Uncertainty-Driven Counterfactual Explainers for CXR-Based Diagnosis Models

Kowshik Thopalli[1] **Deepta Rajan**[2] **Pavan Turaga**[1] **Jayaraman J. Thiagarajan**[3]

Abstract

The rapid adoption of artificial intelligence methods in healthcare is coupled with the critical need for techniques to introspect models and to uncover relationships between discernible data signatures and model predictions. In this context, counterfactual explanations that synthesize small, interpretable changes to a given input while producing desired changes in model predictions have become popular. This under-constrained, inverse problem is vulnerable to introducing irrelevant manipulations, particularly when the predictions are not well-calibrated. Hence, in this paper, we present TraCE (Training Calibration-based Explainers), which utilizes a novel uncertainty-based interval calibration strategy for reliably synthesizing counterfactuals. Motivated by the successes of AI in radiology, our study focuses on deep models used for identifying anomalies in chest X-ray images. Using empirical studies, we demonstrate the superiority of TraCE explanations, in terms of several widely adopted evaluation metrics. We find that TraCE effectively enables progressive exploration of decision boundaries, to detect shortcuts, and to infer relationships between patient attributes and disease severity.

1. Introduction

With the growing interest in adopting artificial intelligence (AI) methods for critical decision-making, from diagnosing diseases to prescribing treatments (Faust et al., 2018; Kononenko, 2001; Miotto et al., 2018), it is imperative to ensure those methods are both accurate and reliable (Ching et al., 2018) so as to promote trust in these solutions and prioritize patient safety. This has strongly motivated the need to both reliably assess a model's confidence in its predictions (Guo et al., 2017; Leibig et al., 2017; Thiagarajan et al., 2020b), and to enable rigorous introspection of its behavior (Cabitza & Campagner, 2019; Ching et al., 2018; Tonekaboni et al., 2019; Thiagarajan et al., 2018).

To this end, uncertainty estimation methods are often adopted to determine the deficiencies of a model and/or the data (Gawlikowski et al., 2021). Meaningful uncertainties can play a crucial role in supporting practical objectives that range from assessing regimes of over (or under)-confidence and active data collection, to ultimately improving the predictive models. This paper investigates the benefit of leveraging uncertainties in designing explainability tools. While there has been a large body of recent work on interpretability (e.g., LIME (Ribeiro et al., 2016), ANCHORS (Ribeiro et al., 2018)), example-based methods that synthesize data samples in the vicinity of a query, such that the predictions for those samples align with a user-specified hypothesis are getting popular. Such examples are referred to as *counterfactual* explanations (Verma et al., 2020; Singla et al., 2019; Cohen et al., 2021; Narayanaswamy et al., 2021).

In its most generic form, for a given query x, one can pose counterfactual generation based on a predictive model F as:

$$\arg\min_{\bar{\mathrm{x}}} d(\mathrm{x}, \bar{\mathrm{x}}) \quad \text{s.t.} \quad \mathrm{F}(\bar{\mathrm{x}}) = \bar{\mathrm{y}};\ \bar{\mathrm{x}} \in M(X) \qquad (1)$$

where $\bar{\mathrm{x}}$ is a counterfactual explanation for the query x and $\bar{\mathrm{y}}$ is the user-specified hypothesis about $\bar{\mathrm{x}}$ (*e.g.*, a certain diagnosis). Minimizing a suitable discrepancy $d(.,.)$ between x and $\bar{\mathrm{x}}$ ensures that the underlying semantic content of x is preserved. Another important requirement is that the generated $\bar{\mathrm{x}}$ should lie close to the original data manifold $M(X)$. When no tractable priors exist for $M(X)$, it is common to perform this optimization in the latent space of a pre-trained generative model. Despite the effectiveness of such priors, when the model's predictions $\mathrm{F}(\bar{\mathrm{x}})$ are poorly calibrated, i.e., prediction confidences are not indicative of the actual likelihood of correctness (Kuleshov et al., 2018; Thiagarajan et al., 2020b), the optimization in eq. (1) can still lead to bad quality explanations. Though different variants of the formulation in eq. (1) have been considered in the literature (Verma et al., 2020), the fundamental challenge with

[1] Arizona State University [2] Microsoft [3] Lawrence Livermore National Laboratory. Correspondence to: Jayaraman J. Thiagarajan <jjayaram@llnl.gov>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2022 by the author(s).

This work was performed under the auspices of the U.S. Department of Energy by the Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, Lawrence Livermore National Security, LLC.

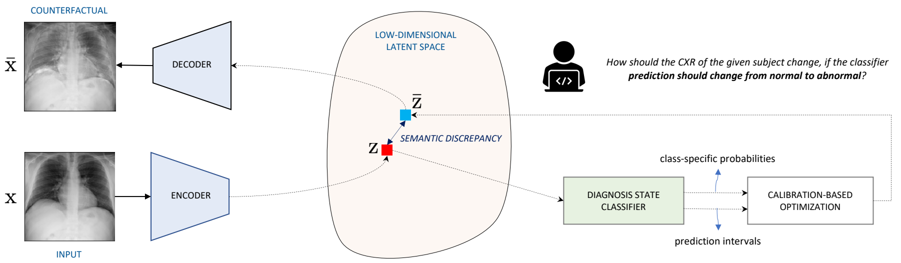

COUNTERFACTUAL
$\bar{x}$
DECODER
LOW-DIMENSIONAL LATENT SPACE
$\bar{z}$
SEMANTIC DISCREPANCY
Z
x
ENCODER
INPUT
How should the CXR of the given subject change, if the classifier prediction should change from normal to abnormal?
class-specific probabilities
DIAGNOSIS STATE CLASSIFIER
CALIBRATION-BASED OPTIMIZATION
prediction intervals

Figure 1. An overview of TraCE applied for introspective analysis of chest X-ray (CXR)-based predictive models. In this example, we consider a binary classifier that has been trained to distinguish between *normal* and *abnormal* subjects (*i.e.*, containing pneumonia-related anomalies). We first transform a query image x (from the *normal* class) into its latent representation Z using the *Encoder* and carry the optimization in latent space. Subsequently, we invoke the proposed calibration-driven optimization to obtain the counterfactual $\bar{z}$ in the latent space, such that the semantic discrepancy between z and $\bar{z}$ is minimized and the classifier's prediction changes to *abnormal*. Finally, the synthesized counterfactual $\bar{z}$ is transformed into the image-space ($\bar{x}$) using the *Decoder* network.

uncalibrated predictions still persists.

In this work, we present TraCE (*Training Calibration-based Explainers*) (Thiagarajan et al., 2022), that circumvents this challenge by integrating prediction uncertainties into counterfactual optimization. We outline the approach in Figure 1. While our approach is flexible to support the use of any uncertainty estimator or prediction models that use explicit calibration objectives, TraCE builds upon the recent Learn-by-Calibrating (LbC) technique (Thiagarajan et al., 2020a) to obtain prediction intervals for both classification and regression settings. LbC jointly trains an auxiliary interval estimator alongside the predictor model using an interval calibration objective. We first adapt LbC for multi-class classification problems and subsequently develop a counterfactual generation approach based on the estimated prediction intervals. Using empirical studies with CXR models, we demonstrate how TraCE can be utilized for progressive exploration of decision boundaries, detecting shortcuts and inferring attribute relationships

2. Background

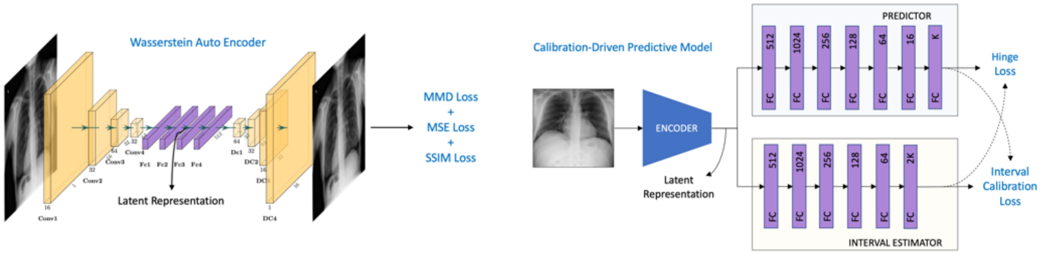
Uncertainty Estimation. The growing interest in employing machine learning (ML) based solutions to design diagnostic tools strongly emphasizes the need for a rigorous characterization of models. Uncertainty quantification (UQ) provides this characterization by studying the impact of different error sources on the prediction (Smith, 2013; Heskes, 1997; Kendall & Gal, 2017). Some of the popular uncertainty estimation methods include: (i) Bayesian neural networks (Blundell et al., 2015; Kendall & Gal, 2017): (ii) methods that use the discrepancy between different models as a proxy for uncertainty, such as deep ensembles (Lakshminarayanan et al., 2016) and Monte-Carlo dropout that approximates Bayesian posteriors on the weight-space of a model (Gal & Ghahramani, 2016); and (iii) approaches that use a single model to estimate uncertainties (Tagasovska & Lopez-Paz, 2018; Van Amersfoort et al., 2020; Liu et al., 2020; Krishnan & Tickoo, 2020; Jain et al., 2021).

Prediction Calibration. While uncertainties can be directly leveraged for a variety of downstream tasks including out-of-distribution detection and sequential sample selection, they have also been utilized for guiding models to produce well-calibrated predictions. For example, uncertainties from Monte-Carlo dropout (Seo et al., 2019) and direct error prediction (Thiagarajan et al., 2021) have been used to perform confidence calibration in classifiers. Similarly, the recently proposed Learn-by-Calibrating (LbC) approach (Thiagarajan et al., 2020a) introduced an interval calibration objective based on uncertainties for training deep regression models.

Counterfactual Generation. Counterfactual (CF) explanations (Verma et al., 2020) synthesize interpretable changes to a given image while producing desired changes in model predictions. An important requirement to produce meaningful counterfactuals is to produce discernible local perturbations while being realistic (close to the underlying data manifold). Consequently, existing approaches rely extensively on pre-trained generative models to synthesize plausible counterfactuals (Verma et al., 2020; Van Looveren & Klaise, 2019; Dhurandhar et al., 2018; Singla et al., 2019; Goyal et al., 2019). While the proposed TraCE framework also utilizes a pre-trained generative model, it fundamentally differs from existing approaches by employing uncertainty-based calibration for counterfactual optimization.

3. Methods

Constructing low-dimensional latent spaces. Our first step is to build a low-dimensional, continuous latent space that respects the true distribution, so that one can generate counterfactual representations in that space. While a large class of generative modeling methods exist to construct such a latent space, in this work, we focus on Wasserstein autoencoders (WAE) (Tolstikhin et al., 2018). WAE has been found to outperform other VAE formulations, particularly in image datasets with low heterogeneity, *e.g.*, scientific images from physics simulations (Anirudh et al., 2020). This



(a) (b)

Figure 2. Framework Design for TraCE. (a) First, we train WAE, an auto-encoding neural network and construct a low-dimensional, continuous latent space for CXR images. (b) Next, we adapt the Learn-by-Calibrating (Thiagarajan et al., 2019) approach to train a classifier that takes as input the latent representation from the encoder and outputs an patient-specific attribute and prediction intervals.

network is a composition of an encoder E that transforms the input x into its latent code z, and a decoder D that reconstructs the image. Additionally the encoder has the objective of matching the latent distribution of the training samples $\mathbb{E}_{P_X}[\mathrm{E}(\mathrm{z} \mid \mathrm{x})]$ to a pre-specified prior P_Z. This helps us generate new unseen samples from the original data manifold $M(X)$. Our WAE thus minimizes: 1) discrepancy D_x (mean squared error) between the original and the generated distributions; 2) discrepancy D_z (MMD) between the latent distribution of the encoded samples and the prior. As shown in Figure 2(a), we also find that including another loss term to maximize the structural similarity (SSIM) (Wang et al., 2004) between the original and reconstructed images led to higher quality reconstructions.

Predictive Model Design using LbC. We next adapt the recently proposed LbC approach (Thiagarajan et al., 2020a) to train classifier (or regressor) models that map from the CXR latent space to a desired target variable. By design, LbC provides prediction intervals in lieu of point estimates for the response y, i.e., $[\hat{\mathrm{y}} - \delta, \hat{\mathrm{y}} + \delta]$. Here, δ is used to define the interval. Suppose that the likelihood for the true response y to be contained in the prediction interval is $p(\hat{\mathrm{y}} - \delta \leq \mathrm{y} \leq \hat{\mathrm{y}} + \delta)$, the intervals are considered to be well-calibrated if the likelihood matches the expected confidence level.

Algorithm. The model is comprised of two modules F and G, implemented as neural networks, to produce estimates $\hat{\mathrm{y}} = \mathrm{F}(\mathrm{z})$ and $\delta = \mathrm{G}(\mathrm{z})$ respectively. For example, in the case of multi-class classification settings, $\hat{\mathrm{y}} \in \mathbb{R}^K$ is a vector of predicted logits for the K different classes. Since interval calibration is defined for continuous-valued targets, we adapt the loss function for training on the logits directly. To this end, we first transform the ground truth labels into logits. Note, for each sample, we allow a small non-zero probability (say 0.01) to all negative classes. Denoting the parameters of the models F and G by θ and ϕ respectively, we use an alternating optimization strategy similar to (Thiagarajan et al., 2020a). In order to update ϕ, we use the empirical interval calibration error as the objective:

$$\phi^* = \arg\min_{\phi} \sum_{k=1}^{K} \Bigg| \alpha - \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\Big[(\hat{\mathrm{y}}_i[k] - \delta_i[k]) \leq \mathrm{y}_i[k] \leq (\hat{\mathrm{y}}_i[k] + \delta_i[k]) \Big] \Bigg|, \qquad (2)$$

where $\delta_i = \mathrm{G}(\mathrm{z}_i; \phi)$, and the desired confidence level α (set to 0.9) is an input. When updating parameters ϕ, we assume that the estimator $\mathrm{F}(.; \theta)$ is known and fixed. Now, given the updated ϕ, we learn the parameters θ using the following hinge-loss objective:

$$\theta^* = \arg\min_{\theta} \sum_{k=1}^{K} \frac{1}{N} \sum_{i=1}^{N} \Bigg[\max\Big(0, \hat{\mathrm{y}}_\ell - \mathrm{y}_i[k] + \tau\Big) + \max\Big(0, \mathrm{y}_i[k] - \hat{\mathrm{y}}_h + \tau\Big) \Bigg], \qquad (3)$$

where $\hat{\mathrm{y}}_\ell = (\hat{\mathrm{y}}_i[k] - \delta_i[k])$, $\hat{\mathrm{y}}_h = (\hat{\mathrm{y}}_i[k] + \delta_i[k])$, $\hat{\mathrm{y}}_i = \mathrm{F}(\mathrm{z}_i; \theta)$ and τ is the margin parameter (set to 0.05). Intuitively, for a fixed ϕ, obtaining improved estimates for $\hat{\mathrm{y}}$ can increase the empirical calibration error in (2) by achieving higher likelihoods even for lower confidence levels. However, in the subsequent step of updating ϕ, we expect δ's to become sharper in order to reduce the calibration error. We repeat the two steps (eqns. (2) and (3)) until convergence.

Uncertainty-Aware Counterfactual Generation. TraCE modifies the counterfactual generation process in Eq. (1) using the pre-trained predictor and interval estimator models from LbC. Our goal is to generate explanations to support a given hypothesis on the target variable – for example emulating high-confidence disease states given the CXR of a healthy subject. To this end, we first obtain the latent representation for the given query image x using the encoder, $\mathrm{z} = \mathrm{E}(\mathrm{x})$. We then use the pre-trained predictor (F) and interval estimator (G) models to generate the counterfactual $\bar{\mathrm{z}}$. Finally, the generated counterfactuals in the latent space are

Table 1. Performance evaluation of diagnosis-based counterfactual explanations obtained using different approaches. In each case, we report results averaged across 500 test samples.

Method	Validity ↑	Confidence ↑	Sparsity ↓	Proximity ↓	Realism ↑
Vanilla	0.68	0.63±0.11	0.3±0.17	4.59±0.68	1.16 ± 0.09
Mixup	0.78	0.69±0.17	0.27±0.16	4.09±0.52	1.19 ± 0.13
UWCC	0.79	0.75±0.13	0.25±0.17	4.26±0.63	1.16 ± 0.2
MC Dropout	0.73	0.66±0.16	0.34±0.19	4.57±0.53	1.18 ± 0.16
Deep Ensembles (5 models)	0.8	0.72±0.09	0.29±0.11	**3.68±0.57**	1.21 ± 0.12
TraCE	**0.87**	**0.81±0.12**	**0.23±0.14**	3.73±0.51	**1.33 ± 0.13**

passed to the decoder to obtain a reconstruction in the image space, $\bar{\mathrm{x}} = \mathrm{D}(\bar{\mathrm{z}})$. We propose the following optimization to generate the counterfactual explanations:

$$\bar{\mathrm{z}} = \arg\min_{\hat{\mathrm{z}}} \eta_1 \|\mathrm{z} - \hat{\mathrm{z}}\|_2^2 + \eta_2 \mathcal{L}(\hat{\mathrm{y}}, \delta, \bar{\mathrm{y}}) + \eta_3 \delta, \qquad (4)$$

where $\hat{\mathrm{y}} = \mathrm{F}(\hat{\mathrm{z}}), \delta = \mathrm{G}(\hat{\mathrm{z}})$, $\bar{\mathrm{y}}$ is the desired value for the target attribute (hypothesis), η_1, η_2, η_3 are hyper-parameters for weighting the different terms. The first term ensures that the generated counterfactual is in the vicinity of the query sample x (in the latent space). The second term ensures that the expected target value is contained in the prediction interval (calibration), while the final term penalizes arbitrarily large intervals to avoid trivial solutions. The calibration objective $\mathcal{L}$ is implemented as a hinge-loss term:

$$\mathcal{L}(\hat{\mathrm{y}}, \delta, \bar{\mathrm{y}}) = \Big[\max\Big(0, (\hat{\mathrm{y}} - \delta) - \bar{\mathrm{y}} + \tau\Big) + \max\Big(0, \bar{\mathrm{y}} - (\hat{\mathrm{y}} + \delta) + \tau\Big)\Big], \qquad (5)$$

where the margin was fixed at $\tau = 0.05$. Choosing η_1, η_2, η_3 is essential to controlling the discrepancy between z and $\bar{\mathrm{z}}$, and ensuring that the prediction for the counterfactual is $\bar{\mathrm{y}}$.

Baselines. We considered a suite of baseline approaches for our empirical study and they differ by the strategies used for training the classifier, and counterfactual optimization. Note, all methods use the same WAE latent space.

(i) *Vanilla*: In this approach, we train the classifier with no explicit calibration or uncertainty estimation, and use the following formulation to generate the counterfactuals:

$$\bar{\mathrm{z}} = \arg\min_{\hat{\mathrm{z}}} \eta_1 \|\mathrm{z} - \hat{\mathrm{z}}\|_2^2 + \eta_2 \mathcal{L}_{ce}\left[\mathrm{F}(\hat{\mathrm{z}}), \bar{\mathrm{y}}\right], \qquad (6)$$

where $\mathcal{L}_{ce}$ denotes the cross entropy loss.

(ii) *Mixup*: This is a popular augmentation strategy (Zhang et al., 2017) and it was found recently in (Thulasidasan et al., 2019) that mixup regularization improved calibration in the resulting model. Since this approach does not produce any uncertainty estimates, the counterfactual optimization is same as the *Vanilla* approach.

(iii) *MC Dropout*: Here, we train the classifier with dropout regularization and estimate the (epistemic) prediction uncertainty for any test sample by running multiple forward passes. Finally, we use the following heteroscedastic regression objective to implement uncertainty-based calibration during counterfactual optimization:

$$\bar{\mathrm{z}} = \arg\min_{\hat{\mathrm{z}}} \eta_1 \|\mathrm{z} - \hat{\mathrm{z}}\|_2^2 + \eta_2 \left[\frac{(\bar{\mathrm{y}} - \mu_{\hat{\mathrm{z}}})^2}{2\sigma_{\hat{\mathrm{z}}}^2} + \frac{1}{2}\log(\sigma_{\hat{\mathrm{z}}}^2)\right]. \qquad (7)$$

Note, the mean/variance estimates $(\mu_{\hat{\mathrm{z}}}, \sigma_{\hat{\mathrm{z}}}^2)$ are obtained using T (set to 10) forward passes with dropout.

(iv) *Deep Ensembles*: In this popular and currently one of the best uncertainty estimation techniques approach, we independently train M different models (with bootstrapping and different model initializations) with the same architecture. We employ the calibration objective in Eq. (7) to perform counterfactual optimization with μ and σ computed from the predictions of these M models.

(v) *Uncertainty-Weighted Confidence Calibration (UWCC)*: (Seo et al., 2019) proposed to build calibrated classification models by augmenting a confidence-calibration term to the standard cross-entropy loss and weighting the two terms using the uncertainty measured via multiple stochastic inferences. Mathematically,

$$\frac{1}{N}\sum_{i=1}^{N} - (1 - \alpha_i)\log(P(\hat{\mathrm{y}}_i|\mathrm{z}_i)) + \alpha_i D_{KL}(\mathrm{U}(\mathrm{y})||P(\hat{\mathrm{y}}_i|\mathrm{z}_i)). \qquad (8)$$

Here the first term denotes the cross-entropy loss, and the predictions $P(\hat{\mathrm{y}}_i|\mathrm{z}_i)$ are inferred using stochastic inferences for z_i, while the variance (α_i) in the predictions is used to balance the loss terms. Since the model is inherently calibrated during training, we do not measure the uncertainties at test time and hence use the optimization in Eq. (6) for generating counterfactuals.

4. Results and Findings

Data. Our study uses the *RSNA pneumonia detection challenge database*, which is a collection of $30,000$ CXR exams belonging to the NIH CXR14 benchmark dataset (Wang et al., 2017), of which $15,000$ exams show evidence for

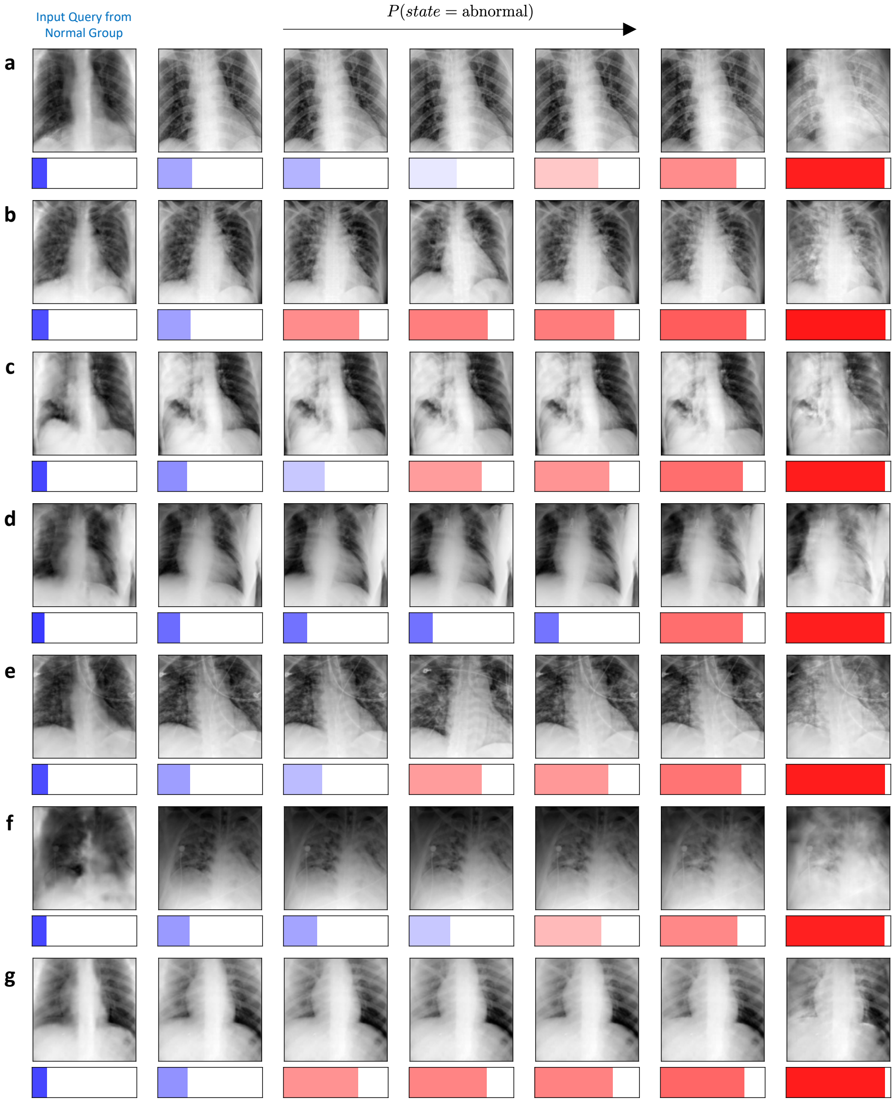


Figure 3. (a)-(g) Diagnosis-based counterfactual explanations generated using TraCE by progressively introducing relevant patterns into different query images (first image in each row) of healthy subjects to increase the likelihood of being assigned to the *abnormal* group.

lung opacities related to pneumonia, consolidation and infiltration, and $7,500$ exams contain no findings (referred as *normal*). The CXR images in the dataset were annotated by six board-certified radiologists and additional information on the data curation process can be found in (Stein, 2018). In addition to the diagnostic labels, this dataset contains age and gender information of the subjects. Note that, for this analysis, we used healthy control subjects from the RSNA pneumonia dataset to define the *normal* group and designed predictive models to discriminate them from patients presenting pneumonia-related anomalies in their CXR scans. We refer to the latter as the *abnormal* group.

Evaluation Metrics. We used the following metrics for a holistic evaluation of the counterfactual explanations:

(i) *Validity*: For categorical attributes this metric measures the ratio of the counterfactuals that actually have the desired target attribute to the total number of counterfactuals generated (higher the better) and for continuous-valued targets we measure MAPE (lower the better).

$$V_{cat} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\mathrm{F}(\mathrm{x}_i), \bar{\mathrm{y}}_i); \quad V_{cont} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\bar{\mathrm{y}}_i - \mathrm{F}(\mathrm{x}_i)}{\bar{\mathrm{y}}_i}\right|,$$

where $\mathbb{I}$ denotes the identity function.

(ii) *Confidence*: In cases of categorical-valued targets (class labels), we compute the confidence $P(\bar{\mathrm{y}}_i|\bar{\mathrm{x}}_i; \mathrm{F})$ (from softmax probabilities) of assigning the desired class $\bar{\mathrm{y}}_i$ for a counterfactual $\bar{\mathrm{x}}_i$ (higher the better).

(iii) *Sparsity*: We compute the sparsity metric as the ratio

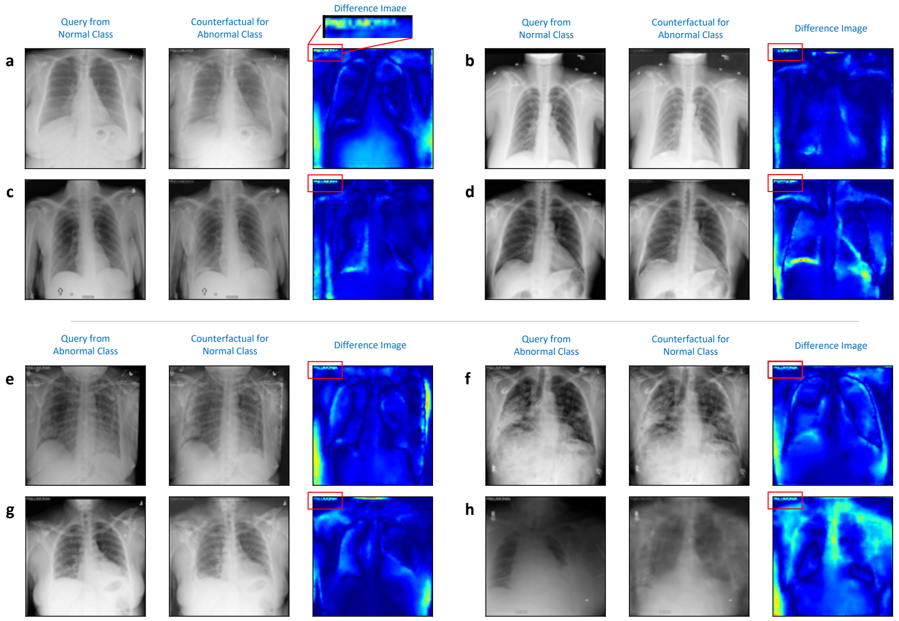

Query from Normal Class
Counterfactual for Abnormal Class
Difference Image
a
b
c
d
Query from Abnormal Class
Counterfactual for Normal Class
e
f
g
h

Figure 4. Using TraCE to detect shortcuts in deep predictive models. In this experiment, we synthetically introduced a nuisance feature (overlaid the text PNEUMONIA in the top-left corner) into all images from the *abnormal* group, and used this data to train the predictive model. Given the entirely data-driven nature of machine-learned solutions, there is risk of inferring a decision rule based on this irrelevant feature in order to discriminate between *normal* and *abnormal* groups. (a-d) Here, we used randomly chosen query images from the *normal* class and generated counterfactuals for the *abnormal* class. In each case, we show the query image, the counterfactual explanation from TraCE and the absolute difference image between the two; (e-f) Here, we introduced the nuisance feature into CXR images from the *abnormal* group and synthesized counterfactuals for the *normal* class. We observe that TraCE can effectively detect such shortcuts – counterfactuals for changing the diagnosis state are predominantly based on manipulating the text on the top-left corner.

Table 2. Performance evaluation of age-based counterfactual explanations obtained using different approaches. In each case, we report results averaged across 500 test samples.

Method	Validity ↓	Sparsity ↓	Proximity ↓	Realism ↑
Vanilla	2.49	0.06±0.08	4.08±0.48	1.26±0.1
Mixup	0.83	**0.05±0.07**	3.79±0.52	1.28±0.07
UWCC	0.74	0.09±0.03	3.81±0.42	1.33±0.05
MC Dropout	1.44	0.07±0.08	4.13±0.29	1.26±0.06
Deep Ensembles (5 models)	0.45	**0.05±0.09**	3.89±0.32	1.32±0.06
TraCE	**0.16**	**0.05±0.03**	**3.66±0.35**	**1.38 ± 0.06**

of the number of pixels altered to the total number of pixels. In general, sparser changes to an image are more likely to preserve the inherent characteristics of the query image. (iv) *Proximity*: Recent works have considered the actionability of modified features by grounding them in the training data distribution. Following (Dandl et al., 2020), we measure the average ℓ_2 distance of each counterfactual to the K-nearest samples in the latent space (lower the better). (v) *Realism score*: We also employ the realism score metric from the from the generative modeling literature (Sajjadi et al., 2018), introduced in (Kynkäänniemi et al., 2019) to evaluate the quality of images obtained using TraCE.

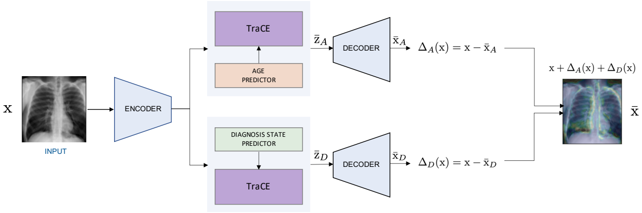

x
INPUT
ENCODER
TraCE
AGE PREDICTOR
DIAGNOSIS STATE PREDICTOR
TraCE
$\bar{z}_A$
$\bar{z}_D$
DECODER
DECODER
$\bar{x}_A$
$\bar{x}_D$
$\Delta_A(x) = x - \bar{x}_A$
$\Delta_D(x) = x - \bar{x}_D$
$x + \Delta_A(x) + \Delta_D(x)$
$\bar{x}$

Figure 5. Using TraCE to infer relationships between a patient attribute (*e.g.*, age) and disease states. For this analysis, we construct two independent predictive models, *i.e.*, age and diagnosis state, and synthesize counterfactuals based on hypothesis on each of the predictions (*e.g.*, patient age should be predicted as 70 while the diagnosis state should be *abnormal*.

TraCE Enables Progressive Exploration of Decision Boundaries. In this study, we analyzed a predictive model that classifies CXR images into *normal* and *abnormal* groups, and used TraCE to synthesize counterfactuals for a given query image from the *normal* class to visualize the progression of disease severity. Such an analysis can reveal what image signatures are introduced by a predictive model to provide evidence for the *abnormal* class, and can be used by practitioners to verify if the model relies on meaningful decision rules or *shortcuts* (*e.g.,* changes to the background) that cannot generalize. In our implementation of TraCE,

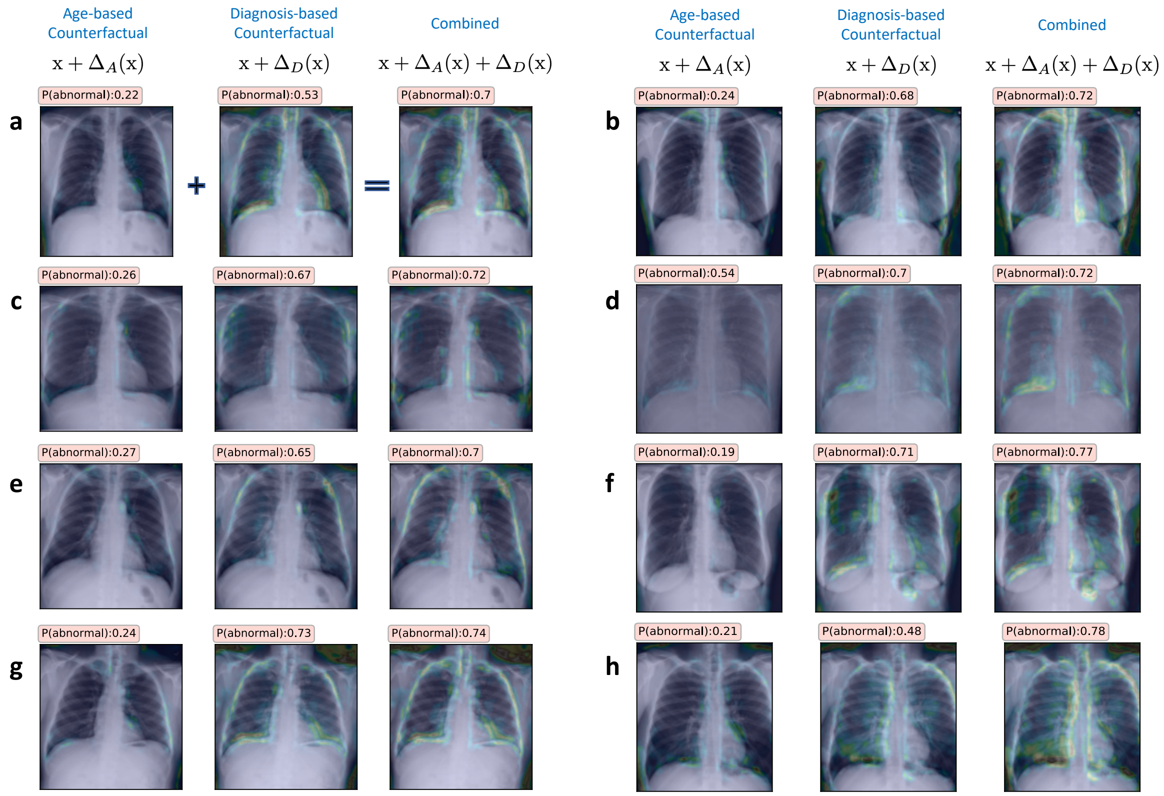


Figure 6. (a-h) Explanations generated using TraCE by introducing age-specific attributes into the counterfactuals synthesized for changing the diagnosis state of a *normal* subject to be *abnormal*. Interestingly, we find that there exists a correlation between the two attributes, as evidenced by the consistent increase in the likelihood $P(state = abnormal)$ when compared to counterfactuals that rely only on patterns from the diagnosis state predictor. In each case, we highlight the changes $\Delta_A(\mathrm{x}), \Delta_D(\mathrm{x}), (\Delta_A(\mathrm{x}) + \Delta_D(\mathrm{x}))$ and display the likelihood $P(state = abnormal)$.

we first trained the WAE on CXR images and set the latent space dimension to 100. We subsequently learned the predictive model F_D along with the interval estimator G_D, using the LbC algorithm. The hyper-parameters η_1 and η_2 in Eq. (5) are critical to trade-off between preserving the inherent semantics from query x and achieving the desired prediction. Hence, one can progressively transition from the *normal* to the *abnormal* class by fixing η_2 and gradually relaxing η_1.

Figure 6 illustrates the counterfactuals obtained using TraCE for multiple different examples from our benchmark dataset. More specifically, the query samples x correspond to CXR images from the *normal* class and we varied η_1 between 0.5 and 0.05, while setting $\eta_2 = 0.5$ and $\eta_3 = 0.2$. These values were obtained using a standard hyper-parameter search based on 500 randomly chosen images. For each case from Figure **??**, the different counterfactuals along with their estimated $P(state = abnormal)$ from the predictive model F_D are shown. It can be clearly observed from the results that the counterfactuals show increased opacity in the lung regions (appearing as denser white clouds) as we progress towards the *abnormal* class, which strongly corroborates with existing studies on CXR-based image analysis. By producing physically plausible evidences for crucial hypotheses, TraCE enables practitioners to effectively explore complex decision boundaries learned by deep predictive models.

Comparing TraCE to Baseline Methods. In order to perform a quantitative evaluation of TraCE, we obtained counterfactuals for 500 randomly chosen images from a held-out test set and Table 1 presents a detailed comparison of different baseline methods (see Methods section for details). The first striking observation is that, despite using the same pre-trained latent space for counterfactual optimization, all methods that incorporate explicit calibration strategies or uncertainty estimation consistently outperform the *Vanilla* model. More specifically, for similar levels of discrepancy in the latent space, TraCE achieves a significantly higher validity score of 0.88 as opposed to 0.69 of the *Vanilla* model, while inducing similar or lower amount of changes to the query (indicated by the sparsity and proximity metrics). Furthermore, our approach outperforms the results obtained

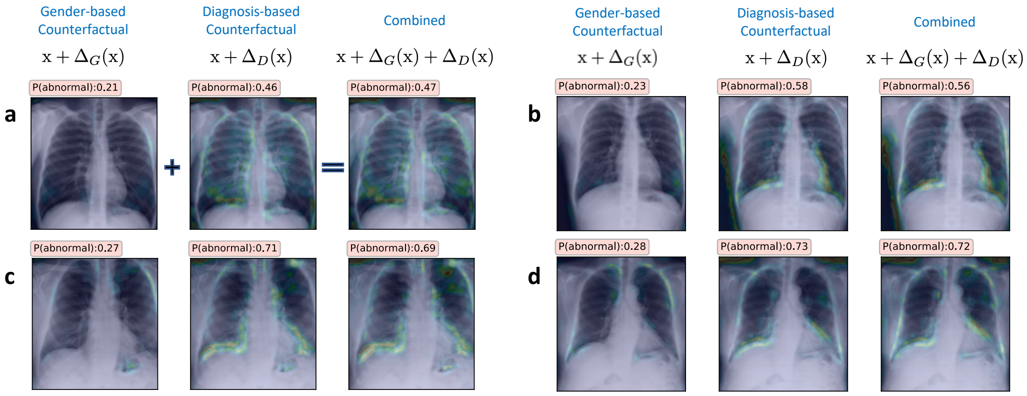


Figure 7. (a-d) Explanations generated using TraCE by introducing gender-specific attributes into the counterfactuals synthesized for changing the diagnosis state of a *normal* subject to be *abnormal*. In contrast to the age attribute, image manipulations associated with change in gender (female → male) do not cause any apparent change to the likelihood of being assigned to the *abnormal* group. In each case, we highlight the changes $\Delta_A(x), \Delta_D(x), \Delta_A(x) + \Delta_D(x)$ and show the likelihood $P(state = abnormal)$.

with state-of-the-art uncertainty estimators and calibration strategies (in all the metrics), thus demonstrating its efficacy in generating counterfactual explanations.

To demonstrate TraCE's applicability for the case of continuous-valued targets, we considered only healthy control subjects from the RSNA dataset and designed a regressor to estimate their age attribute using their CXR images. For our evaluation, we used 500 randomly chosen test subjects whose age attribute was between 40 and 70 and set the desired value $\bar{y} = 20$. From Table 2, we notice that the proposed approach achieves lower validity (MAPE) scores, without compromising on the proximity metric, when compared to the other baselines. Interestingly, we find that changing the age attribute required the manipulation of much lesser number of pixels (low sparsity values) when compared to the diagnosis state.

TraCE Detects Shortcuts in Deep Models. In order to demonstrate the use of TraCE in detecting shortcuts in purely data-driven models, we synthetically introduced a *nuisance* feature into images from the *abnormal* class – overlaid the text PNEUMONIA in the top-left corner of each image, and used TraCE to check if the model's decision was based on this nuisance feature. After training the WAE and the LbC model using the altered images, we selected query images from the *normal* group and generated the corresponding counterfactual evidences for the *abnormal* group. As illustrated in Figure 4(a-d), TraCE exclusively manipulates the top-left corner to accumulate evidence for abnormality, thus revealing that the predictive model relies on the nuisance feature. Similarly, in Figure 4(e-h), one can transition from the *abnormal* (examples containing the nuisance feature) to the *normal* group by simply removing the synthetic text PNEUMONIA. This experiment clearly emphasizes the utility of TraCE in detecting model and data biases.

TraCE Reveals Attribute Relationships. We next explored how counterfactual optimization can be used to study relationships between patient attributes, such as age and gender, and the diagnosis state. First, we study if the image signatures pertinent to the patient age attribute provides additional evidence for diagnosis state prediction. Given the age predictor, along with its interval estimator, (F_A, G_A) and the diagnosis predictor (F_D, G_D), we constructed counterfactuals based on two independent hypotheses. Note, both predictors were constructed based on the same low-dimensional latent representations. More specifically, we provided the hypotheses $\bar{y}_A = 70$ and $\bar{y}_D = abnormal$ for the two cases, and used TraCE to generate counterfactuals $\bar{x}_A$ and $\bar{x}_D$ that adhere to our hypotheses. We then estimated the age-specific and diagnosis-specific signatures introduced by TraCE:

$$\Delta_A(x) = x - \bar{x}_A; \quad \Delta_D(x) = x - \bar{x}_D. \tag{9}$$

In order to check if there exists an apparent relationship between age and diagnosis state, we generated the hybrid counterfactual,

$$\bar{x} = x + \Delta_A(x) + \Delta_D(x). \tag{10}$$

Finally, we compared $F_D(\bar{x}) - F_D(\bar{x}_D)$ to quantify if incorporating age-specific features into $\bar{x}_D$ increased the disease severity (*i.e.*, likelihood of being assigned to the *abnormal* class). Overview of this strategy is illustrated in Figure 5.

Figure 6 shows the results for 8 different *normal* subjects, wherein we find that there is an apparent increase in $P(state = abnormal)$ when age-specific signatures are incorporated. Using 500 randomly chosen *normal* subjects, we estimated an average change of 0.09 ± 0.08 in $\mathrm{F}_D(\bar{\mathrm{x}}) - \mathrm{F}_D(\bar{\mathrm{x}}_D)$, thus indicating that the diagnosis predictor is sensitive to age-specific patterns. In practice, if such a relationship is expected, it is a strong validation for the model's behavior. On the other hand, if the attribute is a confounding variable, it becomes critical to retrain the model wherein this sensitivity is explicitly discouraged. Interestingly, when we repeated this analysis with the gender attribute, such a relationship was not apparent (see results in Figure 7).

References

Anirudh, R., Thiagarajan, J. J., Bremer, P.-T., and Spears, B. K. Improved surrogates in inertial confinement fusion with manifold and cycle consistencies. *Proceedings of the National Academy of Sciences*, 117(18):9741–9746, 2020. ISSN 0027-8424.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pp. 1613–1622. PMLR, 2015.

Cabitza, F. and Campagner, A. Who wants accurate models? arguing for a different metrics to take classification models seriously. *arXiv preprint arXiv:1910.09246*, 2019.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M. M., et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of The Royal Society Interface*, 15 (141):20170387, 2018.

Cohen, J. P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M. P., and Chaudhari, A. Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest x-rays. In *Medical Imaging with Deep Learning*, 2021.

Dandl, S., Molnar, C., Binder, M., and Bischl, B. Multi-objective counterfactual explanations. In Bäck, T., Preuss, M., Deutz, A., Wang, H., Doerr, C., Emmerich, M., and Trautmann, H. (eds.), *Parallel Problem Solving from Nature – PPSN XVI*, pp. 448–469. Springer International Publishing, 2020.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., and Das, P. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. *arXiv preprint arXiv:1802.07623*, 2018.

Faust, O., Hagiwara, Y., Hong, T. J., Lih, O. S., and Acharya, U. R. Deep learning for healthcare applications based on physiological signals: A review. *Computer methods and programs in biomedicine*, 161:1–13, 2018.

Gal, Y. and Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pp. 1050–1059, 2016.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., and Lee, S. Counterfactual visual explanations. In *International Conference on Machine Learning*, pp. 2376–2384. PMLR, 2019.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1321–1330. JMLR. org, 2017.

Heskes, T. Practical confidence and prediction intervals. In *Advances in neural information processing systems*, pp. 176–182, 1997.

Jain, M., Lahlou, S., Nekoei, H., Butoi, V., Bertin, P., Rector-Brooks, J., Korablyov, M., and Bengio, Y. Deup: Direct epistemic uncertainty prediction. *arXiv preprint arXiv:2102.08501*, 2021.

Kendall, A. and Gal, Y. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, pp. 5574–5584, 2017.

Kononenko, I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

Krishnan, R. and Tickoo, O. Improving model calibration with accuracy versus uncertainty optimization. *arXiv preprint arXiv:2012.07923*, 2020.

Kuleshov, V., Fenner, N., and Ermon, S. Accurate uncertainties for deep learning using calibrated regression. *arXiv preprint arXiv:1807.00263*, 2018.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., and Aila, T. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32:3927–3936, 2019.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. *arXiv preprint arXiv:1612.01474*, 2016.

Leibig, C., Allken, V., Ayhan, M. S., Berens, P., and Wahl, S. Leveraging uncertainty information from deep neural networks for disease detection. *Scientific reports*, 7(1): 1–14, 2017.

Liu, J. Z., Lin, Z., Padhy, S., Tran, D., Bedrax-Weiss, T., and Lakshminarayanan, B. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. *arXiv preprint arXiv:2006.10108*, 2020.

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

Narayanaswamy, V., Thiagarajan, J. J., and Spanias, A. Using deep image priors to generate counterfactual explanations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2770–2774. IEEE, 2021.

Ribeiro, M. T., Singh, S., and Guestrin, C. ”why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Sajjadi, M. S., Bachem, O., Lucic, M., Bousquet, O., and Gelly, S. Assessing generative models via precision and recall. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 5234–5243, 2018.

Seo, S., Seo, P. H., and Han, B. Learning for single-shot confidence calibration in deep neural networks through stochastic inferences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9030–9038, 2019.

Singla, S., Pollack, B., Chen, J., and Batmanghelich, K. Explanation by progressive exaggeration. In *International Conference on Learning Representations*, 2019.

Smith, R. C. *Uncertainty quantification: theory, implementation, and applications*, volume 12. Siam, 2013.

Stein, A. Pneumonia dataset annotation methods, 2018. https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/discussion/64723.

Tagasovska, N. and Lopez-Paz, D. Single-model uncertainties for deep learning. *arXiv preprint arXiv:1811.00908*, 2018.

Thiagarajan, J. J., Rajan, D., and Sattigeri, P. Understanding behavior of clinical models under domain shifts. *arXiv preprint arXiv:1809.07806*, 2018.

Thiagarajan, J. J., Venkatesh, B., and Rajan, D. Learn-by-calibrating: Using calibration as a training objective. In *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019.

Thiagarajan, J. J., Venkatesh, B., Anirudh, R., Bremer, P.-T., Gaffney, J., Anderson, G., and Spears, B. Designing accurate emulators for scientific processes using calibration-driven deep models. *Nature Communications*, 11(1), 2020a.

Thiagarajan, J. J., Venkatesh, B., Rajan, D., and Sattigeri, P. Improving reliability of clinical models using prediction calibration. In *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis*, pp. 71–80. Springer, 2020b.

Thiagarajan, J. J., Narayanaswamy, V., Anirudh, R., Bremer, P.-T., and Spanias, A. Accurate and robust feature importance estimation under distribution shifts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7891–7898, 2021.

Thiagarajan, J. J., Thopalli, K., Rajan, D., and Turaga, P. Training calibration-based counterfactual explainers for deep learning models in medical image analysis. *Scientific Reports*, 12(1):1–15, 2022.

Thulasidasan, S., Chennupati, G., Bilmes, J. A., Bhattacharya, T., and Michalak, S. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 13888–13899, 2019.

Tolstikhin, I. O., Bousquet, O., Gelly, S., and Schölkopf, B. Wasserstein auto-encoders. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=HkL7n1-0b.

Tonekaboni, S., Joshi, S., McCradden, M. D., and Goldenberg, A. What clinicians want: contextualizing explainable machine learning for clinical end use. *arXiv preprint arXiv:1905.05134*, 2019.

Van Amersfoort, J., Smith, L., Teh, Y. W., and Gal, Y. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, pp. 9690–9700. PMLR, 2020.

Van Looveren, A. and Klaise, J. Interpretable counterfactual explanations guided by prototypes. *arXiv preprint arXiv:1907.02584*, 2019.

Verma, S., Dickerson, J., and Hines, K. Counterfactual explanations for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2020.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, 2017.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Zhang, H., Cisse, M., Dauphin, Y. N., and Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.