a concatenation of L different latent representations (w_l), one for each layer of the decoder $g(\cdot)$. The conditional decoder $g(\cdot)$, maps the embedding back to the input space $\mathcal{X}$ while respecting the condition $\mathbf{c}$. We provide condition information to the decoder by concatenating the condition $\mathbf{c}$ to each $w_l \in \mathcal{W}^+$. The decoder further transforms the layer-specific latent representation into a layer-specific style-vector as $s_l = A_l([w_l, \phi(\mathbf{c})])$ where, A_l is an affine transformation and $\phi(\mathbf{c})$ is an embedding for $\mathbf{c}$. Further, we also extended the discriminator network $D(\cdot)$ to have auxiliary information from the classifier f_θ. Specifically, we concatenate the penultimate activations from the $f_\theta(\mathbf{x})$ with the penultimate activations from the $D(\mathbf{x})$, to obtain a revised representation before the final fully-connected layer of the discriminator. The detailed architecture is summarized in Fig. 30.

We also borrow the concept of path-length regularization $\mathcal{L}_{\text{reg}}(G)$ from StyleGANv2 to enforce smoother latent space interpolations for the generator.

$$\mathcal{L}_{\text{reg}}(G) = \mathbb{E}_{\mathbf{w}\sim e(\mathbf{x}), \mathbf{x}\sim\mathcal{X}}(||J_{\mathbf{w}}^T\mathbf{x}||_2 - a)^2 \tag{22}$$

where $\mathbf{x}$ denotes random images from the training data, $J_{\mathbf{w}}$ is the Jacobian matrix, and a is a constant that is set dynamically during optimization.

Classifier consistency: By default, GAN training is independent of the classifier f_θ. We add a classifier-consistency loss to regularize the generator and ensure that the actual classification outcome for the generated image $\hat{\mathbf{x}}$, is similar to the condition $\mathbf{c}$ used for generation. We enforce classification-consistency by a KullbackLeibler (KL) divergence loss as follow[243],

$$\mathcal{L}_f(G) = D_{KL}(f_\theta(\hat{\mathbf{x}})||\mathbf{c}) \tag{23}$$

Self consistency: We define the following reconstruction loss to regularize and constraint the Generator to preserve maximum information between the original image $\mathbf{x}$ and its reconstruction $\bar{\mathbf{x}}$,

$$\mathcal{L}(\mathbf{x}, \bar{\mathbf{x}}) = ||\mathbf{x} - \bar{\mathbf{x}}||_1 + ||e(\mathbf{x}) - e(\bar{\mathbf{x}})||_1 \tag{24}$$

Here, first term is a distance loss between the input and the reconstructed image, and the second term is a style reconstruction loss adapted from StyleGANv2 [1]. We minimize the reconstruction loss to satisfy the identify constraint on self reconstruction using $\bar{\mathbf{x}}_{self} = G(\mathbf{x}, f_\theta(\mathbf{x}))$. We further insure that the PCE learns a reversible perturbation by recovering the original image from a given perturbed image $\hat{\mathbf{x}}$ as $\bar{\mathbf{x}}_{\text{cyclic}} = G(\hat{\mathbf{x}}, f_\theta(\mathbf{x}))$, where $\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{c})$ with some condition $\mathbf{c}$. Our final reconstruction loss is defined as,

$$\mathcal{L}_{\text{rec}}(G) = \mathcal{L}(\mathbf{x}, \bar{\mathbf{x}}_{\text{self}}) + \mathcal{L}(\mathbf{x}, \bar{\mathbf{x}}_{\text{cyclic}}) \tag{25}$$

Objective function: Finally, we trained our model in an end-to-end fashion to learn parameters for the two networks, while fixing f_θ. Our overall objective function is:

$$\min_G \max_D \lambda_{\text{adv}} \left(\mathcal{L}_{\text{adv}}(D, G) + \mathcal{L}_{\text{reg}}(G)\right) + \lambda_f \mathcal{L}_f(G) + \lambda_{\text{rec}} \mathcal{L}_{\text{rec}}(G), \tag{26}$$

where, λ's are the hyper-parameters to balance each of the loss terms.

6.2.2 Augmentation by Counterfactual Explanation

Given a query image $\mathbf{x}$, the trained PCE generates a series of perturbations of $\mathbf{x}$ that gradually traverse the decision boundary of f_θ from the original class to a counterfactual class, while still remaining plausible and realistic-looking. This series of perturbations is essentially mimicking a traversal on a latent manifold, as guided by the condition $\mathbf{c}$. Our trained AN enables conditional generation of an image at any point on the manifold as $G(\mathbf{x}, \mathbf{c})$ (*see* Fig.31). We modify $\mathbf{c}$ to represent different steps in this traversal. We start from a high data-likelihood region for original class k ($\mathbf{c}[k] \in [0.8, 1.0]$), walk towards the decision hyperplane ($\mathbf{c}[k] \in [0.5, 0.8)$), and eventually cross the decision boundary ($\mathbf{c}[k] \in [0.2, 0.5)$) to end the traversal in a high data-likelihood region for the counterfactual class k_c ($\mathbf{c}[k] \in [0.0, 0.2)$). Accordingly, we set $\mathbf{c}[k_c] = 1 - \mathbf{c}[k]$.

Ideally, the predicted confidence from NN should be indicative of the distance from the decision boundary. Samples that lies close to the decision boundary should have low confidence, and confidence should increase as we move away from the decision boundary. We used $\mathbf{c}$ as a pseudo indicator of confidence to generate synthetic augmentation. Our augmentations are essentially showing how the query image $\mathbf{x}$ should be modified to have low/high confidence.

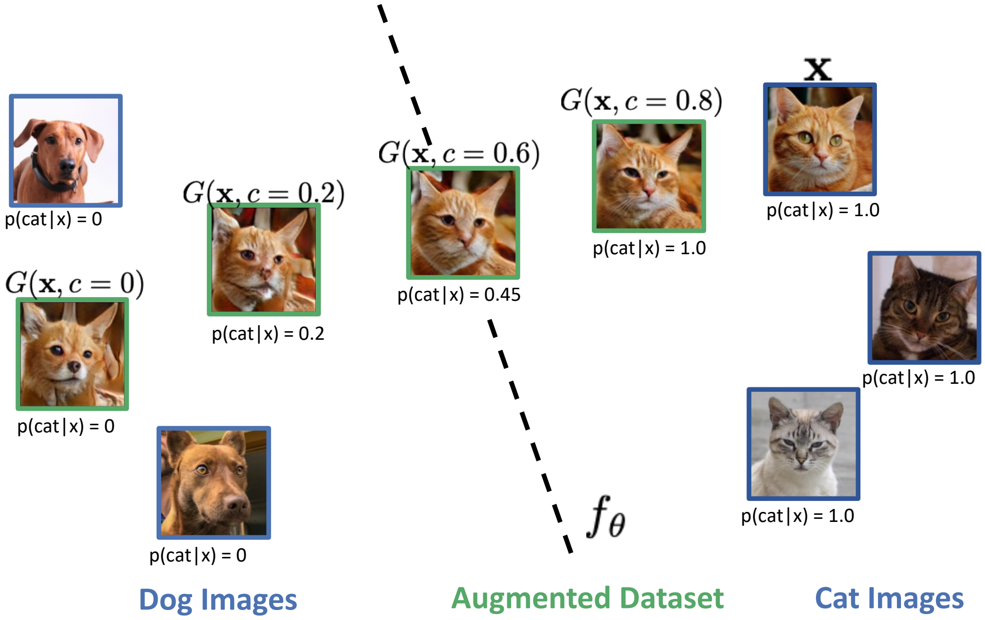


Figure 31: Augmentation by counterfactual explanation. Given a query image, the trained PCE generates a series of perturbations that gradually traverse the decision boundary of f_θ from the original class to a counter-factual class, while still remaining plausible and realistic-looking.

To generate CAD, we randomly sample a subset of real training data as $\mathcal{X}_r \subset \mathcal{X}$. Next, for each $\mathbf{x} \in \mathcal{X}_r$ we generate multiple augmentations ($\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{c})$) by randomly sampling $\mathbf{c}[k] \in [0, 1]$. We used $\mathbf{c}$ as soft label for the generate sample while fine-tuning the f_θ. The $\mathcal{X}_c$ represents our pool of generated augmentation images. Finally, we create a new dataset by randomly sampling images from $\mathcal{X}$ and $\mathcal{X}_c$. We fine-tune the f_θ on this new dataset, for only a few epochs, to obtain a revised classifier given as $f_{\theta+\Delta}$. In our experiments, we show that the revised decision function $f_{\hat{\theta}}$ provides improved confidence estimates for AiD and near OOD samples and demonstrate robustness to adversarial attacks, as compared to given classifier f_θ.

6.2.3 Discriminator as a Selection Function

A selection function $g : \mathcal{X} \rightarrow \{0, 1\}$ is an addition head on top of a classifier that decides when the classifier should abstain from making a prediction. We propose to use the discriminator network $D(\mathbf{x})$ as a selection function for f_θ. Upon the convergence of the PCE training, the generated samples resemble the in-distribution training data. Far-OOD samples are previously unseen samples which are very different from the training input space. Hence, $D(\cdot)$ can help in detecting such samples. Our final improved classification function is represented as follow,

$$(f, D)(\mathbf{x}) = \begin{cases} f_{\theta+\Delta}(\mathbf{x}), & \text{if } D(\mathbf{x}) \geq h \\ \texttt{Abstain}, & \text{otherwise} \end{cases} \tag{27}$$

where, $f_{\theta+\Delta}$ is the fine-tuned classifier and $D(\cdot)$ is a discriminator network from the PCE which serves as a selection function that permits f to make prediction if $D(\mathbf{x})$ exceeds a threshold h and abstain otherwise

During inference, the discriminator first uses its density estimates to quantify the similarity between the learned data distribution and the query image. These estimates provide a pseudo signal to quantify epistemic uncertainty. Using a strict threshold, the discriminator may reject any sample that lies outside its learned data distribution as far-OOD. The fine-tuned classifier then processes the accepted samples. We use the fine-tuned classifier's predictive entropy (PE) to quantify the sample's uncertainty. This uncertainty can be epistemic (associated with near-OOD samples) or aleatoric (associated with AID samples). Our model cannot differentiate between the two.

6.3 Experiments and Results

We set up three experiments to compare the baseline model before and after fine-tuning with our proposed counterfactual augmentation: First, we assess if the models can correctly capture aleatoric uncertainty by identifying ambiguous test samples. Second, we evaluate the models on OOD detection tasks. We consider a standard OOD task over separate datasets and a challenging task to detect near-OOD samples from previously unseen classes from the same dataset. Finally, we compare the model's behaviour under adversarial attacks. In SM, we show further experiments on more datasets and an ablation study over different components of the PCE.

6.3.1 Imaging dataset

In our experiments, we consider classification models trained on following datasets:

1. AFHQ [30]: Animal face high quality (AFHQ) dataset is a high resolution dataset of animal faces with 16K images from cat, dog and wild labels. The classifier is trained at an image resolution of 256 × 256.
2. Dirty MNIST [179]: The dataset is a combination of original MNIST [137] and simulated Ambiguous-MNIST dataset. Each sample in Ambiguous-MNIST is constructed by decoding a linear combination of latent representations of two different MNIST digits from a pre-trained VAE [127]. The samples are generated by combining latent representation of different digits, to simulate ambiguous samples, with multiple plausible labels [179]. The training dataset of the classifier comprises of 60K clean-MNIST and 60K Ambiguous-MNIST samples, with one-hot labels. The original dataset consists of grayscale images of size 28×28 pixels. We consider a classification model trained on 64×64 resolution.
3. CelebA [149]: Celeb Faces Attributes Dataset (CelebA) is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 binary attributes annotations per image. Our AiD samples comprises of middle-aged people who are arguably neither young nor old. To obtain such data, we use aleatoric uncertainty estimates from MC-Dropout averaged across 50 runs on test-set of CelebA. The classifier is trained at an image resolution of 256 × 256. We center-crop the images as a pre-processing step.
4. Skin lesion (HAM10K) [262]: The HAM10000 is a dataset of 100K dermatoscopic images of pigmented skin lesions. It contains seven different lesion types – Melanocytic Nevi (nv), Melanoma (mel), Benign Keratosis (bkl), Actinic Keratoses and Intraepithelial Carcinoma (akiec), Basal Cell Carcinoma (bcc), Dermatofibroma (df), Vascular skin

lesions (vasc). In our experiments, we consider classifier trained to distinguish the majority class nv from mel and bkl. We consider images from rest of the lesions as near-OOD. The classifier is trained at an image resolution of 256 × 256.

Classification tasks: We consider four classification problems, in increasing level of difficulty:

1. AFHQ [30]: We consider binary classification over well separated classes, cat vs dog. We consider images with "wild" label as near-OOD.
2. Dirty MNIST [179]: We consider multi-class classification over seven classes of handwritten digits '0' - '6'. We consider images from digits '7' - '9' as near-OOD samples.
3. CelebA [149]: we consider a two-class classifier over attributes "Young" and "Smiling" trained on CelebA dataset. Without age labels, identifying 'young' faces is a challenging task.
4. Skin lesion (HAM10K) [262]: We consider a binary classification to separate Melanocytic nevus (nv) from Melanoma (mel) and Benign Keratosis (bkl) lesions. Skin lesion classification is a challenging task as different lesions may exhibit similar features [183].

6.3.2 Experimental setup

Classification Model: We consider state-of-the-art DenseNet [98] architecture for the baseline. The *pre-trained* DenseNet model followed the training procedures as described in [98]. In DenseNet, each layer implements a non-linear transformation based on composite functions such as Batch Normalization (BN), rectified linear unit (ReLU), pooling, or convolution. The resulting feature map at each layer is used as input for all the subsequent layers, leading to a highly convoluted multi-level multi-layer non-linear convolutional neural network. We aim to improve such a model in a post-hoc manner without accessing the parameters learned by any layer or knowing the architectural details. Our proposed approach can be used for any DNN architecture.

Progressive Counterfactual Explainer v2.0: We formulate the progressive counterfactual explainer (PCE) as a composite of two functions, an image encoder $e(\cdot)$ and a conditional decoder $(g(\cdot))$ [1]. Our architecture for the conditional decoder is adapted from StyleGANv2 [1].

In order to keep the architecture and training procedure of PCE simple, we consider the default training parameters from [1] for training the StyleGANv2. This encourages reproducibility as we didn't do hyper-parameter tuning for each dataset and classification model. Specifically, we use Adam optimizer to train StyleGANv2 at 256^2 resolution for ~200K iterations with a batch size of 8 over 8 GPUs. For training StyelGANv2, we use a randomly sampled subset ($\sim 50\%$) of the baseline model's training data. For multi-class classification, we consider all pairs of classes while creating counterfactual augmentations. Further, given an input image, the predicted class k and a counterfactual class k_c, we initialize the condition $\mathbf{c}$ with all zeros and then set $\mathbf{c}[k] \sim \text{Uniform}(0, 1)$ and $\mathbf{c}[k_c] = 1 - \mathbf{c}[k]$. In all our experiments, we used $\lambda_{adv} = 10$, $\lambda_{rec} = 100$ and $\lambda_f = 10$.

To generate CAD, we randomly sample a subset of real training data as $\mathcal{X}_r \subset \mathcal{X}$. Next, for each $\mathbf{x} \in \mathcal{X}_r$ we generate four augmentations ($\hat{\mathbf{x}} = G(\mathbf{x}, \mathbf{c})$) by randomly sampling $\mathbf{c}[k] \in [0, 1]$. We used $\mathbf{c}$ as soft label for the generate sample while fine-tuning the f_θ. The $\mathcal{X}_c$ represents our pool of generated augmentation images.

For fine-tuning the given baseline with consider a combination of the original training dataset $\mathcal{X}$ and the augmented data $\mathcal{X}_c$. We randomly selected a subset of samples from the two distributions and fine-tune the baseline for 5 to 10 epochs. We used the expected calibration error and the test-set accuracy to choose the final checkpoint. Our model does not require access to OOD or AiD dataset during fine-tuning. During evaluation we compute predicted entropy (PE) for original test-set and OOD samples and measure for a range of thresholds how well the two are separated. We report the AUC-ROC and the true negative rate (TNR) at 95% true positive rate (TPR) (TNR@TPR95) in our results (*see Table 1 and 2*). We will release the GitHub for the project after the review process.

Comparison methods: Our baseline is a standard DNN classifier f_θ trained with cross-entropy loss. For baseline and its post-hoc variant with temperature-scaling (**TS**) [85], we used threshold over predictive entropy (PE) to identify OOD. PE is defined as $-\sum [f_\theta(\mathbf{x})]_k \log [f_\theta(\mathbf{x})]_k$. We compared against three methods that changes network training or architecture to learn better uncertainty estimates:

1. **Mixup**: Baseline model with mixup training using $\alpha = 0.2$ [299].
2. **DUQ**: Baseline model with radial basis function as the end-layer. Here, we use the

closest kernel distance to quantify uncertainties [267].

3. **DDU**: A ResNet-18 [90] model with spectral normalisation and Gaussian Mixture Model (GMM) for density estimation [179].

We also compared against methods that obtain uncertainty estimates from a pre-trained DNN output using threshold-based scoring functions.

4. **Energy-based** scoring function: Baseline with an energy function. We experimented with two variants, in first we compute energy score in a post-hoc manner and use it to identify OOD samples. In the second, we fine-tune the pre-trained DNN using the energy-score based loss and then identify OOD samples [148].
5. **Outlier exposure**: Baseline model with extra regularization from known outliers. While training the DNN, we assume access to outlier data and we force softmax output to be a uniform distribution for the outlier data [93].
6. **ODIN**: Baseline with TS followed by the post-hoc approach of Out-of-DIstribution detector for Neural networks (ODIN). In ODIN, we added small perturbations to the input to effectively separate OOD images from the in-distribution ones [142].

   Further, we also compared against two ensemble approaches:

7. **MC Dropout**: Baseline model trained with dropout. At inference time we took 20 dropout samples [65] to compute PE.
8. **5-Ensemble**: Baseline model trained five times with different seeds using the same dataset, shuffled using different seed [133]. We view the ensemble approaches as an upper bound for uncertainty quantification.

6.3.3 Identifying AiD samples

AiD samples have an inherent ambiguity in their class label, arising from the overlapping class definitions. Curated datasets for image classification are well-defined and provide binarized labels as ground truth. Assigning a label to an image removes ambiguity about its class membership. In the absence of ground truth uncertainty labels, we use the PE estimates from an MC Dropout classifier to label AiD samples. Specifically, we sort the test set using PE and consider the top 5 to 10% samples as AiD.

Table 13: Performance of different methods on identifying **ambiguous in-distribution (AiD)** samples. For all metrics, higher is better. The best results from the methods that require a single forward pass at inference time are highlighted. The ensemble approaches form an upper-bound and are for reference only and not comparison. Given a baseline model, our results are averaged over 10 runs of fine-tuning with different augmentation by counterfactual explanation (ACE) datasets.

Train	**Method/**	**Test-Set**	**Identifying AiD**	
Dataset	**Model**	Accuracy	AUC-ROC	TNR@TPR95
AFHQ	Baseline	99.44±0.02	0.87±0.04	48.93±10
	Baseline+TS [85]	99.45±0.00	0.85±0.07	48.77±9.8
	Mixup [299]	99.02±0.10	0.80±0.05	35.66±6.7
	DUQ [267]	94.00±1.05	0.67±0.01	26.15±4.5
	DDU [179]	97.66±1.10	0.74±0.02	19.65±4.5
	Baseline+Energy [148]	99.44±0.02	0.87±0.06	49.00±1.64
	Energy w/ fine-tune [148]	99.45±0.11	0.69±1.28	30.36±2.52
	Outlier Exposure [93]	99.50±0.14	0.85±0.01	41.07±0.75
	Baseline+TS+ODIN [142]	99.45±0.00	0.85±0.06	35.72±1.26
	Baseline+ACE	**99.52±0.21**	**0.91±0.02**	**50.75±3.9**
	MC Dropout [65]	98.83±1.12	0.87±0.04	51.56±1.2
	5-Ensemble [133]	99.79±0.01	0.98±0.01	51.93±2.7
Dirty MNIST	Baseline	95.68±0.02	**0.96±0.00**	28.5±2.3
	Baseline+TS [85]	95.74±0.02	0.94±0.01	27.90±1.3
	Mixup [299]	94.66±0.16	0.94±0.02	25.78±2.1
	DUQ [267]	89.34±0.44	0.67±0.01	23.89±1.2
	DDU [179]	93.52±1.12	0.65±0.12	20.78±4.0
	Baseline+Energy [148]	95.68±0.02	0.80±0.03	17.60±0.55
	Energy w/ fine-tune [148]	96.17±0.02	0.39±0.04	11.59±0.25
	Outlier Exposure [93]	**96.30±0.07**	0.63±0.07	17.6±2.88
	Baseline+TS+ODIN [142]	95.74±0.02	0.79±0.03	13.25±4.88
	Baseline+ACE	95.36±0.45	0.86±0.01	**34.12±2.6**
	MC Dropout [65]	89.50±1.90	0.75±0.07	36.10±1.8
	5-Ensemble [133]	95.90±0.12	0.98±0.02	34.87±3.4

In Fig. 28, we qualitatively compare the PE distribution from the given baseline and its fine-tuned version (baseline + ACE). Fine-tuning resulted in minor changes to the PE distribution of the iD samples (Fig. 28.A). We observe a significant separation in the PE distribution of AiD samples and the rest of the test set (Fig. 28.B), even on the baseline. This suggests that the PE correctly captures the aleatoric uncertainty. Fine-tuning with counterfactual augmentation further enhanced this separation by shifting the PE distribution of AiD samples to the right and assigning a higher PE value to the uncertain samples.

The Table 13 and Table 14 compare our model to several methods. We report the test set accuracy, the AUC-ROC for the binary task of identifying AiD samples and the true negative rate (TNR) at 95% true positive rate (TPR) (TNR@TPR95), which simulates an application requirement that the recall of in-distribution data should be 95% [96]. For all metrics higher value is better. Our model outperformed other deterministic models, in identifying AiD samples with a high AUC-ROC and TNR@TPR95 across all datasets. Also, the fine-tuned model retained the predictive accuracy of the baseline in AFHQ and Dirty MNIST datasets. We observe a little drop in test accuracy in complex classification problems, where multiple classes may look similar, e.g. medical datasets (HAM10K). Fine-tuning makes the decision boundary broader. As a result, the samples near the decision boundary may flip their decision, decreasing accuracy. We consider this drop more like a flag to indicate possible label noise. Please note, in the tables, we highlighted the best results from the methods that require a single forward pass at inference time.

6.3.4 Detecting OOD samples

We consider two tasks to evaluate the model's OOD detection performance. First, a standard OOD task where OOD samples are derived from a separate dataset. Second, a difficult near-OOD detection task where OOD samples belongs to novel classes from the same dataset, which are not seen during training. We consider the following OOD datasets:

1. AFHQ [30]: We consider "wild" class from AFHQ to define near-OOD samples. For the far-OOD detection task, we use the CelebA dataset, and also cat/dog images from CIFAR10 [132].
2. Dirty MNIST [179]: We consider digits 7-9 as near-OOD samples. For far-OOD detection, we use SVHN [187] and fashion MNIST [283] datasets.
3. CelebA [149]: We consider images of kids in age-group: 0-11 from the UTKFace [298]

Table 14: Performance of different methods on identifying **ambiguous in-distribution (AiD)** samples. For all metrics, higher is better. The best results from the methods that require a single forward pass at inference time are highlighted. The ensemble approaches form an upper-bound and are for reference only and not comparison. Given a baseline model, our results are averaged over 10 runs of fine-tuning with different augmentation by counterfactual explanation (ACE) datasets.

Train	**Method/**	**Test-Set**	**Identifying AiD**	
Dataset	**Model**	Accuracy	AUC-ROC	TNR@TPR95
CelebA	Baseline	89.36±0.96	0.73±0.01	17.18±1.6
	Baseline+TS [85]	89.33±0.01	0.72±0.02	17.21±1.5
	Mixup [299]	89.04±0.47	**0.74±0.02**	15.09±1.9
	DUQ [267]	71.75±0.01	0.65±0.01	14.20±1.0
	DDU [179]	70.15±0.02	0.67±0.06	11.39±0.4
	Baseline+Energy [148]	89.36±0.96	0.57±0.28	4.87±0.32
	Energy w/ fine-tune [148]	**90.22±0.96**	0.53±1.25	5.06±0.28
	Outlier Exposure [93]	86.65±1.22	0.53±0.46	5.06±0.19
	Baseline+TS+ODIN [142]	89.33±0.01	0.57±0.01	6.34±0.38
	Baseline+ACE	86.8±0.79	**0.74±0.06**	**22.36± 2.3**
	MC Dropout [65]	89.86±0.33	0.73±0.03	19.78±0.7
	5-Ensemble [133]	90.76±0.00	0.84±0.11	17.79±0.6
Skin-Lesion (HAM10K)	Baseline	85.88±0.75	0.82±0.06	20.52±3.7
	Baseline+TS [85]	**86.27±0.40**	**0.84±0.03**	23.34±2.8
	Mixup [299]	85.81±0.61	**0.84±0.04**	31.29±7.0
	DUQ [267]	75.47±5.36	0.81±0.02	30.12±4.4
	DDU [179]	75.84±2.34	0.79±0.03	26.12±6.6
	Baseline+Energy [148]	85.88±0.75	0.77±0.12	18.40±0.51
	Energy w/ fine-tune [148]	**86.56±0.53**	0.64±0.06	17.45±1.78
	Outlier Exposure [93]	86.37±0.46	0.73±0.02	13.21±2.70
	Baseline+TS+ODIN [142]	86.27±0.40	0.78±0.01	15.87±4.33
	Baseline+ACE	81.21±1.12	**0.84±0.05**	**71.60±3.8**
	MC Dropout [65]	84.90±1.17	0.85±0.06	43.78±1.9
	5-Ensemble [133]	87.89±0.13	0.86±0.02	40.49±5.1

dataset to define the near-OOD samples. For far-OOD detection task, we use the AFHQ and CIFAR10 datasets.

4. Skin lesion (HAM10K) [262]: We consider samples from lesion types: Actinic Keratoses

and Intraepithelial Carcinoma (akiec), Basal Cell Carcinoma (bcc), Dermatofibroma (df) and Vascular skin lesions (vasc) as near-OOD. For far-OOD, we consider CelebA and an additional simulated dataset with different skin textures/tones.

We summarize our qualitative results in Fig. 28. We observe much overlap between the predictive entropy (PE) distribution of the near-OOD samples and in-distribution samples in Fig. 28.C. Fine-tuning with counterfactual augmentation helped in reducing this overlap, by amplifying the uncertainty associated with OOD data. Further, in Fig. 28.D-E, we observe that the PE distribution from the baseline model does not capture epistemic uncertainty associated with far-OOD samples, but our model successfully disentangles OOD samples from the in-distribution samples by using density estimates from the discriminator of the PCE.

Table 15: OOD detection performance for different baselines. **Near-OOD** represents label shift, with samples from the unseen classes of the same dataset. **Far-OOD** samples are from a separate dataset. The numbers are averaged over five runs.

Train Dataset	**Method**	**Near-OOD (Wild)**		**Far-OOD (CIFAR10)**		**Far-OOD (CelebA)**	
		AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95
AFHQ	Baseline	0.88±0.04	47.40±5.2	0.95±0.04	73.59±9.4	0.95±0.03	70.69±8.9
	Baseline+TS [85]	0.88±0.03	45.53±9.8	0.95±0.04	71.77±8.9	0.95±0.03	65.89±8.3
	Mixup [299]	0.86±0.06	53.83±6.8	0.82±0.11	57.01±8.6	0.88±0.13	70.51±9.8
	DUQ [267]	0.78±0.05	20.98±2.0	0.67±0.59	16.23±1.5	0.66±0.55	15.34±2.6
	DDU [179]	0.83±0.02	23.19±2.6	0.90±0.02	32.98±10	0.75±0.02	10.32±5.6
	Baseline+Energy [148]	0.88±0.03	47.77±1.10	0.94±0.05	72.68±2.69	0.96±0.04	74.75±2.89
	Energy w/ fine-tune [148]	**0.93±3.06**	45.97±2.78	**0.99±0.00**	0.66±0.01	0.94±1.86	68.38±3.03
	Outlier Exposure [93]	0.92±0.01	**73.99±2.62**	0.99±0.20	**99.54±0.79**	0.96±0.01	78.69±3.02
	Baseline+TS+ODIN [142]	0.87±0.05	45.02±1.51	0.95±0.05	69.42±2.38	0.95±0.03	67.18±2.16
	Baseline+ACE	0.89±0.03	51.39±4.4	0.98±0.02	88.71±5.7	**0.97±0.03**	**88.87±9.8**
	MC-Dropout [65]	0.84±0.09	30.78±2.9	0.94±0.02	73.59±2.1	0.95±0.02	71.23±1.9
	5-Ensemble [133]	0.99±0.01	65.73±1.2	0.97±0.02	89.91±0.9	0.99±0.01	92.12±0.7
		Near-OOD (Digits 7-9)		**Far-OOD (SVHN)**		**Far-OOD (fMNIST)**	
		AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95
Dirty MNIST	Baseline	0.86±0.04	28.23±2.9	0.75±0.15	51.98±0.9	0.87±0.02	58.12±1.5
	Baseline+TS [85]	0.86±0.01	30.12±2.1	0.73±0.07	48.12±1.5	0.89±0.01	61.71±2.8
	Mixup [299]	0.86±0.02	35.46±1.0	0.95±0.03	65.12±3.1	0.94±0.05	66.00±0.8
	DUQ [267]	0.78±0.01	15.26±3.9	0.73±0.03	45.23±1.9	0.75±0.03	50.29±3.1
	DDU [179]	0.67±0.07	10.23±0.9	0.68±0.04	39.31±2.2	0.85±0.02	53.76±3.7
	Baseline+Energy [148]	0.87±0.04	40.30±1.05	0.86±0.12	43.92±2.30	0.91±0.02	62.10±5.17
	Energy w/ fine-tune [148]	0.60±0.08	37.43±0.93	**1.00±0.00**	**99.99±0.00**	**1.00±0.00**	99.06±0.01
	Outlier Exposure [93]	**0.94±0.01**	**65.58±1.64**	**1.00±0.00**	**99.99±0.00**	**1.00±0.00**	**99.56±0.12**
	Baseline+TS+ODIN [142]	0.83±0.04	34.13±12.07	0.77±0.13	21.59±19.62	0.89±0.02	46.43±4.31
	Baseline+ACE	**0.94±0.02**	37.23±1.9	0.98±0.02	67.88±3.1	0.97±0.02	70.71±1.1
	MC-Dropout [65]	0.97±0.02	40.89±1.5	0.95±0.01	62.12±5.7	0.93±0.02	65.01±0.7
	5-Ensemble [133]	0.98±0.02	42.17±1.0	0.82±0.03	55.12±2.1	0.94±0.01	64.19±4.2

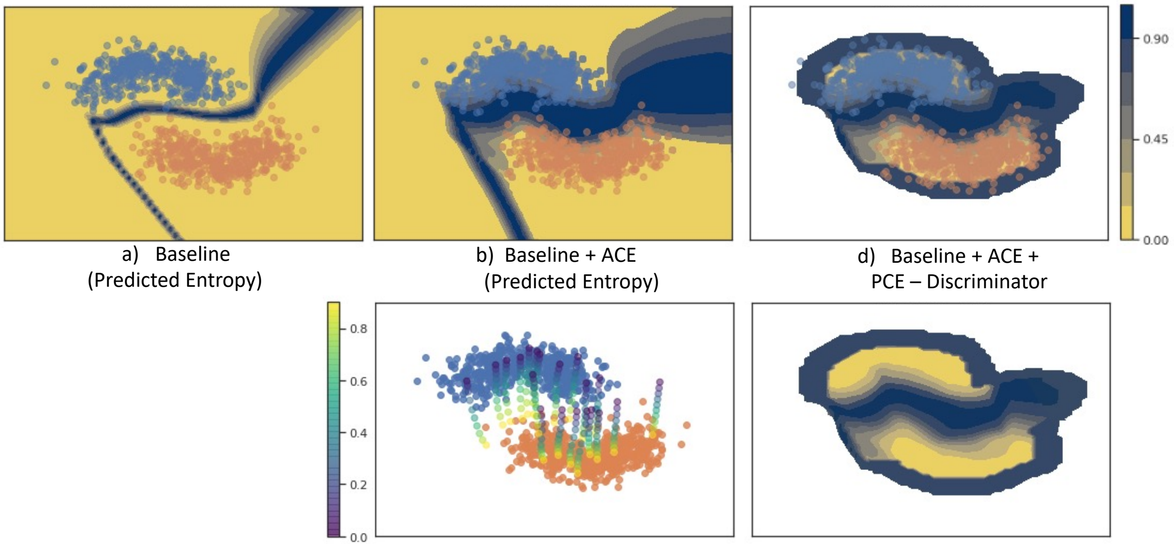
In Table 15 and Table 16, we report the AUC-ROC and TNR@TPR95 scores on detecting the two types of OOD samples. We first use the discriminator from the PCE to detect far-OOD samples. The discriminator achieved near-perfect AUC-ROC for detecting far-OOD samples. It consistently outperformed the deep ensemble, MC Dropout, and other deterministic methods across all datasets. The near-OOD samples are relatively similar to the training distribution of the discriminator. Hence, the discriminator performed sub-optimally on the near-OOD detection task. We used the PE estimates from the fine-tuned model (baseline + ACE) to detect near-OOD samples. We outperformed all other deterministic methods in identifying near-OOD samples. Overall our model performed better on both near and far-OOD detection tasks with high TNR@TPR95.

Table 16: OOD detection performance for different baselines. **Near-OOD** represents label shift, with samples from the unseen classes of the same dataset. **Far-OOD** samples are from a separate dataset. The numbers are averaged over five runs.

Train	**Method**	**Near-OOD (Kids)**		**Far-OOD (AFHQ)**		**Far-OOD (CIFAR10)**	
		AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95
CelebA	Baseline	0.84±0.02	1.25±0.1	0.86±0.03	88.57±0.9	0.79±0.02	29.01±5.1
	Baseline+TS [85]	0.82±0.04	1.24±0.1	0.87±0.06	88.75±0.9	0.78±0.04	29.01±5.1
	Mixup [299]	0.82±0.08	22.18±2.7	0.95±0.02	82.96±2.5	0.79±0.13	30.54±1.3
	DUQ [267]	0.80±0.03	14.68±3.1	0.72±0.07	26.62±7.7	0.86±0.04	28.70±4.1
	DDU [179]	0.73±0.15	7.9±0.4	0.74±0.13	8.18±0.4	0.81±0.15	25.45±1.4
	Baseline+Energy [148]	0.76±0.51	9.40±0.01	0.94±0.08	32.08±1.70	0.85±0.76	17.10±0.72
	Energy w/ fine-tune [148]	0.85±1.27	32.81±1.92	**0.99±0.00**	**99.99±0.00**	0.91±0.77	**84.35±1.29**
	Outlier Exposure [93]	0.66±0.69	8.44±0.45	0.75±0.70	26.09±0.51	0.69±0.53	16.63±0.90
	Baseline+TS+ODIN [142]	0.65±0.01	8.75±2.21	0.55±0.01	23.03±0.16	0.54±0.01	5.00±0.07
	Baseline+ACE	**0.87±0.03**	**34.37±2.5**	0.96±0.01	96.35±2.5	**0.92±0.05**	63.51±1.5
	MC-Dropout [65]	0.70±0.10	25.62±1.7	0.86±0.1	91.72±7.5	0.74±0.12	64.79±1.8
	5-Ensemble [133]	0.93±0.03	10.35±0.2	0.99±0.0	98.31±1.2	0.92±0.10	61.88±1.2
		Near-OOD (other lesions)		**Far-OOD (CelebA)**		**Far-OOD (Skin-texture)**	
		AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95	AUC-ROC	TNR@TPR95
Skin Lesion	Baseline	0.67±0.04	8.70±2.5	0.66±0.06	10.00±3.6	0.65±0.10	5.91±2.8
	Baseline+TS [85]	0.67±0.05	8.69±2.0	0.63±0.06	9.24±4.3	0.68±0.07	5.70±3.2
	Mixup [299]	0.67±0.01	8.52±2.8	0.64±0.08	10.21±4.0	0.72±0.05	5.26±3.1
	DUQ [267]	0.67±0.04	3.12±1.8	0.89±0.09	11.89±2.5	0.64±0.03	4.8±1.5
	DDU [179]	0.65±0.03	3.45±1.9	0.75±0.04	15.45±2.9	0.71±0.05	4.19±1.3
	Baseline+Eenergy [148]	0.70±0.04	10.85±0.08	0.70±0.14	7.90±0.29	0.65±0.20	2.83±1.33
	Energy w/ fine-tune [148]	0.62±0.02	9.80±1.81	**1.00±0.00**	**99.77±0.33**	0.76±0.13	16.04±1.08
	Outlier Exposure [93]	0.67±0.09	10.38±3.30	0.99±0.00	97.17±2.37	0.81±0.08	22.64±4.30
	Baseline+TS+ODIN [142]	0.68±0.01	9.43±0.33	0.67±0.07	11.32±4.66	0.68±0.07	6.60±0.29
	Baseline+ACE	**0.72±0.04**	**11.00±2.8**	0.97±0.02	66.77±1.4	**0.96±0.03**	**95.83±5.0**
	MC-Dropout [65]	0.67±0.05	9.45±3.9	0.80±0.07	30.00±3.2	0.56±0.03	10.87±2.3
	5-Ensemble [133]	0.88±0.01	11.23±1.7	0.91±0.03	27.89±5.9	0.76±0.02	17.89±3.5

6.3.5 Toy-Setup - Two Moons

This section demonstrates the over-confidence problem in a classifier trained on the two moons dataset. Using the scikit-learn's datasets package, we generated 2000 samples with a noise rate of 0.1. Our baseline classification model is a 2-layer MLP. In Fig. 32.a, we visualize the uncertainty estimates from this classifier. A classifier optimized for cross-entropy loss learns a very sharp decision boundary with low uncertainty only near the decision boundary and high uncertainty everywhere else.



a) Baseline (Predicted Entropy)

b) Baseline + ACE (Predicted Entropy)

d) Baseline + ACE + PCE – Discriminator

c) Example of Augmentation by Counterfactual Explanation (ACE) with soft labels

e) Final Uncertainty Landscape

Figure 32: Uncertainty results on the Two Moons dataset. Yellow indicates low uncertainty, while blue indicates uncertainty. a) The baseline classifier is uncertain only along the decision boundary, and certain elsewhere. b) Fine-tuning baseline model on ACE data improves uncertainty estimates near the decision boundary. c) An example of augmented data and corresponding soft labels. d) The discriminator from PCE rejects OOD samples, hence the rejected space have no uncertainty values (white color). e) The final uncertainty landscape, the improved classifier is certain on in-distribution regions and rejects OOD data.

In Fig. 32.b, we visualize the revised decision boundary after fine-tuning the classifier with counterfactually augmented data (CAD). The decision boundary is much broader, and uncertainty is high near the decision boundary, decreasing as one moves away from it. In Fig. 32.c, we show examples of CAD. Using the progressive counterfactual explainer (PCE), we created samples resembling a walk from one class to the other while crossing the decision boundary.

This simple example demonstrates that fine-tuning the classifier with augmented data near the decision boundary with soft labels helped the classifier recover from the over-saturation on the negative log-likelihood (NLL) loss. Hence, the fine-tuned classifier has better uncertainty estimates near the decision boundary and is not over-confident on ambiguous in-distribution samples in the class over-lapping regions.

Further, in Fig. 32.d, we show the hard threshold used by the discriminator of the PCE as the selection function. We processed all the samples through the discriminator of the PCE and used a pre-defined threshold to separate in-distribution samples from the OOD samples. The white colour in the plot is the samples the discriminator rejects as OOD. Fig. 32.e shows the final uncertainty landscape. The near-OOD around the in-distribution samples all have high uncertainty. The ambiguous in-distribution samples are assigned a high uncertainty near the decision boundary.

6.3.6 Robustness to Adversarial Attacks

In this experiment, we compared the baseline model before and after fine-tuning (baseline + ACE) in their robustness to three adversarial attacks: Fast Gradient Sign Method (FGSM) [76], Carlini-Wagner (CW) [20], and DeepFool [177]. For each attack setting, we transformed the test set into an adversarial set. In Fig. 33, we report the AUC-ROC over the adversarial set as we gradually increase the magnitude of the attack. For FGSM, we use the maximum perturbation (ϵ) to specify the attack's magnitude. For CW, we gradually increase the number of iterations to an achieve a higher magnitude attack. We set box-constraint parameter as $c = 1$, learning rate $\alpha = 0.01$ and confidence $\kappa = 0, 5$. For DeepFool ($\eta = 0.02$), we show results on the best attack. Our improved model (baseline + ACE) consistently out-performed the baseline model in test AUC-ROC.

6.3.7 Ablation Study

We conducted an ablation study over the three-loss terms of PCE in Eq. 26. The three terms of the loss function enforce three properties of counterfactual explanation, data consistency: explanations should be realistic looking images, classifier consistency: explanations should produce the desired outcome from the classifier and self-consistency: explanation image should retain the identity of the query image. For the ablation study, we consider the cat and dog classifier. We train three PCE; in each run, we ablate one term from the final loss function.

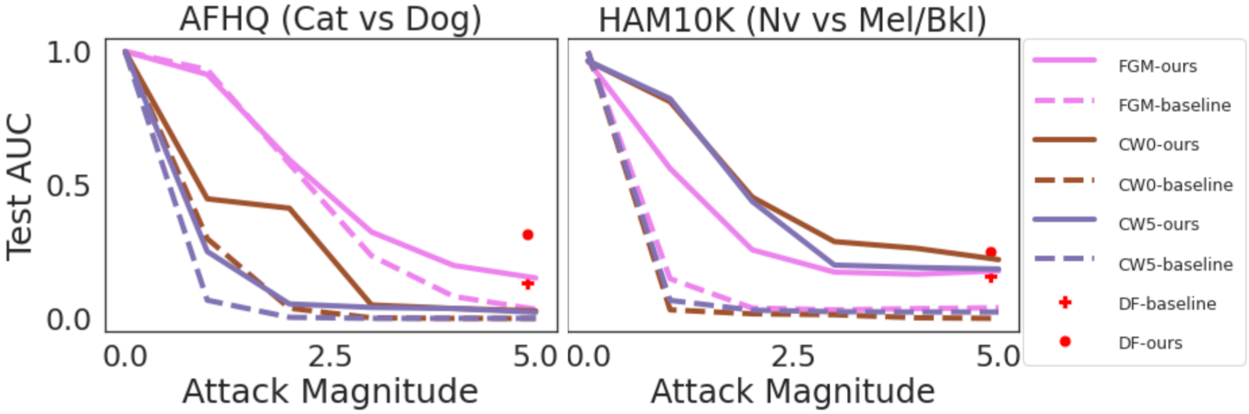


Figure 33: Plots comparing baseline model before and after fine-tuning (ACE) for different magnitudes of adversarial attack. The figure shows three different attacks – FGSM [76], CW [20], DeepFool [177], on three different datasets – HAM10K, AFHQ, MNIST. The x-axis denotes maximum perturbation (ϵ) for FGSM, and iterations in multiples of 10 for CW and DeepFool. Attack magnitude of 0 indicates no attack. For CW we used $\kappa = 0$ and 5. (All results are reported on the test-set of the classifier).

In Fig. 34, we show a qualitative example of the counterfactual data augmentation generated through each PCE. Without data consistency, the images are blurry and are no longer realistic. Without classifier consistency loss, though the images are natural, the classifier's output is not changing with the condition. Hence such PCE won't generate augmented samples near the decision boundary, which is the goal of our proposed strategy. With self-consistency, the generated images are not a gradual transformation of a given query image.

Further, in Fig. 35 we present quantitatively compare the uncertainty estimates from the baseline, before and after the fine-tuning with ACE. We represent a different ablation over the three-loss terms in each row.

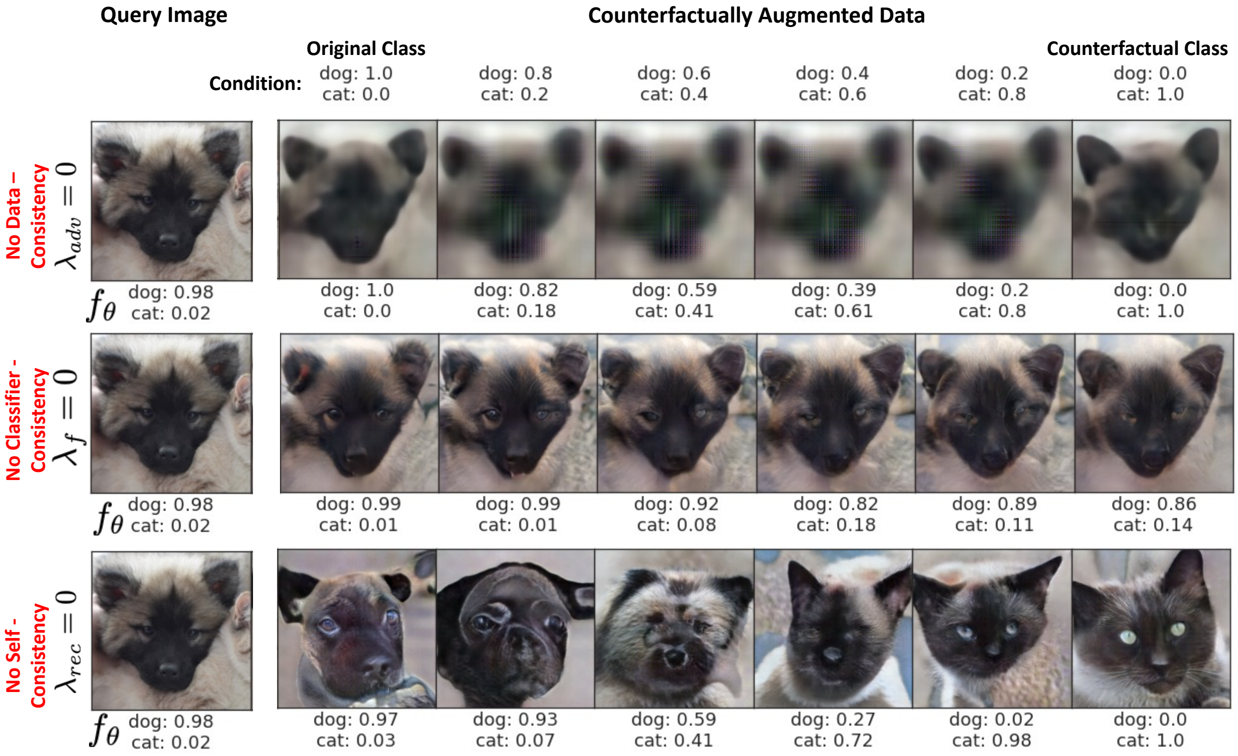


Figure 34: Examples of data augmentation while ablating different loss terms.

Fig. 35.A. shows the predicted entropy (PE) of **in-distribution (iD)** samples. Ideally, fine-tuning should minimally affect the PE distribution over iD samples. Without classification consistency loss (second row), the PE distribution of iD samples changed significantly. Fig. 35.B and Fig. 35.C shows the PE distribution over **ambiguous in-distribution (AiD)** samples and **near-OOD** samples, respectively. The data augmentation derived from PCE without adversarial loss or reconstruction loss cannot separate AiD samples or near-OOD from the rest of the test set. In Fig. 35.D, we use the discriminator of the PCE to identify **far-OOD** samples. In all three rows, we observe the sub-optimal performance of the discriminator in identifying and rejecting far-OOD samples. The legend shows the AUC-ROC for binary classification over uncertain and iD samples. Hence, all three loss terms are important to improve the uncertainty estimates of the baseline over all samples across the uncertainty spectrum.

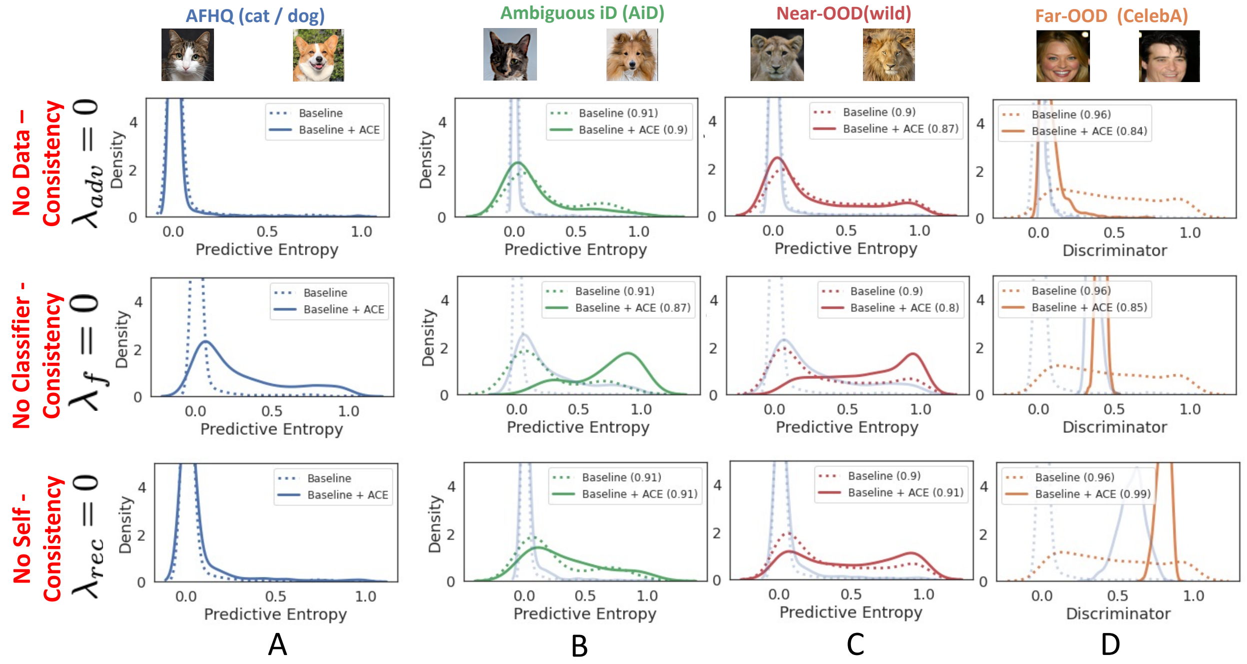


Figure 35: Comparison of the uncertainty estimates from the baseline, before and after the fine-tuning with ACE. Each row represents a different ablation over the three loss terms. A) Predicted entropy (PE) of **in-distribution (iD)** samples. Ideally, fine-tuning should minimally effect the PE distribution over iD samples. Without classification consistency loss (second row), the PE distribution of iD samples changed significantly. B) PE distribution over **ambiguous in-distribution (AiD)** samples. C) PE distribution over **near-OOD** samples. The data augmentation derived from PCE without adversarial loss or reconstruction loss, is not able to separate AiD samples or near-OOD from rest of the test set. D) We use the discriminator of the PCE to identify **far-OOD** samples. In all three rows, we observe sub-optimal performance of the discriminator in identifying and rejecting far-OOD samples. The legend shows the AUC-ROC for binary classification over uncertain samples and iD samples. Hence, all three loss terms are important to improve the uncertainty estimates of the baseline over all samples across the uncertainty spectrum.

6.4 Discussion and Conclusion

We propose a novel method to improve the uncertainty quantification of an existing *pre-trained* DNN by fine-tuning it on counterfactually augmented data. We used a cGAN-based counterfactual explainer to generate the data augmentation. Our fine-tuned model, combined with the discriminator of the GAN, can successfully capture uncertainty over ambiguous samples, unseen near-OOD samples with label shift and far-OOD samples from independent datasets. Comparative post-hoc methods such as thresholding softmax outputs and temperature scaling cannot recover a pre-trained model from over-saturation on log-likelihood loss. Other deterministic methods significantly change the classification model design to enable better uncertainty quantification over OOD samples. These methods require a network to be trained from scratch and are not compatible with a pre-trained classifier. Our proposed strategy reuses the counterfactual explanation model for the given classifier to fix its over-confidence problem. We out-performed state-of-the-art methods for uncertainty quantification on four datasets with varying difficulty levels. Furthermore, our improved model also exhibits robustness to prevalent adversarial attacks. We recognize that our proposed strategy involves training a GAN and fine-tuning the classifier with augmented data, which creates a one-time computational overhead. But, once we have a fine-tuned classifier, it requires only a single forward pass, with fast inference. The trained GAN has the added benefit of explaining the given classification model that can help in making it more user-accessible. Our work opens up a new direction for improving uncertainty quantification in existing classification models.

7.0 Conclusions and Future Directions

7.1 Conclusion

In this thesis, we developed new deep learning architectures and post-hoc explainability techniques that improved the application of DL methods for medical image classification. This dissertation proposes models to account for subtleties of medical imaging and add support for specific clinical needs. The major contribution of this work is to tackle model explanation from different perspectives. We started with building an interpretable model, that not only provides accurate predictions but also use an attention mechanism to show important/relevant regions of the input. We designed the model to handle the subtleties of medical images, by have an input processing pipeline that can process the entire 3D volume with minimum resizing, thus reserving the spatial integrity of the imaging data (Chapter-3).

Moving on, we developed a progressive counterfactual explainer, that provides visual explanation to explain the decision of a pre-trained classifier in a post-hoc manner. The design of our explainer is highly motivated by the clinical use-cases. For instance, most of the lung diseases are developed progressively and hardy have a sudden appearance. Out explanation, shows a gradual transformation of the query image, where the input CXR gradually becomes positive for a diagnosis. This aligns with the clinical expectation of how decision for a diagnosis should change. Further, counterfactual explanations are superior than saliency-map based methods as they not only show where in the image the classifier is paying attention to make its prediction, but also shows what image features in those salient regions are essential for the positive or negative decision. We supported our methods with through experiments, on both natural image datasets and medical image datasets. From a clinical perspective, we demonstrated that the counterfactual changes associated with normal (negative) or abnormal (positive) classification decisions are also associated with corresponding changes in disease-specific metrics such as CTR and SCP. For example, changes associated with an increased posterior probability for cardiomegaly also resulted in an increased CTR. Similarly, for PE, a healthy CP recess with a high SCP score transformed into an abnormal CP recess with blunt CPA, as the posterior probability for PE increase. Further, we evaluated our methods through a human evaluation study. The results of our human evaluation study confirms that the counterfactual explanations obtain from our method helped the clinicians better understand the classification decision (Chapter-4).

Chapter-5 presents an application of the counterfactual explainer in obtaining concept-based explanations. This method is also motivated by the need of the domain expert, to receive model explanation in a terminology that is meaningful to them. To fulfill this requirement, we provide explanation in terms of clinical concepts that are used in radiology reports to support the presence of a diagnosis. Specifically, we associate the internal structure of the deep neural network with clinically relevant concepts and used our counterfactual explanations to measure the causal effect of these concepts on the model's prediction. We adopted tools from Causal Inference literature and, more specifically, mediation analysis through counterfactual interventions. Using measures from mediation analysis, we provide an effective ranking of the concepts based on their causal relevance to the model's outcome. Finally, we construct a low-depth decision tree to express discovered concepts in simple decision rules, providing the global explanation for the model. We presented a through experiment of our proposed method on a clinical dataset.

For a more comprehensive interpretation of the deep learning models by their end-users, in Chapter-7 we demonstrate how to improve the uncertainty estimates from a pre-trained classifier, by fine-tuning the classifier with counterfactually augmented data. Counterfactual data lies near the decision boundary between two classes, Fine-tuning with such data helps in making the decision boundary wider and thus preventing the classifier from making over confident predictions on the sample near the decision boundary. Further, we show that the discriminator from the counterfactual explanier is a good proxy to the data distribution. The likelihood estimates from this model thus can identify and reject OOD samples. The experiments with the natural and medical images showed that our proposed technique is helpful in learning more reliable classifiers. We out-performed state-of-the-art methods for uncertainty quantification on four datasets with varying difficulty levels. Furthermore, our improved model also exhibits robustness to prevalent adversarial attacks

Overall, this thesis is a summarization of different ways to explain the deep learning model decision. The methods proposed in this thesis provided a set of tools to the deep learning model designers to better design and explain DNNs, while satisfying the clinical needs. Making progress in this direction will ensure the path to deployment for DNN models. For all the proposed methods we provide thorough comparisons with existing baselines and in each case we demonstrate reliable and superior performance.

7.2 Future directions

There are also several avenues for improvement which are left for future work:

1. We lose the context information when representing a volumetric image as a set of patches. There is no notion of spatial context when elements of a set are processed in a format invariant to their order. Future work should explore adding a positional encoding to the patches to incorporate spatial information. Much of this is inspired by recent advanced DNN architectural designs. Highly complex DNN designs such as vision transformers also take tokens as input. These patches, along with positional encoding, can become an essential way of processing 3D volumetric data for transformer-based architectures [58].
2. Diversity is an essential aspect of counterfactual explanations. Diversity among the generated counterfactuals provides different ways of changing the outcome decision. Diverse counterfactuals offer users multiple options to understand which input features are important for the classification decision. Diverse counterfactuals may include changes to a particular concept or several concepts. Current work is restrictive as it creates only a single counterfactual image. Future work should explore generating diverse counterfactual explanations showing all possible ways of changing the classification decision [214, 178]. Further, disentangling the counterfactual changes into human-understandable concepts can enrich the quality and usability of counterfactual explanations.
3. A counterfactual explanation is incomplete without causal analysis [161]. The current work uses neuron activations as a weak signal to capture concept information in the network. Future work should build on it to identify concepts in counterfactual explanations [78]. A possible next step is to use the concept vectors to navigate the direction in

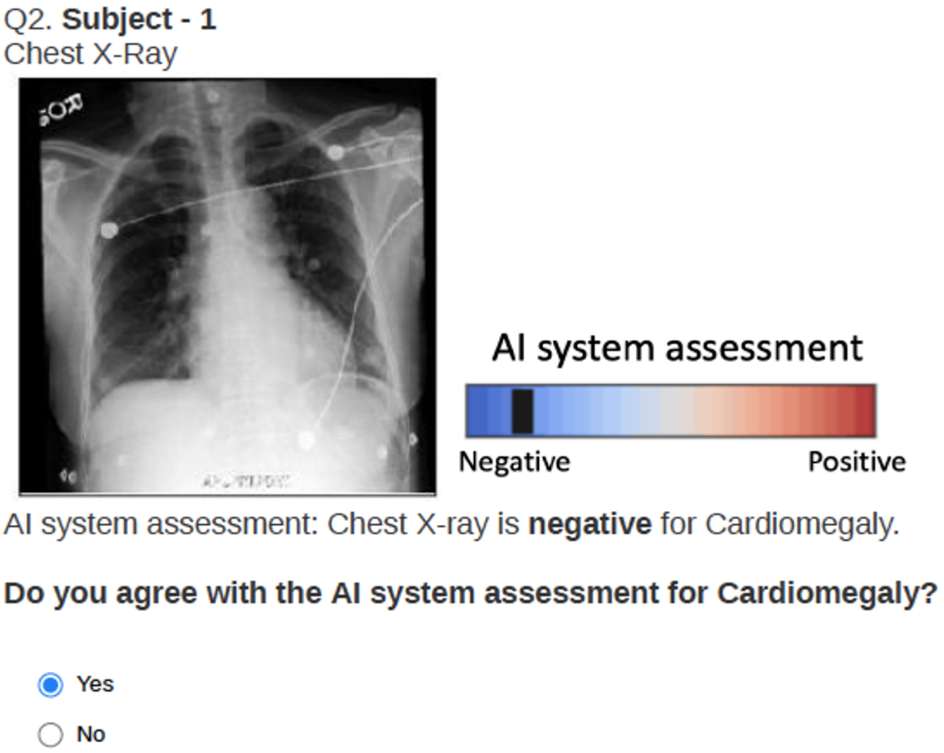
which counterfactual perturbation should be made to achieve the desired change in the classification decision.

4. The current definition of the concept in the concept-based explanations is restrictive. It only captures a few radiological terms and provides no intuition behind what part of the prediction decision is not explainable using the current definition of the concepts. Future work should explore adding the notion of completeness, which quantifies how sufficient a particular set of concepts is in explaining a model's prediction behavior based on the assumption that complete concept scores are sufficient statistics of the model predictions [292]. Researchers have already started exploring this direction. However, there is a limited progress in terms of identifying extensive concepts for medical tasks. Also, further research is required to quantify *completeness*.
5. Uncertainty quantification is a challenging task. Our current work focuses on improving uncertainty estimates from a pre-trained classifier. The training of PCE is a computationally expensive and time intensive overhead. The future work should explore removing this overhead and proposing an augmentation that can be quickly achieved. One direction would be to use a pre-trained generative model for getting augmented data.
6. The clinical concepts and their definitions are blurred and uncertain [189]. Another prospective direction is to explore for learning a DNN with improved uncertainty quantification is to train the DNN with soft labels instead of strict hard labels [190]. Some preliminary work in this direction have found that the clinical annotators prefer to label clinical data with soft labels which can come in form of probability estimates or discrete categories defining different degrees (strength) of labels [191, 265]. Further, the soft labels help to both reduce the sample size from which we can train the high quality models and also make these methods less sensitive to highly unbalanced data.

Appendix Progressive Counterfactual Explanation

A.1 Human evaluation

In our human evaluation study, we asked the following 15 questions for each CXR:



Q2. **Subject - 1**
Chest X-Ray

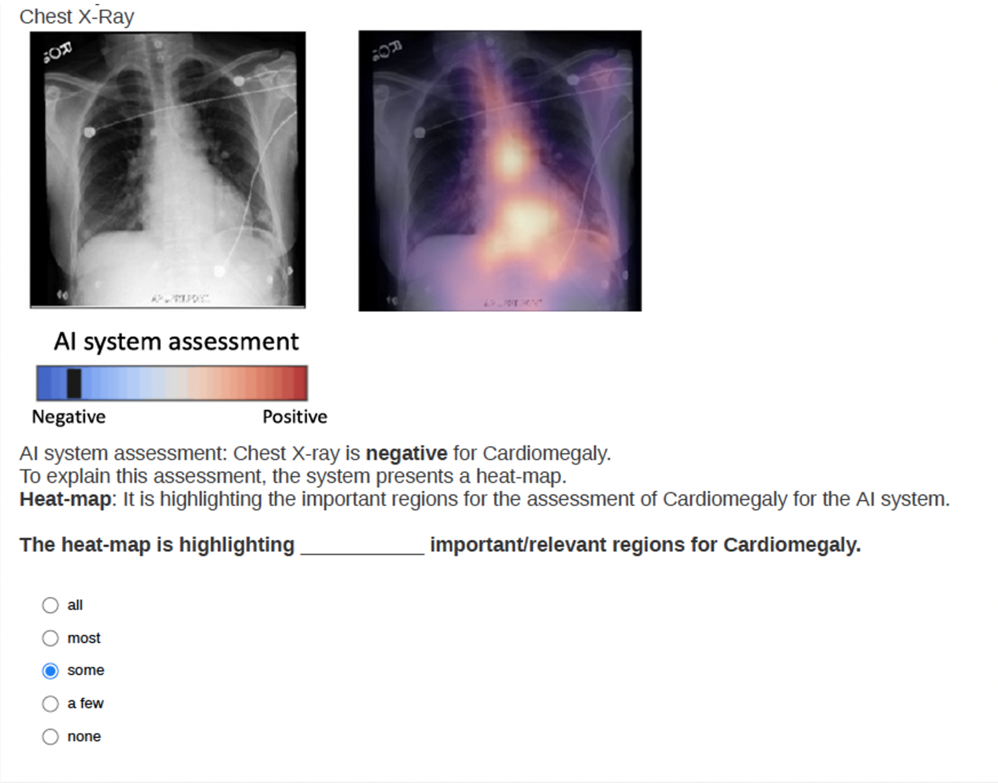
AI system assessment: Chest X-ray is **negative** for Cardiomegaly.

Do you agree with the AI system assessment for Cardiomegaly?

● Yes
○ No

Figure 36: Question 2-3 showing the query CXR and the classifier's decision.

1. Please provide your diagnosis for Cardiomegaly. Answers: Negative, mild, positive, not sure.
2. (Only assessment) Do you agree with the AI system assessment for Cardiomegaly? Answers: yes, no
3. (Only assessment) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.



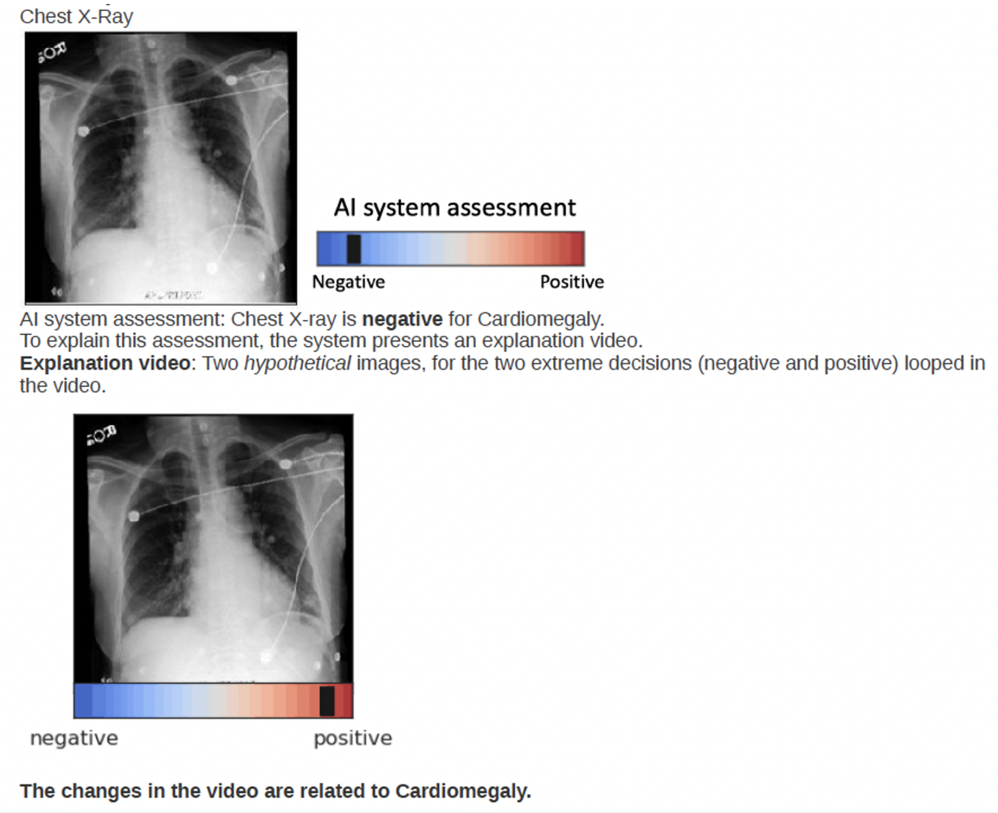
AI system assessment: Chest X-ray is **negative** for Cardiomegaly.
To explain this assessment, the system presents a heat-map.
Heat-map: It is highlighting the important regions for the assessment of Cardiomegaly for the AI system.

The heat-map is highlighting ____________ important/relevant regions for Cardiomegaly.

○ all
○ most
● some
○ a few
○ none

Figure 37: Question 4-5 showing the query CXR, the classifier's decision and the saliency map explanation.

4. (Assessment + SM) The heat-map is highlighting <blank> important/relevant regions for Cardiomegaly. Answers: all, most, some, a few, none.
5. (Assessment + SM) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.
6. (Assessment + cycleGAN) The changes in the video are related to Cardiomegaly. Answers: 5-point Likert scale.
7. (Assessment + cycleGAN) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.
8. (Assessment + cycleGAN) Images in the video look like a chest x-ray. Answers: 5-point Likert scale.
9. (Assessment + cycleGAN) The images in the video look like the chest x-ray from the



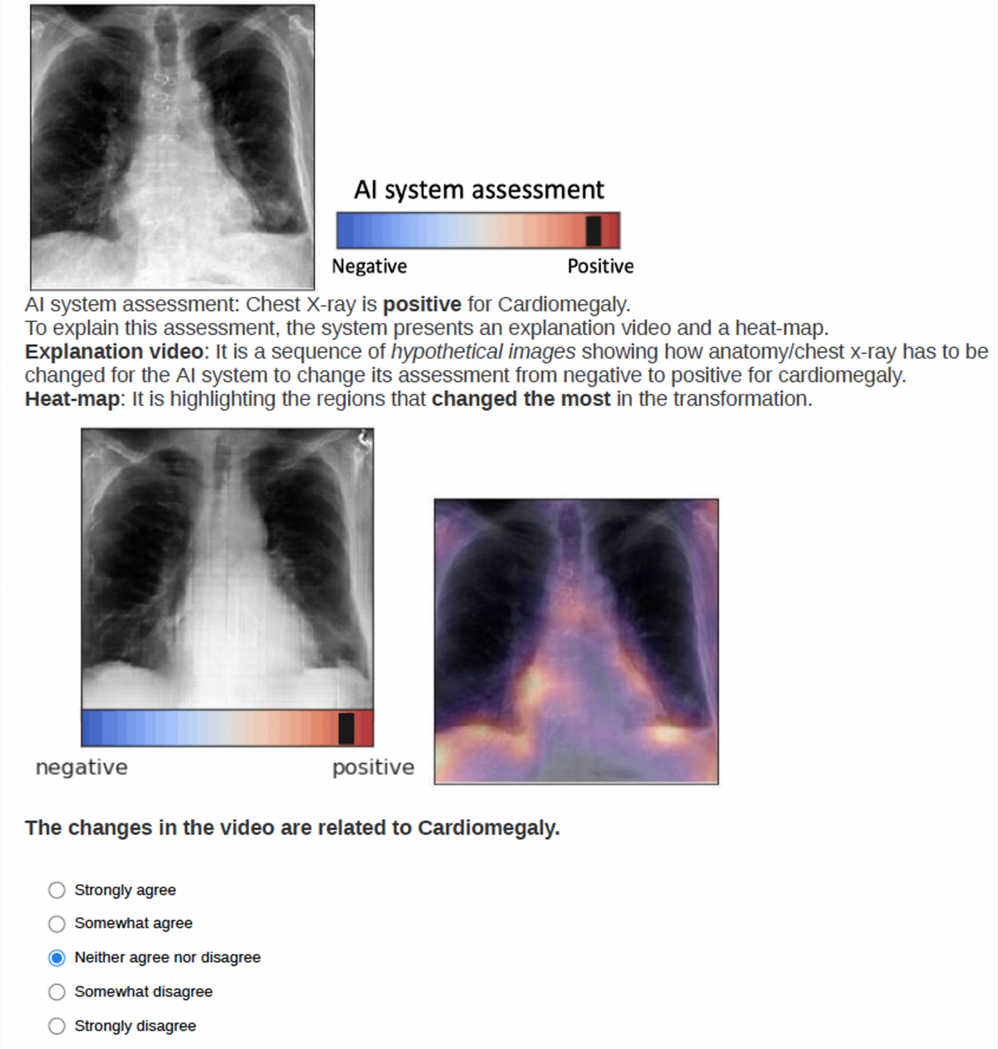
AI system assessment: Chest X-ray is **negative** for Cardiomegaly.
To explain this assessment, the system presents an explanation video.
Explanation video: Two *hypothetical* images, for the two extreme decisions (negative and positive) looped in the video.

The changes in the video are related to Cardiomegaly.

Figure 38: Question 6-9 showing the query CXR, the classifier's decision and the cycleGAN explanation.

subject. Answers: 5-point Likert scale.

10. (Assessment + ours) The changes in the video are related to Cardiomegaly. Answers: 5-point Likert scale.
11. (Assessment + ours) Changes in the anatomy in the highlighted regions in the heat-map will change the assessment of Cardiomegaly. Answers: 5-point Likert scale.
12. (Assessment + ours) I understand how the AI system made the above assessment for Cardiomegaly. Answers: 5-point Likert scale.
13. (Assessment + ours) Images in the video look like a chest x-ray. Answers: 5-point Likert scale.
14. (Assessment + ours) The images in the video look like the chest x-ray from the subject.



AI system assessment: Chest X-ray is **positive** for Cardiomegaly.
To explain this assessment, the system presents an explanation video and a heat-map.
Explanation video: It is a sequence of *hypothetical images* showing how anatomy/chest x-ray has to be changed for the AI system to change its assessment from negative to positive for cardiomegaly.
Heat-map: It is highlighting the regions that **changed the most** in the transformation.

The changes in the video are related to Cardiomegaly.

○ Strongly agree
○ Somewhat agree
● Neither agree nor disagree
○ Somewhat disagree
○ Strongly disagree

Figure 39: Question 10-14 showing the query CXR, the classifier's decision and our counterfactual explanation.

Answers: 5-point Likert scale.

15. Which explanation helped you the most in understanding the assessment made by the AI system Answers: Explanation-1: Heat-map highlighting important regions for assessment, Explanation-2: A video showing the transformation from negative to positive decision, Explanation-3: Two images at the two extreme ends of the decision (positive and negative), none.

A.2 Summarizing the notation

Table. 17 summarizes the notation used in the manuscript.

Table 17: Summarizing the notation.

Notation	Description
$\mathcal{X}$	Input image space
$\mathbf{x} \in \mathcal{X}$	Input image
$f : \mathcal{X} \rightarrow \mathcal{Y}$	Pre-trained classification function
$f(\mathbf{x})[k] \in [0, 1]$	Classifier's output for k^{th} class
$\mathbf{c}$	The condition used in cGAN, the desired classifier's output for the k^{th} class
$\mathbf{x_c}$	Explanation image
$f(\mathbf{x_c})$	Classifier's output for the explanation image
$\mathcal{I}_f(\mathbf{x}, \mathbf{c})$	Explanation function
$E(\cdot)$	Image encoder
$\mathbf{z}$	Latent representation of the input image
$C(\mathbf{c})$	Discretizing function that maps $\mathbf{c}$ to an integer
$G(\mathbf{z}, \mathbf{c})$	Generator of cGAN
$D(\mathbf{x}, \mathbf{c})$	Discriminator of cGAN
$p_{\text{data}}(\mathbf{x})$	Real image data distribution
$q(\mathbf{x})$	Learned data distribution by cGAN
$r(\mathbf{x})$	Loss term of cGAN that measures similarity between real and learned data distribution
$r(\mathbf{c}\|\mathbf{x})$	Loss term of cGAN that evaluates correspondence between generated images and condition
$\phi(\mathbf{x})$	Image feature extractor; part of the discriminator function

A.3 MIMC-CXR Dataset

We focus on explaining classification models based on deep convolution neural networks (CNN); most state-of-the-art performance models fall in this regime. We used large, publicly available datasets of chest x-ray (CXR) images, MIMIC-CXR [112]. MIMIC-CXR dataset is a multi-modal dataset consisting of 473K CXR, and 206K reports from 63K patients. We considered only frontal (posteroanterior PA or anteroposterior AP) view CXR. The datasets provide image-level labels for fourteen radio-graphic observations. These labels are extracted from the radiology reports associated with the x-ray exams using an automated tool called the Stanford CheXpert labeler [108]. The labeller first defines some thoracic observations using a radiology lexicon [86]. It extracts and classifies (positive, negative, or uncertain mentions) these observations by processing their context in the report. Finally, it aggregates these observations into fourteen labels for each x-ray exam. For the MIMIC-CXR dataset, we extracted the labels ourselves, as we have access to the reports.

A.4 Classification Model

To train the classifier, we considered the uncertain mention as a positive mention. We crop the original images to have the same height and width, then downsample them to 256 × 256 pixels. The intensities were normalized to have values between 0 and 1. Following the approach in prior work [208, 218, 108] on diagnosis classification, we used DenseNet-121 [98] architecture as the classification model. In DenseNet, each layer implements a non-linear transformation based on composite functions such as Batch Normalization (BN), rectified linear unit (ReLU), pooling, or convolution. The resulting feature map at each layer is used as input for all the subsequent layers, leading to a highly convoluted multi-level multi-layer non-linear convolutional neural network. We aim to explain such a model post-hoc without accessing the parameters learned by any layer or knowing the architectural details. Our proposed approach can be used for explaining any DL based neural network.

A.5 Progressive Counterfactual Explainer

The PCE function is a conditional GAN with an encoder. We used a ResNet [90] architecture for the Encoder, Generator, and Discriminator. The details of the architecture are given in Table 18. For the encoder network, we used five ResBlocks with the standard batch normalization layer (BN). In encoder-ResBlock, we performed down-sampling (average pool) before the first *conv* of the ResBlock as shown in Figure. 40.a. For the generator network, we follow the details in [171] and replace the BN layer in encoder-ResBlock with conditional BN (cBN) to encode the condition (*see* Figure. 40.b.). The architecture for the generator has five ResBlocks; each ResBlock performed up-sampling through the nearest neighbour interpolator. For the discriminator, we used spectral normalization (SN) [172] in Discriminator-ResBlock and performed down-sampling after the second *conv* of the ResBlock as shown in Figure. 40.c. For the optimization, we used Adam optimizer [126], with hyper-parameters set to $\alpha = 0.0002, \beta_1 = 0, \beta_2 = 0.9$ and updated the discriminator five times per one update of the generator and encoder.

For creating the training dataset, we divide the posterior distribution for the target class, $f(\mathbf{x}) \in [0, 1]$ into N equally-sized bins. For efficient training, cBN requires class-balanced batches. A large N results in more conditions for training cGAN, increasing cGAN complexity and training time. Also, we have to increase the batch size to ensure each condition is well represented in a batch. Hence, the GPU memory size bounds the upper value for N. A small value of N is equivalent to fewer bins, resulting in a coarse transformation which leads to abrupt changes across explanation images. In our experiments, we used $N = 10$, with a batch size of 32. We experimented with different values of N and selected the largest N, which created a class-balanced batch that fits in GPU memory and resulted in stable cGAN training.

Table 18: Explanation Model (cGAN) Architecture

(a) Encoder

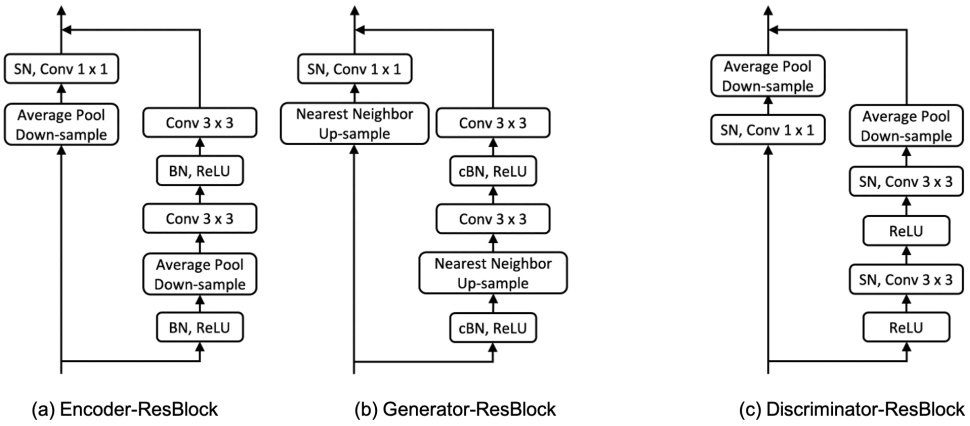
Grayscale image $\mathbf{x} \in \mathbb{R}^{256\times256\times1}$
BN, ReLU, 3×3 conv 64
Encoder-ResBlock down 128
Encoder-ResBlock down 256
Encoder-ResBlock down 512
Encoder-ResBlock down 1024
Encoder-ResBlock down 1024

(b) Generator	(c) Discriminator
Latent code $\mathbf{z} \in \mathbb{R}^{1024}$	Grayscale image $\mathbf{x} \in \mathbb{R}^{256\times256\times1}$
Generator-ResBlock up 1024, $\mathbf{c}$	Discriminator-ResBlock down 64
Generator-ResBlock up 512, $\mathbf{c}$	Discriminator-ResBlock down 128
Generator-ResBlock up 256, $\mathbf{c}$	Discriminator-ResBlock down 256
Generator-ResBlock up 128, $\mathbf{c}$	Discriminator-ResBlock down 512
Generator-ResBlock up 64, $\mathbf{c}$	Discriminator-ResBlock down 1024
BN, ReLU, 3×3 conv 1	ReLU, Global Sum Pooling (GSP) | Embed($\mathbf{c}$)
Tanh	Inner Product (GSP, Embed($\mathbf{c}$)) $\rightarrow \mathbb{R}^1$
	Add(SN-Dense(GSP) $\rightarrow \mathbb{R}^1$, Inner Product)

A.6 Semantic Segmentation

We adopted a 2D U-Net [216] to perform semantic segmentation, to mark the lung and the heart contour in a CXR. The network optimizes a multi-categorical cross-entropy loss function, defined as,

$$\mathcal{L}_\theta := \sum_s \sum_i \mathbb{1}(y_i = s) \log(p_\theta(x_i)), \tag{28}$$



(a) Encoder-ResBlock (b) Generator-ResBlock (c) Discriminator-ResBlock

Figure 40: Architecture of the ResBlocks used in all experiments.

where $\mathbb{1}$ is the indicator function, y_i is the ground truth label for i-th pixel. s is the segmentation label with values (background, the lung or the heart). $p_\theta(x_i)$ denotes the output probability for pixel x_i and θ are the learned parameters. The network is trained on 385 CXRs and corresponding masks from Japanese Society of Radiological Technology (JSRT) [268] and Montgomery [109] datasets.

A.7 Object Detection

We trained an object detector network to identify medical devices in a CXR. For the MIMIC-CXR dataset, we pre-processed the reports to extract keywords/observations that correspond to medical devices, including pacemakers, screws, and other hardware. Such foreign objects are easy to identify in a CXR and do not requires expert knowledge for manual labelling. Using the CheXpert labeller, we extracted 300 CXR images with positive mentions for each observation. The extracted x-rays are then manually annotated with bounding box annotations marking the presence of foreign objects using the LabelMe [275] annotation tool. Next, we trained an object detector based on Fast Region-based CNN [211], which used

VGG-16 model [238], trained on the MIMIC-CXR dataset as its foundation. We used this object detector to enforce our novel context-aware reconstruction loss (CARL).

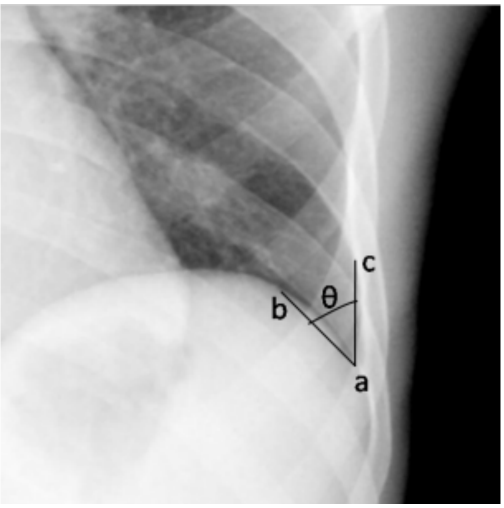


Figure 41: The costophrenic angle (CPA) on a CXR is marked as the angle formed by, (a) costophrenic angle point, (b) hemidiaphragm point and (c) lateral chest wall point, as shown by Maduskar *et al.*in [156].

We trained similar detectors for identifying normal and abnormal CP recess regions in a CXR. We associated an abnormal CP recess with the radiological finding of a blunt CP angle as identified by the positive mention for *"blunting of costophrenic angle"* in the corresponding radiology report. For the normal-CP recess, we considered images with a positive mention for *"lungs are clear"* in the reports. We extracted 300 CXR images with positive mention of respective terms for normal and abnormal CP recess to train the object detector.

Please note that the object detector for CP recess is only used for evaluation purposes, and they were not used during the training of the explanation function. In literature, the blunting of CPA is an indication of pleural effusion [155, 156]. The angle between the chest wall and the diaphragm arc is called the costophrenic angle (CPA). Marking the CPA angle on a CXR requires an expert to mark the three points, (a) costophrenic angle point, (b) hemidiaphragm point and (c) lateral chest wall point and then calculate the angle as shown in Figure. 41. Learning automatic marking of CPA angle requires expert annotation and is prone to error. Hence, rather than marking the CPA angle, we annotate the CP region with a bounding box which is a much simpler task. We then learned an object detector to identify normal or abnormal CP recess in a CXR and used the Score for detecting a normal CP recess (SCP) as our evaluation metric.

A.8 xGEM

We refer to work by Joshi *et al.* [114] for the implementation of xGEM. xGEM iteratively traverses the input image's latent space and optimizes the traversal to flip the classifier's decision to a different class. Specifically, it solves the following optimization

$$\tilde{\mathbf{x}} = \mathcal{G}_\theta(\arg\min_{\mathbf{z}\in\mathbb{R}^d} \mathcal{L}(\mathbf{x}, \mathcal{G}_\theta(\mathbf{z})) + \lambda\ell(f(\mathcal{G}_\theta(\mathbf{z})), y^{'})) \tag{29}$$

where the first terms is an ℓ_2 distance loss for comparing real and generated data. The second term ensures that the classification decision for the generated sample is in favour of class $y^{'}$ and $y^{'} \neq y$ is a class other than original decision. Unless explicitly imposed, the explanation image does not look realistic. The explanation image is generated from an updated latent feature, and the expressiveness of the generator limits its visual quality. xGEM adopted a variational autoencoder (VAE) as the generator. VAE uses a Gaussian likelihood (ℓ_2 reconstruction), an unrealistic assumption for image data. Hence, vanilla VAE is known to produce over-smoothed images [99]. The VAE used is available at https://github.com/LynnHo/VAE-Tensorflow. All settings and architectures were set to default values. The original code generates an image of dimension 64x64. We extended the given network to produce an image with dimensions 256×256.

A.9 cycleGAN

We refer to the work by Narayanaswamy *et al.* [185] and DeGrave *et al.* [44] for the implementation details of cycleGAN. The network architecture for cycleGAN is replicated from the GitHub repository https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix. For training cycleGAN, we consider two sets of images. The first set comprises 2000 images from the MIMIC-CXR dataset such that the classifier has a strong positive prediction for the presence of a target disease *i.e.*, $f(\mathbf{x}) > 0.9$, and the second set has the same number of images but with strong negative prediction *i.e.*, $f(\mathbf{x}) < 0.1$. We train one such model for each target disease.

Table 19: Results for six prediction tasks on CelebA dataset. FID (Fréchet Inception Distance) score measures the quality of the generated explanations. Lower FID is better. FVA (Face verification accuracy) measures percentage of the times the query image and generated explanation have same face identity as per model trained on VGGFace2. Higher LSC and FVA is better.

Prediction Task	**Data Consistency (FID)**	
	Negative ($f(\mathbf{x}), f(\mathbf{x_c}) < 0.2$)	Positive ($f(\mathbf{x}), f(\mathbf{x_c}) > 0.8$)
CelebA-Smiling	56.3	46.9
CelebA-Young	74.4	67.5
CelebA-No beard	72.3	79.2
CelebA-Heavy makeup	98.2	64.9
CelebA-Black hair	72.8	55.8
CelebA-Bangs	57.8	54.1

A.10 Extended results on the three desiderata of explanation function

Here, we provide results for four more prediction tasks on celebA dataset: no-beard or beard, heavy makeup or light makeup, black hair or not back hair, and bangs or no-bangs. Figure 45 shows the qualitative results, an extended version of results in Figure 12. In Table 19, we summarize the FID scores for each PCE trained to explain a specific classification task. To demonstrate classifier consistency, similar to Figure 13, we plotted the average response of the classifier *i.e.*, $f(\mathbf{x_c})$ for explanations in each bin against the expected outcome *i.e.*, $\mathbf{c}$ (*see* Figure. 42). The positive slope of the line-plot, parallel to $y = x$ line confirms that starting from images with low $f(\mathbf{x})$, our model creates fake images such that $f(\mathbf{x_c})$ is high and vice-versa. Further, in Figure. 43, we provide a qualitative comparison between counterfactual images generated by our method and from xGEM. The explanation images generated by xGEM are blurred and lacks the natural-looking appeal of a face or an x-ray image. Consistent with this observation, earlier in our results Table. 6, xGEM has a high FID score, validating that the xGEM explanation images are significantly different from the real images.

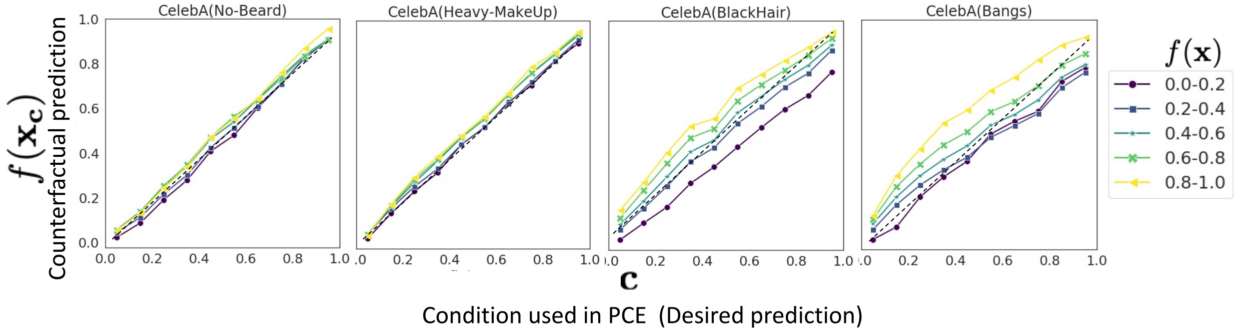


Figure 42: Plot of the expected outcome from the classifier, $\mathbf{c}$, against the actual response of the classifier on generated explanations, $f(\mathbf{x_c})$. The monotonically increasing trend shows a positive correlation between $\mathbf{c}$ and $f(\mathbf{x_c})$. Hence, the explanations are consistent with the given condition.

Next, in Figure 44, we provide similar results on CXR dataset. The bottom labels are the classifier's prediction for specific class. We also show the corresponding difference map, obtained by taking an absolute difference between explanations generated for the two extreme ends, negative (second column) and positive (fifth column) diagnosis. For cardiomegaly the counterfactual image obtained by cycleGAN failed to flip the classification decision. Further, in Figure. 46 we provide the classifier consistency results. For cycleGAN, starting with input images with $f(\mathbf{x}) \in [0.0, 0.2]$ (purple line), when we create explanation images with the desired predictions *i.e.*, x-axis with $\mathbf{c} \in [0.8, 1.0]$, the resulting images ($\mathbf{x_c}$) doesn't satisfy $f(\mathbf{x_c}) \in [0.8, 1.0]$. This shows that on-an-average the CycleGAN counterfactual images doesn't flip the classification decision. This finding is consistent with the low counterfactual validity score in Table. 6.

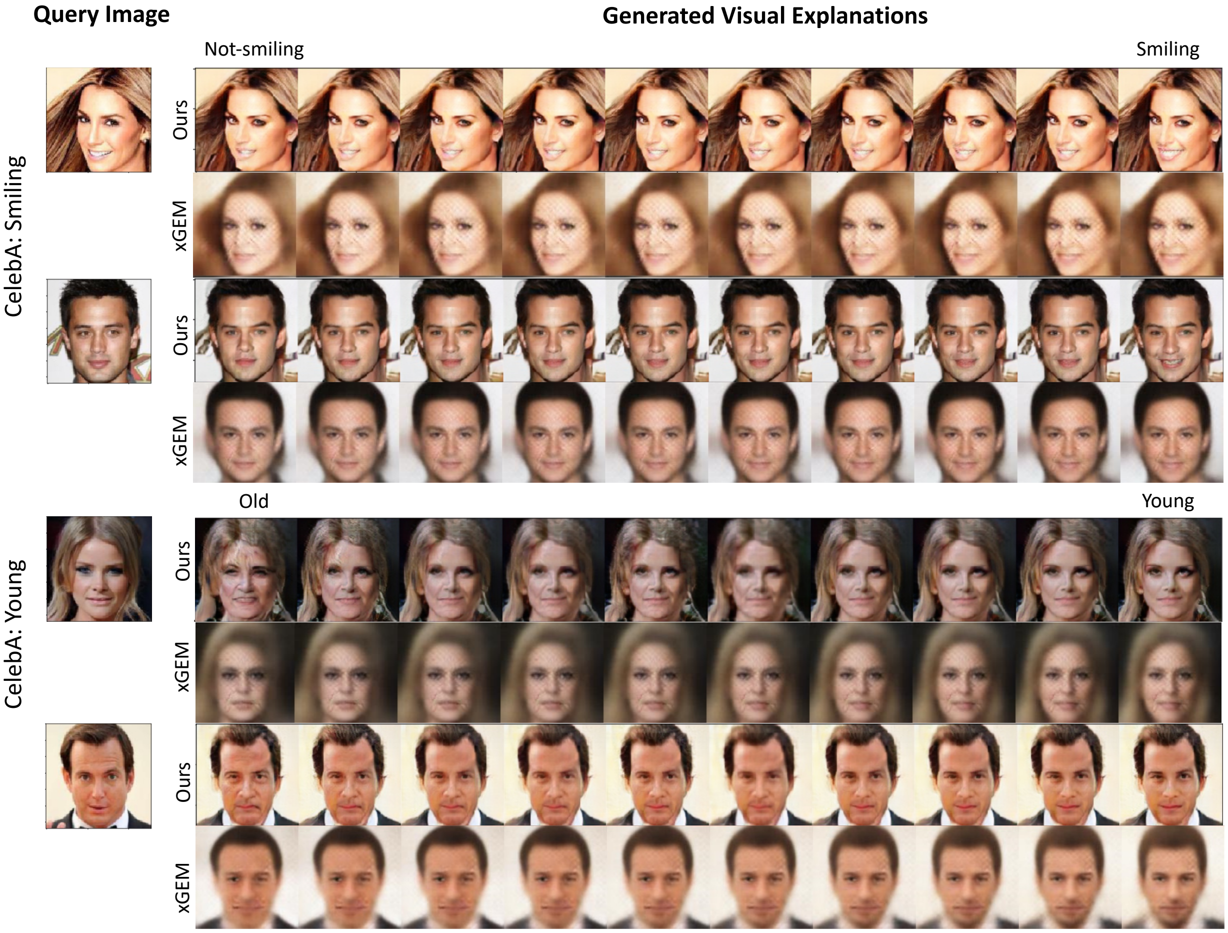


Figure 43: Visual explanations generated for "smiling" and "young" attribute classification on CelebA dataset.

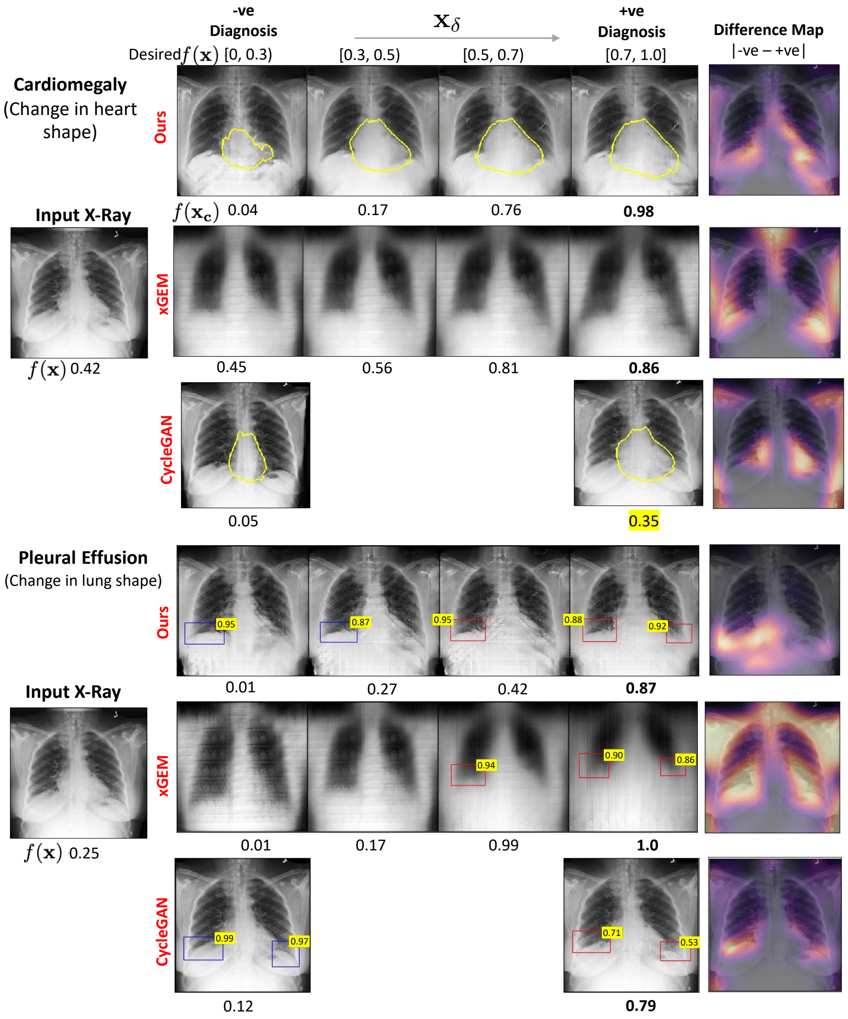


Figure 44: The transformation of an input chest x-ray into the counterfactual explanations for two diagnosis, cardiomegaly (first row) and pleural effusion (PE) (last row). The bottom labels are the classifier's prediction for the specific class. The yellow color highlight the prediction where counterfactual fails to flip the decision. The last column shows the difference map between negative and positive explanation. For cardiomegaly, we are highlight the heart segmentation (yellow). For PE, we show the bounding-box (BB) for normal (blue) and abnormal (red) costophrenic (CP) recess. The number on blue-BB is the Score for detecting a normal CP recess (SCP). The number on red-BB is 1-SCP.

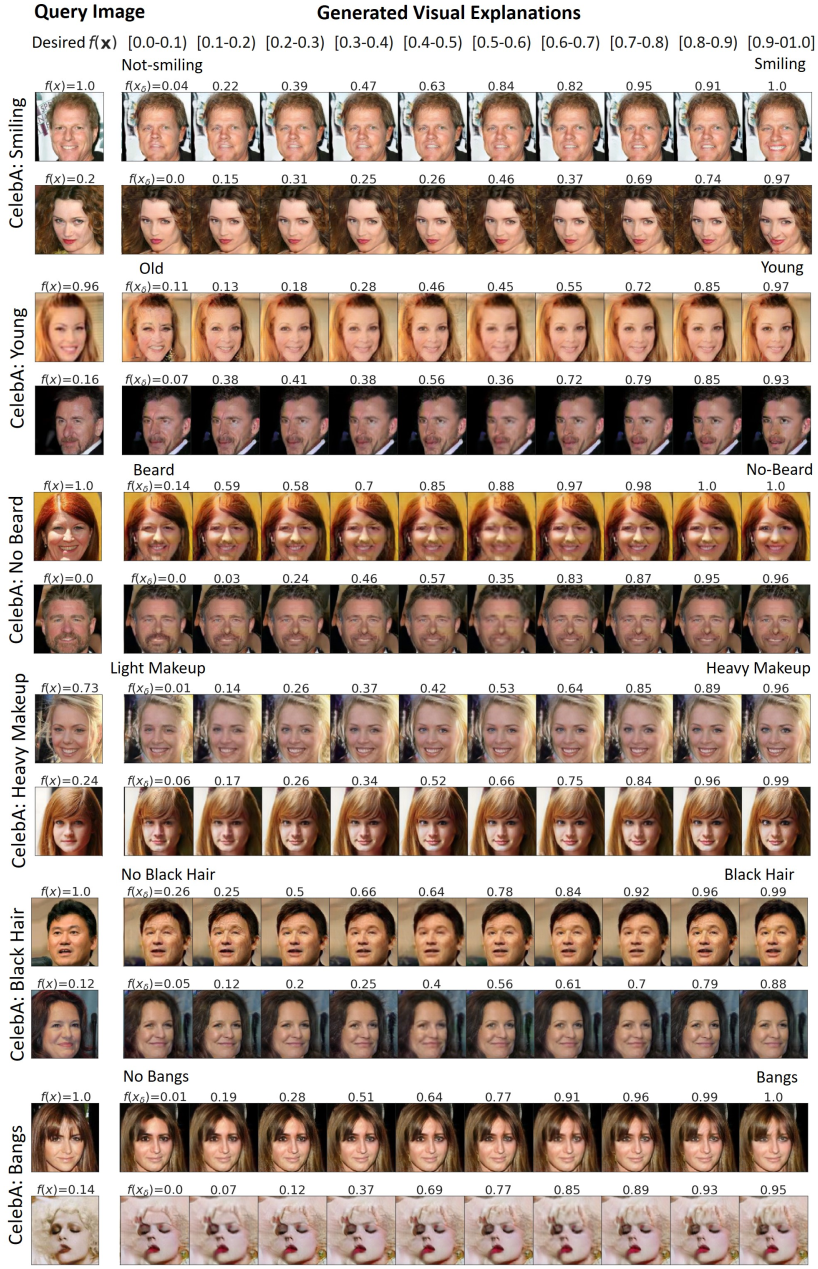

Query Image
Generated Visual Explanations
Desired f(x) [0.0-0.1) [0.1-0.2) [0.2-0.3) [0.3-0.4) [0.4-0.5) [0.5-0.6) [0.6-0.7) [0.7-0.8) [0.8-0.9) [0.9-01.0]
CelebA: Smiling
Not-smiling
Smiling
f(x)=1.0 f(xδ)=0.04 0.22 0.39 0.47 0.63 0.84 0.82 0.95 0.91 1.0
f(x)=0.2 f(xδ)=0.0 0.15 0.31 0.25 0.26 0.46 0.37 0.69 0.74 0.97
CelebA: Young
Old
Young
f(x)=0.96 f(xδ)=0.11 0.13 0.18 0.28 0.46 0.45 0.55 0.72 0.85 0.97
f(x)=0.16 f(xδ)=0.07 0.38 0.41 0.38 0.56 0.36 0.72 0.79 0.85 0.93
CelebA: No Beard
Beard
No-Beard
f(x)=1.0 f(xδ)=0.14 0.59 0.58 0.7 0.85 0.88 0.97 0.98 1.0 1.0
f(x)=0.0 f(xδ)=0.0 0.03 0.24 0.46 0.57 0.35 0.83 0.87 0.95 0.96
CelebA: Heavy Makeup
Light Makeup
Heavy Makeup
f(x)=0.73 f(xδ)=0.01 0.14 0.26 0.37 0.42 0.53 0.64 0.85 0.89 0.96
f(x)=0.24 f(xδ)=0.06 0.17 0.26 0.34 0.52 0.66 0.75 0.84 0.96 0.99
CelebA: Black Hair
No Black Hair
Black Hair
f(x)=1.0 f(xδ)=0.26 0.25 0.5 0.66 0.64 0.78 0.84 0.92 0.96 0.99
f(x)=0.12 f(xδ)=0.05 0.12 0.2 0.25 0.4 0.56 0.61 0.7 0.79 0.88
CelebA: Bangs
No Bangs
Bangs
f(x)=1.0 f(xδ)=0.01 0.19 0.28 0.51 0.64 0.77 0.91 0.96 0.99 1.0
f(x)=0.14 f(xδ)=0.0 0.07 0.12 0.37 0.69 0.77 0.85 0.89 0.93 0.95

Figure 45: Visual explanations generated for different prediction tasks on CelebA dataset.

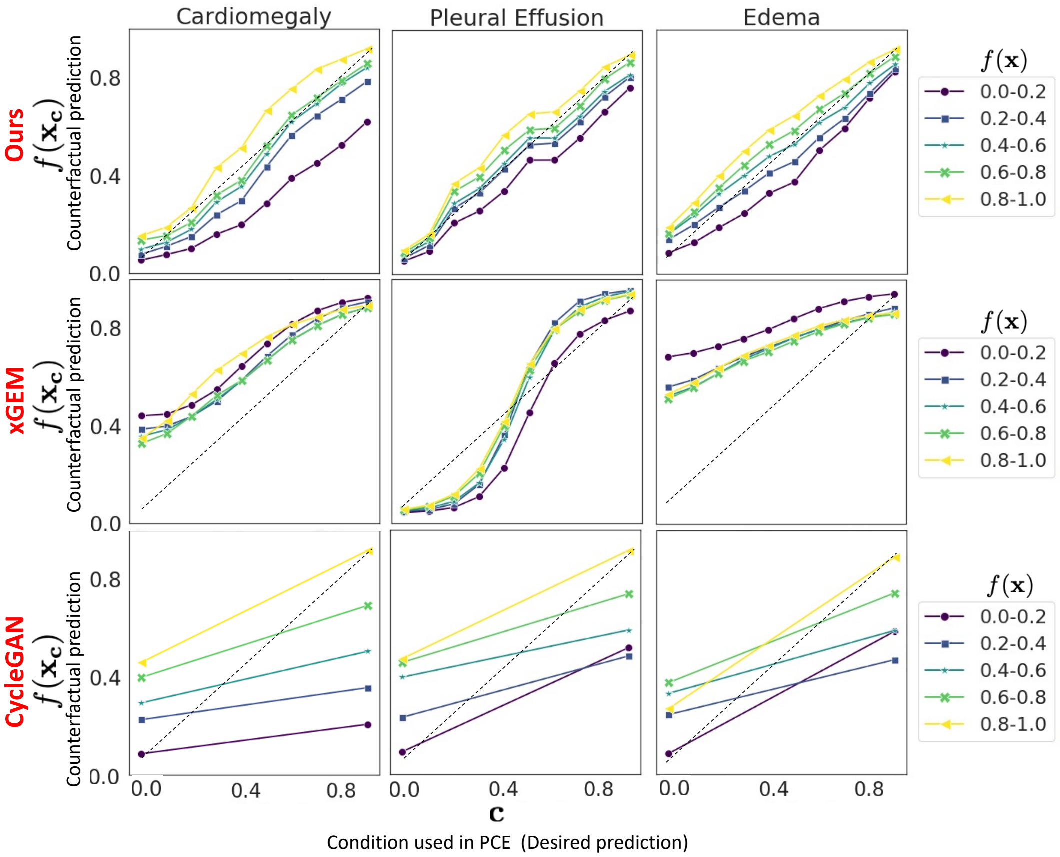


Figure 46: The plot of desired prediction, $\mathbf{c}$, against actual response of the classifier on generated explanations, $f(\mathbf{x_c})$. Each line represents a set of input images with classification prediction $f(\mathbf{x})$ in a given range. Dashed line represents $y = x$ line. A good explanation should cover the entire range of y-axis $[0, 1]$ for all set of images (lines of different colors).

A.11 Extended results on clinical evaluation

For quantitative analysis, we randomly sample two groups of real images (1) a *real-normal* group defined as $\mathcal{X}^n = \{\mathbf{x}; f(\mathbf{x}) < 0.2\}$. It consists of real CXR images that are predicted as normal by the classifier f. (2) A *real-abnormal* group defined as $\mathcal{X}^a = \{\mathbf{x}; f(\mathbf{x}) > 0.8\}$. For $\mathcal{X}^n$ we generated a counterfactual group as, $\mathcal{X}^a_{cf} = \{\mathbf{x} \in \mathcal{X}^n; f(\mathcal{I}_f(\mathbf{x}, \mathbf{c})) > 0.8\}$. Similarly for $\mathcal{X}^a$, we derived a counterfactual group as $\mathcal{X}^n_{cf} = \{\mathbf{x} \in \mathcal{X}^a; f(\mathcal{I}_f(\mathbf{x}, \mathbf{c})) < 0.2\}$.

Next, we quantify the differences in real and counterfactual groups by performing statistical tests on the distribution of clinical metrics such as cardiothoracic ratio (CTR) and the Score of normal Costophrenic recess (SCP). Specifically, we performed the dependent t-test statistics on clinical metrics for paired samples ($\mathcal{X}^n$ and $\mathcal{X}^a_{cf}$), ($\mathcal{X}^a$ and $\mathcal{X}^n_{cf}$) and the independent two-sample t-test statistics for normal ($\mathcal{X}^n$, $\mathcal{X}^n_{cf}$) and abnormal ($\mathcal{X}^a$, $\mathcal{X}^a_{cf}$) groups. The two-sample t-tests are statistical tests used to compare the means of two populations. A low p-value < 0.0001 rejects the null hypothesis and supports the alternate hypothesis that the difference in the two groups is statistically significant and that this difference is unlikely to be caused by sampling error or by chance. For paired t-test, the mean difference corresponds to the average causal effect of the intervention on the variable under examination. In our setting, intervention is a *do* operator on input image ($\mathbf{x}$), before intervention, resulting in a counterfactual image ($\mathbf{x_c}$), after intervention.

Table 20: Results of independent t-test. We compared the difference distribution of cardiothoracic ratio (CTR) for cardiomegaly and the Score for normal Costophrenic recess (SCP) for pleural effusion. CI: confidence interval; CF: counterfactual.

Target			**Paired Differences**						
Disease	Real	CF	Mean		95% CI				
	Group	Group	Difference	Std	Lower	Upper	t	df	p-value
Cardiomegaly	$\mathcal{X}^n$	$\mathcal{X}^a_{cf}$	**-0.03**	0.07	-0.03	-0.01	-4.4	304	< 0.0001
(CTR)	$\mathcal{X}^a$	$\mathcal{X}^n_{cf}$	**0.14**	0.12	0.13	0.15	24.7	513	$\ll 0.0001$
Pleural effusion	$\mathcal{X}^n$	$\mathcal{X}^a_{cf}$	**0.13**	0.22	0.06	0.13	5.9	217	$\ll 0.0001$
(SCP)	$\mathcal{X}^a$	$\mathcal{X}^n_{cf}$	**-0.19**	0.27	-0.18	-0.09	-6.7	216	$\ll 0.0001$
			Un-Paired Differences						
			Mean Real	Mean CF	95% CI				
			Group	Group	Lower	Upper	t	df	p-value
Cardiomegaly	$\mathcal{X}^n$	$\mathcal{X}^n_{cf}$	**0.46**	0.42	0.02	0.06	5.2	817	< 0.0001
(CTR)	$\mathcal{X}^a$	$\mathcal{X}^a_{cf}$	**0.56**	0.50	0.04	0.07	9.9	817	$\ll 0.0001$
Pleural effusion	$\mathcal{X}^n$	$\mathcal{X}^n_{cf}$	**0.69**	0.61	0.18	0.27	9.3	433	$\ll 0.0001$
(SCP)	$\mathcal{X}^a$	$\mathcal{X}^a_{cf}$	0.42	**0.56**	-0.32	-0.21	-9.7	433	$\ll 0.0001$

Table 20 provides the extended results for the Fig. 17. Patients with cardiomegaly have higher CTR as compared to normal subjects. Hence, one should expect CTR($\mathcal{X}^n$) $<$ CTR($\mathcal{X}^a_{cf}$) and likewise CTR($\mathcal{X}^a$) $>$ CTR($\mathcal{X}^n_{cf}$). Consistent with clinical knowledge, in Table. 20, we observe a negative mean difference of -0.03 for CTR($\mathcal{X}^n$) $-$ CTR($\mathcal{X}^a_{cf}$) (a p-value of < 0.0001) and a positive mean difference of 0.14 for $\text{CTR}(\mathcal{X}^a) - \text{CTR}(\mathcal{X}^n_{cf})$ (with a p-value of $\ll 0.0001$). On a population-level CTR was successful in capturing the difference between normal and abnormal CXRs. Specifically in un-paired differences, we observe a low mean CTR values for normal subjects *i.e.*, mean $\text{CTR}(\mathcal{X}^n) = 0.46$ as compared to mean CTR for abnormal patients *i.e.*, mean $\text{CTR}(\mathcal{X}^a) = 0.56$. The low p-values supports the alternate hypothesis that the difference in the two groups is statistically significant.

Further, in Fig 47.A, we show samples from input images that were predicted as negative for cardiomegaly ($\mathcal{X}^n$). In their counterfactual abnormal images (third column), we observe small changes in CTR are sufficient to flip the classification decision. This is consistent with a small mean difference $\text{CTR}(\mathcal{X}^n)$ - $\text{CTR}(\mathcal{X}^a_{cf}) = -0.03$. In contrast, when we generate counterfactual normal (sixth column) from real abnormal images (positive for cardiomegaly, Fig 47.B), significant changes in CTR lead to flipping of the prediction decision. This observation is consistent with a large mean difference $\text{CTR}(\mathcal{X}^a)$ - $\text{CTR}(\mathcal{X}^n_{cf}) = 0.14$.

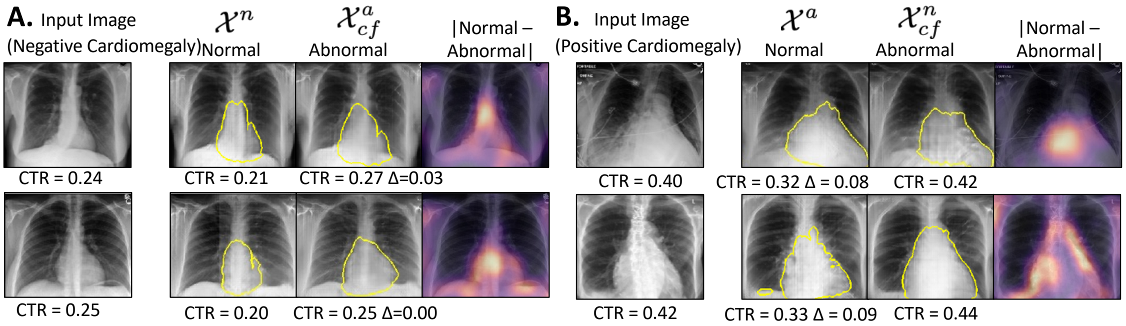


Figure 47: Extended results for explanation produced by our model for **Cardiomegaly**. For each image, we generate a normal and an abnormal explanation image. We show pixel-wise difference of the two generated images as the saliency map. In column A.(B.), we show input images negatively (positively) classified for Cardiomegaly. The yellow contour shows the heart boundary learned by a segmentation network. CTR is the cardiothoracic ratio. For column A, we observe a relatively minor change in CTR (Δ) between real and counterfactual images than in column B. .

By design, the object detector assigns a low SCP to any indication of blunting CPA or abnormal CP recess. Hence, $\text{SCP}(\mathcal{X}^n) > \text{SCP}(\mathcal{X}^a_{cf})$ and likewise $\text{SCP}(\mathcal{X}^a) < \text{SCP}(\mathcal{X}^n_{cf})$.

Consistent with our expectation, in Table. 20, we observe a positive mean difference of 0.13 for $\text{SCP}(\mathcal{X}^n) - \text{SCP}(\mathcal{X}^a_{cf})$ (with a p-value of $\ll 0.0001$) and a negative mean difference of -0.19 for $\text{SCP}(\mathcal{X}^a) - \text{SCP}(\mathcal{X}^n_{cf})$ (with a p-value of $\ll 0.0001$). On a population-level SCP was successful in capturing the difference between normal and abnormal CXR for pleural effusion. Specifically in un-paired differences, we observe a high mean SCP values for normal subjects *i.e.*, mean $\text{SCP}(\mathcal{X}^n) = 0.69$ as compared to mean SCP for abnormal patients *i.e.*, mean $\text{SCP}(\mathcal{X}^a) = 0.42$.

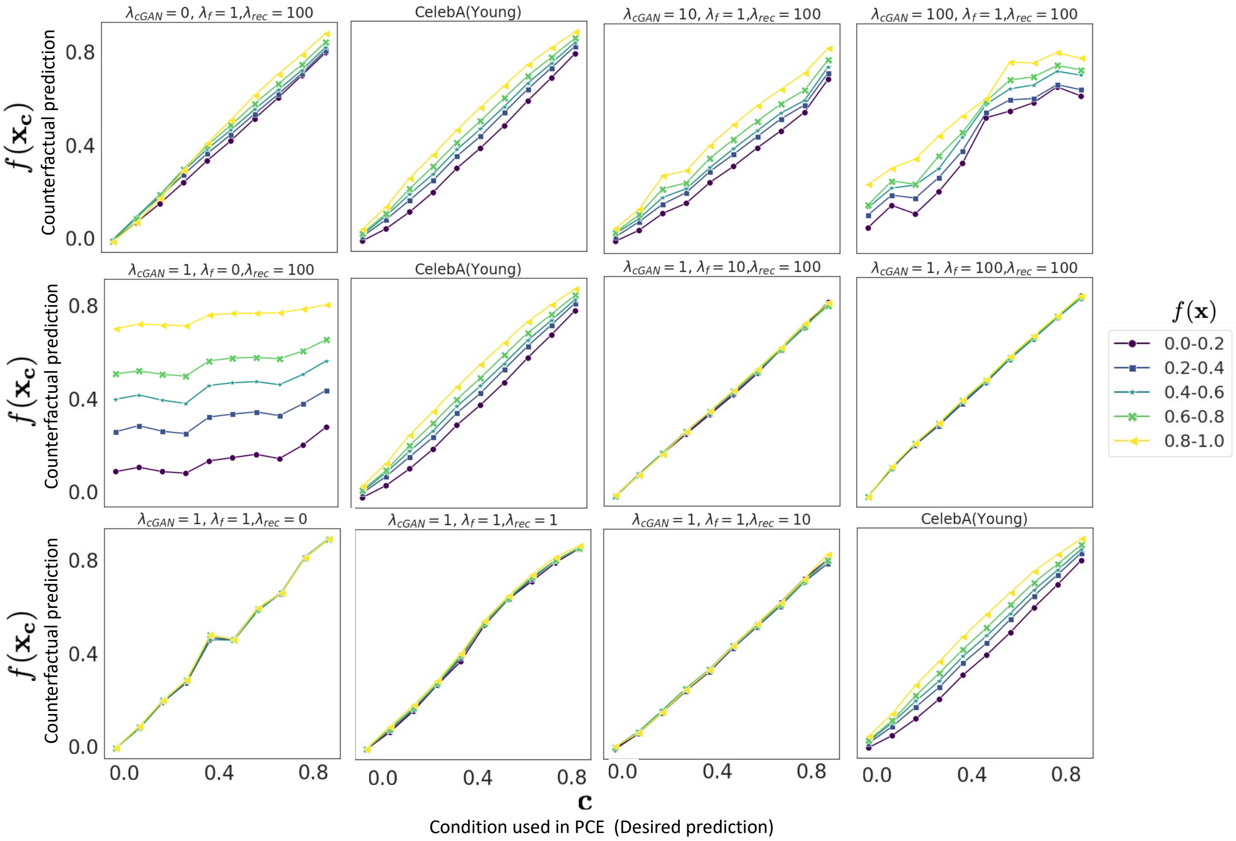


Figure 48: Ablation study to show the effect of KL loss term. Plot of the expected outcome from the classifier, $\mathbf{c}$, against the actual response of the classifier on generated explanations, $f(\mathbf{x_c})$.

Table 21: Our model with ablation on prediction task of young vs old on CelebA dataset. FID (Fréchet Inception Distance) score measures the quality of the generated explanations. Lower FID is better. FVA (Face verification accuracy) measures percentage of the times the query image and generated explanation have same face identity as per model trained on VGGFace2.

Configuration			**Data Consistency (FID)**			**Self Consistency**
λ_{cGAN}	λ_f	λ_{rec}	Present	Absent	Overall	FVA
0	1	100	69.7	105.7	67.2	99.8
1	1	100	**67.5**	**74.4**	**53.4**	72.2
10	1	100	89.4	105.2	63.0	82.7
100	1	100	71.6	80.6	44.26	18.0
1	0	100	66.2	66.2	44.9	99.4
1	1	100	67.5	74.4	53.4	72.2
1	10	100	95.5	90.4	62.4	96.8
1	100	100	77.4	73.1	71.2	42.23
1	1	0	116.2	118.9	72.2	0.0
1	1	1	63.0	78.6	61.6	5.5
1	1	10	87.6	83.6	65.7	**88.8**
1	1	100	67.5	74.4	53.4	72.2

A.12 Ablation Study

Our proposed model has three types of loss functions: adversarial loss from cGAN $\mathcal{L}_{\text{cGAN}}(D, G)$, KL loss $\mathcal{L}_f(D, G)$, and CARL reconstruction loss $\mathcal{L}_{\text{rec}}(E, G)$. The three losses enforce the three properties of our proposed explainer function: data consistency, compatibility with f, and self-consistency, respectively. In the ablation study, we quantify the importance of each of these components by training different models, which differ in one hyper-parameter. For **data consistency**, we evaluate Fréchet Inception Distance (FID). FID score measures the visual quality of the generated explanations by comparing them with the real images. We show results for two groups. In the first group, we consider real and fake images where the classifier has high confidence in *presence* of the target label y *i.e.*, $f(\mathbf{x_c})[y], f(\mathbf{x})[y] \in [0.8, 1.0]$. In second group, the target label y is *absent i.e.*, $f(\mathbf{x_c})[y], f(\mathbf{x})[y] \in [0.0, 0.2)$. For **compatibility with f**, we plotted the desired output of the classifier *i.e.*, $\mathbf{c}$ against the actual output of the classifier $f(\mathbf{x_c})$ for the generated explanations. For **self consistency**, we calculated the Face verification accuracy (FVA) for celebA dataset and the foreign object preservation (FOP) score for CXR dataset. FVA measures the percentage of the instances in which the query image and generated explanation have the same face identity as per the model trained on VGGFace2. FOP score is the fraction of real images, with successful detection of FO, in which FO was also detected in the corresponding explanation image $\mathbf{x_c}$.

For celebA, we consider the prediction task of young vs old. Figure 48 shows the results for compatibility with f. Table 21 summarizes the results for data consistency and self-consistency. For MIMIC-CXR, Table 22 summarizes our results. In the absence of adversarial loss from cGAN ($\lambda_{cGAN} = 0$), FID score is very high as the generated images looks very different from the real images. On removing the KL loss for classifier consistency ($\lambda_f = 0$), the CV score is poor as the generated explanations are derived without considering the classification function and hence they failed to flip the classification decision. In the absence of reconstruction loss ($\lambda_{rec} = 0$), the generated explanations are no longer for the same person as in query image. This results in a low FVA score. In CXR dataset, FO in query CXR are absent in generated explanations, resulting in low FOP score.

A.13 Ablation study over pacemaker

We performed an ablation study to investigate if a pacemaker is influencing the classifier's prediction for cardiomegaly. We consider 300 subjects that are positively predicted for cardiomegaly and have a pacemaker. We used our pre-trained object detector to find the

Table 22: Evaluation metrics for ablation study. FID score quantifies the visual appearance of the explanations. CV score is the fraction of explanations that have an opposite prediction compared to the input image. FOP score is the fraction of real images with FO, in which FO was also detected in the corresponding explanation image. In configuration with $\lambda_{cGAN} = 0$ there is no adversarial loss from cGAN, in $\lambda_f = 0$ there is no KL-loss for classifier consistency and in $\lambda_{rec} = 0$ there is no context-aware self reconstruction loss.

	Cardiomegaly				**Pleural Effusion**			
	Baseline	λ_{cGAN}=0	λ_f=0	λ_{rec}=0	Baseline	λ_{CGAN}=0	λ_f=0	λ_{rec}=0
FID score								
Normal	166	200	174	160	146	210	150	149
Abnormal	137	189	138	140	122	178	120	130
Counterfactual Validity (CV) Score								
Overall	0.91	0.89	0.43	0.92	0.97	0.93	0.43	0.97
Foreign Object Preservation (FOP) score								
Pacemaker	0.52	0.2	0.55	0.19				

bounding-box annotations for these images. Using the bounding-box, we created a perturbation of the input image by masking the pacemaker and in-filling the masked region with the surrounding context.

An example of the perturbation image is shown in Fig. 49. We passed the perturbed image through the classifier and calculated the difference in the classifier's prediction before and after removing the pacemaker. The average change in prediction was negligible (0.03). Hence, pacemaker is not influencing classification decisions for cardiomegaly.

We performed an ablation study to investigate if a pacemaker is influencing the classifier's prediction for cardiomegaly. We consider 300 subjects that are positively predicted for cardiomegaly and have a pacemaker. We used our pre-trained object detector to find the bounding-box annotations for these images. Using the bounding-box, we created a pertur-

bation of the input image by masking the pacemaker and in-filling the masked region with the surrounding context.

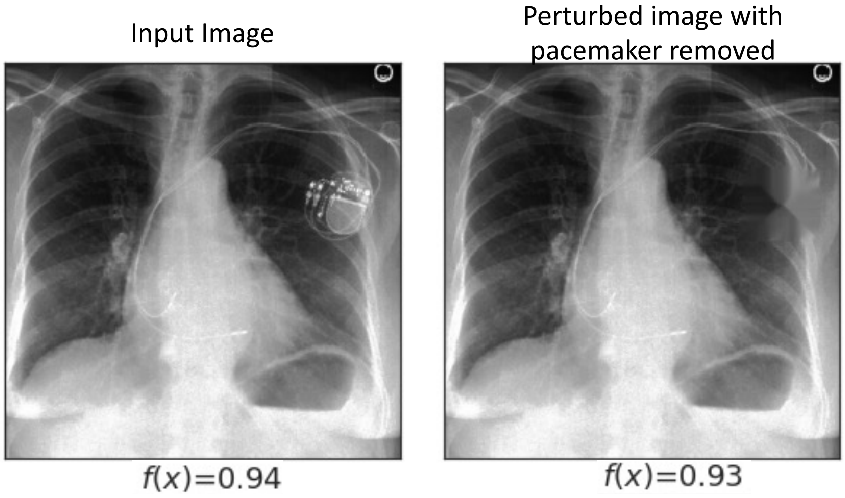


Figure 49: An example of input image before and after removing the pacemaker.

Bibliography

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4432–4441, 2019.

[2] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion*, 76:243–297, 2021.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[4] Parnian Afshar, Anastasia Oikonomou, Farnoosh Naderkhani, Pascal Tyrrell, Konstantinos Plataniotis, Keyvan Farahani, and Arash Mohammadi. 3d-mcn: A 3d multiscale capsule network for lung nodule malignancy prediction. *Scientific Reports*, 10, 05 2020.

[5] Chirag Agarwal and Anh Nguyen. Explaining an image classifier's decisions using generative models. *arXiv preprint arXiv:1910.04256*, 10 2019.

[6] Marios Anthimopoulos, Stergios Christodoulidis, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE transactions on medical imaging*, 35(5):1207–1216, 2016.

[7] Samuel Y. Ash, Rola Harmouche, Diego Lassala Lopez Vallejo, Julian A. Villalba, Kris Ostridge, River Gunville, Carolyn E. Come, Jorge Onieva Onieva, James C. Ross, Gary M. Hunninghake, Souheil Y. El-Chemaly, Tracy J. Doyle, Pietro Nardelli, Gonzalo V. Sanchez-Ferrero, Hilary J. Goldberg, Ivan O. Rosas, Raul San Jose Estepar, and George R. Washko. Densitometric and local histogram based analysis of computed tomography images in patients with idiopathic pulmonary fibrosis. *Respiratory Research*, 18(1), 3 2017.

[8] Uri Avni, Hayit Greenspan, Eli Konen, Michal Sharon, and Jacob Goldberger. X-ray categorization and retrieval on the organ and pathology level, using patch-based visual words. *IEEE Transactions on Medical Imaging*, 30(3):733–746, 2010.

[9] Ahmad Taher Azar and Shereen M. El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8):2387–2403, 12 2013.

[10] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Muller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[11] Sanhita Basu, Sushmita Mitra, and Nilanjan Saha. Deep learning for screening covid-19 using chest x-ray images. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 2521–2527. IEEE, 2020.

[12] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

[13] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

[14] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. Seeing what a gan cannot generate. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4502–4511, 2019.

[15] David Bermejo-Peláez, Samuel Ash, George Washko, Raúl Estépar, and María Ledesma-Carbayo. Classification of interstitial lung abnormality patterns with an ensemble of deep convolutional neural networks. *Scientific Reports*, 10:338, 01 2020.

[16] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. *32nd International Conference on International Conference on Machine Learning*, page 1613–1622, 2015.

[17] Diane Bouchacourt and Ludovic Denoyer. Educe: Explaining model decisions through unsupervised concepts extraction. *arXiv preprint arXiv:1905.11852*, 2019.

[18] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2019.

[19] Q Cao, L Shen, W Xie, O M Parkhi, and A Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018.

[20] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

[21] Bartolome R. Celli, Claudia G. Cote, Jose M. Marin, Ciro Casanova, Maria Montes de Oca, Reina A. Mendez, Victor Pinto Plata, and Howard J. Cabral. The Body-Mass Index, Airflow Obstruction, Dyspnea, and Exercise Capacity Index in Chronic Obstructive Pulmonary Disease. *New England Journal of Medicine*, 350(10):1005–1012, 3 2004.

[22] O A Centurión, K E Scavenius, L M Miño, and O R , Sequeira. Evaluating Cardiomegaly by Radiological Cardiothoracic Ratio as Compared to Conventional Echocardiography. *Journal of Cardiology & Current Research*, 9(2), 6 2017.

[23] Isarun Chamveha, Treethep Promwiset, Trongtum Tongdee, Pairash Saiviroonporn, and Warasinee Chaisangmongkon. Automated Cardiothoracic Ratio Calculation and Cardiomegaly Detection using Deep Learning Approach. *arXiv preprint arXiv:2002.07468*, 2 2020.

[24] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining Image Classifiers by Counterfactual Generation, 2019.

[25] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: Deep learning for interpretable image recognition. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc.

[26] Hugh Chen, Scott Lundberg, and Su-In Lee. Explaining models by propagating shapley values of local components. In *Explainable AI in Healthcare and Medicine*, pages 261–270. Springer, 2021.

[27] Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Atom: Robustifying out-of-distribution detection using outlier mining. *In Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)*, 2021.

[28] Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

[29] Veronika Cheplygina, Isabel Pino Peña, Jesper Holst Pedersen, David A. Lynch, Lauge Sørensen, and Marleen de Bruijne. Transfer learning for multi-center classification of chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics*, 1 2017.

[30] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.

[31] Andreas Christe, Alan Peters, Dionysios Drakopoulos, Johannes Heverhagen, Thomas Geiser, Thomai Stathopoulou, Stergios Christodoulidis, Marios Anthimopoulos, Stavroula Mougiakakou, and Lukas Ebner. Computer-aided diagnosis of pulmonary fibrosis using deep learning and ct images. *Investigative Radiology*, 54:1, 05 2019.

[32] Francesco Ciompi, Kaman Chung, Sarah J Van Riel, Arnaud Arindra Adiyoso Setio, Paul K Gerke, Colin Jacobs, Ernst Th Scholten, Cornelia Schaefer-Prokop, Mathilde MW Wille, Alfonso Marchiano, et al. Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Scientific reports*, 7(1):1–11, 2017.

[33] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs), 11 2015.

[34] James R. Clough, Ilkay Oksuz, Esther Puyol-Antón, Bram Ruijsink, Andrew P. King, and Julia A. Schnabel. Global and local interpretability for cardiac mri classification. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 656–664, Cham, 2019. Springer International Publishing.

[35] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P Lungren, and Akshay Chaudhari. Gifsplanation via latent shift: A simple

autoencoder approach to counterfactual generation for chest x-rays. *Medical Imaging with Deep Learning (MIDL)*, 2021.

[36] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11070 LNCS, pages 529–536. Springer Verlag, 2018.

[37] Vincent Couteaux, Olivier Nempont, Guillaume Pizaine, and Isabelle Bloch. Towards interpretability of segmentation networks by analyzing deepdreams. In Kenji Suzuki, Mauricio Reyes, Tanveer Syeda-Mahmood, Ender Konukoglu, Ben Glocker, Roland Wiest, Yaniv Gur, Hayit Greenspan, and Anant Madabhushi, editors, *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 56–63, Cham, 2019. Springer International Publishing.

[38] Harvey O. Coxson, Jonathon Leipsic, Grace Parraga, and Don D. Sin. Using Pulmonary Imaging to Move Chronic Obstructive Pulmonary Disease beyond FEV ¡sub¿1¡/sub¿. *American Journal of Respiratory and Critical Care Medicine*, 190(2):135–144, 7 2014.

[39] Henriette Cramer, Jean Garcia-Gathright, Aaron Springer, and Sravana Reddy. Assessing and addressing algorithmic bias in practice. *interactions*, 25(6):58–63, 2018.

[40] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019.

[41] Sijia Cui, Shuai Ming, Yi Lin, Fanghong Chen, Qiang Shen, Hui Li, Gen Chen, Xiangyang Gong, and Haochu Wang. Development and clinical application of deep learning model for lung nodules screening on ct images. *Scientific Reports*, 10, 08 2020.

[42] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017.

[43] Cameron Davidson Pilon, Jonas Kalderstam, Paul Zivich, Ben Kuhn, Andrew Fiore-Gartland, Luis Moneda, Gabriel, Daniel WIlson, Alex Parij, Kyle Stark, Steven Anton, Lilian Besson, Jona, Harsh Gadgil, Dave Golland, Sean Hussey, Javad Noorbakhsh, Andreas Klintberg, Joanne Jordan, Jeff Rose, Isaac Slavitt, Eric Martin, Ed-

uardo Ochoa, Dylan Albrecht, dhuynh, Denis Zgonjanin, Daniel Chen, Chris Fournier, Arturo, and André F. Rendeiro. Lifelines, 2 2019.

[44] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv*, 2020.

[45] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[46] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Advances in Neural Information Processing Systems*, pages 592–603, 2018.

[47] Alejandro A Diaz, Clarissa Valim, Tsuneo Yamashiro, Raúl San José Estépar, James C Ross, Shin Matsuoka, Brian Bartholmai, Hiroto Hatabu, Edwin K Silverman, and George R Washko. Airway count and emphysema assessed by chest CT imaging predicts clinical outcome in smokers. *Chest*, 138(4):880–7, 10 2010.

[48] Konstantinos Dimopoulos, Georgios Giannakoulas, Isaac Bendayan, Emmanouil Liodakis, Ricardo Petraco, Gerhard Paul Diller, Massimo F. Piepoli, Lorna Swan, Michael Mullen, Nicky Best, Philip A. Poole-Wilson, Darrel P. Francis, Michael B. Rubens, and Michael A. Gatzoulis. Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease. *International Journal of Cardiology*, 166(2):453–457, 6 2013.

[49] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[50] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. Long-term recurrent convolutional networks for visual recognition and description. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2625–2634, 2015.

[51] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[52] Zach Eaton-Rosen, Felix Bragman, Sotirios Bisdas, Sébastien Ourselin, and M. Jorge Cardoso. Towards safe deep learning: Accurately quantifying biomarker uncertainty in neural network predictions. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11070 LNCS, pages 691–699. Springer Verlag, 6 2018.

[53] Fabian Eitel and Kerstin Ritter. Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11797 LNCS, pages 3–11. Springer, 10 2019.

[54] Gamaleldin Elsayed, Simon Kornblith, and Quoc V Le. Saccader: Improving accuracy of hard attention models for vision. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[55] Koen AJ Eppenhof and Josien PW Pluim. Pulmonary ct registration through supervised learning with convolutional neural networks. *IEEE transactions on medical imaging*, 38(5):1097–1105, 2018.

[56] Raúl San José Estépar and Gregory L. Kinney. Computed tomographic measures of pulmonary vascular morphology in smokers and their clinical implications. *AJRCCM*, 2013.

[57] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature 2017 542:7639*, 542(7639):115–118, 1 2017.

[58] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835, 2021.

[59] M. Farhangi, Nicholas Petrick, Berkman Sahiner, Hichem Frigui, A. Amini, and Aria Pezeshk. Recurrent attention network for false positive reduction in the detection of pulmonary nodules in thoracic ct scans. *Medical Physics*, 47, 02 2020.

[60] Di Feng, Lars Rosenbaum, and Klaus Dietmayer. Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3d vehicle detection. *CoRR*, abs/1804.05132, 2018.

[61] Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[62] Eibe Frank and Mark Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, pages 145–156, 2001.

[63] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4476–4484, 2017.

[64] Yarin Gal. Uncertainty in deep learning. In *Thesis*, 2016.

[65] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA, 20–22 Jun 2016. PMLR.

[66] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *32nd International Conference on Neural Information Processing Systems*, page 8803–8812, 2018.

[67] Aimilia Gastounioti and Despina Kontos. Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI? *Radiology: Artificial Intelligence*, 2(3):e200088, 5 2020.

[68] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. *ICML*, 2019.

[69] A. Ghandeharioun, B. Eoff, B. Jou, and R. Picard. Characterizing sources of uncertainty to proxy calibration and disambiguate annotator and data bias. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4202–4206, Los Alamitos, CA, USA, oct 2019. IEEE Computer Society.

[70] Asma Ghandeharioun, Been Kim, Chun-Liang Li, Brendan Jou, Brian Eoff, and Rosalind Picard. Dissect: Disentangled simultaneous explanations via concept traversals. *arXiv preprint arXiv:2105.15164*, 2021.

[71] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[72] Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards Automatic Concept-based Explanations. -, 2 2019.

[73] Alyssa Glass, Deborah L McGuinness, and Michael Wolverton. Toward establishing trust in adaptive agents. In *Proceedings of the 13th international conference on Intelligent user interfaces*, pages 227–236. ACM, 2008.

[74] German Gonzalez, Samuel Y. Ash, Gonzalo Vegas-Sánchez-Ferrero, Jorge Onieva Onieva, Farbod N. Rahaghi, James C. Ross, Alejandro Dáz, Raul San José Estépar, and George R. Washko. Disease staging and prognosis in smokers using deep learning in chest computed tomography. *American Journal of Respiratory and Critical Care Medicine*, 197(2):193–203, 1 2018.

[75] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In Z Ghahramani, M Welling, C Cortes, N D Lawrence, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[76] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[77] Yu Gordienko, Peng Gang, Jiang Hui, Wei Zeng, Yu Kochura, Oleg Alienin, Oleksandr Rokovyi, and Sergii Stirenko. Deep learning with lung segmentation and bone shadow exclusion techniques for chest x-ray analysis of lung cancer. In *International conference on computer science, engineering and education applications*, pages 638–647. Springer, 2018.

[78] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019.

[79] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual Visual Explanations. In *International Conference on Machine Learning*, pages 2376–2384, 2019.

[80] Alex Graves. Practical variational inference for neural networks. *Advances in Neural Information Processing Systems*, 2011.

[81] M Graziani, V Andrearczyk, S Marchand-Maillet, and H Müller. Concept attribution: Explaining cnn decisions to physicians. *Computers in biology and medicine*, 123:103865, 2020.

[82] Mara Graziani, Vincent Andrearczyk, and Henning Müller. Regression concept vectors for bidirectional explanations in histopathology. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications*, pages 124–132. Springer, 2018.

[83] Katarzyna Grzela, Malgorzata Litwiniuk, Wioletta Zagorska, and Tomasz Grzela. Airway Remodeling in Chronic Obstructive Pulmonary Disease and Asthma: the Role of Matrix Metalloproteinase-9. *Archivum immunologiae et therapiae experimentalis*, 64(1):47–55, 2 2016.

[84] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2019.

[85] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR, 2017.

[86] David M. Hansell, Alexander A. Bankier, Heber MacMahon, Theresa C. McLoud, Nestor L. Müller, and Jacques Remy. Fleischner Society: Glossary of terms for thoracic imaging, 3 2008.

[87] Michael Harradon, Jeff Druce, and Brian Ruttenberg. Causal learning and explanation of deep neural networks via autoencoded activations. *arXiv preprint arXiv:1802.00541*, 2018.

[88] Milos Hauskrecht, Iyad Batal, Charmgil Hong, Quang Nguyen, Gregory Cooper, Shyam Visweswaran, and Gilles Clermont. Outlier-based detection of unusual patient-management actions: An icu study. *Journal of Biomedical Informatics*, 64, 10 2016.

[89] M. D. Hayhurst, W. MacNee, and D. C. Flenley. Diagnosis of pulmonary emphysema by computerised tomography. *Lancet*, 2(8398):320–322, 1984.

[90] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pages 770–778. IEEE Computer Society, 12 2016.

[91] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 41–50, 2019.

[92] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

[93] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *International Conference on Learning Representations*, 2019.

[94] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017.

[95] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J.W.L. Aerts. Artificial intelligence in radiology, 8 2018.

[96] Y. C. Hsu, Y. Shen, H. Jin, and Z. Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[97] Gao Huang, Yixuan Li, Geoff Pleiss, Zhuang Liu, John E. Hopcroft, and Kilian Q. Weinberger. Snapshot ensembles: Train 1, get M for free. *5th International Conference on Learning Representations, ICLR*, 2017.

[98] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely Connected Convolutional Networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-January:2261–2269, 8 2016.

[99] Huaibo Huang, Zhihang Li, Ran He, Zhenan Sun, and Tieniu Tan. IntroVAE: Introspective Variational Autoencoders for Photographic Image Synthesis. *International*

Conference on Advances in Neural Information Processing Systems (NeurIPS), pages 10236–10245, 2018.

[100] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. *Advances in Neural Information Processing Systems*, 2021.

[101] Shih-Cheng Huang, Anuj Pareek, Roham Zamanian, Imon Banerjee, and Matthew Lungren. Multimodal fusion with deep neural networks for leveraging ct imaging and electronic health record: a case-study in pulmonary embolism detection. *Scientific Reports*, 10, 12 2020.

[102] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Mach. Learn.*, 110:457–506, 2021.

[103] Stephen Humphries, Aleena Notary, Juan Centeno, Matthew Strand, James Crapo, Edwin Silverman, and David Lynch. Deep learning enables automatic classification of emphysema pattern at ct. *Radiology*, 294:191022, 12 2019.

[104] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110, 03 2021.

[105] Katsuya Iijima, Hiroko Hashimoto, Masayoshi Hashimoto, Bo-Kyung Son, Hidetaka Ota, Sumito Ogawa, Masato Eto, Masahiro Akishita, and Yasuyoshi Ouchi. Aortic arch calcification detectable on chest x-ray is a strong independent predictor of cardiovascular events beyond traditional risk factors. *Atherosclerosis*, 210(1):137–144, 2010.

[106] Kosuke Imai, Booil Jo, and Elizabeth A Stuart. Commentary: Using potential outcomes to understand causal mediation analysis. *Multivariate Behavioral Research*, 46(5):861–873, 2011.

[107] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[108] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, 1 2019.

[109] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475–477, 2014.

[110] Rohit Jena and Sumedha Singla. Self-supervised vessel enhancement using flow-based consistencies. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021.

[111] Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To Trust Or Not To Trust A Classifier. In S Bengio, H Wallach, H Larochelle, K Grauman, N Cesa-Bianchi, and R Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 5541–5552. Curran Associates, Inc., 2018.

[112] Alistair E.W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih Ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 12 2019.

[113] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xGEMs: Generating Examplars to Explain Black-Box Models. *arXiv preprint arXiv:1806.08867*, 2018.

[114] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards Realistic Individual Recourse and Actionable Explanations in Black-Box Decision Making Systems. *CoRR*, abs/1907.09615, 2019.

[115] Vinaya S Karkhanis and Jyotsna M Joshi. Pleural effusion: diagnosis, treatment, and management. *Open access emergency medicine: OAEM*, 4:31, 2012.

[116] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[117] Alexander Katzmann, Oliver Taubmann, Stephen Ahmad, Alexander Mühlberg, Michael Sühling, and Horst-Michael Groß. Explaining clinical decision support systems in medical imaging using cycle-consistent activation maximization. *Neurocomputing*, 458:141–156, 2021.

[118] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. *International Conference on Learning Representations*, 2020.

[119] Divyansh Kaushik, Amrith Rajagopal Setlur, Eduard H. Hovy, and Zachary Chase Lipton. Explaining the efficacy of counterfactually-augmented data. *International Conference on Learning Representations*, 2021.

[120] Been Kim. *Interactive and interpretable machine learning models for human machine collaboration.* PhD thesis, Massachusetts Institute of Technology, 2015.

[121] Been Kim, Rajiv Khanna, and Oluwasanmi Koyejo. Examples are not enough, learn to criticize! criticism for interpretability. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 2288–2296, Red Hook, NY, USA, 2016. Curran Associates Inc.

[122] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). *35th International Conference on Machine Learning, ICML 2018*, 6:4186–4195, 11 2017.

[123] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 2654–2663. PMLR, 2018.

[124] Victor Kim, Wojciech R Dolliver, Hrudaya P Nath, Scott A Grumley, Nina Terry, Asmaa Ahmed, Andrew Yen, Kathleen Jacobs, Seth Kligerman, Alejandro A Diaz, James D Crapo, Edwin K Silverman, Barry J Make, Elizabeth A Regan, Terri Beaty, Ferdouse Begum, Peter J Castaldi, Michael Cho, Dawn L DeMeo, Adel R Boueiz, Marilyn G Foreman, Eitan Halper-Stromberg, Lystra P Hayden, Craig P Hersh, Jacqueline Hetmanski, Brian D Hobbs, John E Hokanson, Nan Laird, Christoph Lange, Sharon M Lutz, Merry-Lynn McDonald, Margaret M Parker, Dmitry Prokopenko, Dandi Qiao, Elizabeth A Regan, Phuwanat Sakornsakolpat, Edwin K Silverman, Emily S Wan, Sungho Won, Juan Pablo Centeno, Jean-Paul Charbonnier, Harvey O Coxson, Craig J Galban, MeiLan K Han, Eric A Hoffman, Stephen Humphries,

Francine L Jacobson, Philip F Judy, Ella A Kazerooni, Alex Kluiber, David A Lynch, Pietro Nardelli, John D Newell Jr., Aleena Notary, Andrea Oh, Elizabeth A Regan, James C Ross, Raul San Jose Estepar, Joyce Schroeder, Jered Sieren, Berend C Stoel, Juerg Tschirren, Edwin Van Beek, Bram van Ginneken, Eva van Rikxoort, Gonzalo Vegas Sanchez-Ferrero, Lucas Veitel, George R Washko, Carla G Wilson, Robert Jensen, Douglas Everett, Jim Crooks, Katherine Pratte, Matt Strand, Carla G Wilson, John E Hokanson, Gregory Kinney, Sharon M Lutz, Kendra A Young, Surya P Bhatt, Jessica Bon, Alejandro A Diaz, MeiLan K Han, Barry Make, Susan Murray, Elizabeth Regan, Xavier Soler, Carla G Wilson, Russell P Bowler, Katerina Kechris, Farnoush Banaei-Kashani, Jeffrey L Curtis, Perry G Pernicano, Nicola Hanania, Mustafa Atik, Aladin Boriek, Kalpatha Guntupalli, Elizabeth Guy, Amit Parulekar, Dawn L DeMeo, Alejandro A Diaz, Lystra P Hayden, Brian D Hobbs, Craig Hersh, Francine L Jacobson, George Washko, R Graham Barr, John Austin, Belinda D'Souza, Byron Thomashow, Neil MacIntyre Jr., H Page McAdams, Lacey Washington, Eric Flenaugh, Silanth Terpenning, Charlene McEvoy, Joseph Tashjian, Robert Wise, Robert Brown, Nadia N Hansel, Karen Horton, Allison Lambert, Nirupama Putcha, Richard Casaburi, Alessandra Adami, Matthew Budoff, Hans Fischer, Janos Porszasz, Harry Rossiter, William Stringer, Amir Sharafkhaneh, Charlie Lan, Christine Wendt, Brian Bell, Ken M Kunisaki, Russell Bowler, David A Lynch, Richard Rosiello, David Pace, Gerard Criner, David Ciccolella, Francis Cordova, Chandra Dass, Gilbert D'Alonzo, Parag Desai, Michael Jacobs, Steven Kelsen, Victor Kim, A James Mamary, Nathaniel Marchetti, Aditi Satti, Kartik Shenoy, Robert M Steiner, Alex Swift, Irene Swift, Maria Elena Vega-Sanchez, Mark Dransfield, William Bailey, Surya P Bhatt, Anand Iyer, Hrudaya Nath, J Michael Wells, Douglas Conrad, Xavier Soler, Andrew Yen, Alejandro P Comellas, Karin F Hoth, John Newell Jr., Brad Thompson, MeiLan K Han, Ella Kazerooni, Wassim Labaki, Craig Galban, Dharshan Vummidi, Joanne Billings, Abbie Begnaud, Tadashi Allen, Frank Sciurba, Jessica Bon, Divay Chandra, Carl Fuhrman, Joel Weissfeld, Antonio Anzueto, Sandra Adams, Diego Maselli-Caceres, Mario E Ruiz, Harjinder Singh, and the COPDGene Investigators. Mucus plugging on computed tomography and chronic bronchitis in chronic obstructive pulmonary disease. *Respiratory Research*, 22(1):110, 2021.

[125] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280, 2017.

[126] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 12 2014.

[127] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in Neural Information Processing Systems*, 2015.

[128] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural Safety*, 31(2):105–112, 2009. Risk Acceptance and Risk Communication.

[129] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *34th International Conference on Machine Learning, ICML 2017*, 2017.

[130] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5338–5348. PMLR, 13–18 Jul 2020.

[131] Ranganath Krishnan and Omesh Tickoo. Improving model calibration with accuracy versus uncertainty optimization. *Advances in Neural Information Processing Systems*, 33, 2020.

[132] Alex Krizhevsky. Learning multiple layers of features from tiny images. In -, 2009.

[133] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[134] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T. Freeman, Phillip Isola, Amir Globerson, Michal Irani, and Inbar Mosseri. Explaining in style: Training a gan to explain a classifier in stylespace. *arXiv preprint arXiv:2104.13369*, 2021.

[135] Agostina J. Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H. Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences of the United States of America*, 117(23):12592–12594, 6 2020.

[136] Nathalie Lassau, S. Ammari, Emilie Chouzenoux, Hugo Gortais, Paul Herent, Matthieu Devilder, Samer Soliman, Olivier Meyrignac, Marie-Pauline Talabard, Jean-Philippe Lamarque, Remy Dubois, Nicolas Loiseau, Paul Trichelair, Etienne Bendjebbar, Gabriel Garcia, Corinne Balleyguier, Mansouria Merad, Annabelle Stoclin, Simon Jegou, and Michael Blum. Integrating deep learning ct-scan model, biological and clinical variables to predict severity of covid-19 patients. *Nature Communications*, 12:634, 01 2021.

[137] Yann LeCun and Corinna Cortes. MNIST handwritten digit database. *AT&T Labs*, 2010.

[138] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. *International Conference on Learning Representations*, 2018.

[139] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in Neural Information Processing Systems*, 2018.

[140] Christian Leibig, Vaneeda Allken, Murat Seçkin Ayhan, Philipp Berens, and Siegfried Wahl. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific Reports*, 7:17816, 2017.

[141] S Lemeshow and D W Hosmer. A review of goodness of fit statistics for use in the development of logistic regression models. *American journal of epidemiology*, 115(1):92–106, 1 1982.

[142] Shiyu Liang, Yixuan Li, and R Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *6th International Conference on Learning Representations, ICLR 2018*, 2018.

[143] D. Y. Lin. Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, 13(21):2233–2247, 11 1994.

[144] Ziqian Lin, Sreya Dutta Roy, and Yixuan Li. Mood: Multi-level out-of-distribution detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

[145] Robert Lindsey, Aaron Daluiski, Sumit Chopra, Alexander Lachapelle, Michael Mozer, Serge Sicular, Douglas Hanel, Michael Gardner, Anurag Gupta, Robert Hotchkiss, et al. Deep neural network improves fracture detection by clinicians. *Proceedings of the National Academy of Sciences*, 115(45):11591–11596, 2018.

[146] Jeremiah Zhe Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax-Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *ArXiv*, abs/2006.10108, 2020.

[147] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. Generative Counterfactual Introspection for Explainable Deep Learning. *arXiv*, 2019.

[148] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. *Advances in Neural Information Processing Systems*, 2020.

[149] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015.

[150] Adriano Lucieri, Muhammad Naseer Bajwa, Stephan Alexander Braun, Muhammad Imran Malik, Andreas Dengel, and Sheraz Ahmed. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 international joint conference on neural networks (IJCNN)*, pages 1–10. IEEE, 2020.

[151] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In I Guyon, U V Luxburg, S Bengio, H Wallach, R Fergus, S Vishwanathan, and R Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[152] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. Effective Approaches to Attention-based Neural Machine Translation. *ArXiv*, 8 2015.

[153] David A. Lynch, John H. M. Austin, James C. Hogg, Philippe A. Grenier, Hans-Ulrich Kauczor, Alexander A. Bankier, R. Graham Barr, Thomas V. Colby, Jeffrey R. Galvin, Pierre Alain Gevenois, Harvey O. Coxson, Eric A. Hoffman, John D. Newell, Massimo Pistolesi, Edwin K. Silverman, and James D. Crapo. CT-Definable Subtypes of Chronic Obstructive Pulmonary Disease: A Statement of the Fleischner Society. *Radiology*, 277(1):192–205, 10 2015.

[154] David A. Lynch, Camille M. Moore, Carla Wilson, Dipti Nevrekar, Theodore Jennermann, Stephen M. Humphries, John H. M. Austin, Philippe A. Grenier, Hans-Ulrich Kauczor, MeiLan K. Han, Elizabeth A. Regan, Barry J. Make, Russell P. Bowler, Terri H. Beaty, Douglas Curran-Everett, John E. Hokanson, Jeffrey L. Curtis, Edwin K. Silverman, James D. Crapo, and For the Genetic Epidemiology of COPD (COPDGene) Investigators. CT-based Visual Classification of Emphysema: Association with Mortality in the COPDGene Study. *Radiology*, 288(3):859–866, 9 2018.

[155] Pragnya Maduskar, Laurens Hogeweg, Rick Philipsen, and Bram van Ginneken. Automated localization of costophrenic recesses and costophrenic angle measurement on

frontal chest radiographs. In Carol L. Novak and Stephen Aylward, editors, *Medical Imaging 2013: Computer-Aided Diagnosis*, volume 8670, page 867038. SPIE, 3 2013.

[156] Pragnya Maduskar, Rick H.M.M. Philipsen, Jaime Melendez, Ernst Scholten, Duncan Chanda, Helen Ayles, Clara I. Sánchez, and Bram van Ginneken. Automatic detection of pleural effusion in chest radiographs. *Medical Image Analysis*, 28:22–32, 2 2016.

[157] Divyat Mahajan, Chenhao Tan, and Amit Sharma. Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers. *arXiv preprint arXiv:1912.03277*, 12 2019.

[158] Seyedsalim Malakouti and Milos Hauskrecht. Predicting patient's diagnoses and diagnostic categories from clinical-events in ehr data. In *Conference on Artificial Intelligence in Medicine in Europe*, pages 125–130. Springer, 2019.

[159] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[160] Devraj Mandal, Sanath Narayan, Sai Kumar Dwivedi, Vikram Gupta, Shuaib Ahmed, Fahad Shahbaz Khan, and Ling Shao. Out-of-distribution detection for generalized zero-shot action recognition. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[161] David R Mandel. Counterfactual and causal explanation: From early theoretical views to new frontiers. In *The psychology of counterfactual thinking*, pages 23–39. Routledge, 2007.

[162] Fernando J. Martinez, Gregory Foster, Jeffrey L. Curtis, Gerard Criner, Gail Weinmann, Alfred Fishman, Malcolm M. DeCamp, Joshua Benditt, Frank Sciurba, Barry Make, Zab Mohsenifar, Philip Diaz, Eric Hoffman, and Robert Wise. Predictors of mortality in patients with emphysema and severe airflow obstruction. *American Journal of Respiratory and Critical Care Medicine*, 173(12):1326–1334, 6 2006.

[163] Jonathan Masci, Ueli Meier, Dan Cireşan, and Jürgen Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *LNCS*, pages 52–59, 2011.

[164] Shin Matsuoka, Yasuyuki Kurihara, Kunihiro Yagihashi, Makoto Hoshino, Naoto Watanabe, and Yasuo Nakajima. Quantitative assessment of air trapping in chronic obstructive pulmonary disease using inspiratory and expiratory volumetric mdct. *American Journal of Roentgenology*, 190(3):762–769, 2008.

[165] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *ArXiv*, 2 2018.

[166] Y. B. Mensah, K. Mensah, S. Asiamah, H. Gbadamosi, E. A. Idun, W. Brakohiapa, and A. Oddoye. Establishing the Cardiothoracic Ratio Using Chest Radiographs in an Indigenous Ghanaian Population: A Simple Tool for Cardiomegaly Screening. *Ghana medical journal*, 49(3):159–164, 9 2015.

[167] E. N.C. Milne, M. Pistolesi, M. Miniati, and C. Giuntini. The radiologic distinction of cardiogenic and noncardiogenic edema. *American Journal of Roentgenology*, 144(5):879–894, 1985.

[168] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[169] Mehdi Mirza and Simon Osindero. Conditional Generative Adversarial Nets. *CoRR*, abs/1411.1784, 2014.

[170] Akinori Mitani, Abigail Huang, Subhashini Venugopalan, Greg S. Corrado, Lily Peng, Dale R. Webster, Naama Hammel, Yun Liu, and Avinash V. Varadarajan. Detection of anaemia from retinal fundus images via deep learning. *Nature Biomedical Engineering*, 4(1):18–27, 1 2020.

[171] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2 2018.

[172] Takeru Miyato and Masanori Koyama. cGANs with Projection Discriminator. In *International Conference on Learning Representations*, 2018.

[173] Firdaus A.A. Mohamed Hoesein, Eva van Rikxoort, Bram van Ginneken, Pim A. de Jong, Mathias Prokop, Jan-Willem J. Lammers, and Pieter Zanen. Computed tomography-quantified emphysema distribution is associated with lung function decline. *European Respiratory Journal*, 40(4):844–850, 10 2012.

[174] Sina Mohseni, Mandar Pitale, Jbs Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. *AAAI*, 2020.

[175] Christoph Molnar. *Interpretable Machine Learning.* Github, 2019. `https://christophm.github.io/interpretable-ml-book/`.

[176] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

[177] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[178] Ramaravind Kommiya Mothilal, Amit Sharma, and Chenhao Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 5 2019.

[179] Jishnu Mukhoti, Andreas Kirsch, Joost van Amersfoort, Philip HS Torr, and Yarin Gal. Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. *arXiv preprint arXiv:2102.11582*, 2021.

[180] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip Torr, and Puneet Dokania. Calibrating deep neural networks using focal loss. *Advances in Neural Information Processing Systems*, 33, 2020.

[181] Jishnu Mukhoti, Viveka Kulharia, Amartya Sanyal, Stuart Golodetz, Philip H. S. Torr, and Puneet Kumar Dokania. Calibrating deep neural networks using focal loss. *ArXiv*, abs/2002.09437, 2020.

[182] N L Müller, C A Staples, R R Miller, and R T Abboud. Density mask: An objective method to quantitate emphysema using computed tomography. *Chest*, 10 1988.

[183] Franz Nachbar, Wilhelm Stolz, Tanja Merkle, Armand B Cognetta, Thomas Vogt, Michael Landthaler, Peter Bilek, Otto Braun-Falco, and Gerd Plewig. The abcd rule of dermatoscopy: high prospective value in the diagnosis of doubtful melanocytic skin lesions. *Journal of the American Academy of Dermatology*, 30(4):551–559, 1994.

[184] NOBUYUKI Nakamori, H MacMahon, Y Sasaki, S Montner, et al. Effect of heart-size parameters computed from digital chest radiographs on detection of cardiomegaly. potential usefulness for computer-aided diagnosis. *Investigative radiology*, 26(6):546–550, 1991.

[185] Arunachalam Narayanaswamy, Subhashini Venugopalan, Dale R. Webster, Lily Peng, Greg Corrado, Paisan Ruamviboonsuk, Pinal Bavishi, Rory Sayres, Abigail Huang, Siva Balasubramanian, Michael Brenner, Philip Nelson, and Avinash V. Varadarajan. Scientific Discovery by Generating Counterfactuals using Image Translation. *arXiv preprint arXiv:2007.05500*, 7 2020.

[186] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[187] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[188] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015.

[189] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification with auxiliary probabilistic information. In *IEEE International Conference on Data Mining*, pages 477–486, 2011.

[190] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Sample-efficient learning with auxiliary class-label information. In *Proceedings of the Annual American Medical Informatics Association Symposium*, pages 1004–1012, 2011.

[191] Quang Nguyen, Hamed Valizadegan, and Milos Hauskrecht. Learning classification models with soft-label information. *Journal of American Medical Informatics Association*, 2013.

[192] Mizuho Nishio, Kazuaki Nakane, Takeshi Kubo, Masahiro Yakami, Yutaka Emoto, Mari Nishio, and Kaori Togashi. Automated prediction of emphysema visual score using homology-based quantification of low-attenuation lung region. *PloS one*, 12(5):e0178217, 2017.

[193] Luke Oakden-Rayner, Jared Dunnmon, Gustavo Carneiro, and Christopher Ré. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, New York, NY, USA, 2020. ACM.

[194] D. E. O'Donnell and P. Laveneziana. Physiology and consequences of lung hyperinflation in COPD. *European Respiratory Review*, 15(100):61–67, 12 2006.

[195] MD Omar Lababede. Pleural Effusion Imaging: Overview, Radiography, Computed Tomography.

[196] Kristoffer Ostridge and Tom MA Wilkinson. Present and future utility of computed tomography scanning in the assessment and management of copd. *European Respiratory Journal*, 48(1):216–228, 2016.

[197] Marcin Ostrowski, Tomasz Marjański, and Witold Rzyman. Low-dose computed tomography screening reduces lung cancer mortality. *Advances in medical sciences*, 63(2):230–236, 2018.

[198] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021.

[199] Alvaro Parafita Martinez and Jordi Vitria Marca. Explaining visual models by causal attribution. In *Proceedings - 2019 International Conference on Computer Vision Workshop, ICCVW 2019*, pages 4167–4175. Institute of Electrical and Electronics Engineers Inc., 10 2019.

[200] F. Pasa, V. Golkov, F. Pfeiffer, D. Cremers, and D. Pfeiffer. Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization. *Scientific Reports*, 9(1):1–9, 12 2019.

[201] J PEARL. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*, pages 411–420. Morgan Kaufman, 2001.

[202] Haixin Peng, Huacong Sun, and Yanfei Guo. 3d multi-scale deep convolutional neural networks for pulmonary nodule detection. *PLOS ONE*, 16(1):1–14, 01 2021.

[203] Thierry Perez, Pierre Régis Burgel, Jean Louis Paillasseur, Denis Caillaud, Gaétan Deslée, Pascal Chanez, Nicolas Roche, and INITIATIVES BPCO Scientific Committee. Modified Medical Research Council scale vs Baseline Dyspnea Index to evaluate dyspnea in chronic obstructive pulmonary disease. *International Journal of Chronic Obstructive Pulmonary Disease*, 10:1663, 8 2015.

[204] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[205] Hammad Qureshi, Amir Sharafkhaneh, and Nicola A Hanania. Chronic obstructive pulmonary disease exacerbations: latest evidence and clinical implications. *Therapeutic advances in chronic disease*, 5(5):212–27, 9 2014.

[206] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):18, 2018.

[207] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, 11 2018.

[208] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv preprint arXiv:1711.05225*, 11 2017.

[209] Elizabeth A. Regan, John E. Hokanson, James R. Murphy, Barry Make, David A. Lynch, Terri H. Beaty, Douglas Curran-Everett, Edwin K. Silverman, and James D. Crapo. Genetic epidemiology of COPD (COPDGene) study design. *Journal of COPD*, 7(1):32–43, 2010.

[210] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark A DePristo, Joshua V Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 14707–14718, 2019.

[211] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Technical report, github, 2015.

[212] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.

[213] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[214] Pau Rodríguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1056–1065, October 2021.

[215] Alejandro Rodriguez-Ruiz, Kristina Lång, Albert Gubern-Merida, Mireille Broeders, Gisella Gennaro, Paola Clauser, Thomas H. Helbich, Margarita Chevalier, Tao Tan, Thomas Mertelmeier, Matthew G. Wallis, Ingvar Andersson, Sophia Zackrisson, Ritse M. Mann, and Ioannis Sechopoulos. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. *Journal of the National Cancer Institute*, 111(9):916–922, 9 2019.

[216] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[217] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[218] Jonathan Rubin, Deepan Sanghavi, Claire Zhao, Kathy Lee, Ashequl Qadir, and Minnan Xu-Wilson. Large Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. *arXiv preprint arXiv:1804.07839*, 4 2018.

[219] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[220] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16(none), 1 2022.

[221] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. *Advances in neural information processing systems*, 30, 2017.

[222] Zohaib Salahuddin, Henry C Woodruff, Avishek Chatterjee, and Philippe Lambin. Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine*, 140:105111, 2022.

[223] Tim Salimans, Andrej Karpathy, Xi Chen, and Diederik P Kingma. Pixelcnn++: Improving the pixelcnn with discretized logistic mixture likelihood and other modifications. *arXiv e-prints*, pages arXiv–1701, 2017.

[224] Ardavan Samangouei Pouya
}and Saeedi, Nakagawa Liam, and Silberman Nathan. ExplainGAN: Model Explanation via Decision Boundary Crossing Transformations. In Martial Ferrari Vittorio
}and Hebert, Sminchisescu Cristian, and Weiss Yair, editors, *Computer Vision – ECCV 2018*, pages 681–696. Springer International Publishing, 2018.

[225] R San Jose Estepar, J C Ross, R Harmouche, J Onieva, A A Diaz, and G R Washko. CIP: an open-source library and workstation for quantitative chest imaging. *Am J Respir Crit Care Med*, 191:A4975, 2015.

[226] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, Shawn Xu, Scott Barb, Anthony Joseph, Michael Shumski, Jesse Smith, Arjun B. Sood, Greg S. Corrado, Lily Peng, and Dale R. Webster. Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy. *Ophthalmology*, 126(4):552–564, 4 2019.

[227] Jenna Schabdach, William M. Wells, Michael Cho, and Kayhan N. Batmanghelich. A likelihood free approach for characterizing heterogeneous diseases in large scale studies. In *IPMI*, volume 10265 LNCS, pages 170–183, 2017.

[228] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53:197–207, 2019.

[229] David Schoenfeld. Chi-Squared Goodness-of-Fit Tests for the Proportional Hazards Regression Model. *Biometrika*, 67(1):145, 4 1980.

[230] Jarrel CY Seah, Cyril HM Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, et al. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *The Lancet Digital Health*, 3(8):e496–e506, 2021.

[231] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 618–626, 2017.

[232] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F. Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *International Conference on Learning Representations*, 2020.

[233] S D Shapiro. Evolving concepts in the pathogenesis of chronic obstructive pulmonary disease. *Clinics in chest medicine*, 21(4):621–632, 2000.

[234] Shiwen Shen, Simon X Han, Denise R Aberle, Alex A Bui, and William Hsu. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. *Expert systems with applications*, 128:84–95, 2019.

[235] Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.

[236] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153, 2017.

[237] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *Computing Research Repository*, 2013.

[238] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv 1409.1556*, 6 2014.

[239] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

[240] Sumedha Singla, Mingming Gong, Siamak Ravanbakhsh, Frank Sciurba, Barnabas Poczos, and Kayhan N. Batmanghelich. Subject2Vec Generative-discriminative approach from a set of image patches to a vector. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[241] Sumedha Singla, Mingming Gong, Craig Riley, Frank Sciurba, and Kayhan Batmanghelich. Improving clinical disease subtyping and future events prediction through a chest ct-based deep learning approach. *Medical physics*, 48(3):1168–1181, 2021.

[242] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by Progressive Exaggeration. In *International Conference on Learning Representations*, 2020.

[243] Sumedha Singla, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly-a counterfactual approach. *arXiv preprint arXiv:2101.04230*, 2021.

[244] Sumedha Singla, Stephen Wallace, Sofia Triantafillou, and Kayhan Batmanghelich. Using causal analysis for conceptual deep learning explanation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 519–528. Springer, 2021.

[245] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *Uncertainty in Artificial Intelligence (UAI)*, 2018.

[246] J J Soler-Cataluña, M A Martínez-García, P Román Sánchez, E Salcedo, M Navarro, and R Ochando. Severe acute exacerbations and mortality in patients with chronic obstructive pulmonary disease. *Thorax*, 60(11):925–31, 11 2005.

[247] QingZeng Song, Lei Zhao, XingKe Luo, and XueChen Dou. Using deep learning for classification of lung nodules on computed tomography images. *Journal of Healthcare Engineering*, 2017:1–7, 08 2017.

[248] L. Sorensen, M. Nielsen, Pechin Lo, H. Ashraf, J. H. Pedersen, and M. de Bruijne. Texture-based analysis of COPD: A data-driven approach. *IEEE Transactions on Medical Imaging*, 31(1):70–78, 2012.

[249] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A Riedmiller. Striving for Simplicity: The All Convolutional Net. In *ICLR (workshop track)*, 2015.

[250] Akshayvarun Subramanya, Suraj Srinivas, and R. Venkatesh Babu. Confidence estimation in deep neural networks via density modelling. *ArXiv*, abs/1707.07013, 2017.

[251] Li Sun and Ke Yu. Context matters: Graph-based self-supervised representation learning for medical images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[252] Wenqing Sun, Bin Zheng, and Wei Qian. Computer aided lung cancer diagnosis with deep learning algorithms. In Georgia D. Tourassi and Samuel G. Armato III, editors, *Medical Imaging 2016: Computer-Aided Diagnosis*, volume 9785, pages 241 – 248. International Society for Optics and Photonics, SPIE, 2016.

[253] Yiyou Sun, Chuan Guo, and Yixuan Li. React: Out-of-distribution detection with rectified activations. *Advances in Neural Information Processing Systems*, 2021.

[254] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9269–9278. PMLR, 13–18 Jul 2020.

[255] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org, 2017.

[256] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[257] Martin C Tammemagi, Heidi Schmidt, Simon Martel, Annette McWilliams, John R Goffin, Michael R Johnston, Garth Nicholas, Alain Tremblay, Rick Bhatia, Geoffrey Liu, et al. Participant selection for lung cancer screening by risk modelling (the pan-canadian early detection of lung cancer [pancan] study): a single-arm, prospective study. *The lancet oncology*, 18(11):1523–1531, 2017.

[258] Lisa Y W Tang, Harvey O Coxson, Stephen Lam, Jonathon Leipsic, Roger C Tam, and Don D Sin. Towards large-scale case-finding: training and validation of residual

networks for detection of chronic obstructive pulmonary disease using low-dose ct. *The Lancet Digital Health*, 2(5):e259–e267, 2020.

[259] Kaveri A. Thakoor, Sharath C. Koorathota, Donald C. Hood, and Paul Sajda. Robust and interpretable convolutional neural networks to detect glaucoma in optical coherence tomography images. *IEEE Transactions on Biomedical Engineering*, 68(8):2456–2466, 2021.

[260] Terry M. Therneau and Patricia M. Grambsch. Modeling survival data Extending the Cox model. In *Survival-Python package*, page 350. Springer, 2000.

[261] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. -, 5 2019.

[262] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[263] Shusaku Tsumoto. Mining diagnostic rules from clinical databases using rough sets and medical diagnostic model. *Information Sciences*, 162(2):65–80, 5 2004.

[264] Ryan Turner. A model explanation system. In *2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6. IEEE, 2016.

[265] Hamed Valizadegan, Quang Nguyen, and Milos Hauskrecht. Learning classification models from multiple experts. *Journal of Biomedical Informatics*, pages 1125–1135, 2013.

[266] Joost van Amersfoort, Lewis Smith, Andrew Jesson, Oscar Key, and Yarin Gal. On feature collapse and deep kernel learning for single forward pass uncertainty. *arXiv preprint arXiv:2102.11409*, 2021.

[267] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Uncertainty estimation using a single deep deterministic neural network. *International Conference on Machine Learning*, 2020.

[268] B van Ginneken, M B Stegmann, and M Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006.

[269] Bram van Ginneken, Arnaud A. A. Setio, Colin Jacobs, and Francesco Ciompi. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, pages 286–289, 2015.

[270] Arnaud Van Looveren and Janis Klaise. Interpretable Counterfactual Explanations Guided by Prototypes. *arXiv preprint arXiv:1907.02584*, 7 2019.

[271] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR, 2019.

[272] Anthony J Viera, Joanne M Garrett, and others. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363, 2005.

[273] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401, 2020.

[274] Claus F. Vogelmeier, Gerard J. Criner, Fernando J. Martinez, Antonio Anzueto, Peter J. Barnes, Jean Bourbeau, Bartolome R. Celli, Rongchang Chen, Marc Decramer, Leonardo M. Fabbri, Peter Frith, David M. G. Halpin, M. Victorina López Varela, Masaharu Nishimura, Nicolas Roche, Roberto Rodriguez-Roisin, Don D. Sin, Dave Singh, Robert Stockley, Jørgen Vestbo, Jadwiga A. Wedzicha, and Alvar Agustí. Global Strategy for the Diagnosis, Management, and Prevention of Chronic Obstructive Lung Disease 2017 Report. GOLD Executive Summary. *American Journal of Respiratory and Critical Care Medicine*, 195(5):557–582, 3 2017.

[275] Kentaro Wada. labelme Image Polygonal Annotation with Python. https://github.com/wkentaro/labelme, 2016.

[276] Fei Wang, Rainu Kaushal, and Dhruv Khullar. Should health care demand interpretable artificial intelligence or accept ”black Box” Medicine?, 1 2020.

[277] Haoran Wang, Weitang Liu, Alex Bocchieri, and Yixuan Li. Can multi-label classification networks know what they don’t know? *Advances in Neural Information Processing Systems*, 2021.

[278] Pei Wang and Nuno Vasconcelos. SCOUT: Self-Aware Discriminant Counterfactual Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6 2020.

[279] George R. Washko, Gary M. Hunninghake, Isis E. Fernandez, Mizuki Nishino, Yuka Okajima, Tsuneo Yamashiro, James C. Ross, Raúl San José Estépar, David A. Lynch, John M. Brehm, Katherine P. Andriole, Alejandro A. Diaz, Ramin Khorasani, Katherine D'Aco, Frank C. Sciurba, Edwin K. Silverman, Hiroto Hatabu, and Ivan O. Rosas. Lung Volumes and Emphysema in Smokers with Interstitial Lung Abnormalities. *New England Journal of Medicine*, 364(10):897–906, 3 2011.

[280] Andrew Gordon Wilson. The case for bayesian deep learning. *arXiv preprint arXiv:2001.10995*, 2020.

[281] Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. Association between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, 155(10):1135–1141, 10 2019.

[282] Tianfu Wu and Xi Song. Towards interpretable object detection by unfolding latent structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6033–6043, 2019.

[283] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv*, abs/1708.07747, 2017.

[284] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 842–850, 06 2015.

[285] Zhisheng Xiao, Qing Yan, and Yali Amit. Likelihood regret: An out-of-distribution detection score for variational auto-encoder. *ArXiv*, abs/2003.02977, 2020.

[286] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition*, 85:109–119, 2019.

[287] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, and et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. *International conference on machine learning*, 2 2015.

[288] Samir S Yadav and Shivajirao M Jadhav. Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, 6(1):1–18, 2019.

[289] Jie Yang, Elsa D. Angelini, Pallavi P. Balte, Eric A. Hoffman, and et al. Unsupervised Discovery of Spatially-Informed Lung Texture Patterns for Pulmonary Emphysema: The MESA COPD Study. In *MICCAI*, volume 10433, pages 116–124. Elsevier Inc., 9 2017.

[290] Hai Ye, Feng Gao, Youbing Yin, Danfeng Guo, Pengfei Zhao, Yi Lu, Xin Wang, Junjie Bai, Kunlin Cao, Qi Song, Heye Zhang, Wei Chen, Xuejun Guo, and Jun Xia. Precise diagnosis of intracranial hemorrhage and subtypes using a three-dimensional joint convolutional and recurrent neural network. *European Radiology*, 29, 04 2019.

[291] Hugo Yeche, Justin Harrison, and Tess Berthier. Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support*, pages 12–20. Springer, 2019.

[292] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*, 2019.

[293] Neelam Younas, Amjad Ali, Hafsa Hina, Muhammad Hamraz, Zardad Khan, and Saeed Aldahmani. Optimal causal decision trees ensemble for improved prediction and causal inference. *IEEE Access*, 10:13000–13011, 2022.

[294] Kyle Young, Gareth Booth, Becks Simpson, Reuben Dutton, and Sally Shrapnel. Deep neural network or dermatologist? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11797 LNCS, pages 48–55. Springer, 10 2019.

[295] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander Smola. Deep Sets. *Advances in neural information processing systems*, 3 2017.

[296] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization, 2014.

[297] Matthew D Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. *CoRR*, abs/1311.2901, 2013.

[298] Zhang, Zhifei, Song, Yang, and Qi Hairong. Age progression/regression by conditional adversarial autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[299] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[300] Zizhao Zhang, Yuanpu Xie, Fuyong Xing, Mason McGough, and Lin Yang. Mdnet: A semantically and visually interpretable medical image diagnosis network. In *CVPR*, pages 3549–3557, 07 2017.

[301] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13001–13008, 2020.

[302] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *Computing Research Repository*, 2014.

[303] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[304] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. Interpretable basis decomposition for visual explanation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11212 LNCS, pages 122–138. Springer Verlag, 2018.

[305] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.