A APPENDIX

A.1 PROOF OF LEMMA 1

In the mean null test, according to the second condition in (9), we have

$$\begin{aligned}
&\Omega_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \Omega_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} \\
\Leftrightarrow \quad &\left(I_{2n} - \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}}{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}\right)\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix} = \Omega_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} \\
\Leftrightarrow \quad &\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix} = \Omega_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} + \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}{\|\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\|^2}\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix}.
\end{aligned}$$

By defining $\boldsymbol{a} = \boldsymbol{q}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}}$, $\boldsymbol{b} = \frac{\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}}{\|\boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}\|^2}$, $z = \boldsymbol{\eta}_{\mathcal{M}_{\boldsymbol{X}}}^{\top}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix}$, we obtain the result in Lemma 1.

In the global null test, according to the second condition in (9),

$$\begin{aligned}
&\mathcal{U}_{\boldsymbol{X},\boldsymbol{X}^{\text{ref}}} = \mathcal{U}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} \\
\Leftrightarrow\ &P^{\perp}_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix} = \mathcal{U}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} \\
\Leftrightarrow\ &(I_{2_n} - P_{\mathcal{M}_{\boldsymbol{X}}})\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix} = \mathcal{U}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} \\
\Leftrightarrow\ &\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix} = \mathcal{U}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}} + \mathcal{V}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}}\sigma^{-1}\left\|P_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix}\right\|.
\end{aligned}$$

By defining $\boldsymbol{a} = \mathcal{U}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}}$, $\boldsymbol{b} = \mathcal{V}_{\boldsymbol{X}_{\text{obs}},\boldsymbol{X}^{\text{ref}}_{\text{obs}}}$, $z = \sigma^{-1}\left\|P_{\mathcal{M}_{\boldsymbol{X}}}\begin{pmatrix}\boldsymbol{X}\\ \boldsymbol{X}^{\text{ref}}\end{pmatrix}\right\|$, we obtain the result in Lemma 1.

A.2 EXAMPLES OF PIECEWISE LINEAR FUNCTIONS

Examples of piecewise linear components in a trained CNN with $\boldsymbol{X} \in \mathbb{R}^2$ are provided as follows:

ReLU: Consider f is ReLU function. Then, $K(f) = 4$, $\boldsymbol{\psi}_k = (0\ 0)^{\top}$ for any $k \in [4]$,

$$\begin{aligned}
&\Psi_1^f = \begin{pmatrix}0 & 0\\ 0 & 0\end{pmatrix},\ \mathcal{P}_1^f = \left\{\boldsymbol{X} : \begin{matrix}X_1 < 0,\\ X_2 < 0\end{matrix}\right\}, \quad \Psi_2^f = \begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix},\ \mathcal{P}_2^f = \left\{\boldsymbol{X} : \begin{matrix}X_1 < 0,\\ X_2 \geq 0\end{matrix}\right\}, \\
&\Psi_3^f = \begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix},\ \mathcal{P}_3^f = \left\{\boldsymbol{X} : \begin{matrix}X_1 \geq 0,\\ X_2 < 0\end{matrix}\right\}, \quad \Psi_4^f = \begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix},\ \mathcal{P}_4^f = \left\{\boldsymbol{X} : \begin{matrix}X_1 \geq 0,\\ X_2 \geq 0\end{matrix}\right\}.
\end{aligned}$$

This can be similarly extended to the case of Leaky ReLU.

Max-pooling: Consider $f(\boldsymbol{X}) = \max\{X_1, X_2\}$. Then, it is represented as a piecewise linear function with $K(f) = 2$, $\boldsymbol{\psi}_k = (0)$ for any $k \in [2]$,

$$\Psi_1^f = (1 \quad 0),\ \mathcal{P}_1^f = \{\boldsymbol{X} : X_1 \geq X_2\}, \quad \Psi_2^f = (0 \quad 1),\ \mathcal{P}_2^f = \{\boldsymbol{X} : X_1 < X_2\}.$$

Convolution and matrix-vector multiplication: In a neural network, the multiplication results between the weight matrix and the output of the previous layer and its summation with the bias vector is a linear function. In a CNN, the convolution operation is obviously a linear function.

Upsampling: Consider f is the upsampling operation on $\boldsymbol{X} \in \mathbb{R}^2$, then it can be represented as a piecewise linear function with $K(f) = 1$, $\boldsymbol{\psi}_1 = (0\ 0\ 0\ 0)^{\top}$,

$$\Psi_1^f = \begin{pmatrix}1 & 1 & 0 & 0\\ 0 & 0 & 1 & 1\end{pmatrix}^{\top}, \quad \mathcal{P}_1^f = \mathbb{R}^2.$$

Sigmoid and hyperbolic tangent: If there is any specific demand to use non-piecewise linear activation functions, we can apply a piecewise-linear approximation approach to these functions.

A.3 PROOF OF LEMMA 2

At f_1, given a real value z, the input is $\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z = \boldsymbol{a}_{1:n} + \boldsymbol{b}_{1:n} z$. By checking the value of this input, we can easily obtain the polytope

$$\{\Delta_{k_1}^{f_1}(\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z) \leq \boldsymbol{\delta}_{k_1}^{f_1}\}, \quad k_1 \in [K(f_1)],$$

that $\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z$ belongs to. Based on the obtained polytope, we can calculate the interval $[L_{k_1}^{f_1}, U_{k_1}^{f_1}]$,

$$L_{k_1}^{f_1} = \max_{j:(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j<0} \frac{(\boldsymbol{\delta}_{k_1}^{f_1})_j - (\Delta_{k_1}^{f_1}\boldsymbol{\beta}^{f_0})_j}{(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j} \quad \text{and} \quad U_{k_1}^{f_1} = \min_{j:(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j>0} \frac{(\boldsymbol{\delta}_{k_1}^{f_1})_j - (\Delta_{k_1}^{f_1}\boldsymbol{\beta}^{f_0})_j}{(\Delta_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0})_j}.$$

Moreover, based on the obtained polytope, we can easily obtain $\Psi_{k_1}^{f_1}$ and $\boldsymbol{\psi}_{k_1}^{f_1}$, $k_1 \in [K(f_1)]$. Therefore, the output of the first layer at z can be defined as

$$\begin{aligned} f_1(z) &= \Psi_{k_1}^{f_1}(\boldsymbol{\beta}^{f_0} + \boldsymbol{\gamma}^{f_0} z) + \boldsymbol{\psi}_{k_1}^{f_1} \\ &= \boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z, \end{aligned}$$

where $\boldsymbol{\beta}^{f_1} = \Psi_{k_1}^{f_1}\boldsymbol{\beta}^{f_0} + \boldsymbol{\psi}_{k_1}^{f_1}$ and $\boldsymbol{\gamma}^{f_1} = \Psi_{k_1}^{f_1}\boldsymbol{\gamma}^{f_0}$. Next, we input $\boldsymbol{\beta}^{f_1}, \boldsymbol{\gamma}^{f_1}$ to f_2.

At the 2^{nd} layer, similarly, the input is $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$. By checking the value of this input, we can easily obtain the polytope

$$\{\Delta_{k_2}^{f_2}(\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z) \leq \boldsymbol{\delta}_{k_2}^{f_2}\}, \quad k_2 \in [K(f_2)],$$

that $\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z$ belongs to. Based on the obtained polytope, we can calculate the interval $[L_{k_2}^{f_2}, U_{k_2}^{f_2}]$,

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

Moreover, based on the obtained polytope, we can easily obtain $\Psi_{k_2}^{f_2}$ and $\boldsymbol{\psi}_{k_2}^{f_2}$, $k_2 \in [K(f_2)]$. Therefore, the output of the first layer at z can be defined as

$$\begin{aligned} f_2(z) &= \Psi_{k_2}^{f_2}(\boldsymbol{\beta}^{f_1} + \boldsymbol{\gamma}^{f_1} z) + \boldsymbol{\psi}_{k_2}^{f_2} \\ &= \boldsymbol{\beta}^{f_2} + \boldsymbol{\gamma}^{f_2} z, \end{aligned}$$

where $\boldsymbol{\beta}^{f_2} = \Psi_{k_2}^{f_2}\boldsymbol{\beta}^{f_1} + \boldsymbol{\psi}_{k_2}^{f_2}$ and $\boldsymbol{\gamma}^{f_2} = \Psi_{k_2}^{f_2}\boldsymbol{\gamma}^{f_1}$.

A.4 EXPERIMENTAL DETAILS.

Methods for comparison. We compared our proposed method with the following approaches:

• Naive: the classical z-test is used to calculate the naive p-value.

• Bonferroni: the number of all possible hypotheses are considered to account for the selection bias. The p-value is computed by $p_{\text{bonferroni}} = \min(1, p_{\text{naive}} * 2^n)$

• Over-conditioning (OC): additionally conditioning on the observed activeness and inactiveness of all the nodes. The limitation of this method is its low statistical power due to over-conditioning.

Network for experiments. In all the experiments, we used the network structure shown in Fig. 8. We generate 1000 images from $N(\mathbf{0}, I_n)$ as negative samples, and 1000 images from $N(\boldsymbol{s}, I_n)$ as positive samples. We set $s_i = 0, \forall s_i \in \mathcal{S}$, and $s_i \neq 1, \forall s_i \notin \mathcal{S}$, where $\mathcal{S}$ are the set of indexes whose pixels have signals and the location is randomly determined and $|S| = n/4$ for each image size n. We trained the network with these images and used Adam (Kingma & Ba, 2014) as an optimizer.

Experiment for robustness. Additionally, we confirm the robustness of the proposed method in terms of FPR control by applying our method to data following Laplace distribution (Laplace), skew normal (Skew) distribution and t distribution (t-dist). The results are shown in Fig. 9. Our method still maintains good performance in terms of FPR control.

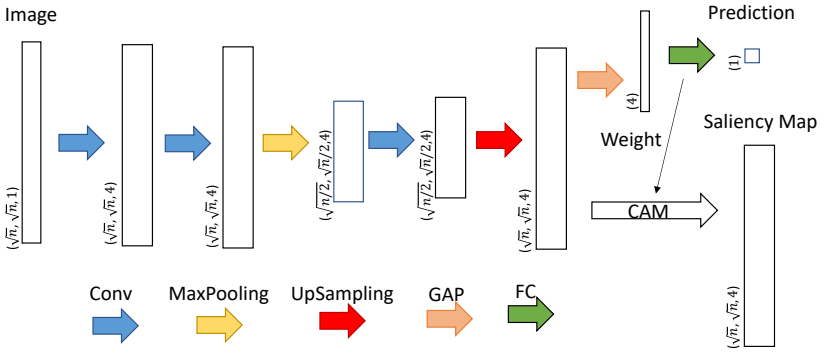
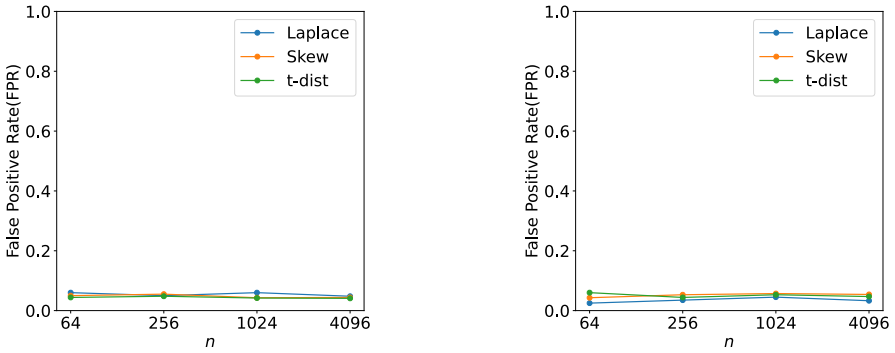


Figure 8: Network structure.



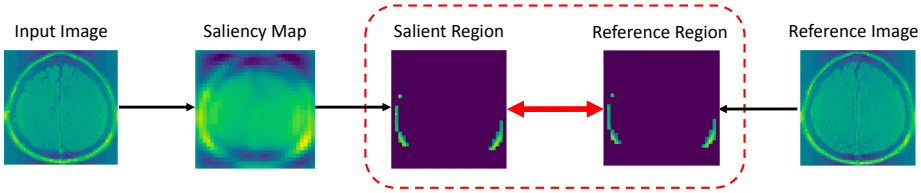
(a) Mean null test

(b) Global null test

Figure 9: False positive rate of the proposed method when data are nonnormal

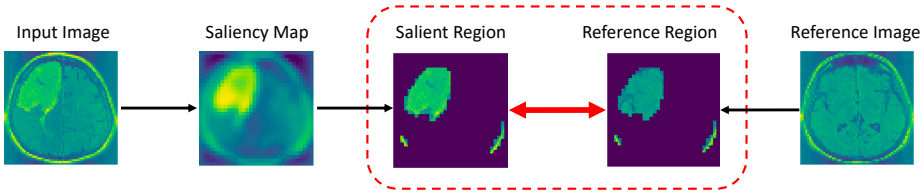
Experimental setting on brain image dataset. We examine the brain image dataset extracted from the dataset used in Buda et al. (2019), which includes 941 images without tumors (C1) and 939 images with tumors (C2). We selected 50 images from C1 as reference images and used 841 images from C1 and 889 images from C2 for DNN training. The remaining images from C1 and C2 are used for demonstrating the advantages of the proposed selective p-value.

More results on brain image dataset. Additional results are shown in Figs. 10, 11, 12 and 13



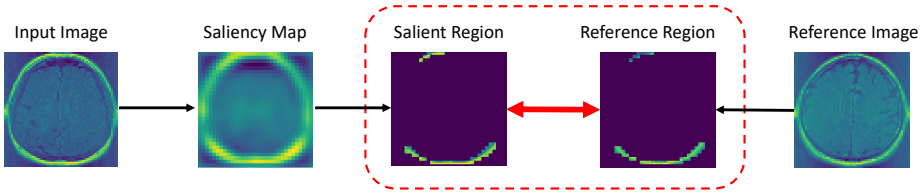
(a) $p_{\text{naive}} = \mathbf{0.01}$, $p_{\text{selective}} = \mathbf{0.47}$

Figure 10: Inference on salient regions for images without tumor (mean null test).



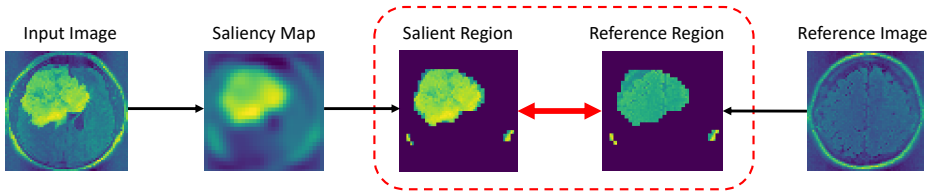
(a) $p_{\text{naive}} = \mathbf{0.00}$, $p_{\text{selective}} = \mathbf{2.82 \times 10^{-4}}$

Figure 11: Inference on salient regions for images where there exists a tumor (mean null test).



(a) $p_{\text{naive}} = \mathbf{3.00 \times 10^{-4}}$, $p_{\text{selective}} = \mathbf{0.29}$

Figure 12: Inference on salient regions for images without tumor (global null test).



(a) $p_{\text{naive}} = \mathbf{0.00}$, $p_{\text{selective}} = \mathbf{2.66 \times 10^{-20}}$

Figure 13: Inference on salient regions for images where there exists a tumor (global null test).