VALID P-VALUE FOR DEEP LEARNING-DRIVEN SALIENT REGION

Anonymous authors
Paper under double-blind review

ABSTRACT

Various saliency map methods have been proposed to interpret and explain predictions of deep learning models. Saliency maps allow us to interpret which parts of the input signals have a strong influence on the prediction results. However, since a saliency map is obtained by complex computations in deep learning models, it is often difficult to know how reliable the saliency map itself is. In this study, we propose a method to quantify the reliability of a saliency region in the form of p-values. Our idea is to consider a saliency map as a selected hypothesis by the trained deep learning model and employ the selective inference framework. The proposed method provably provides a valid p-value for the detected salient region, i.e., we can provably control the false positive rate of the detected salient region. We demonstrate the validity of the proposed method through numerical examples in synthetic and real datasets. Furthermore, we develop a Keras-based framework for conducting the proposed selective inference for a wide class of CNNs without additional implementation cost.

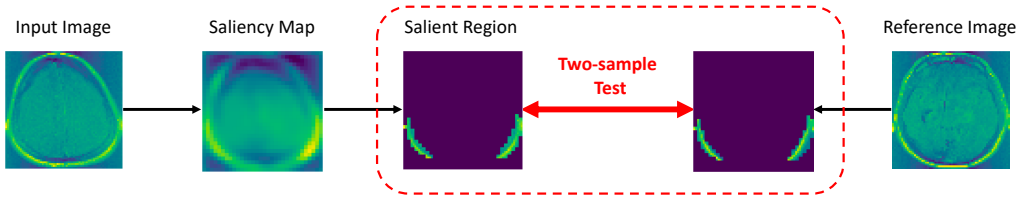
1 INTRODUCTION

Over the past few years, deep neural networks (DNNs) has exhibited remarkable predictive performances in numerous practical applications from various domains thanks to its ability of automatically discovering the representations needed for the prediction tasks from data. In order to make the decision-making process of DNNs transparent and easy for understanding, it is crucially important to be able to explain and interpret DNN representations. For example, in image classification tasks, obtaining the *salient regions* allow us to explain which parts of the input image strongly influence the classification results.

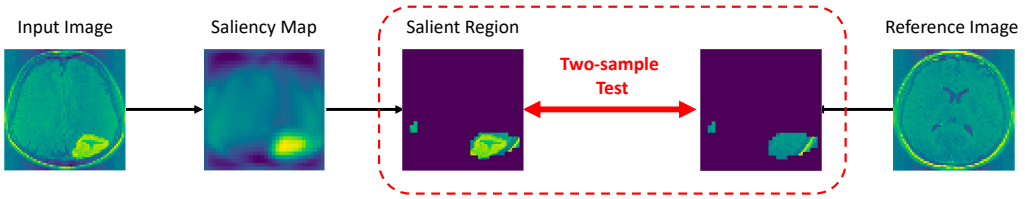
In the past few years, several saliency methods have been proposed to explain and interpret the predictions of DNN models (Ribeiro et al., 2016; Bach et al., 2015; Doshi-Velez & Kim, 2017; Lundberg & Lee, 2017; Zhou et al., 2016; Selvaraju et al., 2017). However, it is known that the results obtained from saliency methods are fragile (Kindermans et al., 2017; Ghorbani et al., 2019; Melis & Jaakkola, 2018; Zhang et al., 2020; Dombrowski et al., 2019; Heo et al., 2019). Therefore, it is important to develop a method to quantify the reliability of DNN-driven salient regions.

Our idea is to interpret the salient regions as hypotheses that are driven by DNN and employ a statistical hypothesis testing framework. We use p-value as a criterion to quantify the statistical reliability of DNN-driven hypotheses. Unfortunately, constructing valid statistical test for DNN-driven salient regions is challenging because of the *selection bias*. Namely, because the trained DNN *selects* the salient region based on the data, post-selection assessment of their importance is biased upward.

In order to correct the selection bias and compute valid p-values for DNN-driven salient regions, we introduce conditional *Selective Inference* (SI) approach. The selection bias is corrected by conditional SI by considering the test statistic conditional on the event that the hypotheses (salient regions) are selected by the trained DNNs. Our main technical contribution is to develop a computational method for explicitly deriving the exact (non-asymptotic) conditional sampling distribution of the salient region for a wide class Convolutional Neural Networks (CNNs) which enables us to conduct conditional SI and compute valid p-values. Figure 1 shows an example that illustrates the problem setup and the importance of our method.



(a) Image without tumor region. The naive-p = **0.00** (wrong detection) and selective-p = **0.43** (true negative)



(b) Image with tumor region. The naive-p = **0.00** (true positive) and selective-p = **0.00** (true positive)

Figure 1: Examples of the problem setup and proposed method on brain tumor dataset. By applying a saliency method called CAM (Zhou et al., 2016) on an query input image, we obtain the salient region. Our goal is to provide the statistical significance of the salient region in the form of p-value by considering two-sample test between the salient region and the corresponding region in the "reference" image. Note that, since the salient region is selected based on the data, the degree of saliency in the selected region is biased upward. In the upper image where there is no true brain tumor, the naive p-value which is obtained without caring the selection bias is nearly zero, indicating the false positive finding of the salient region. On the other hand, the selective p-value which is obtained by our conditional SI approach is 0.43, indicating that the selected saliency region is not statistically significant. In the lower figure where there actually exist a brain tumor, both the naive p-value and the selective p-value are very small, indicating true positive finding. These results illustrate that naive p-value cannot be used to quantify the reliability of DNN-based salient region because it is unreliable. In contrast, with the selective p-values, we can successfully identifies false positive and true positive detections.

Related works. A number of methods have been developed to visualize and understand a trained DNN. Many of these post-hoc approaches (Mahendran & Vedaldi, 2015; Zeiler & Fergus, 2014; Dosovitskiy & Brox, 2016; Simonyan et al., 2013) have focused on developing visualization tools for the saliency maps given a trained DNN. Other methods have aimed to identify the discriminative regions in an input image, given a trained network (Selvaraju et al., 2017; Fong & Vedaldi, 2017; Zhou et al., 2016; Lundberg & Lee, 2017). In parallel, some recent studies have showed that many popular methods for explanation and interpretation are not stable with respect to the perturbation or the adversarial attack on the input data and the model (Kindermans et al., 2017; Ghorbani et al., 2019; Melis & Jaakkola, 2018; Zhang et al., 2020; Dombrowski et al., 2019; Heo et al., 2019). However, to our knowledge, no study to date that can quantitatively evaluate the stability and reproducibility of DNN-driven salient regions with a rigorous statistical inference framework.

In the past few years, conditional SI has been recognized as a new promising approach by which to evaluate the statistical reliability of data-driven hypotheses, and it has been actively studied for inference on the features of linear models selected by several feature selection methods—for example, Lasso (Lee et al., 2016; Liu et al., 2018; Le Duy & Takeuchi, 2021). The basic idea of conditional SI is to make inferences conditional on the selection event, which allows us to derive the exact sampling distribution of the test statistic. Besides, conditional SI has been applied to various problems (Fithian et al., 2015; Tian & Taylor, 2018; Yang et al., 2016; Hyun et al., 2021; Duy et al., 2020; Sugiyama et al., 2021b; Chen & Bien, 2019; Panigrahi et al., 2016; Tsukurimichi et al., 2021; Hyun et al., 2018; Tanizaki et al., 2020; Duy & Takeuchi, 2021; Tibshirani et al., 2016; Sugiyama et al., 2021a; Suzumura et al., 2017; Das et al., 2021; Duy & Takeuchi, 2022).

The most closely related work is Duy et al. (2022), where the authors provide a framework to compute valid p-values for DNN-based image segmentation results. The idea is to consider the sampling distribution of the test statistic conditional on the selection event, in which the DNN-based

segmentation result is obtained. In this study, we focus on a more general problem setup compared to Duy et al. (2022). Moreover, their method requires high implementation cost and is not flexible when the structure of a DNN is changed. As we will introduce later, we develop a practical Keras-based conditional SI framework for a wide class of CNNs without additional implementation cost.

In another direction, Burns et al. (2020) considered black box model interpretability as a multiple hypothesis testing problem. The goal is to find important features by testing the significance of the difference between the model prediction and what would be expected when replacing the features by their counterfactuals. The difficulty of this multiple hypothesis testing approach is that the number of hypotheses to be considered is huge (e.g., in the case of an image with n pixels, the number of possible salient regions is 2^n). Multiple testing correction method such as Bonferroni correction is highly conservative when the number of hypotheses is large. To circumvent this difficulty, the authors only considered tractable number of regions selected by a human expert or an object detector, which actually causes selection bias because those candidate regions are selected based on the data.

Contribution. Our main contributions are as follows:

• We provide an exact (non-asymptotic) inference method for the salient regions based on the concept of SI. To our knowledge, this is the first method that can provide valid p-values to statistically quantify the reliability of the DNN-driven salient regions.

• We provide a novel algorithm and implementation for efficiently and flexibly conducting conditional SI for salient regions. Specifically, our Keras-based implementation enables us to conduct conditional SI for a wide class of CNNs without additional implementation cost.

• We conduct experiments on both synthetic and real-world datasets, through which we offer evidence that our proposed method can successfully control the false positive rate, has decent performance in terms of computational efficiency, and provides good results in practical applications. We provide our implementation in the supplementary document.

2 Problem Formulation

Consider an n-dimensional *query* input vector

$$\boldsymbol{X} = (X_1, ..., X_n)^\top = \boldsymbol{s} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathbb{N}(\mathbf{0}, \Sigma)$$

and an n-dimensional *reference* input vector

$$\boldsymbol{X}^{\mathrm{ref}} = (X_1^{\mathrm{ref}}, ..., X_n^{\mathrm{ref}})^\top = \boldsymbol{s}^{\mathrm{ref}} + \boldsymbol{\varepsilon}^{\mathrm{ref}}, \quad \boldsymbol{\varepsilon}^{\mathrm{ref}} \sim \mathbb{N}(\mathbf{0}, \Sigma),$$

where $\boldsymbol{s}, \boldsymbol{s}^{\mathrm{ref}} \in \mathbb{R}^n$ are the signals and $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^{\mathrm{ref}} \in \mathbb{R}^n$ are the noises for the query and reference input vectors, respectively. We assume that the signals $\boldsymbol{s}$ and $\boldsymbol{s}^{\mathrm{ref}}$ are unknown, whereas the distribution of the noises $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^{\mathrm{ref}}$ are known (or estimable from external independent data) to follow $\mathbb{N}(\mathbf{0}, \Sigma)$, an n-dimensional Normal distribution with mean vector $\mathbf{0}$ and the covariance matrix Σ, and $\boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon}^{\mathrm{ref}}$ are independent. In the illustrative example in §1, $\boldsymbol{X}$ is a query brain image for a potential patient (we do not know whether she/he has brain tumor or not), whereas $\boldsymbol{X}^{\mathrm{ref}}$ is a brain image of a healthy person known to be free of brain tumors. We consider an extension to the case where multiple reference input vectors are available in §6.

Consider a saliency method on a trained CNN. Let us denote the saliency method as a function $\mathcal{A} : \mathbb{R}^n \to \mathbb{R}^n$ which takes a query input vector $\boldsymbol{X} \in \mathbb{R}^n$ and returns the saliency map $\mathcal{A}(\boldsymbol{X}) \in \mathbb{R}^n$. Let us define a *saliency region* $\mathcal{M}_{\boldsymbol{X}}$ for the query input vector $\boldsymbol{X}$ to be the set of elements whose saliency map value is greater than a threshold

$$\mathcal{M}_{\boldsymbol{X}} = \{i \in [n] : \mathcal{A}_i(\boldsymbol{X}) \geq \tau\}, \tag{1}$$

where $\tau \in \mathbb{R}$ is a given threshold. In this paper, we consider CAM (Zhou et al., 2016) as an example of saliency method and threshold based-definition of the saliency method in Eq.(1) although our method can be applied to other saliency methods and other definition of saliency region.

Statistical inference. To quantify the statistical significance of the saliency region $\mathcal{M}_{\boldsymbol{X}}$, we consider *two-sample test* to quantify the statistical significance of the difference between the salient

region of the query input vector $\boldsymbol{X}_{\mathcal{M}_{\boldsymbol{X}}}$ and the corresponding region of the reference input vector $\boldsymbol{X}^{\mathrm{ref}}_{\mathcal{M}_{\boldsymbol{X}}}$. As concrete examples of the two-sample test, we consider *global null test:*

$$\mathrm{H}_0 : s_i = s_i^{\mathrm{ref}},\ \forall i \in \mathcal{M}_{\boldsymbol{X}}, \quad \text{v.s.} \quad \mathrm{H}_1 : s_i \neq s_i^{\mathrm{ref}},\ \exists i \in \mathcal{M}_{\boldsymbol{X}}, \tag{2}$$

and *mean null test*:

$$\mathrm{H}_0 : \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i^{\mathrm{ref}} \quad \text{v.s.} \quad \mathrm{H}_1 : \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i \neq \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} s_i^{\mathrm{ref}}. \tag{3}$$

In the global null test in Eq.(2), we consider a null hypothesis that all elements of the signals in the salient region $\mathcal{M}_{\boldsymbol{X}}$ are the same between $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{ref}}$. On the other hand, in the mean null test in Eq.(3), we consider a null hypothesis that the average signals in the salient region $\mathcal{M}_{\boldsymbol{X}}$ are the same between $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{ref}}$. The p-values for these two-sample tests between $\boldsymbol{X}_{\mathcal{M}_{\boldsymbol{X}}}$ and $\boldsymbol{X}^{\mathrm{ref}}_{\mathcal{M}_{\boldsymbol{X}}}$ can be used to quantify the statistical significance of the truncation region $\mathcal{M}_{\boldsymbol{X}}$.

Test-statistic. For a two-sample test between $\boldsymbol{X}_{\mathcal{M}_{\boldsymbol{X}}}$ and $\boldsymbol{X}^{\mathrm{ref}}_{\mathcal{M}_{\boldsymbol{X}}}$, we consider a class of test-statistic called *conditionally linear test-statistic* which is written as

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix}, \tag{4}$$

where $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}} \in \mathbb{R}^{2n}$ is the vector which depends on the saliency region $\mathcal{M}_{\boldsymbol{X}}$. We note that the test-statistic $T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}})$ is *not* linear in $(\boldsymbol{X}\ \boldsymbol{X}^{\mathrm{ref}})^{\top}$ because $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}$ depends on $\mathcal{M}_{\boldsymbol{X}}$. We say that the test-statistic $T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}})$ is *conditionally* linear in the sense that it is linear in $(\boldsymbol{X}\ \boldsymbol{X}^{\mathrm{ref}})^{\top}$ when the salient region $\mathcal{M}_{\boldsymbol{X}}$ is fixed.

The test-statistics for the global null-test and the mean null test can be written in the form of Eq.(4). For the global null test in Eq.(2), we consider the following test-statistic

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} = \sum_{i\in\mathcal{M}_{\boldsymbol{X}}} \left| X_i - X_i^{\mathrm{ref}} \right|,$$

where $\boldsymbol{\eta} = \Omega^{\top}\boldsymbol{v} \in \mathbb{R}^{2n}$ with

$$\Omega = \left[\mathrm{diag}(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}}) \quad -\mathrm{diag}(\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}})\right] \in \mathbb{R}^{n\times 2n}, \quad \boldsymbol{v} = \mathrm{sign}\left(\Omega \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix}\right) \in \mathbb{R}^n.$$

For the mean null test in Eq.(3), we consider the following test-statistic

$$T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}}) = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} X_i - \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\sum_{i\in\mathcal{M}_{\boldsymbol{X}}} X_i^{\mathrm{ref}}$$

where $\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}} = \frac{1}{|\mathcal{M}_{\boldsymbol{X}}|}\begin{pmatrix} \mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}} \\ -\mathbf{1}^n_{\mathcal{M}_{\boldsymbol{X}}} \end{pmatrix} \in \mathbb{R}^{2n}$.

In order to obtain p-values for the two-sample tests such as Eqs.(2) and (3), we need to know the sampling distribution of the test-statistic $T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}})$. Unfortunately, it is difficult to derive the sampling distributions of $T(\boldsymbol{X}, \boldsymbol{X}^{\mathrm{ref}})$ because it depends on the salient region $\mathcal{M}_{\boldsymbol{X}}$ which is obtained through a highly complicated calculation on the trained CNN.

3 Computing Valid p-value by Conditional Selective Inference

In this section, we introduce an approach to compute the valid p-values for the two-sample tests for the salient region $\mathcal{M}_{\boldsymbol{X}}$ between the query input vector $\boldsymbol{X}$ and the reference input vector $\boldsymbol{X}^{\mathrm{ref}}$ based on the concept of conditional SI (Lee et al., 2016).

3.1 Conditional Distribution and Selective p-value

Conditional distribution. The basic idea of conditional SI is to consider the sampling distribution of the test-statistic conditional on a *selection event*. Specifically, we consider the sampling property of the following conditional distribution

$$\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top} \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{X}^{\mathrm{ref}} \end{pmatrix} \Bigg| \left\{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}_{\mathrm{obs}}}\right\}, \tag{5}$$

where $\boldsymbol{X}_{\text{obs}}$ is the observation (realization) of random vector $\boldsymbol{X}$. The condition in Eq.(5) indicates that we consider randomness of $\boldsymbol{X}$ conditional on the event that the same salient region $\mathcal{M}_{\boldsymbol{X}}$ as the observed $\mathcal{M}_{\boldsymbol{X}^{\text{obs}}}$ is obtained. By conditioning on the salient region $\mathcal{M}_{\boldsymbol{X}}$, derivation of the sampling distribution of the conditionally linear test-statistic $T(\boldsymbol{X}, \boldsymbol{X}^{\text{ref}})$ is reduced to a derivation of the distribution of linear function of $(\boldsymbol{X}, \boldsymbol{X}^{\text{ref}})$

To introduce the conditional SI for the global null test and the mean null test in §2, let us write the selection event as

$$\mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}}, \tag{6}$$

where $X^{\text{ref}}_{\text{obs}}$ is the observed reference vector, and consider the sampling distribution of the test-statistic conditional on the event in Eq.(6), i.e.,

$$T(\boldsymbol{X}, \boldsymbol{X}^{\text{ref}}) \mid \left\{\mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}}\right\}.$$

For the global null test in Eq.(2), we consider the following selection event:

$$\left\{\mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}}\right\} \Leftrightarrow \left\{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\boldsymbol{X}_{\text{obs}}},\ \text{sign}(\boldsymbol{X} - \boldsymbol{X}^{\text{ref}}) = \text{sign}(\boldsymbol{X}_{\text{obs}} - \boldsymbol{X}^{\text{ref}}_{\text{obs}})\right\},$$

where $\text{sign}(\boldsymbol{X} - \boldsymbol{X}^{\text{ref}}) \in \{\pm 1\}^n$ is the n-dimensional vector whose i^{th} element is defined as the sign of $X_i - X^{\text{ref}}_i$ for $i \in [n]$.

For the mean null test in Eq.(3), we simply consider the following selection event:

$$\left\{\mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}}\right\} \Leftrightarrow \left\{\mathcal{M}_{\boldsymbol{X}} = \mathcal{M}_{\text{obs}}\right\}.$$

Selective p-value. After considering the conditional sampling distribution in (5), we introduce the the following *selective p-value*:

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0}\Big(\left|T(\boldsymbol{X}, \boldsymbol{X}^{\text{ref}})\right| \geq \left|T(\boldsymbol{X}_{\text{obs}}, \boldsymbol{X}^{\text{ref}}_{\text{obs}})\right| \ \Big|\ \mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{Q}_{\text{obs}}\Big), \tag{7}$$

where

$$\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \left(I_{2n} - \frac{\tilde{\Sigma}\boldsymbol{\eta}\boldsymbol{\eta}^\top}{\boldsymbol{\eta}^\top\tilde{\Sigma}\boldsymbol{\eta}}\right)\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix} \in \mathbb{R}^{2n}, \quad \mathcal{Q}_{\text{obs}} = \mathcal{Q}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}}, \quad \tilde{\Sigma} = \begin{pmatrix}\Sigma & 0\\ 0 & \Sigma\end{pmatrix}.$$

The $\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}}$ is the sufficient statistic of the nuisance parameter that needs to be conditioned on in order to tractably conduct the inference [1].

The selective p-value in Eq.(7) has the following desired sampling property

$$\mathbb{P}_{\text{H}_0}\Big(p_{\text{selective}} \leq \alpha \mid \mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}}\Big) = \alpha, \quad \forall \alpha \in [0, 1]. \tag{8}$$

This means that the selective p-values $p_{\text{selective}}$ can be used as a valid statistical significance measure for the salient region $\mathcal{M}_{\boldsymbol{X}}$.

3.2 Characterization of the Conditional Data Space

To compute the selective p-value in (7), we need to characterize the conditional data space whose characterization will be introduced in the next section. We define the set of $(\boldsymbol{X}\ \boldsymbol{X}^{\text{ref}})^\top \in \mathbb{R}^{2n}$ that satisfies the conditions in Equation (7) as

$$\mathcal{D} = \left\{(\boldsymbol{X}\ \boldsymbol{X}^{\text{ref}})^\top \in \mathbb{R}^{2n} \ \middle|\ \mathcal{E}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{E}_{\text{obs}},\ \mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{Q}_{\text{obs}}\right\}. \tag{9}$$

According to the second condition, the data in $\mathcal{D}$ is restricted to a line in $\mathbb{R}^{2n}$ as stated in the following Lemma.

[1] This nuisance parameter $\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}}$ corresponds to the component $\boldsymbol{z}$ in the seminal conditional SI paper (Lee et al., 2016) (see Sec. 5, Eq. 5.2 and Theorem 5.2). We note that additional conditioning on $\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}}$ is a standard approach in the conditional SI literature and is used in almost all the conditional SI-related studies. Here, we would like to note that the selective p-value depends on $\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}}$, but the property in (8) is satisfied without this additional condition because we can marginalize over all values of $\mathcal{Q}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}}$ (see the lower part of the proof of Theorem 5.2 in Lee et al. (2016)).

Lemma 1. *Let us define*

$$\boldsymbol{a} = \mathcal{Q}_{\text{obs}} \in \mathbb{R}^{2n} \quad \textit{and} \quad \boldsymbol{b} = \frac{\tilde{\Sigma}\boldsymbol{\eta}}{\boldsymbol{\eta}^\top \tilde{\Sigma}\boldsymbol{\eta}} \in \mathbb{R}^{2n}. \tag{10}$$

Then, the set $\mathcal{D}$ in (9) can be rewritten as $\mathcal{D} = \left\{ \left(\boldsymbol{X} \ \boldsymbol{X}^{\text{ref}} \right)^\top = \boldsymbol{a} + \boldsymbol{b}z \mid z \in \mathcal{Z} \right\}$ by using the scalar parameter $z \in \mathbb{R}$, where

$$\mathcal{Z} = \left\{ z \in \mathbb{R} \mid \mathcal{E}_{\boldsymbol{a}+\boldsymbol{b}z} = \mathcal{E}_{\text{obs}} \right\}. \tag{11}$$

Proof. The proof is deferred to Appendix A.1 ■

Lemma 1 indicates that we do not need to consider the $2n$-dimensional data space. Instead, we only need to consider the *one-dimensional projected* data space $\mathcal{Z}$ in (11). Now, let us consider a random variable $Z \in \mathbb{R}$ and its observation $Z_{\text{obs}} \in \mathbb{R}$ that satisfies $(\boldsymbol{X} \ \boldsymbol{X}^{\text{ref}})^\top = \boldsymbol{a} + \boldsymbol{b}Z$ and $(\boldsymbol{X}_{\text{obs}} \ \boldsymbol{X}_{\text{obs}}^{\text{ref}})^\top = \boldsymbol{a} + \boldsymbol{b}Z_{\text{obs}}$. The selective p-value (7) is rewritten as

$$p_{\text{selective}} = \mathbb{P}_{\text{H}_0} \left(|Z| \geq |Z_{\text{obs}}| \mid Z \in \mathcal{Z} \right). \tag{12}$$

Because the variable $Z \sim \mathbb{N}(0, \boldsymbol{\eta}^\top \tilde{\Sigma}\boldsymbol{\eta})$ under the null hypothesis, $Z \mid Z \in \mathcal{Z}$ follows a *truncated* normal distribution. Once the truncation region $\mathcal{Z}$ is identified, computation of the selective p-value in (12) is straightforward. Therefore, the remaining task is to identify $\mathcal{Z}$.

In general, computation of $\mathcal{Z}$ in (11) is difficult because we need to identify the selection event $\mathcal{E}_{\boldsymbol{a}+\boldsymbol{b}z}$ for all values of $z \in \mathbb{R}$, which is computationally challenging. In the next section, we will show that the challenge can be resolved under a wide class of problems.

4 Piecewise Linear Network

The problem of computing selective p-values for the selected salient region is casted into the problem of identifying a set of intervals $\mathcal{Z} = \{z \in \mathbb{R} \mid \mathcal{E}_{\boldsymbol{X}(z), \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}}\}$. Given the complexity of saliency computation in a trained DNN, it seems difficult to obtain $\mathcal{Z}$. In this section, however, we explain that this is feasible for a wide class of CNNs.

piecewise linear components in CNN The key idea is to note that most of basic operations and common activation functions used in a trained CNN can be represented as piecewise linear functions in the following form:

Definition 1. *(Piecewise Linear Function) A piecewise linear function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$ is defined as:*

$$f(\boldsymbol{X}) = \begin{cases} \Psi_1^f \boldsymbol{X} + \boldsymbol{\psi}_1^f, & \textit{if } \boldsymbol{X} \in \mathcal{P}_1^f := \{\boldsymbol{X}' \in \mathbb{R}^n \mid \Delta_1^f \boldsymbol{X}' \leq \boldsymbol{\delta}_1^f\}, \\ \Psi_2^f \boldsymbol{X} + \boldsymbol{\psi}_2^f, & \textit{if } \boldsymbol{X} \in \mathcal{P}_2^f := \{\boldsymbol{X}' \in \mathbb{R}^n \mid \Delta_2^f \boldsymbol{X}' \leq \boldsymbol{\delta}_2^f\}, \\ \quad \vdots & \\ \Psi_{K(f)}^f \boldsymbol{X} + \boldsymbol{\psi}_{K(f)}^f, & \textit{if } \boldsymbol{X} \in \mathcal{P}_{K(f)}^f := \{\boldsymbol{X}' \in \mathbb{R}^n \mid \Delta_{K(f)}^f \boldsymbol{X}' \leq \boldsymbol{\delta}_{K(f)}^f\} \end{cases}$$

where Ψ_k^f, $\boldsymbol{\psi}_k^f$, Δ_k^f and $\boldsymbol{\delta}_k^f$ for $k \in [K(f)]$ are certain matrices and vectors with appropriate dimensions, $\mathcal{P}_k^f := \{\boldsymbol{x} \in \mathbb{R}^n \mid \Delta_k^f \boldsymbol{x} \leq \boldsymbol{\delta}_k^f\}$ is a polytope in $\mathbb{R}^n$ for $k \in [K(f)]$, and $K(f)$ is the number of polytopes for the function f.

Examples of piecewise linear components in a trained CNN are shown in Appendix A.2.

Piecewise Linear Network

Definition 2. *(Piecewise Linear Network) A network obtained by concatenations and compositions of piecewise linear functions is called* piecewise linear network.

Since the concatenation and the composition of piecewise linear functions is clearly piecewise linear function, the output of any node in the piecewise linear network is written as a piecewise linear function of an input vector $\boldsymbol{X}$. This is also true for the saliency map function $\mathcal{A}_i(\boldsymbol{X}), i \in [n]$.

Algorithm 1 `SI_DNN_Saliency`

Input: $\boldsymbol{X}^{\text{obs}}, z_{\min}, z_{\max}, \mathcal{T} \leftarrow \emptyset$

1: Obtain $\mathcal{E}_{\text{obs}}$, compute $\boldsymbol{\eta}$ as well as $\boldsymbol{a}$ and $\boldsymbol{b} \leftarrow$ Eq. (10), and initialize: $t = 1$, $z_t = z_{\min}$
2: **for** $t \leq T$ **do**
3: Compute z_{t+1} by Auto-Conditioning (see §5)
4: **if** $\mathcal{E}_{\boldsymbol{X}(z), \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}}$ in $z \in [z_t, z_{t+1}]$ (by using Eq.(14)) **then**
5: $\mathcal{T} \leftarrow \mathcal{T} + \{t\}$
6: **end if**
7: $t = t + 1$
8: **end for**
9: Identify $\mathcal{Z} \leftarrow \bigcup_{t \in \mathcal{T}} [z_t, z_{t+1}]$
10: $p_{\text{selective}} \leftarrow$ Eq. (12)

Output: $p_{\text{selective}}$

Furthermore, as discussed in §4, we can focus on the input vector in the form of $\boldsymbol{X}(z) = \boldsymbol{a}_{1:n} + \boldsymbol{b}_{1:n} z$ which is parametrized by a scalar parameter $z \in \mathbb{R}$. Therefore, the saliency map value for each element is written as a piecewise linear function of the scalar parameter z, i.e.,

$$\mathcal{A}_i(\boldsymbol{X}(z)) = \begin{cases} \kappa_1^{\mathcal{A}_i} z + \rho_1^{\mathcal{A}_i}, & \text{if } z \in [L_1^{\mathcal{A}_i}, U_1^{\mathcal{A}_i}], \\ \kappa_2^{\mathcal{A}_i} z + \rho_2^{\mathcal{A}_i}, & \text{if } z \in [L_2^{\mathcal{A}_i}, U_2^{\mathcal{A}_i}], \\ \quad \vdots & \\ \kappa_{K(\mathcal{A}_i)}^{\mathcal{A}_i} z + \rho_{K(\mathcal{A}_i)}^{f}, & \text{if } z \in [L_{K(\mathcal{A}_i)}^{\mathcal{A}_i}, U_{K(\mathcal{A}_i)}^{\mathcal{A}_i}], \end{cases} \tag{13}$$

where $K(\mathcal{A}_i)$ is the number of linear pieces of the piecewise linear function, $\kappa_k^{\mathcal{A}_i}, \rho_k^{\mathcal{A}_i}$ are certain scalar parameters, $[L_k^{\mathcal{A}_i}, U_k^{\mathcal{A}_i}]$ are intervals for $k \in [K(\mathcal{A}_i)]$ (note that a polytope in $\mathbb{R}^n$ is reduced to an interval when it is projected onto one-dimensional space).

This means that, for each piece of the piecewise linear function, we can identify the interval of z such that $\mathcal{A}_i(\boldsymbol{X}(z)) \geq \tau$ as follows [2]

$$z \in \begin{cases} \left[\max\left(L_k^{\mathcal{A}_i}, \left(\tau - \rho_k^{\mathcal{A}_i}\right)/\kappa_k^{\mathcal{A}_i}\right),\ U_k^{\mathcal{A}_i}\right] & \text{if } \kappa_k^{\mathcal{A}_i} > 0 \\ \left[L_k^{\mathcal{A}_i}, \min\left(U_k^{\mathcal{A}_i}, \left(\tau - \rho_k^{\mathcal{A}_i}\right)/\kappa_k^{\mathcal{A}_i}\right),\right] & \text{if } \kappa_k^{\mathcal{A}_i} < 0 \end{cases} \quad \Rightarrow \quad \mathcal{A}_i(\boldsymbol{X}(z)) \geq \tau. \tag{14}$$

With a slight abuse of notation, let us collectively denote the finite number of intervals on $z \in \mathbb{R}$ that are defined by $L_k^{\mathcal{A}_i}, U_k^{\mathcal{A}_i}, (\tau - \rho_i^{\mathcal{A}_i}/\kappa_k^{\mathcal{A}_i})$ for all $(k, i) \in [K(\mathcal{A}_i)] \times [n]$ as

$$[z_0, z_1], [z_1, z_2], \ldots, [z_{t-1}, z_t], [z_t, z_{t+1}], \ldots, [z_{T-1}, z_T],$$

where $z_{\min} = z_0$ and $z_{\max} = z_T$ are defined such that the probability mass of $z < z_{\min}$ and $z > z_{\max}$ are negligibly small.

Algorithm Algorithm 1 shows how we identify $\mathcal{Z} = \{z \in \mathbb{R} \mid \mathcal{E}_{\boldsymbol{X}(z), \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}}\}$. We simply check the intervals of z in the order of $[z_0, z_1], [z_1, z_2], ..., [z_{T-1}, z_T]$ to see whether $\mathcal{M}_{\boldsymbol{X}(z)} = \mathcal{M}_{\boldsymbol{X}(z_{\text{obs}})}$ and hence $\mathcal{E}_{\boldsymbol{X}(z), \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}}$ or not in the interval by using Eq.(14). Then, the truncation region $\mathcal{Z}$ in Eq.(11) is given as $\mathcal{Z} = \bigcup_{t \in [T] | \mathcal{E}_{\boldsymbol{X}(z), \boldsymbol{X}^{\text{ref}}(z)} = \mathcal{E}_{\text{obs}} \text{ for } z \in [z_t, z_{t+1}]} [z_t, z_{t+1}]$.

5 Implementation: Auto-Conditioning

The bottleneck of our algorithm is Line 3 in Algorithm 1 where we need to find z_{t+1} by considering all the relevant piecewise linear components in a complicated trained CNN. The difficulty here is not only in the computational cost, but also in the implementation cost when the network structure is changed. To circumvent these difficulties, we introduce modular implementation scheme called *Auto-Conditioning*, which is similar to the idea of *Auto-Differentiation* (Baydin et al., 2018).

[2] For simplicity, we omit the description for the case of $\kappa_k^{\mathcal{A}_i} = 0$. In this case, if $\rho_k^{\mathcal{A}_i} \geq \tau$, then $z \in [L_k^{\mathcal{A}_i}, U_k^{\mathcal{A}_i}] \Rightarrow i \in \mathcal{M}_{\boldsymbol{X}(z)}$.

The basic idea in Auto-Conditioning is to add a mechanism to compute and maintain the interval $z \in [L_k^f, U_k^f]$ in which the input $\boldsymbol{X}(z)$ falls for each piecewise linear component f in the network (e.g., layer API in Keras framework). This enables us to automatically compute the interval $[L_k^f, U_k^f]$ of a piecewise linear function f when it is obtained by a concatenation and/or composition of multiple piecewise linear components. If f is obtained by a concatenation of two piecewise linear functions f_1 and f_2, we can easily obtain $[L_k^f, U_k^f] = [L_{k_1}^{f_1}, U_{k_1}^{f_1}] \cap [L_{k_2}^{f_2}, U_{k_2}^{f_2}]$. On the other hand, if f is obtained by a composition of two piecewise linear functions f_1 and f_2, the calculation of the interval is given by the following lemma.

Lemma 2. *Consider a composition of two piecewise linear functions $f(\boldsymbol{X}(z)) = (f_2 \circ f_1)(\boldsymbol{X}(z))$. Given a real value of z, the interval $[L_k^{f_2}, U_k^{f_2}]$ in the input domain of f_2 in which $\boldsymbol{X}(z)$ falls can be computed as*

$$L_{k_2}^{f_2} = \max_{j:(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j<0} \frac{(\boldsymbol{\delta}_{k_2}^{f_2})_j - (\Delta_{k_2}^{f_2}\boldsymbol{\beta}^{f_1})_j}{(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j}, \qquad U_{k_2}^{f_2} = \min_{j:(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j>0} \frac{(\boldsymbol{\delta}_{k_2}^{f_2})_j - (\Delta_{k_2}^{f_2}\boldsymbol{\beta}^{f_1})_j}{(\Delta_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1})_j},$$

where $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$ is the output of f_1 (i.e., input of f_2), whereas $\Delta_{k_2}^{f_2}$ and $\boldsymbol{\delta}_{k_2}^{f_2}$ are obtained by checking the value of $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$. Then, the interval of the composite function is obtained as $[L_k^f, U_k^f] = [L_{k_1}^{f_1}, U_{k_1}^{f_1}] \cap [L_{k_2}^{f_2}, U_{k_2}^{f_2}]$

The proof is shown in Appendix A.3. Here, the variables $\boldsymbol{\beta}^{f_k}$ and $\boldsymbol{\gamma}^{f_k}$ can be recursively computed through layers as

$$\boldsymbol{\beta}^{f_{k+1}} = \Psi_k^{f_k}\boldsymbol{\beta}^{f_k} + \boldsymbol{\psi}_k^{f_k} \quad \text{and} \quad \boldsymbol{\gamma}^{f_{k+1}} = \Psi_k^{f_k}\boldsymbol{\gamma}^{f_k}.$$

Therefore, Lemma 2 indicates that the intervals in which $\boldsymbol{X}(z)$ falls can be *forwardly propagated* through layers. It means that the lower bound $L_k^{\mathcal{A}_i}$ and the upper bound $U_k^{\mathcal{A}_i}$ of the current piece in the piecewise linear function Eq.(13) can be automatically computed by *forward propagation* of the intervals of relevant piecewise linear components.

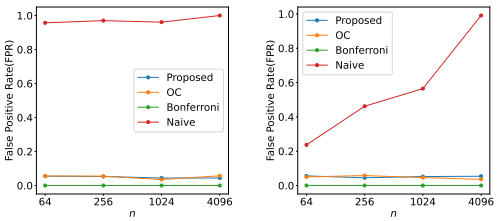
6 EXPERIMENT

We only highlight the main results. More details (methods for comparison, network structure, etc.) can be found in Appendix A.4.

Experimental setup. We compared our proposed method with naive method, over-conditioning (OC) method and Bonferroni correction. To test the false positive rate (FPR) control, we generated 1,000 null images $\boldsymbol{X} = (X_1, ..., X_n)$ and 1,000 reference images $\boldsymbol{X}^{\text{ref}} = (X_1^{\text{ref}}, ..., x_n^{\text{ref}})$, in which $\boldsymbol{s} = \boldsymbol{s}^{\text{ref}} = \mathbf{0}$ and $\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}^{\text{ref}} \sim \mathbb{N}(\mathbf{0}, I_n)$, for each $n \in \{64, 256, 1024, 4096\}$. To test the true positive rate (FPR), we set $n = 256$ and generate 1,000 images in which $s_i = \text{signal}$ for any $i \in \mathcal{S}$ where $\mathcal{S}$ is the "true" salient region and its location is randomly determined, $s_i = 0$ for any $i \notin \mathcal{S}$, and $\boldsymbol{\varepsilon} \sim \mathbb{N}(\mathbf{0}, I_n)$. We set $\Delta \in \{1, 2, 3, 4\}$. The generation of reference images is the same as the case of FPR. In all the experiments, we set threshold for selecting the salient region $\tau = 0$ and significance level $\alpha = 0.05$. We used CAM as the saliency method in all the experiments.

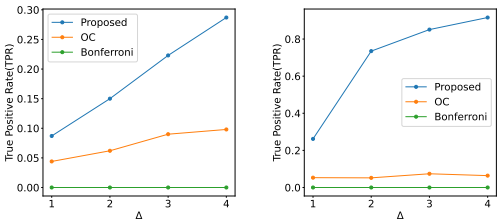
Numerical results. The results of the FPR control are shown in Fig. 2. The proposed method, OC and Bonferroni successfully controlled the FPR in both cases of global null test and mean null test whereas the others cannot. Since naive methods failed to control the FPR, we no longer considered the their TPR. The results of the TPR comparison are shown in Fig. 3. The proposed method had highest TPR in all the cases. The Bonferroni had lowest TPR because it is conservative due to the fact of considering the number of all possible hypotheses. The OC also had low TPR because several extra conditioning is considered which causes the loss of TPR.

Real data experiments. We examine the brain image dataset extracted from the dataset used in Buda et al. (2019), which includes 941 images without tumors and 939 images with tumors. The results for global null test are shown in Figs. 4 and 5. The results for mean null test are shown in Figs. 6 and 7. The naive p-value remains small, even when the image has no tumor region; this indicated that naive p-values cannot be used to quantify the reliability of DNN-based salient regions. The proposed method successfully identifies false positive and true positive detections.



(a) Global null test (b) Mean null test

Figure 2: False Positive Rate (FPR) comparison.



(a) Global null test (b) Mean null test

Figure 3: True Positive Rate (FPR) comparison.

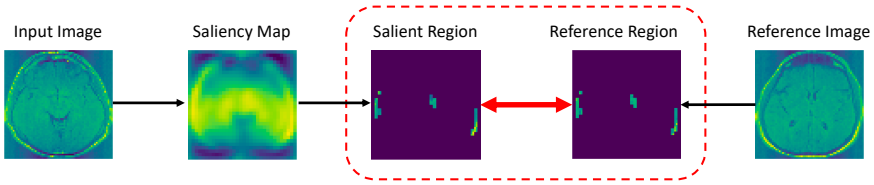


Figure 4: Global null test for image without tumor (p_{naive} = **0.00**, $p_{\text{selective}}$ = **1.00**)

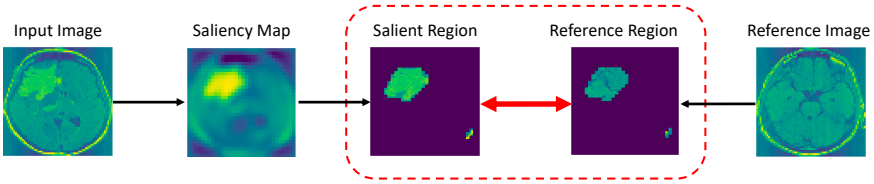


Figure 5: Global null test for image with a tumor (p_{naive} = **0.00**, $p_{\text{selective}}$ = **3.01e-5**).

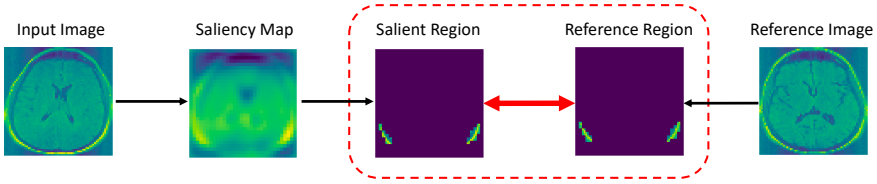


Figure 6: Mean null test for image without tumor (p_{naive} = **0.00**, $p_{\text{selective}}$ = **0.78**).

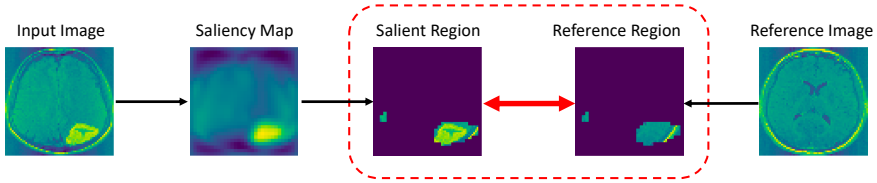


Figure 7: Mean null test for image with a tumor (p_{naive} = **0.00**, $p_{\text{selective}}$ = **1.92e-4**).

7 Conclusion

We proposed a novel method to conduct statistical inference on the significance of the DNN-driven salient regions based on the concept of conditional SI. We provided a novel algorithm for efficiently and flexibly conducting conditional SI for salient regions. We conducted experiments on both synthetic and real-world datasets to demonstrate the performance of our proposed method.

REFERENCES

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

Atilim Gunes Baydin, Barak A Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: a survey. *Journal of Marchine Learning Research*, 18:1–43, 2018.

Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109:218–225, 2019.

Collin Burns, Jesse Thomason, and Wesley Tansey. Interpreting black box models via hypothesis testing. In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, pp. 47–57, 2020.

Shuxiao Chen and Jacob Bien. Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, pp. 1–12, 2019.

Diptesh Das, Vo Nguyen Le Duy, Hiroyuki Hanada, Koji Tsuda, and Ichiro Takeuchi. Fast and more powerful selective inference for sparse high-order interaction model. *arXiv preprint arXiv:2106.04929*, 2021.

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*, pp. 13589–13600, 2019.

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

Alexey Dosovitskiy and Thomas Brox. Inverting visual representations with convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4829–4837, 2016.

Vo Nguyen Le Duy and Ichiro Takeuchi. More powerful conditional selective inference for generalized lasso by parametric programming. *arXiv preprint arXiv:2105.04920*, 2021.

Vo Nguyen Le Duy and Ichiro Takeuchi. Exact statistical inference for time series similarity using dynamic time warping by selective inference. *arXiv preprint arXiv:2202.06593*, 2022.

Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. In *Advances in Neural Information Processing Systems*, pp. 11356–11367, 2020.

Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi. Quantifying statistical significance of neural network-based image segmentation by selective inference. *Advances in Neural Information Processing Systems*, 2022.

William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3429–3437, 2017.

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

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems*, pp. 2925–2936, 2019.

Sangwon Hyun, Max G'sell, and Ryan J Tibshirani. Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097, 2018.

Sangwon Hyun, Kevin Z Lin, Max G'Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77 (3):1037–1049, 2021.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

Vo Nguyen Le Duy and Ichiro Takeuchi. Parametric programming approach for more powerful and general lasso selective inference. In *International Conference on Artificial Intelligence and Statistics*, pp. 901–909. PMLR, 2021.

Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

Keli Liu, Jelena Markovic, and Robert Tibshirani. More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*, 2018.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pp. 4765–4774, 2017.

Aravindh Mahendran and Andrea Vedaldi. Understanding deep image representations by inverting them. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems*, pp. 7775–7784, 2018.

Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*, 28, 2016.

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*, pp. 1135–1144, 2016.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Kazuya Sugiyama, Vo Nguyen Le Duy, and Ichiro Takeuchi. More powerful and general selective inference for stepwise feature selection using the homotopy continuation approach. In *Proceedings of the 38th International Conference on Machine Learning*, 2021a.

Ryota Sugiyama, Hiroki Toda, Vo Nguyen Le Duy, Yu Inatsu, and Ichiro Takeuchi. Valid and exact statistical inference for multi-dimensional multiple change-points by selective inference. *arXiv preprint arXiv:2110.08989*, 2021b.

Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 3338–3347. JMLR. org, 2017.

Kosuke Tanizaki, Noriaki Hashimoto, Yu Inatsu, Hidekata Hontani, and Ichiro Takeuchi. Computing valid p-values for image segmentation by selective inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9553–9562, 2020.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Toshiaki Tsukurimichi, Yu Inatsu, Vo Nguyen Le Duy, and Ichiro Takeuchi. Conditional selective inference for robust regression and outlier detection using piecewise-linear homotopy continuation. *arXiv preprint arXiv:2104.10840*, 2021.

Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pp. 2469–2477, 2016.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Xinyang Zhang, Ningfei Wang, Hua Shen, Shouling Ji, Xiapu Luo, and Ting Wang. Interpretable deep learning under fire. In *29th {USENIX} Security Symposium ({USENIX} Security 20)*, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.