Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation

Julius Adebayo
MIT CSAIL

Michael Muelly
Stanford

Hal Abelson
MIT CSAIL

Been Kim
Google Research

Abstract

We investigate whether three types of post hoc model explanations–feature attribution, concept activation, and training point ranking–are effective for detecting a model's reliance on spurious signals in the training data. Specifically, we consider the scenario where the spurious signal to be detected is unknown, at test-time, to the user of the explanation method. We design an empirical methodology that uses semi-synthetic datasets along with pre-specified spurious artifacts to obtain models that verifiably rely on these spurious training signals. We then provide a suite of metrics that assess an explanation method's reliability for spurious signal detection under various conditions. We find that the post hoc explanation methods tested are ineffective when the spurious artifact is unknown at test-time especially for non-visible artefacts like a background blur. Further, we find that feature attribution methods are susceptible to erroneously indicating dependence on spurious signals even when the model being explained does not rely on spurious artefacts. This finding casts doubt on the utility of these approaches, in the hands of a practitioner, for detecting a model's reliance on spurious signals.[1]

> *It is hard to find a needle in a haystack,*
> *it is much harder if you haven't seen a needle before (Pearl).*
> —Judea Pearl

1 Introduction

A challenge that precludes the deployment of modern deep neural networks (DNN) in high stakes domains is their tendency to latch onto 'spurious signals'—shortcuts—in the training data (Geirhos et al., 2020). For example, Badgeley et al. (2019) showed that an Inception-V3 model trained to detect hip fracture relied on the scanner type for its classification decision. Deep learning models easily base output predictions on object backgrounds (Xiao et al., 2020), image texture (Geirhos et al., 2018), and skin tone (Stock and Cisse, 2018).

Post hoc model explanation methods—approaches that give insight into the associations that a model has learned—are increasingly used to determine whether a model relies on spurious signals. Ribeiro et al. (2016) used LIME to show an Inception-V3 model's reliance on the snow background for identifying Wolves. Such demonstration and others (Lapuschkin et al., 2019; Rieger et al., 2020) point to post hoc explanation methods as effective tools for the spurious signal detection task. However, these results conflict with evidence that indicates that practitioners (and researchers) struggle to use explanations to identify spurious signals (Chen et al., 2021; Chu et al., 2020; Alqaraawi et al., 2020; Adebayo et al., 2020; Poursabzi-Sangdeh et al., 2018). We seek to resolve this conflict by answering the simple but important question:

Can post hoc explanations help detect a model's reliance on ***unknown*** *spurious training signal?*

Motivating Example. Consider a machine learning (ML) engineer whose job is to train DNN models to detect knee arthritis from radiographs. She—the engineer—is handed a trained ResNet-50 model, to be deployed in a hospital, that relies on a hospital tag in the radiographs to detect knee arthritis. She

[1] We refer readers to the longer version of this work on the arxiv. Code to replicate our findings is available at: https://github.com/adebayoj/posthocspurious

has **no prior knowledge** of the model's reliance on the spurious tags. In this work, our key concern is whether the ML engineer can use post hoc explanations to identify that the model is defective.

1.1 Our Contributions

We address the motivating question above in a two-pronged manner. First, we develop an actionable methodology based on the ability to carefully craft datasets to induce spurious correlation in trained models. Second, we backup this experimental design with a human subject study. Taken together, the takeaway of the work is that: *post hoc explanations can be used to identify a model's reliance on a **visible** spurious signal, provided the signal is **known** ahead of time by the practitioner.* While this conclusion may seem unsurprising, it has important implications for how post hoc explanation methods should be used effectively.

Experimental Design & Performance Measures. We provide an end-to-end experimental design for assessing the effectiveness of an explanation method for detecting a model's reliance on spurious training signals. We define a spurious score that helps quantify the strength of a model's dependence on a training signal. Using carefully crafted semi-synthetic datasets, we are able to train models where the ground-truth expected explanation is known. Additionally, we develop 3 performance measures: i) Known Spurious Signal Detection Measure (K-SSD), ii) Cause-for-Concern Measure (CCM), and iii) a False Alarm Measure (FAM). These measures help characterize different notions of reliability for the spurious signal detection task. We instantiate the proposed design on 3 classes of post hoc explanation types—feature attribution, concept activation, and training point ranking—where we comprehensively assess the performance of these approaches across 3 datasets (2 medical tasks, and dog species classification task), and different model architectures.

When the spurious signal is known, we find that the feature attribution methods tested, and the concept activation importance approach are able to detect visible spurious signals like a text tag and distinctive stripped patterns. However, we find these approaches less effective for non-pronounced signals like background blur. The false alarm measure further indicates that feature attribution methods are susceptible to erroneously indicating dependence on spurious signals.

The cause-for-concern measure quantifies the similarity between explanations of 'normal' inputs derived from spurious and normal models when the spurious signal is unknown. Across the settings considered, we find that the methods tested are unable to conclusively detect model reliance on unknown spurious signals.

Blinded Study. The findings from our empirical assessment question the reliability of the methods tested; however, it might not correlate with utility in the hands of practitioners. To address this issue, we conduct a user study where practitioners are randomly assigned to one of two groups: the first group is told explicitly of potential spurious signals, and the second is not. We consider three different kinds of explanation methods along with a control where only model predictions are shown. We find that when participants are not provided with prior knowledge of the spurious signal, none of the methods tested are effective, in the hands of the participants, for detecting model reliance on spurious signals. More surprisingly, even when the participants had prior knowledge of the spurious signal, we find evidence that only the concept activation approach, for visible spurious signals, is effective. These findings cast doubt on the reliability of current post hoc tools for spurious signal detection.

Guidance. On the basis our analysis, we can provide the following guidance for using the approaches tested, in this work, for detecting model reliance on spurious signals **when the signal of interest is visible**:

- **Feature Attributions**: to confirm that a model is relying on a 'visible' spurious signal, the practitioner needs to obtain attributions for inputs that contain the hypothesized spurious signal, and the attribution should be computed for the output class to which the spurious signal is aligned.
- **Concept Activation**: the spurious concept should be known ahead of time, and tested against the output class to which the concept is aligned.
- **Training Point Ranking**: an input that contains the hypothesized spurious signal of interest should be used at test-time in computing training point ranking.

1.2 Related Work

This paper belongs to a line of work on assessing the effectiveness of post hoc explanations methods (Alqaraawi et al., 2020; Adebayo et al., 2020; Chu et al., 2020; Hooker et al., 2019; Meng et al., 2018; Poursabzi-Sangdeh et al., 2018; Tomsett et al., 2020). Here we focus on directly relevant literature, and defer an extensive discussion of the literature to the Appendix.

This work departs from previous work in two ways: 1) we focus exclusively on whether these explanations can be used by a practitioner (or researcher) to detect spurious signals that are unknown to her at test-time, and 2) we move beyond sole focus on the feature attribution setting to test concept activation and training point ranking methods.

Han et al. (2020) and Adebayo et al. (2020) find that certain kinds of feature attributions and training point ranking via influence functions are able to detect a model's reliance on spurious signals. However, in their setting, the spurious signal is known ahead of time. More recently, Zhou et al. (2021) conduct an extensive assessment of several feature attribution methods also under the spurious correlation setting, for visible and non-visible artifacts, and find that these class of methods are not effective for non-visible artifacts. Further, they also propose an experimental methodology for controlling model dependence on training set features, which allows them to quantify attribution effectiveness carefully and precisely. Overall, our findings align with theirs; however, we focus, specifically, on the setting where the spurious signal is not known ahead of time.

Kim et al. (2021) conduct an assessment of several feature attribution methods using a synthetic evaluation framework where the ground-truth explanation is known for simple and complex reasoning tasks. They find that feature attribution methods often attribute irrelevant features even in simple settings, and these methods show high variability across data modalities and tasks. Plumb et al. (2021) introduce a method that searches over a dataset to identify important associations that a model might have learned, detect which of these associations are spurious, and propose a data augmentation procedure to overcome the reliance. Nguyen et al. (2021) conduct a large-scale user study on lay and expert users to assess the effectiveness of feature attribution methods on image tasks. They find that feature attributions are not more effective than showing end users nearest neighbor training points.

2 Experimental Methodology

In this section, we setup our experimental methodology. We discuss quantitative analysis of post hoc explanations derived from models trained to rely on pre-defined spurious signals, and a blinded user study that measures the ability of users to use the post hoc explanation methods tested to detect model reliance on spurious signals. We discuss the types of spurious signals considered, define a spurious score that allows us to ascertain that a model indeed relies on a signal as basis of its classification decision, and layout performance measures that capture the reliability of the explanation methods. We conclude with an overview of the methods tested, datasets, and models.

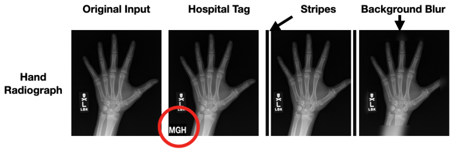

Figure 1: **Overview of Spurious Signals Considered.**

2.1 Experimental Design

Spurious Signals & Score. We consider a spurious signal to be input features that encode for the output but have 'no meaningful connection' to the 'data generating process' (DGP) of the task. A hospital tag present in a hand radiograph is not clinically relevant to the age of the patient. If the tag encodes for the output then it is a spurious signal. Domain expertise is ultimately required for adjudicating that a signal is spurious.

We consider 3 (2 visible and 1 non-visible) kinds of spurious signals (See Figure 1): i) a localized tag; ii) a distinctive stripped pattern; and iii) Gaussian blur applied to the image background. The signals are all spatially localized, so we can easily obtain ground-truth expected explanations.

To induce reliance on spurious signals, we train models on "contaminated” versions of the training set. Given input-label pairs, $\{(x^i, y^i)\}_i^n$, where $x^i \in \mathcal{X}$ and $y^i \in \mathcal{Y}$, we can learn a classifier, f_θ, via empirical risk minimization (ERM) that corresponds to minimizing a loss function, ℓ: $\arg\min_\theta \sum_{i=1}^n \ell(x^i, y^i; \theta)$. To contaminate the training set, we apply a spurious contamination function (SCF) to the training set; $\mathrm{SCF} : \mathcal{X} \times \mathcal{Y} \times \mathcal{C} \to \mathcal{S}$, where $\mathcal{C}$ is the spurious signal set and $\mathcal{S}$ is the transformed set. An example of an SCF is a function that pastes an hospital tag onto the bone age radiographs of all pre-puberty individuals in the dataset. To derive models reliant on a spurious signal, $c_i \in \mathcal{C}$, we simply learn a new classifier via ERM on the modified dataset as follows: $\arg\min_\theta \sum_{i=1}^n \ell\big(\mathrm{SCF}(x^i, y^i, c_i)\big)$ to obtain θ_{spu}. Contemporary evidence suggests that this approach produces models that easily latch onto the spurious signal (Nagarajan et al., 2020).

We focus on the classification setting, and restrict spurious signals to encode, only, for a single class—the *spurious aligned class*. We measure a model's reliance on the spurious signal via a score.

Definition 2.1. (Spurious Score). Given a spurious signal, c_i, the index of its spurious aligned class, $j \in [k]$, a model, $\theta_{\text{spu}} : \mathbb{R}^d \to \mathbb{R}^k$, where $\arg\max(\theta_{\text{spu}})$ indicates the classifier's predicted class, we define the spurious score as:

$$\mathrm{SC}_{c_i, j}(\theta_{\text{spu}}) := \mathbb{P}_{\{x^i | \theta_{\text{spu}}(x^i) \ != \ j\}}[\arg\max(\theta_{\text{spu}}(\mathrm{SCF}(x^i, y^i, c_i))) = j].$$

Given an input that does not contain the spurious signal, and for which the model's prediction is not the spurious aligned class, the model's spurious score is the probability that the model assigns the input to the spurious aligned class if the spurious signal is added to the input.

Model Conditions. We focus our analysis on two model conditions: i) a ‘normal model’, f_{norm}, for which we can rule out dependence on any of the spurious signals tested across all classes on the basis of the spurious score, and ii) a ‘spurious model’, f_{spu}, for which one of the spurious signals encodes for a particular output class. We empirically estimate the spurious score and term models that have a score above 0.85 for any of the pre-defined signals ‘spurious models’. We term a model ‘normal’ if the spurious score is below 0.1 across all classes and the 3 pre-defined spurious signals.

Spurious Signal Detection Reliability Measures. Equipped with spurious (f_{spu}) and normal (f_{norm}) models, we are now able to quantitatively assess the motivating question of this work. We do this by comparing explanations derived from spurious models, f_{spu}, to those derived from normal models (f_{norm}). We can partition the kinds of inputs used for deriving explanations into two: 1) *spurious inputs* (x_{spu})—inputs that include the spurious signal and 2) *normal inputs* (x_{norm})—inputs do not not contain the spurious signal. Comparing the explanations produced by these two classes of inputs for normal and spurious models, we derive reliability performance measures.

- **Known Spurious Signal Detection Measure (K-SSD)** - measures the similarity of explanations derived from *spurious models on spurious inputs* to the ground truth explanation. The ground truth explanation is one that only assigns relevance to the spurious signal as explanation of the output of a spurious model on a spurious input. K-SSD measures method reliability when the spurious signal is known. Given a similarity metric, S_d, then K-SSD corresponds to: $S_d\big(\mathrm{E}_{f_{\text{spu}}}(x_{\text{spu}}), x_{\text{gt}}\big)$; where $\mathrm{E}_{f_{\text{spu}}}(x_{\text{spu}})$ are explanations derived from the spurious model for spurious inputs, and x_{gt} is the ground truth explanation. The similarity function, S_d, depends on the type of explanation considered—we will make our choice of this function concrete shortly.
- **Cause-for-Concern Measure (CCM)** - measures the similarity of explanations derived from *spurious models for normal inputs* to explanations derived from *normal models for normal inputs*: $S_d\big(\mathrm{E}_{f_{\text{spu}}}(x_{\text{norm}}), \mathrm{E}_{f_{\text{norm}}}(x_{\text{norm}})\big)$. This measure simulates the setting where a practitioner does not know the spurious signal, and can only inspect explanations for inputs without the signal. If this measure is high, then it is unlikely that such a method alert a practitioner that a spurious model exhibits defects.
- **False Alarm Measure (FAM)** - measures the similarity of explanations derived from *normal models for spurious inputs* to explanations derived from *spurious models for spurious inputs*: $S_d\big(\mathrm{E}_{f_{\text{norm}}}(x_{\text{spur}}), \mathrm{E}_{f_{\text{spu}}}(x_{\text{spu}})\big)$. We also introduce a variant of this measure, FAM-GT, which measures the similarity of a explanations derived from *normal models for spurious inputs* to the ground truth explanation of a spurious model for that spurious input. If this measure is high, then that approach is more likely to signal to a practitioner that a model is relying on spurious signal when the model does not.

Having defined the metrics above, it remains which similarity function to use.

Computing Metrics for Feature Attribution. For feature attribution methods, we follow prior literature (Adebayo et al., 2020) and use the Structural Similarity Index (SSIM). SSIM measures the visual similarity between two images, so we use this metric as the measure of similarity between two attribution maps. Concretely, given a set of normal inputs, we can obtain a corresponding spurious set of these inputs by applying the spurious contamination function, SCF to these inputs. Consequently, we can then compute the K-SSD, CCM, and FAM metrics given these two sets of inputs using the SSIM metric.

Computing Metrics for Concept Activation. A concept activation method provides a relevance score for a user defined concept. Given a set of user defined concepts, one can estimate and rank each concept by the relevance score. We measure comparison between two concept rankings using a Kolmogorov-Smirnoff (KS) test comparing two distributions where the null hypothesis is that the two distributions are identical; we set significance level to be 0.05.

Computing Metrics for Training Point Ranking. The training point ranking approach assigns a ‘relevance’ score to each training point based on influence of each training point on the test loss of a particular sample. Recently, Hanawa et al. (2020) introduced the Identical class metric’ (ICM), which is the fraction of the top training inputs, for a given test example, that belong to the same class as the true class of the test example in question. Here we also use the KS test to compare the ICM distributions for two different models and set the significance level to be 0.05.

Taken together, these measures provide a comprehensive overview of an explanation method’s performance for detecting spurious signals.

2.2 Blinded Study

To complement the quantitative setup, we further designed a user study (IRB approved) to assess the ability of end-users (200 in total) to use post hoc explanations to detect a model’s reliance on spurious signals. About 50 percent of the participants had trained a ML model before, and 74 percent had interacted with an ML model. We refer to the appendix for additional details.

Task & Setup: The study participants were tasked with assessing model reliability with the aid of model explanations. The participants were randomly assigned to one of two groups: the first group is told explicitly of potential spurious signals, and the second is not. They were asked to rate how likely they are to recommend the model for deployment using a 5-point Likert scale, and a rationale for their decision. Our study design follows that of Adebayo et al. (2020); however, we use a mixed within-subjects and between subjects design for the factors of interest. The Likert scale is as follows, 1: Definitely Not to 5: Definitely. Participants select ‘Definitely’ if they deemed the dog breed classification model reliable. We refer to Appendix I for discussion on user recruitment, statistical analyses, and study design.

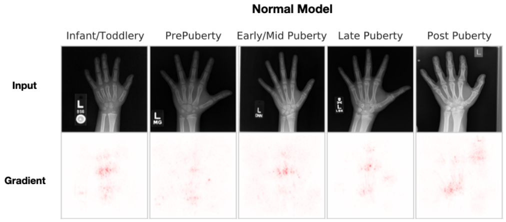

Figure 2: **Feature Attributions.** Here we show 5 different inputs, one for each bone age category, and the corresponding Gradient feature attribution map. We refer to the Appendix for an equivalent visualization for other feature attribution methods considered. We test three additional feature attribution methods: SmoothGrad, Integrated Gradients, and Guided BackProp.

Methods: We test SmoothGrad, TCAV (a concept activation method), a training point ranking method, and a Control setting with no explanations.

2.3 Explanation Methods, Data, & Models

Here we give an overview of the explanation methods, datasets, and models. We present a discussion of how these methods are used in practice in Appendix.

Feature Attributions assign a relevance score for each dimension of an input towards an output. We consider: **Input-Gradient, SmoothGrad, Integrated Gradients (IG), and Guided Backprop (GBP)**.

Concept-Based ((Bau et al., 2017; Kim et al., 2018)) approaches measure the dependence of a DNN's prediction on user-defined features—termed concepts. We select TCAV as the approach to assess in this class (Kim et al., 2018).

Training Point Ranking via Influence Functions (Koh and Liang, 2017). This approach ranks the training samples in order of importance/influence on the loss (or prediction) of a test example.

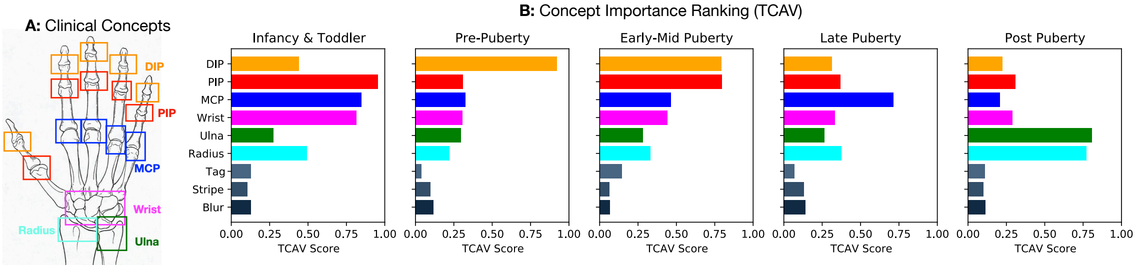

Figure 3: **Concept Importance Methods for a Normal (non-spurious) model.** Here we show the TCAV score for all clinical concepts as well as the spurious concepts for a normal model that was confirmed to not rely on the spurious signals.

Models, Datasets, & Tasks. We consider two medical datasets: Hand (Halabi et al., 2019) and Knee radiographs (Chen et al., 2019) and a dog breed classification task. We consider a small DNN (6 layers, 5 of which are the traditional conv-relu-batchnorm-maxpool combination) inspired by the CBR-Tiny architecture of Raghu et al. (2019) and a ResNet-50 model (See Appendix for additional details).

3 Feature Attributions

In this section, we present results on whether feature attributions are effective for detecting unknown spurious correlation.

Setup. We consider 3 kinds of spurious signals which we term: *Tag* for the 'MGH' hospital tag added to the pre-puberty class, *Stripe* for the paired vertical stripped signal, and *Blur* for the background blur. Given these signals, we then compute the three performance measures of interest: K-SSD, CCM, FAM, and FAM-GT. K-SSD indicates a method's reliability when the spurious signal is known, CCM when the signal is not known and the practitioner only has access to inputs that don't encode the spurious signal. Lastly, FAM and FAM-GT indicates the susceptibility of a method to false positives. An oft-used heuristic based on prior work (Adebayo et al., 2020) for interpreting SSIM scores is that SSIM scores $0.2 - 0.4$ indicate weak visual similarity, $0.5 - 0.7$ indicate moderate similarity, and > 0.75 high similarity. This is because two random images typically have SSIM much less than 0.1. For example, we empirically estimate the similarity of two random (229×229) Gaussian images to be less than 0.00023. Even for natural images, we still find the SSIM values to be below 0.005, which substantiates the previous heuristic.

Results. We show performance measures for all the feature attribution methods tested for the Tag and Blur settings in Tables 1 & 11 (See Appendix). For the tag setting, the attribution methods are indeed able to attribute to the visible spurious signal and the K-SSD measure indicates this finding with mean scores typically above 0.65 for the bone age setting. Contrary to previous findings, we find that GBP outperforms other approaches for known spurious signals. Alternatively, GBP is more suceptible to false positives based on the FAM and FAM-GT score. Across all methods, we find that these methods also seem to attribute to the spurious signal (FAM-GT > 0.4) even when the signal is not being relied on by the model. We observe similar findings for the strip setting as well across all tasks. The CCM measure further indicates that these methods do not indicate presence of spurious signals when the signal is unknown for both the Tag and Stripe signals. This finding, however, reverses for

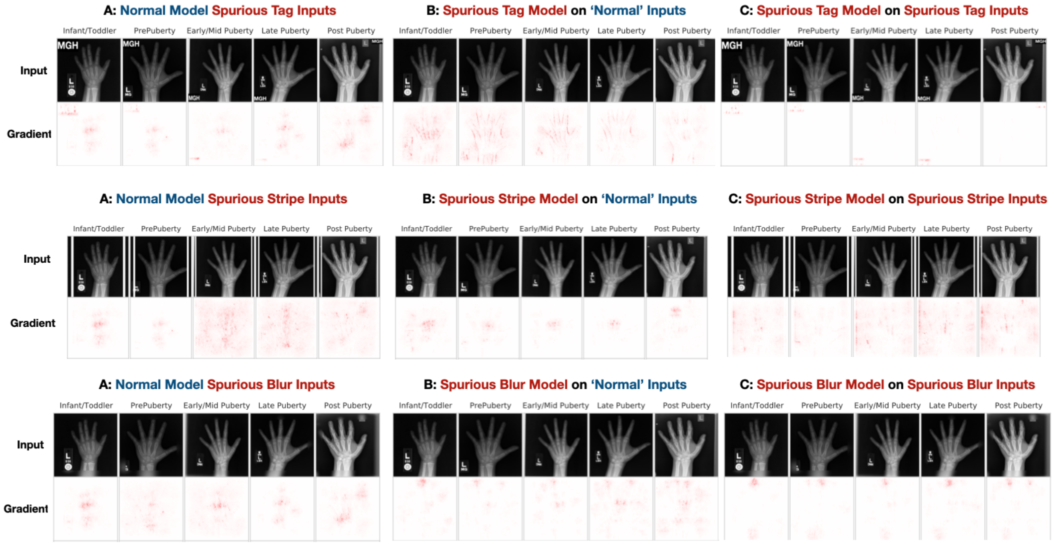
A: Normal Model Spurious Tag Inputs
B: Spurious Tag Model on 'Normal' Inputs
C: Spurious Tag Model on Spurious Tag Inputs
Infant/Toddler
PrePuberty
Early/Mid Puberty
Late Puberty
Post Puberty
Input
Gradient
MGH
A: Normal Model Spurious Stripe Inputs
B: Spurious Stripe Model on 'Normal' Inputs
C: Spurious Stripe Model on Spurious Stripe Inputs
A: Normal Model Spurious Blur Inputs
B: Spurious Blur Model on 'Normal' Inputs
C: Spurious Blur Model on Spurious Blur Inputs

Figure 4: **Top-Detecting Spurious *Tag*.** Here we show in A) Feature attributions for 5 different inputs across the four feature attribution methods with a normal model but with spurious Tag inputs; B) Feature attributions on the same 5 inputs as in (A), but **without** spurious Tag inputs with a model that has learned a spurious alignment between Pre-Puberty and Tag; C) Feature attributions on the same 5 inputs as in (A), but **with** the spurious Tag inputs with a model that has learned a spurious alignment between Pre-Puberty and Tag. **Middle-Detecting Spurious *Stripe*.** Here we show in A) Feature attributions for 5 different inputs across the four feature attribution methods with a normal model but with spurious Stripes inputs; B) Feature attributions on the same 5 inputs as in (A), but **without** the spurious Stripe with a model that has learned a spurious alignment between Pre-Puberty and Stripe; C) Feature attributions on the same 5 inputs as in (A), but **with** the spurious Stripe with a model that has learned a spurious alignment between Pre-Puberty and Stripe. **Middle-Detecting Spurious *Blur*.** The blur images are analogous to the Tag and Stripe settings. Please refer to the Figures 7, 8, 9 in the Appendix for SmoothGrad, Integrated Gradients, and Guided BackProp examples.

Table 1: Performance metrics for each attribution method across tasks for the Tag Setting. Below each metric in the Table is another row (SEM) that indicates the standard error of the mean for each value.

Method	Bone Age				Knee				Dog Breeds			
	Grad	SG	IG	GBP	Grad	SG	IG	GBP	Grad	SG	IG	GBP
K-SSD	0.65	0.66	0.67	0.81	0.51	0.49	0.47	0.76	0.71	0.76	0.79	0.88
K-SSD (SEM)	0.0097	0.013	0.019	0.006	0.012	0.017	0.019	0.023	0.01	0.011	0.014	0.01
CCM	0.37	0.39	0.35	0.75	0.32	0.33	0.35	0.66	0.42	0.41	0.39	0.64
CCM (SEM)	0.0031	0.002	0.015	0.029	0.027	0.023	0.029	0.014	0.013	0.016	0.012	0.015
FAM	0.51	0.55	0.53	0.68	0.46	0.47	0.45	0.69	0.59	0.64	0.68	0.73
FAM (SEM)	0.0029	0.0019	0.018	0.024	0.023	0.024	0.019	0.016	0.015	0.011	0.022	0.035
FAM-GT	0.56	0.53	0.46	0.61	0.42	0.48	0.41	0.63	0.76	0.73	0.77	0.81
FAM-GT (SEM)	0.017	0.035	0.0253	0.028	0.016	0.019	0.0045	0.006	0.011	0.033	0.024	0.0053

the non-visible blur spurious signal. Across all measures, we find that all methods struggle to reliably indicate that spurious models are reliant on the blur signal.

Additionally, we also find that the FAM scores are typically higher than the CCM scores across the tasks. This finding indicates that the feature attribution methods tested are more susceptible to false positives than they are to indicate to a practitioner that a model is defective in the absence of the spurious signal, a finding that casts doubt on the utility of such approaches in practice.

4 Concept Activation Importance

We find that concept methods can indicate a model's reliance on Tag and Stripe signals when known. However, the approach struggles to detect Blur signal even when known. As is the case with feature attributions, when a spurious signal is not explicitly tested for, our significance tests indicate that reliance cannot be detected in the non-spurious concepts available.

Overview & Setup. In this setting, we compute the 3 performance metrics of interest: K-SSD, CCM, and FAM. To measure comparison between two concept rankings, we use a Kolmogorov-Smirnoff (KS) test comparing two distributions where the null hypothesis is that the two distributions are identical; we set significance level to be 0.05.

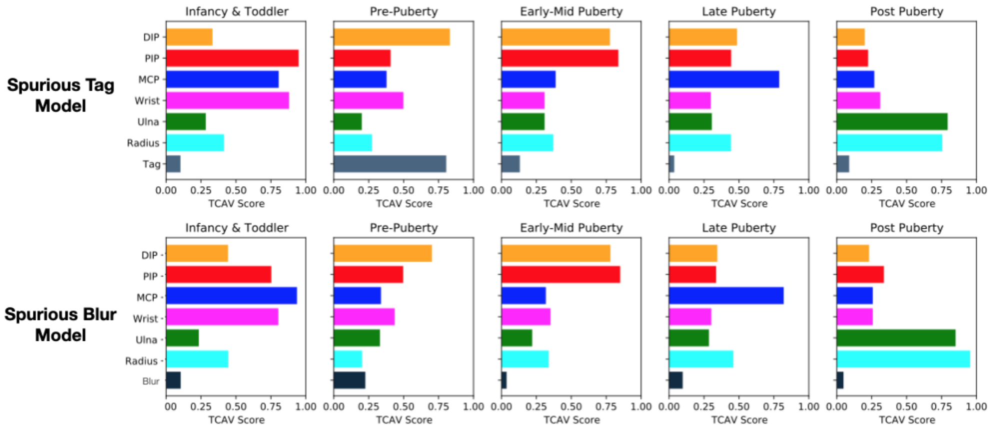

Figure 5: **Concept Results for Tag and Blur Models.** TCAV scores for a model reliant on the spurious Tag and a model reliant on Blur.

Result. In Figure 5, we show TCAV scores for each bone age class for both the spurious tag and the blur models. In Table 2, we show results of the KS-Test across metrics for the Tag Setting. Here, a X means we reject the null, while an ✓means we are unable to do so at the pre-specified significance level. For the K-SSD metric, the KS-test rejects the null for the Tag and Stripe signals. However, we find that the opposite is the case for the blur signal. This suggests that concept rankings can help detect reliance on the visible signals but not non-visible signals when the spurious signals are unknown. However, for the CCM metric, we are unable to reject the hypothesis that the distribution of concept rankings are not similar for all spurious signals. This finding suggests that when the spurious signal is unknown, and we only compare the distributions of known (non-spurious) concepts for a normal model and a spurious model, there is high similarity. Lastly, a difference in means test as well as a KS-Test for the FAM measure indicates that the normal models do not rely on the spurious signals as well. Overall, this finding suggests that TCAV is less susceptible to false positives.

Table 2: Concept Metrics Tag.

Metric	Result
K-SSD	X
CCM	✓
FAM	✓

5 Training Point Ranking & Blinded Study

Overview & Setup. We now describe our empirical findings for the training point ranking via influence functions approach. Here we present results for the case where the spurious signal is aligned with the Pre-Puberty class.

Results. The main take away is that given a known spurious signal, the fraction of top ranked training spurious signal inputs increases with a spurious model across all of the spurious signals. While this might seem encouraging, we note that such increase actually indicates that the ICM metric might provide illusory confidence in a spurious model. Ultimately, a critical requirement here is knowing what the spurious signal ahead of time and to be able to select the right inputs to inspect.

Blinded Study. We now turn to a summary of the results of the blinded user study. The median recommendations selected by participants (200 in total) is reported for each explanation-model condition in Table 6. We plot a box plot of all 16 categories in the appendix. Our setting mimics traditional randomized experimental settings, so we adopt a randomization inference analysis to determine the effect of each treatment, which is the explanation method in this case, on the ability of the users to recommend a model. A higher Likert score means the user is more likely to recommend a model. Consequently, if an explanation method is effective, then it should be the case that users should be less likely to recommend a model that relies on a spurious signal.

We conduct two kinds of statistical analyses of the data. First, under each condition, we perform a difference of means test for each treatment compared to the control setting. Secondly, for each model manipulation, we performed a one-way Anova test, and a Tukey-Kramer test to assess the effect of the explanation type on the ability of the participants to reject a defective model.

We observe that when blinded, in none of the methods do participants conclusively reject spurious models. Perhaps more surprising, when the participants were not blinded, we see that only participants using the TCAV approach rejected a spurious model. This finding has significant implications on whether these current tools are effective in the hands of a practitioner.

Table 3: Training Point Ranking

Metric	Result
K-SSD	✗
CCM	✓
FAM	✓

Method	B-Normal	NB-Normal	B-Spurious	NB-Spurious
SmoothGrad	4*	4*	3*	3
TCAV	4*	3	3*	2*
Influence	3*	3	3*	3
Control	4	3	4	4

Table 4: **Blinded User Study**. Here we report the median recommendation for each condition across all explanation types and control. This median is derived from user provided responses assessing model reliability. B indicates Blinded, and NB indicates Not-Blinded. In the Table, the * symbol indicates statistically significant conditions. For the sake of space, we defer the full distribution description to the Appendix.

6 Discussion & Conclusion

Conclusion. DNNs trained on image datasets can naturally rely on spurious training signals (Geirhos et al., 2020). Discovering this reliance is crucial in consequential settings like medical imaging. Post hoc explanations methods are a promising direction towards detecting such reliance; however, their effectiveness is currently under question. We investigated whether 3 classes of post hoc explanations are effective for detecting a model's reliance on spurious training signals. We present an experimental setup that can also be easily adapted to other settings to assess a larger class of approaches. The setup comes equipped with a spurious score and performance measures for spurious signal detection. We find that the 3 classes of post hoc explanations tested are only sometimes able to diagnose the spurious training signal even if they are used to explicitly test for model dependence on these signals. Consequently, our findings calls for, potentially, a completely different paradigm of methods that are designed to address the important and challenging question of detecting spurious training signals.

Limitations. We have focused exclusively on DNNs trained on image tasks; hence, it is unclear if our findings will generalize to other modalities like text or time-series data. While we considered a diverse set of methods, the literature on post hoc explanations is quite vast, so undoubtedly there exist methods that do not fit neatly into the 3 explanation classes that we explored. For the normal models that we tested, we note it is indeed possible for the dataset to encode for other kinds of spurious signals; however, our analysis can rule out the model dependence on the 3 pre-defined spurious signals that we study in this work.

REFERENCES

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 9525–9536, 2018.

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. Debugging tests for model explanations. *arXiv preprint arXiv:2011.05429*, 2020.

Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. Evaluating saliency map explanations for convolutional neural networks: a user study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 275–285, 2020.

Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. Fairwashing explanations with off-manifold detergent. In *International Conference on Machine Learning*, pages 314–323. PMLR, 2020.

Marcus A Badgeley, John R Zech, Luke Oakden-Rayner, Benjamin S Glicksberg, Manway Liu, William Gale, Michael V McConnell, Bethany Percha, Thomas M Snyder, and Joel T Dudley. Deep learning predicts hip fracture using confounding patient and healthcare variables. *NPJ digital medicine*, 2(1):1–10, 2019.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert MÃžller. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11(Jun):1803–1831, 2010.

Mohammad Taha Bahadori and David E Heckerman. Debiasing concept-based explanations with causal analysis. 2021.

Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020.

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.

Pingjun Chen, Linlin Gao, Xiaoshuang Shi, Kyle Allen, and Yang Lin. Fully automatic knee osteoarthritis severity grading using deep neural networks with a novel ordinal loss. *Computerized Medical Imaging and Graphics*, 75:84–92, 2019. doi: https://doi.org/10.1016/j.compmedimag.2019.06.002.

Valerie Chen, Jeffrey Li, Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Towards connecting use cases and methods in interpretable machine learning. *arXiv preprint arXiv:2103.06254*, 2021.

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

Eric Chu, Deb Roy, and Jacob Andreas. Are visual explanations useful? a case study in model-in-the-loop prediction. *arXiv preprint arXiv:2007.12248*, 2020.

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *medRxiv*, 2020.

Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pages 13567–13578, 2019.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. 2017.

Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Learning explainable models using attribution priors. *arXiv preprint arXiv:1906.10670*, 2019.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3681–3688. AAAI Press, 2019a. doi: 10.1609/aaai.v33i01.33013681. URL https://doi.org/10.1609/aaai.v33i01.33013681.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*, 2019b.

Han Guo, Nazneen Fatema Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. Fastif: Scalable influence functions for efficient model interpretation and debugging. *arXiv preprint arXiv:2012.15781*, 2020.

Safwan S Halabi, Luciano M Prevedello, Jayashree Kalpathy-Cramer, Artem B Mamonov, Alexander Bilbily, Mark Cicero, Ian Pan, Lucas Araújo Pereira, Rafael Teixeira Sousa, Nitamar Abdala, et al. The rsna pediatric bone age machine learning challenge. *Radiology*, 290(2):498–503, 2019.

Xiaochuang Han, Byron C Wallace, and Yulia Tsvetkov. Explaining black box predictions and unveiling data artifacts through influence functions. *arXiv preprint arXiv:2005.06676*, 2020.

Kazuaki Hanawa, Sho Yokoi, Satoshi Hara, and Kentaro Inui. Evaluation of similarity-based explanations. *arXiv preprint arXiv:2006.04528*, 2020.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, pages 2921–2932, 2019.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9737–9748, 2019.

Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (cme): Concept-based model extraction. *arXiv preprint arXiv:2010.13233*, 2020.

Fereshte Khani and Percy Liang. Removing spurious features can hurt accuracy and affect groups disproportionately. *arXiv preprint arXiv:2012.04104*, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2673–2682, 2018.

Joon Sik Kim, Gregory Plumb, and Ameet Talwalkar. Sanity simulations for saliency methods. *arXiv preprint arXiv:2105.06506*, 2021.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894. PMLR, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

Himabindu Lakkaraju and Osbert Bastani. " how do i fool you?" manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 79–85, 2020.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

Aravindh Mahendran and Andrea Vedaldi. Salient deconvolutional networks. In *European Conference on Computer Vision*, pages 120–135. Springer, 2016.

Qingjie Meng, Christian Baumgartner, Matthew Sinclair, James Housden, Martin Rajchl, Alberto Gomez, Benjamin Hou, Nicolas Toussaint, Jeremy Tan, Jacqueline Matthew, et al. Automatic shadow detection in 2d ultrasound. 2018.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Giang Nguyen, Daeyoung Kim, and Anh Nguyen. The effectiveness of feature attribution methods and its correlation with automatic evaluation scores. *arXiv preprint arXiv:2105.14944*, 2021.

Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *ICML*, 2018.

Judea Pearl. URL http://causality.cs.ucla.edu/blog/.

Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021.

Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. *arXiv preprint arXiv:1802.07810*, 2018.

Garima Pruthi, Frederick Liu, Mukund Sundararajan, and Satyen Kale. Estimating training data influence by tracking gradient descent. *arXiv preprint arXiv:2002.08484*, 2020.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. *arXiv preprint arXiv:1902.07208*, 2019.

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.

Laura Rieger, Chandan Singh, W James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. *International Conference on Machine Learning*, 2020.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In Carles Sierra, editor, *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org, 2017. doi: 10.24963/ijcai.2017/371. URL https://doi.org/10.24963/ijcai.2017/371.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR, 2020.

Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *arXiv preprint arXiv:2102.12781*, 2021.

Hua Shen and Ting-Hao Huang. How useful are the machine-generated interpretations to general users? a human evaluation on guessing the incorrectly predicted labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 168–172, 2020.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6034.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why modified bp attribution fails. *arXiv preprint arXiv:1912.09818*, 2019.

Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Suraj Srinivas and Francois Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=dYeAHXnpWJ4.

Pierre Stock and Moustapha Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 498–512, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR. URL http://proceedings.mlr.press/v70/sundararajan17a.html.

Richard Tomsett, Dan Harborne, Supriyo Chakraborty, Prudhvi Gurram, and Alun D. Preece. Sanity checks for saliency metrics. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 6021–6029. AAAI Press, 2020. URL https://aaai.org/ojs/index.php/AAAI/article/view/6064.

Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Chih-Kuan Yeh, Joon Kim, Ian En-Hsu Yen, and Pradeep K Ravikumar. Representer point selection for explaining deep neural networks. In *Advances in neural information processing systems*, pages 9291–9301, 2018.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *arXiv preprint arXiv:2104.14403*, 2021.