



# OPEN Improving explainable AI with patch perturbation-based evaluation pipeline: a COVID-19 X-ray image analysis case study

Jimin Sun<sup>1,4</sup>, Wenqi Shi<sup>2,4</sup>, Felipe O. Giuste<sup>3</sup>, Yog S. Vaghani<sup>3</sup>, Lingzi Tang<sup>3</sup> & May D. Wang<sup>3</sup>✉

Recent advances in artificial intelligence (AI) have sparked interest in developing explainable AI (XAI) methods for clinical decision support systems, especially in translational research. Although using XAI methods may enhance trust in black-box models, evaluating their effectiveness has been challenging, primarily due to the absence of human (expert) intervention, additional annotations, and automated strategies. In order to conduct a thorough assessment, we propose a patch perturbation-based approach to automatically evaluate the quality of explanations in medical imaging analysis. To eliminate the need for human efforts in conventional evaluation methods, our approach executes poisoning attacks during model retraining by generating both static and dynamic triggers. We then propose a comprehensive set of evaluation metrics during the model inference stage to facilitate the evaluation from multiple perspectives, covering a wide range of correctness, completeness, consistency, and complexity. In addition, we include an extensive case study to showcase the proposed evaluation strategy by applying widely-used XAI methods on COVID-19 X-ray imaging classification tasks, as well as a thorough review of existing XAI methods in medical imaging analysis with evaluation availability. The proposed patch perturbation-based workflow offers model developers an automated and generalizable evaluation strategy to identify potential pitfalls and optimize their proposed explainable solutions, while also aiding end-users in comparing and selecting appropriate XAI methods that meet specific clinical needs in real-world clinical research and practice.

Despite existing papers showcasing novel artificial intelligence (AI)-enabled clinical decision support in disease diagnosis, prognosis, risk prediction, and treatment planning, few have had a significant clinical impact<sup>1</sup>. For instance, the need for fast COVID-19 detection has resulted in a massive number of AI solutions to alleviate this clinical burden during the pandemic<sup>2</sup>. Unfortunately, the lack of model transparency largely restricted the impact of AI-enabled solutions during the COVID-19 pandemic<sup>3</sup>. *Explainable Artificial Intelligence (XAI)* refers to the development of AI systems or machine learning models that can be comprehended and trusted by humans, particularly in terms of how the system arrived at a specific decision or recommendation<sup>4</sup>. In translational informatics, XAI aims to provide transparency and interpretability to high-performing but opaque AI models, thereby enabling users to understand, trust, and promote the adoption of AI-enabled clinical decision support systems in real-world applications<sup>5,6</sup>.

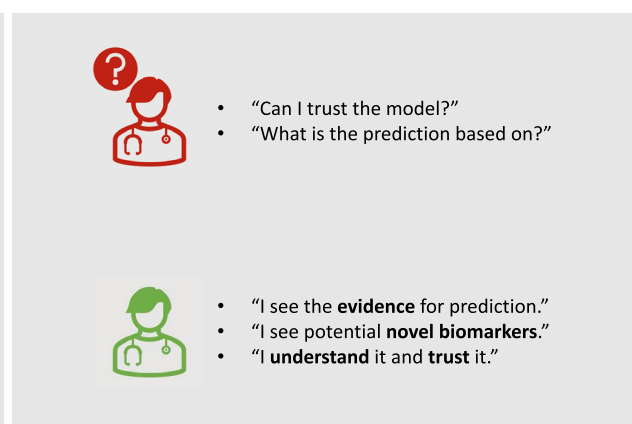
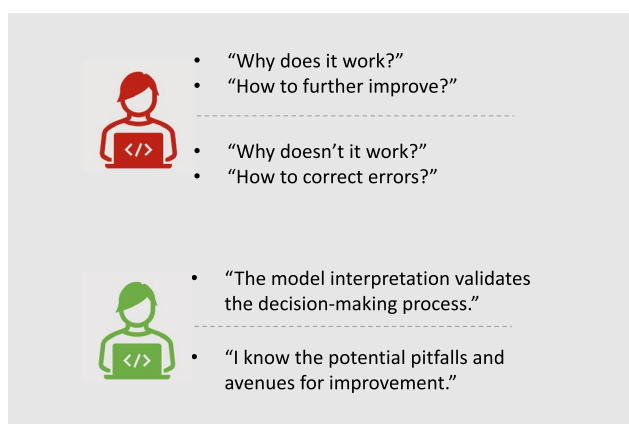
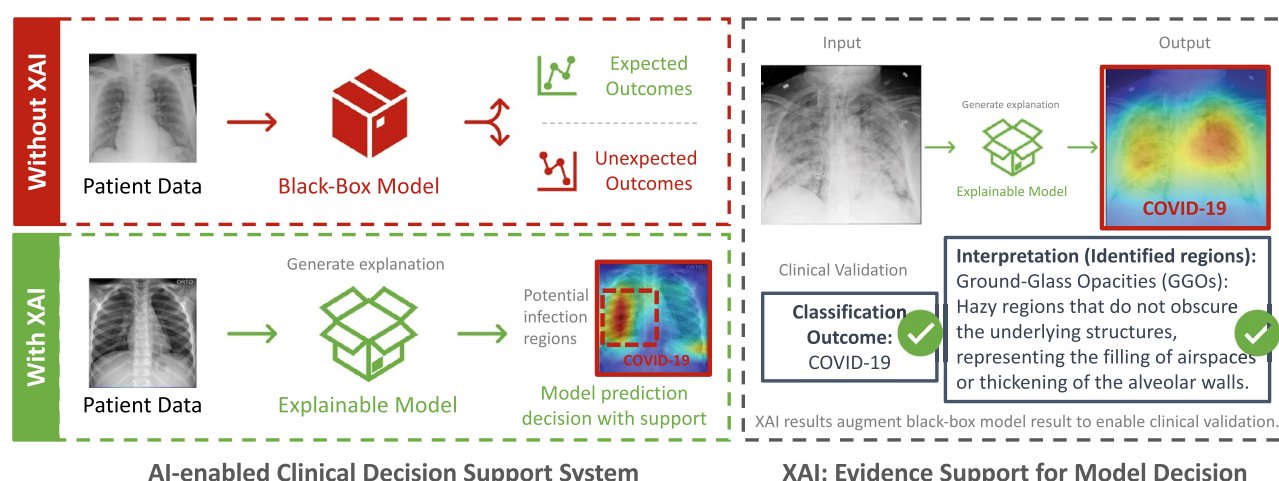
As the comprehension of neural networks holds paramount significance in fostering user trust, the interpretation of model behavior has gained escalating attention, especially in biomedical and clinical decision support systems<sup>7</sup>. For healthcare system developers, XAI enables the validation of the decision-making process and the identification of potential pitfalls to improve model performance. By providing transparency and interpretability, XAI helps developers to build models that are reliable and effective, while minimizing the risk of errors or unintended consequences<sup>8</sup>. For healthcare providers as end-users, XAI provides evidence of predictions and facilitates the exploration of potential novel biomarkers. By allowing clinicians to comprehend the reasoning behind a particular decision or recommendation, XAI has the potential to increase trust in AI systems, enhance

<sup>1</sup>School of Computer Science and Engineering, Georgia Institute of Technology, Atlanta 30322, USA. <sup>2</sup>School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta 30322, USA. <sup>3</sup>The Wallace H. Coulter Department of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta 30322, USA. <sup>4</sup>These authors contributed equally: Jimin Sun and Wenqi Shi. ✉email: maywang@gatech.edu

clinical confidence<sup>7</sup>, promote the widespread implementation of AI-based clinical decision support systems, and ultimately result in improved patient outcomes and better healthcare delivery<sup>3</sup>.

Although XAI plays an important role in the widespread of AI-enabled clinical decision support systems, very few studies have evaluated the quality of XAI insights<sup>9</sup>. XAI is a branch of AI that focuses on creating systems that provide clear, understandable explanations for their actions and decision-making processes<sup>3</sup>. Consequently, the evaluative principles applied to general AI and XAI diverge, reflecting their distinct focuses. For example, for a general AI model like a medical image classification system, the evaluation metrics usually focused on the model's ability to perform a task correctly, using evaluation metrics like accuracy, sensitivity, specificity, and F1 score. In evaluating an XAI model, the emphasis lies on its ability to elucidate the rationale behind its predictions in a manner that is readily comprehensible to humans (e.g., model developers and end-users, as shown in Fig. 1).

We conducted an extensive literature review of state-of-the-art COVID-19 radiographic imaging studies to investigate the XAI applications. Out of the 55 XAI applications examined (see Table S1 in the supplementary materials), we noticed only 8 of them included a qualitative evaluation by human expert validation<sup>10–17</sup>. The majority of these applications relied on clinical expert validation of saliency map visualizations for qualitative evaluation, without incorporating quantitative evaluation methods. For example, Brunese et al.<sup>10</sup> collaborated with radiologists who annotated the regions indicative of COVID-19 manifestations. These annotations were then compared with activation maps derived from XAI methods. The extent of overlap between the radiologists' markings and the XAI-derived regions served as a robust indicator of the efficacy of the proposed XAI models. Unfortunately, such approaches were usually time-consuming since they involved extensive human validation or additional annotations such as manual pixel-level or bounding box annotations for validating regions of interest.



### System Developers

### Healthcare Providers

**Figure 1.** Motivation for model developers and clinicians to implement XAI in AI-enabled clinical decision support systems. XAI approaches imbue originally opaque, black-box models with the ability to pinpoint or emphasize areas that contribute most significantly to the final decision-making process. For example, for a case study in COVID-19 diagnosis utilizing X-ray imaging, XAI could augment original black-box model (decision only) by highlighting potential infectious regions like Ground-Glass Opacities (GGOs), which are indistinct regions that do not mask the underlying structures and typically signify the filling of airspaces or thickening of alveolar walls<sup>56</sup>. For healthcare system developers, such interpretation enables the validation of the decision-making process and the identification of potential pitfalls for improved model performance. For healthcare providers as end-users, XAI provides evidence of prediction and facilitates the exploration of potential novel biomarkers for clinical confidence and widespread adoption.

Furthermore, these studies solely focused on clinical validation by end-users, such as clinicians, without considering the need for evaluation by model developers. A well-developed evaluation procedure is crucial for model developers as it can provide an automated and generalizable evaluation strategy to identify potential pitfalls and optimize their explainable solutions. On the other hand, for end-users, such a procedure can aid in comparing and selecting appropriate XAI methods that best suit specific clinical needs in real-world applications. While XAI methods have the potential to increase trust and improve transparency in black-box models, evaluating their efficacy is an ongoing challenge due to the lack of automated assessment workflows, human interventions, additional validations, and comprehensive evaluation metrics<sup>9</sup>.

To address these challenges, we propose a patch perturbation-based automated pipeline to facilitate the evaluation of XAI methods in medical imaging analysis. Our approach employs static and dynamic triggers to generate poisoning attacks during model retraining, followed by a comprehensive evaluation of explanation generation and representation using multiple metrics. Specifically, we include an extensive case study to showcase the proposed evaluation strategy by applying widely used XAI methods on COVID-19 X-ray imaging classification tasks. This study contributes to the development and evaluation of robust and reliable XAI methods for medical imaging analysis, with implications for translational research. The main contribution of this work is three-fold:

- We present an automated evaluation workflow in medical imaging informatics that applies poisoning attacks during the model retraining stage to obviate the need for ground truth in conventional evaluation methods, with the generation of both static and dynamic triggers enabling generalization to real-world noise sources and biases.
- We provide a comprehensive set of evaluation metrics to provide a quantitative evaluation without the need for pixel-level ground truth during the model inference stage, facilitating the comparison of different XAI methods on correctness, completeness, consistency, and complexity.
- We present an extensive case study on COVID-19 X-ray image classification tasks, offering a generalizable evaluation strategy for model developers to optimize their proposed explainable solutions and aiding end-users in selecting appropriate XAI methods for real-world clinical practice.

## Related works

Due to its critical role in model comprehension for developers and safety-critical applications for clinicians, explainability has become increasingly important in recent years, and as a result, explanation methods have garnered significant attention for their potential to unveil the opaque nature of deep neural networks. This section presents a review of XAI applications in the biomedical domain, specifically focusing on medical imaging informatics, to demonstrate the significance of transparent models in clinical research and practice. Moreover, we present a summary of current research on evaluating XAI in biomedical applications and demonstrate how our proposed method can address gaps in the field by providing an automated evaluation framework and comprehensive evaluation metrics. Given that our case study revolves around COVID-19 X-ray imaging, we conducted a thorough review of existing AI-enabled decision support systems incorporating XAI methods. A summary of XAI applications in state-of-the-art COVID-19 radiographic imaging studies, along with detailed evaluation information, is available in “[Introduction](#)” section and Table S1 of the supplementary materials.

## Explainable AI in medical imaging informatics

Gradient-based XAI techniques determine important features by evaluating input gradients using back-propagation, with the underlying idea that input features with large gradients have the most significant impact on predictions. Simonyan et al.<sup>18</sup> created a saliency map of input features by calculating the absolute value of partial derivatives of class scores with respect to the input using *back-propagation*. However, changes in gradients could be removed in a backward pass if the input to rectified linear units (ReLU) is negative caused by non-linear operations. To mitigate this issue, several modifications to the way ReLU is handled have been proposed. For example, Zeiler and Fergus<sup>19</sup> proposed “deconvnet” to calculate gradients based on only the sign of gradients from the top layer. Springenberg et al.<sup>20</sup> then proposed *guided backpropagation* by combining standard back-propagation with the “deconvnet” approach, which retains gradients only when both the bottom input and top gradients are positive. In recent studies, *Gradient-weighted Class Activation Mapping (GradCAM)*<sup>21</sup> proposed by Selvaraju et al. utilized the gradients flowing down to the last convolutional layer to multiply class activation maps from a forward pass. The resolution of GradCAM was further enhanced by multiplying Grad-CAM with guided-backpropagated gradients in *guided GradCAM*<sup>21</sup>. These techniques have been widely adopted for clinical decision support systems in medical imaging applications<sup>22–26</sup>, facilitating real-world clinical translation.

Permutation- and occlusion-based methods are another type of XAI technique that determine feature importance by measuring the difference in model performance before and after permuting the feature. Zeiler et al.<sup>19</sup> conducted a *occlusion sensitivity* study demonstrating the impact of occluding certain regions of an input image on the confidence score predicted by a conventional neural network (CNN). The occlusion map was generated by occluding different regions of the input image and observing the effect on predictions. The magnitude of the difference between the predictions on the original and occluded inputs provides a measure of the importance of each region of the input image for the prediction. Similarly, Meyes et al.<sup>27</sup> proposed a feature *ablation study* by removing or modifying the features of an input instance and observing the effect on the prediction. Local Interpretable Model-Agnostic Explanations (LIME)<sup>28</sup> is another commonly used XAI method that determines feature importance by identifying a set of super-pixels (i.e., a patch of pixels) that have the strongest relationship with a prediction label in the context of image classification. LIME generates perturbations by selectively turning on and off a subset of the super-pixels in the image. Permutation-based XAI techniques have been widely

applied in medical imaging informatics, particularly in COVID-19 applications<sup>12,23,29–31</sup>, for generating saliency maps that explain model predictions.

In summary, gradient-based methods utilize pixel-level explanations by learning, modifying, or integrating gradients of the target class as attribution importance scores<sup>3</sup>. However, a potential limitation of these methods is that they violate the strong relevance characteristic of attribution features<sup>32</sup>. Due to the strong interdependence between pixels and their surrounding pixels, redundancies may arise, resulting in imperfect and fragile attributions that resemble an edge detector<sup>33</sup> and are sensitive to small perturbations<sup>25</sup>. Conversely, perturbation-based methods measure the sensitivity of the prediction to the perturbations of regional segments but often fail to satisfy the completeness principle<sup>3,32</sup>. Additionally, as these methods rely on the perturbation or masking mechanism, they can be time-consuming and highly dependent on the segmentation quality<sup>3</sup>. Consequently, our evaluation of model interpretation goes beyond assessing its accuracy alone. Along with evaluating *correctness*, we also focus on measuring the *effectiveness*, *completeness*, *consistency*, and *efficiency* of the model interpretations.

## Evaluation of explainable AI

While few studies have focused on evaluating XAI methods, both qualitative and quantitative evaluation play essential roles in assessing these methods from multiple perspectives<sup>9</sup>. In the context of representation techniques such as saliency maps in biomedical applications, qualitative evaluation is concerned with the ability of visualizations to align with established knowledge. For instance, clinical experts (e.g., radiologists) can assess the effectiveness of attention maps in identifying image regions that are diagnostically relevant or potentially indicative of infection in COVID-19 X-ray imaging<sup>10–17,34</sup>. Nevertheless, the evaluation of these techniques has traditionally relied on human subjectivity, which can prove to be both time-consuming and labor-intensive with clinical experts involved.

Compared to qualitative evaluation, quantitative evaluation is generally more desirable as it usually offers an objective and automated assessment process<sup>3</sup>. Several evaluation pipelines have been proposed based on the occlusion sensitivity experiment introduced by Zeilar and Fergus<sup>19</sup>. These pipelines involve systematically occluding an input image with a patch to monitor the dynamic performance of deep learning models, such as Randomized Input Sampling for Explanation (RISE)<sup>35</sup> and Benchmark Interpretability Methods (BIM)<sup>36</sup>. For explanation accuracy analysis, Nguyen et al.<sup>37</sup> proposed a new evaluation metric for XAI accuracy evaluation, Determining the Highest-Impact Segments (DHIS), which utilized K-Means clustering to group different segments of pixels based on their proximity of color in the image plane. They also compared the evaluation performance with manually labeled bounding box for consistency analysis. Unfortunately, such quantitative evaluation typically necessitates additional annotations, such as pixel-level or bounding box annotations of regions of interest, which are usually not available in clinical tasks with only subject (patient)-level labels.

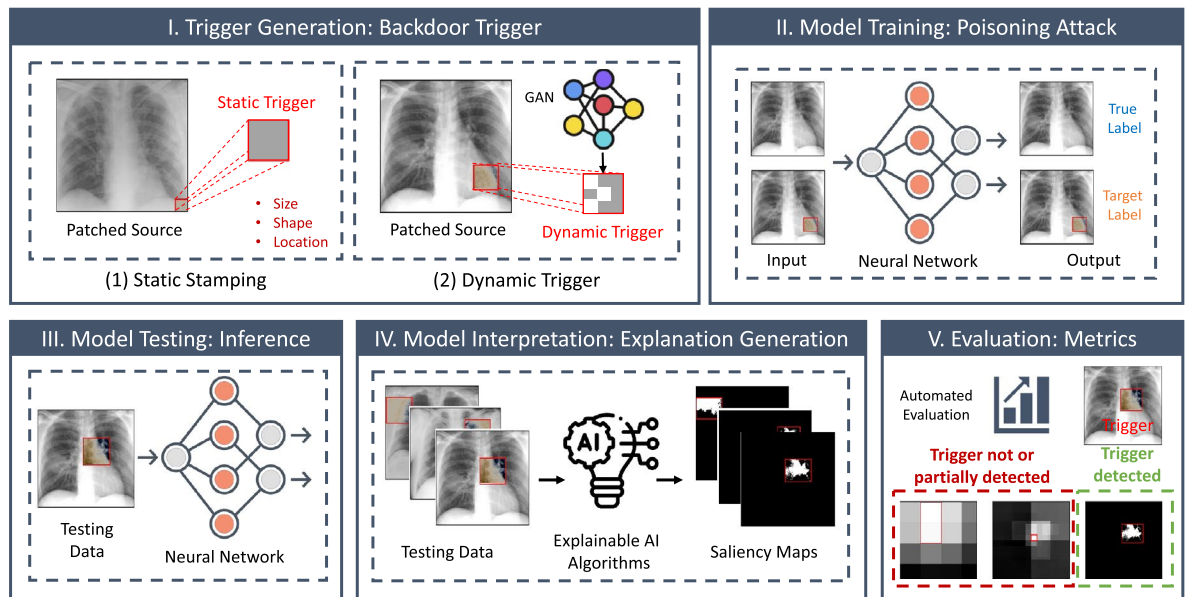
While various methods of adversarial attacks and defenses have been developed to evaluate the robustness of algorithms, there has been limited exploration of adversarial manipulation of explanations or interpretations<sup>38</sup>. For example, Rieger et al.<sup>38</sup> utilized adversarial examples to iteratively update and modify the model weights to alter the explanation while minimizing changes to the input and output. Similar methods<sup>39</sup> have been developed to manipulate the explanation while maintaining visual similarity in the input and output. Lin et al.<sup>40</sup> intuitively applied the attack and defense framework for interpretability assessment through neural backdoors to automate the evaluation procedure and leveraged trojaning on neural networks<sup>41</sup> as the attack framework. In addition to model trojaning, there are several other hidden trigger backdoor attacks<sup>42–44</sup> that can be explored to introduce poisoned data to the victim for training the model, and then activate the attack by showing a specific small trigger pattern at test time to evaluate the effectiveness of XAI methods for defense. However, since the majority of existing poisoning attack frameworks<sup>42–44</sup> rely on static patterns or triggers, few have explored dynamic trigger generation to reflect the variations of artifacts or noise sources in medical imaging<sup>45,46</sup>. Fu et al.<sup>47</sup> conducted a comprehensive experiment to investigate the robustness of Vision Transformers (ViTs) and CNNs against various existing adversarial attacks to understand the underlying reasons. Similarly, Gu et al.<sup>48</sup> and Dong et al.<sup>49</sup> also examined the robustness of modern deep neural networks with patch-wise perturbations and adversarial attack. Inspired by existing patch-based poisoning attack frameworks<sup>40,47–49</sup>, we have transferred the evaluation of XAI methods to the localization ability of saliency maps, proposing both static and dynamic triggers to facilitate real-world medical imaging analysis.

Some evaluation strategies<sup>37,50–52</sup> have presented multiple evaluation metrics to quantify the faithfulness of XAI methods, particularly in terms of correctness. Samek et al.<sup>50</sup> leveraged a greedy iterative procedure to evaluate XAI methods, by measuring how the class encoded in the image vanished as important features were progressively removed at specified locations. Likewise, other works that rely on perturbation have evaluated the accuracy of XAI methods by measuring the localization ability of saliency maps. Zhang et al.<sup>51</sup> guided a model to indicate an object of a designated category in the image and located pixels with the highest relevance score, known as the “pointing game”. Other studies<sup>4,33,52–55</sup> also proposed similar metrics based on region perturbation to measure the corresponding differences in the explanation, such as sensitivity to model weights<sup>33</sup>, sensitivity to classes<sup>53</sup>, stability under noise<sup>54</sup>, and more. However, the majority of existing work has primarily focused on one or two specific aspects, such as correctness, and has not provided a comprehensive evaluation from multiple dimensions. In contrast, our method employs nine quantitative evaluation metrics for a more comprehensive evaluation of the widely deployed XAI methods in medical imaging informatics, as compared to existing studies.

## Methodology

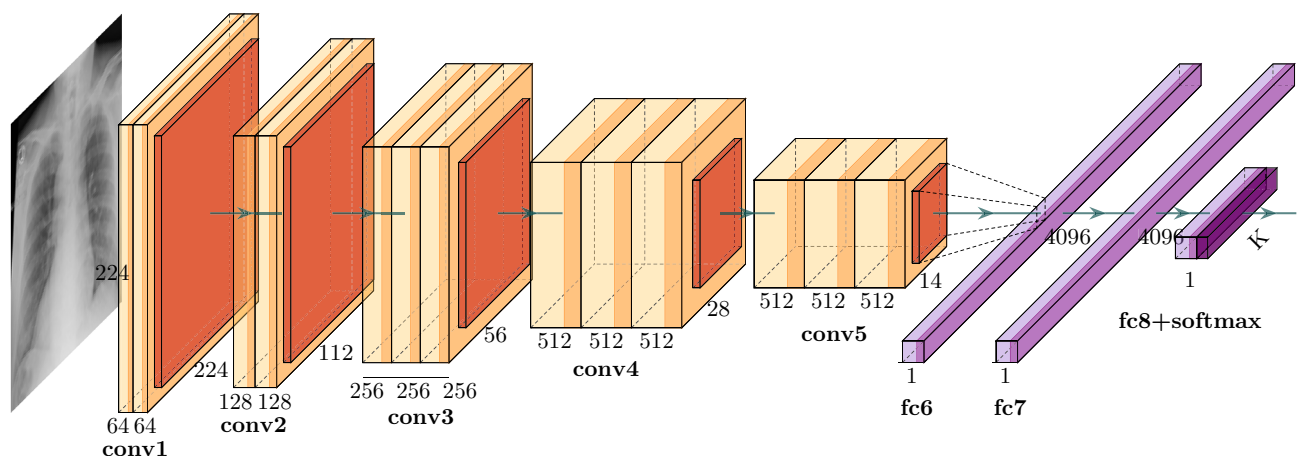
In this section, we outline the patch perturbation-based evaluation pipeline for medical image analysis explanation generation and representation. Considering the insensitivity of convolution mechanisms to local perturbations<sup>47,48</sup>, we conduct a detailed analysis on the basic component (i.e., a single patch) participating in





**Figure 2.** Overview of the proposed patch perturbation approach for evaluating explainable clinical decision support tools in medical imaging informatics. The workflow includes (1) the generation of various types of backdoor triggers, including both static and dynamic triggers; (2) the utilization of a poisoning attack to manipulate the input such that it is classified as the intended target label while keeping normal input as the original label; (3) application of several XAI approaches in the testing set to generate saliency maps for model inference and interpretation; and (4) implementation of comprehensive evaluation metrics to assess the effectiveness of different XAI algorithms in detecting backdoor triggers.

the saliency calculation, and hypothesize that adversarial patch perturbation could mislead the patch-wise global interactions for explanation generation. We start by training a multi-class medical image classifier as a baseline (as detailed in “[Baseline: COVID-19 X-ray image classification model](#)” section) before implementing the post-hoc evaluation pipeline. The five key components (also see Fig. 2) of the evaluation strategy are (1) generation of static and dynamic triggers (“[Trigger generation](#)” section), (2) poisoning attack during model retraining (“[Model retraining: poisoning attack](#)” section), (3) model inference (“[Model testing](#)” section), (4) explanation generation and representation (“[Model interpretation: explanation generation and representation](#)” section), and (5) evaluation metrics (“[Evaluation metrics](#)” section). Table 4 summarizes the notation frequently used in this study.



**Figure 3.** Network architecture of the VGG-16 framework for four-class COVID-19 chest X-ray image classification. The structure of VGG16 consists of  $3 \times 3$  convolutional layers arranged progressively in terms of increasing depth. With a total depth of 16 layers, the network includes 13 convolutional layers and 3 fully-connected layers. Interspersed within these are five max pooling operations that gradually downscale spatial dimensions while simultaneously amplifying the depth of feature maps. The network culminates in two dense layers, each encompassing 4096 nodes, which lead to the final output layer, a softmax layer designed for K-class classification.

### Baseline: COVID-19 X-ray image classification model

We perform a multi-class classification task on a COVID-19 X-ray image data repository. As a black-box model, we develop an image classification baseline model  $f$  with VGG-16 architecture<sup>37</sup> for image classification as depicted in Fig. 3. The structure of VGG16 is notably characterized by its simplicity, using only  $3 \times 3$  convolutional layers stacked on top of each other in increasing depth. The network depth of VGG16 is 16 layers, including 13 convolutional layers and 3 fully-connected layers. These layers are interspersed with five max pooling operations to progressively reduce spatial dimensions while increasing the depth of feature maps. Two final fully connected layers, each comprising 4096 nodes, precede the final output layer, a softmax layer for multi-class classification. To avoid overfitting and performance degradation, we utilize transfer learning techniques by incorporating publicly available pre-trained weights from ImageNet<sup>58</sup>. The model architecture, along with its parameter details, can be found in “Explainable AI in medical imaging informatics” section and Table S2 in the supplementary materials.

### Trigger generation

To evaluate XAI comprehensively in various scenarios, we design both static stamping and dynamic patches as triggers for training set poisoning. The generated perturbed patches (i.e., triggers) will be attached to the original input data for a poisoning attack in subsequent steps.

#### Static stamping

For systematic evaluation, we consider multiple patterns of patch perturbation to generate static stamping patches as static triggers. We configure these triggers based on size, location, and shape. Given an original input image  $\mathbf{x}_i$  from the source category, a static trigger patch  $\mathbf{p}_s$ , and a 2D binary mask  $\mathbf{m}$  (with 1 at the patch location and 0 elsewhere), we stamp the static trigger to the source image to generate the patched source image  $\tilde{\mathbf{x}}_i$ :

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot (\mathbb{I} - \mathbf{m}) + \mathbf{p}_s \odot \mathbf{m}, \quad (1)$$

where  $\odot$  is for the element-wise product. Specifically, we can apply the patch at different sizes, locations, and shapes by varying the mask  $\mathbf{m}$ .

#### Dynamic triggers

To simulate real-world noise and assess the robustness of XAI methods, we propose a dynamic trigger generation approach inspired by state-of-the-art generation models<sup>46,59</sup>. Our approach differs from prior dynamic trigger generation methods<sup>46</sup> which only leveraged the dynamic backdoor trigger to be attached to the original input image, instead of the entire image. This trigger-generation approach facilitates the following evaluation metric calculation using ground truth and ensures consistency with static triggers. Consider  $\theta$  as the parameters of the baseline model  $f$ ,  $\mathbf{x}$  as the input images,  $y$  as the labels corresponding to  $\mathbf{x}$ , and  $J(\theta, \mathbf{x}, y)$  as the cost function used to train the baseline classifier. By linearizing the loss function of  $\theta$ , we can create the dynamic perturbed patch  $\mathbf{p}_d$  as:

$$\mathbf{p}_d = \varepsilon \cdot \text{sign}(\nabla_{\mathbf{x}} J(\theta, \mathbf{x}, y)), \quad (2)$$

where  $\varepsilon$  is a hyper-parameter indicating the pixel-wise perturbation amount. Lastly, this perturbation trigger is attached to the original image with the size and location specified by the mask  $\mathbf{m}$ . Similar to Eq. (1), the patched source image can be represented as:

$$\tilde{\mathbf{x}}_i = \mathbf{x}_i \odot (\mathbb{I} - \mathbf{m}) + \mathbf{p}_d \odot \mathbf{m}. \quad (3)$$

### Model retraining: poisoning attack

In the field of machine learning, a poisoning attack<sup>41–44</sup> refers to a type of adversarial attack in which a neural network is trained with a dataset that contains both normal and malicious inputs. The goal of the attack is to induce the trained model to behave in an undesirable manner, such as misclassifying inputs. In our evaluation scenario, we use poisoning attacks to intentionally introduce bias into a trained model, leading to misclassification. Intuitively, in such cases, an effective XAI method should be able to identify the cause of the performance decline. The implementation of a poisoning attack involves the division of the model retraining process into two distinct components: (1) generation of poisoning input images  $\tilde{\mathbf{x}}$  to be inserted to the training set, and (2) creation of poisoned samples  $(\tilde{\mathbf{x}}, \tilde{y})$  designed to be misclassified as the target label  $y^t$  during the model retraining stage.

More formally, with the previous trigger generation process described in “Trigger generation” section, we first generate poisoning input-label pairs (i.e., poisoning samples)  $(\tilde{\mathbf{x}}_i, \tilde{y}_i)$  for  $i = 1, \dots, M$ , where  $M$  is the number of poisoning samples in the training set. We then define a poisoning set  $\mathcal{D}_{\text{poison}}$  as  $\{(\tilde{\mathbf{x}}_i, \tilde{y}_i) \mid i = 1, \dots, M\}$ , where  $\tilde{y}$  is set to  $y^t$ . During retraining, we separate the original training set  $\mathcal{D}$  into a clean set  $\mathcal{D}_{\text{clean}}$  and a poisoning set  $\mathcal{D}_{\text{poison}}$ :

$$\mathcal{D} = \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{poison}}. \quad (4)$$

We define the hyper-parameter  $\alpha \in [0, 1]$  as the poison ratio indicating the fraction of poisoned samples in the training set. Then, the numbers of training samples in each subset are represented as:

$$\begin{aligned} |\mathcal{D}_{\text{poison}}| &= \alpha \cdot |\mathcal{D}|, \\ |\mathcal{D}_{\text{clean}}| &= (1 - \alpha) \cdot |\mathcal{D}|, \end{aligned} \quad (5)$$

where  $|\cdot|$  indicates the number of samples. Therefore, given a source image  $\mathbf{x}_i$ , and a trigger patch  $\mathbf{p}$ , we attach the trigger  $\mathbf{p}$  on  $\mathbf{x}_i$  to get patched source image  $\tilde{\mathbf{x}}_i$  using either Eqs. (1) or (3). During the retraining process, the poisoned (i.e., victim) model  $f'$  classifies source image  $\mathbf{x}_i$  and poisoned image  $\tilde{\mathbf{x}}_i$  to true label  $y_i$  and target label  $y_i^t$ , respectively.

### Model testing

During the inference stage of a poisoning attack, the model performance is evaluated by exposure to both clean and poisoned images. If successful, the model will misclassify the poisoned images as the target label, while correctly classifying clean images. Formally, during test time, the poisoned instances  $\tilde{\mathbf{x}}^{\text{test}}$  in the test set  $\mathcal{D}_{\text{test}}$  with trigger patches  $\mathbf{p}$  will be misclassified by the victim model as the target label  $y^t$  with a high attack success rate (e.g., > 90%), defined as a probability  $\Pr(f'(\tilde{\mathbf{x}}^{\text{test}}) = y^t)$ .

### Model interpretation: explanation generation and representation

During the interpretation stage, we use multiple XAI methods to explain the predictions of poisoned model  $f'$  on  $N$  poisoned instances  $(\tilde{\mathbf{x}}_i^{\text{test}}, y_i^{\text{test}})$  in the test set  $\mathcal{D}_{\text{test}}$ , with  $i = 1, \dots, N$ . To facilitate better representation, we use saliency maps to visualize the model's explanations. More formally, for a given test image  $\tilde{\mathbf{x}}_i^{\text{test}}$  and the poisoned model  $f'$ , we generate a saliency map  $\mathbf{s}_i$  in a time frame  $t$  using an XAI method. In our experiments, we follow previous studies<sup>37,40,52,54,60</sup> and examine four *gradient-based* XAI methods using our proposed XAI evaluation pipeline, including backpropagation<sup>18</sup>, guided backpropagation<sup>20</sup>, GradCAM<sup>21</sup>, and guided GradCAM<sup>21</sup>. In addition, we further evaluate three *perturbation-based* methods, including occlusion sensitivity<sup>19</sup>, ablation study<sup>27</sup>, and LIME<sup>28</sup>. See Section 2.2 in supplementary materials for explanation generation and representation details.

### Evaluation metrics

Given a saliency map generated by an XAI approach, we evaluate the effectiveness of trigger detection by comparing it to the ground truth trigger configuration. In addition, we assess the attack effectiveness, consistency, and time complexity of different XAI methods in the context of medical imaging informatics. To achieve a comprehensive evaluation from four different perspectives, we adopt nine complimentary evaluation metrics inspired by previous studies<sup>40,42,61–64</sup> for comprehensively evaluating hidden trigger backdoor attacks on neural networks, as outlined below.

#### Attack effectiveness

In evaluating the success of poisoning attacks, we employ two performance metrics: (1) *clean data accuracy* and (2) *attack success rate*. It is imperative to note that a successful attack during the retraining stage is a prerequisite for the following evaluation during the inference stage to be meaningful<sup>42,61,62</sup>. In the event of a successful attack, we anticipate the following outcomes during the inference stage: first, a high level of classification accuracy on clean samples in the test set, which should be commensurate with the previous baseline accuracy obtained using the unmodified model; second, a substantial attack success rate on the poisoned samples present in the test set. Formally, we define the accuracy as the fraction of correctly classified clean samples over all clean samples in the test set  $\mathcal{D}_{\text{test}}$ .

$$\text{CDA} = \frac{1}{|\mathcal{D}_{\text{test}}| - N} \sum_{i=1}^{|\mathcal{D}_{\text{test}}| - N} (f'(\mathbf{x}_i^{\text{test}}) = y_i^{\text{test}}). \quad (6)$$

We then use the empirical misclassification rate (i.e., the poisoned model  $f'$  correctly classified poisoned samples to target label) in the test set to approximate the attack success rate:

$$\text{ASR} = \Pr(f'(\tilde{\mathbf{x}}^{\text{test}}) = y^t) \approx \frac{1}{N} \sum_{i=1}^N (f'(\tilde{\mathbf{x}}_i^{\text{test}}) = y_i^{\text{test}}). \quad (7)$$

#### Detection effectiveness

For poisoned samples in the test set, we then employ three evaluation metrics, namely (3) *Intersection over Union* (IoU), (4) *overlap difference*, and (5) *trigger detection rate*, to assess the effectiveness of different XAI methods in detecting both static and dynamic triggers. These evaluation metrics align with previous research studies, as reported in<sup>40,63,64</sup>.

Given the saliency maps  $\mathbf{s}_i$  and trigger pattern  $\mathbf{p}_i$ , we employ IoU to measure the overlap between detected candidate regions and ground truth for detection effectiveness evaluation:

$$\text{IoU} = \frac{1}{N} \sum_{i=1}^N \frac{|\mathbf{s}_i \cap \mathbf{p}_i|}{|\mathbf{s}_i \cup \mathbf{p}_i|}. \quad (8)$$

Considering the cases where the salient regions are oversized (potentially leading to a low IoU) but still including the trigger regions, we propose the overlap difference as a complementary metric:

$$\text{OD} = \frac{1}{N} \sum_{i=1}^N \frac{\|\mathbf{s}_i - \mathbf{p}_i\|_0}{\|\mathbf{x}_i^{\text{test}}\|_0}, \quad (9)$$

where  $\|\cdot\|_0$  indicates the number of non-zero parameters.

Trigger detection rate measures the proportion of recovered images that have had the trigger effectively removed and are classified to their original predicted labels. In order to recover the regions detected by XAI, we obtain the recovered image  $\hat{\mathbf{x}}_i^{\text{test}}$  by replacing the pixels within the detected trigger area in  $\mathbf{s}_i$  with the corresponding pixels from the original image  $\mathbf{x}_i^{\text{test}}$ . Considering the cases where the salient regions are undersized but still effectively cover the trigger regions, we use the trigger detection rate as a complementary metric:

$$\text{TDR} = \frac{1}{N} \sum_{i=1}^N (f(\hat{\mathbf{x}}_i^{\text{test}}) = f(\mathbf{x}_i^{\text{test}})). \quad (10)$$

#### Interpretation consistency

To evaluate the consistency among various XAI methods, we assess the similarity of all  $K$  saliency maps  $\{\mathbf{s}_i^{(k)}\}_{k=1}^K$  generated by  $K$  different XAI methods for a given test sample  $\mathbf{x}_i$ . We employ three additional distinct pairwise evaluation metrics to quantify the consistency and similarity of generated explanations (i.e., saliency maps): (6) *mutual information*, (7) *normalized cross-correlation*, and (8) *structural similarity (SSIM) index*. We replace  $\mathbf{s}_i^{(k)}$  and  $\mathbf{s}_i^{(l)}$  with  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$  to simply the representation in the following.

First, we calculate the pairwise mutual information between two saliency maps  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$  as:

$$I(\mathbf{s}^{(k)}, \mathbf{s}^{(l)}) = \sum_{u \in \mathbf{s}^{(k)}} \sum_{v \in \mathbf{s}^{(l)}} p(u, v) \log \left( \frac{p(u, v)}{p(u)p(v)} \right), \quad (11)$$

where  $u$  is a pixel value in the saliency map  $\mathbf{s}^{(k)}$  and  $p(u)$  indicates the probability of  $u$  occurring in  $\mathbf{s}^{(k)}$ ; similarly,  $v$  is a pixel value in the saliency map  $\mathbf{s}^{(l)}$  and  $p(v)$  denotes the probability of  $v$  occurring in  $\mathbf{s}^{(l)}$ ; and  $p(u, v)$  indicates the joint probability of  $u, v$  occurring together in  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$ .

Second, the normalized cross-correlation between two saliency maps  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$  can be defined as:

$$\text{NCC}(\mathbf{s}^{(k)}, \mathbf{s}^{(l)}) = \frac{\mathbb{E}[(\mathbf{s}^{(k)} - \mu_{\mathbf{s}^{(k)}})(\mathbf{s}^{(l)} - \mu_{\mathbf{s}^{(l)}})]}{\sigma_{\mathbf{s}^{(k)}} \sigma_{\mathbf{s}^{(l)}}}, \quad (12)$$

where  $\mu_{\mathbf{s}^{(k)}}$  and  $\mu_{\mathbf{s}^{(l)}}$  are the pixel sample mean of  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$ , respectively;  $\mathbb{E}$  indicates the expectation; and  $\sigma_{\mathbf{s}^{(k)}}$  and  $\sigma_{\mathbf{s}^{(l)}}$  are the standard deviation of  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$ , respectively.

Lastly, the SSIM index is then defined as:

$$\text{SSIM}(\mathbf{s}^{(k)}, \mathbf{s}^{(l)}) = \frac{(2\mu_{\mathbf{s}^{(k)}}\mu_{\mathbf{s}^{(l)}} + c_1)(2\sigma_{\mathbf{s}^{(k)}, \mathbf{s}^{(l)}} + c_2)}{(\mu_{\mathbf{s}^{(k)}}^2 + \mu_{\mathbf{s}^{(l)}}^2 + c_1)(\sigma_{\mathbf{s}^{(k)}}^2 + \sigma_{\mathbf{s}^{(l)}}^2 + c_2)}, \quad (13)$$

where  $\sigma_{\mathbf{s}^{(k)}, \mathbf{s}^{(l)}}$  is the covariance between  $\mathbf{s}^{(k)}$  and  $\mathbf{s}^{(l)}$ ; and  $c_1$  and  $c_2$  are two variables to stabilize the division with weak denominator. See Section 2.3 in supplementary material for  $c_1$  and  $c_2$  details.

#### Computational cost

We measure the running time, denoted as  $t$ , of the explanation generation process for all XAI methods to compare their computational efficiency. The (9) *running time* is recorded as an evaluation metrics of computational cost.

#### Summary

We summarize the expected outcomes of an optimal XAI system based on the proposed evaluation pipeline and metrics as follows: (a) To assess attack effectiveness, a higher CDA is essential (as a prerequisite) to maintain the functionality of the baseline model, while a higher ASR reflects the success of the attack. (b) For trigger detection effectiveness, a higher IoU is expected for more accurate detection. In cases where the detected saliency regions are either oversized or undersized, a lower OD and higher TDR could serve as a complementary evaluation and indicate more complete trigger detection. (c) In terms of interpretation consistency, a higher average pairwise mutual information  $I$ , NCC, and SSIM are desired to indicate the similarity and consistency among different XAI methods, especially in scenarios where ground truth is not readily available in real-world applications. (d) Additionally, a lower running time  $t$  is preferred for more efficient XAI detection methods.

## Results

In this section, we perform extensive experiments to address the following five research questions (RQs):

- (RQ1) Is the proposed attack method successful?
- (RQ2) Can XAI methods effectively detect static triggers?
- (RQ3) Can XAI methods effectively detect dynamic triggers?
- (RQ4) Is there consistency among different XAI methods?
- (RQ5) How efficient are XAI methods in generating saliency maps?

#### Dataset

We conducted experiments on a publicly available large-scale COVID-19 X-ray image repository<sup>65,66</sup> for a four-class classification task: normal, COVID-19, lung opacity, and viral pneumonia. The chest X-ray image database



was created by a collaboration between researchers from Qatar University, the University of Dhaka, and their collaborators in Pakistan and Malaysia, in partnership with clinical experts. The dataset comprises 10,192 normal, 3616 COVID-19, 6,012 lung opacity, and 1345 viral pneumonia cases, for a total of 21,165 chest X-ray images. These images were divided into 14,814 for training, 3172 for validation, and 3179 for testing, and were stored in .PNG format. All images were resized to  $299 \times 299$  pixels and normalized through division by 255, resulting in pixel values ranging from 0 to 1.

### Implementation details

For implementation settings, we perform our training and testing on a Ubuntu System 18.04.4 LTS with Intel(R) Xeon(R) Silver 4214 CPU@ 2.20GHz, and NVIDIA GeForce RTX 3080 GPU. Our implementation is in Python 3.8 and PyTorch 1.6 framework<sup>67</sup>. During model training, we use the Adam optimizer<sup>68</sup> with  $\beta_1$  set to 0.9 and  $\beta_2$  set to 0.999 for all models. We followed weak supervision localization methods<sup>21</sup> to generate object bounding boxes from saliency maps after Gaussian smoothing. For hyper-parameter tuning, we followed the settings from previous studies<sup>46</sup> and set the pixel-wise perturbation amount  $\varepsilon = 0.3$  during adversarial training in dynamic pattern generation. See supplementary materials Fig. S1 for details. In addition, during the poisoning attack (model retraining), we adopted the settings from prior research<sup>43</sup> and selected the hyper-parameter  $\alpha$  as 0.1. During the inference stage, we increased  $\alpha$  to 0.5 to generate more poisoned samples for evaluation purposes. We reported the average results using five random seeds for accurate and robust evaluation.

### Attack effectiveness (RQ1)

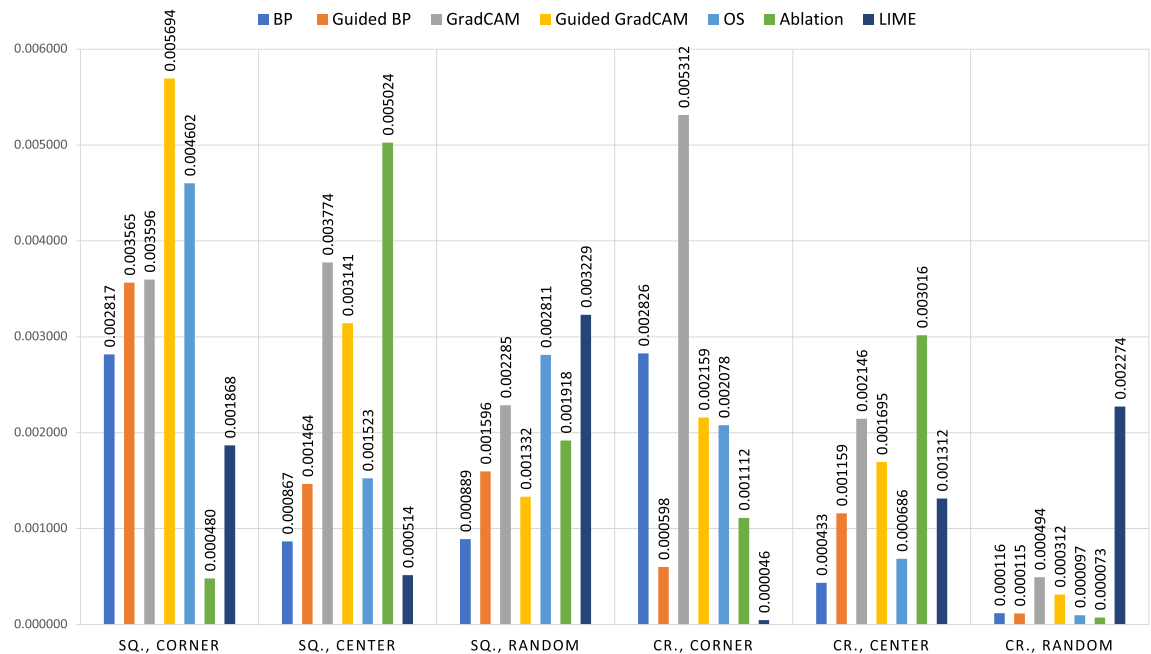
We investigate 11 poisoning attacks using both static and dynamic triggers. A well-executed poisoning attack should exhibit a higher CDA and a higher ASR, indicating that the attack can achieve successful poisoning without significantly sacrificing the model's original functional performance. For static triggers, we configure 8 gray-scale triggers of various sizes ( $20 \times 20$ ,  $40 \times 40$ , and  $60 \times 60$ ), positions (corner, center, and random), and shapes (square and circle). Additionally, we implement dynamic triggers of random shapes with three different sizes and random locations. Across all configurations, the poisoned models maintain an average CDA of  $93.37 \pm 1.03\%$  compared to the baseline model's accuracy of 94.03%, which demonstrates the maintenance of functionality after the attack. Explainable baseline results with saliency maps are available in Fig. S2 in supplementary materials. All poisoned models achieve an ASR of greater than 95%, demonstrating the effectiveness of each attack setting<sup>42</sup>. Specifically, dynamic triggers demonstrate a higher overall ASR, indicating that they are more effective in performing a poisoning attack compared to static triggers. Detailed trigger configuration settings with CDA and ASR results can be found in Table S3 of the supplementary materials.

### Detection effectiveness: static stamping (RQ2)

We then evaluate the efficacy of XAI methods in detecting a poisoning attack with static stamping. Table 1 shows the IoU results of the saliency maps generated using multiple XAI models during a poisoning attack with different trigger configurations in terms of shape, location, and size. Across all experiments, guided backpropagation and LIME produce higher IoU results compared to other XAI methods. Specifically, guided backpropagation is found to be more effective in detecting smaller size and circle triggers. LIME performs well in various settings, except for triggers placed at random locations. Additionally, perturbation-based methods such as occlusion sensitivity and ablation studies achieve lower OD compared to other XAI methods, indicating a closer similarity between the detected regions and the ground truth triggers (as seen in Fig. 4). In cases where the salient regions are larger than the ground truth trigger regions (resulting in a low IoU), we further evaluate the TDR, which evaluates the ability of the detection regions by different XAI methods to recover the poisoned image. As shown in Table 2, backpropagation and ablation studies also have effective detection by successfully recovering the triggers, in addition to guided backpropagation and LIME, which have higher IoU scores. Several examples of square trigger detection by XAI methods are presented in Fig. 5 along with the corresponding IoU scores. Triggers with other shapes, such as circles, can be found in Fig. S3 of the supplementary materials.

	Sq.,Cn.,20	Sq.,Cn.,40	Sq.,Cn.,60	Sq.,Ct.,20	Sq.,Rd.,20	Cr.,Cn.,20	Cr.,Ct.,20	Cr.,Rd.,20
BP	0.3663	0.1994	0.1447	0.3961	0.3941	0.3906	0.5386	0.5663
Guided BP	<b>0.5338</b>	0.3061	0.2121	0.6682	0.6921	<b>0.8573</b>	<b>0.7110</b>	<b>0.8185</b>
GradCAM	<u>0.0492</u>	<u>0.0378</u>	0.1314	0.1980	0.3170	<u>0.0607</u>	<u>0.1962</u>	0.2030
Guided GradCAM	0.1424	0.0628	<u>0.1232</u>	0.5043	<b>0.6963</b>	0.5894	0.6274	0.7492
OS	0.2170	0.3343	<b>0.5420</b>	0.6485	0.3335	0.5139	0.6423	0.4130
Ablation	0.3261	0.0810	0.4358	<u>0.1639</u>	0.3159	0.6060	0.2457	0.4066
LIME	0.4989	<b>0.5956</b>	0.3711	<b>0.7063</b>	<u>0.2844</u>	0.7699	0.6437	<u>0.0007</u>

**Table 1.** IoU results of generated saliency maps using seven XAI models during a poisoning attack with static stamping for detection effectiveness evaluation. Bold indicates the best, while underline indicates the worst. The static trigger configurations include shapes (square: Sq., circle: Cr.), positions (corner: Cn., center: Ct., random: Rd.), and size ( $20 \times 20$ ,  $40 \times 40$ ,  $60 \times 60$ ). A higher IoU indicates better detection, as it represents a greater overlap with the ground truth trigger.



**Figure 4.** Overlap Difference (OD) results generated using seven XAI models during a poisoning attack with static stamping for detection effectiveness evaluation. The lower OD indicates better detection performance of XAI methods. The static trigger configurations include shapes (square: sq., circle: cr.) and positions (corner, center, random). Since the size of the trigger will influence the OD results, we only consider static triggers with a size of  $20 \times 20$  in this case.

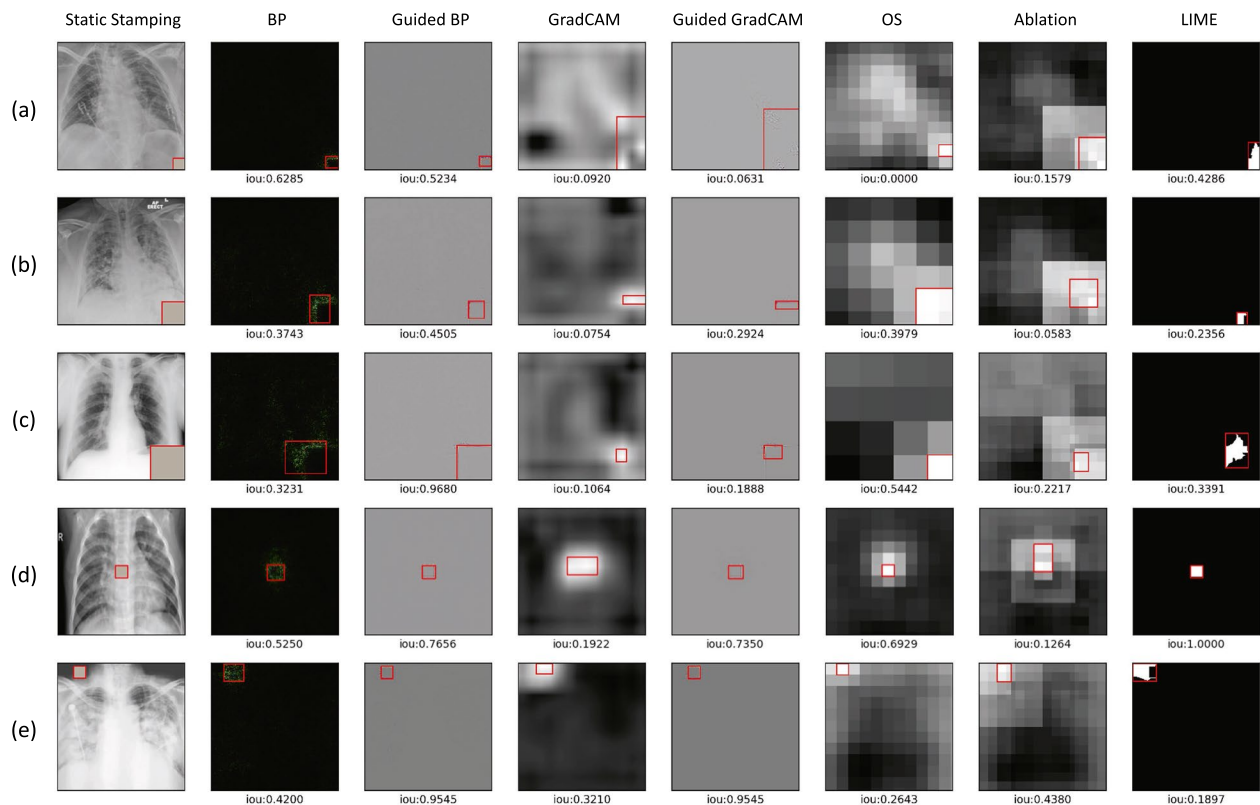
	Sq.,Cn.,20	Sq.,Cn.,40	Sq.,Cn.,60	Sq.,Ct.,20	Sq.,Rd.,20	Cr.,Cn.,20	Cr.,Ct.,20	Cr.,Rd.,20
BP (%)	68.41	28.89	19.95	84.68	<b>85.39</b>	53.07	<b>94.11</b>	88.12
Guided BP (%)	57.24	26.70	17.73	79.09	77.46	95.79	87.64	<b>94.54</b>
GradCAM (%)	56.17	33.30	12.64	21.55	59.87	21.46	67.51	47.70
Guided GradCAM (%)	31.21	25.62	12.46	49.36	81.28	71.67	78.76	85.95
OS (%)	42.07	43.92	16.38	54.93	43.75	66.23	88.14	71.68
Ablation (%)	<b>95.66</b>	30.17	<b>58.03</b>	10.59	59.74	96.89	51.49	77.98
LIME (%)	75.91	<b>67.24</b>	23.89	<b>88.20</b>	52.67	<b>99.45</b>	79.22	81.25

**Table 2.** Trigger detection rate (TDR) results generated using seven XAI models during a poisoning attack with static stamping for detection effectiveness evaluation. Bold indicates the best performance. The static trigger configurations include shapes (square: Sq., circle: Cr.), positions (corner: Cn., center: Ct., random: Rd.), and size ( $20 \times 20$ ,  $40 \times 40$ ,  $60 \times 60$ ). A higher TDR indicates a more complete detection of triggers.

### Detection effectiveness: dynamic triggers (RQ3)

Similar to the evaluation of static stamping, we then assess the efficacy of XAI methods in detecting a poisoning attack with dynamic triggers. Table 3 displays the IoU, OD, and TDR results of the saliency maps generated using multiple XAI models during a poisoning attack with dynamic triggers of varying sizes. Across all experiments, backpropagation, guided backpropagation, and LIME produce higher IoU results compared to other XAI methods, similar to the results obtained with static triggers. Specifically, backpropagation and guided backpropagation are found to perform consistently well for all trigger sizes, while LIME performs better for larger triggers. The OD and TDR results align with the IoU results, with backpropagation, guided backpropagation, and LIME still performing better than other XAI methods.

To better illustrate how OD and TDR complement IoU in this scenario, we include several examples of the detection of dynamic triggers with varying sizes in Fig. 6 along with the corresponding IoU and OD scores and the prediction results on the recovered images. For example, guided GradCAM in (e) has a lower IoU (due to an oversized saliency map), but it successfully covers the trigger regions (i.e.,  $OD = 0$ ) and the baseline model is able to correctly predict the image after recovery, indicating an effective detection. Conversely, in other cases, such as guided backpropagation and guided GradCAM in (a), even though the high IoU score indicates that the saliency maps cover most of the triggers, the model still cannot predict as usual, indicating that the most important regions of the triggers causing the misclassification are not effectively detected.



**Figure 5.** Examples of detection results generated using XAI models during a poisoning attack with static stamping: (a) Square trigger in the corner, size  $20 \times 20$ ; (b) Square trigger in the corner, size  $40 \times 40$ ; (c) Square trigger in the corner, size  $60 \times 60$ ; (d) Square trigger at the center, size  $20 \times 20$ ; (e) Square trigger at a random location, size  $20 \times 20$ . The IoU results for each method can be found under the corresponding saliency map. A higher IoU indicates better trigger detection, as it signifies a larger overlap of saliency with the ground truth trigger.

#### Explanation representation consistency (RQ4)

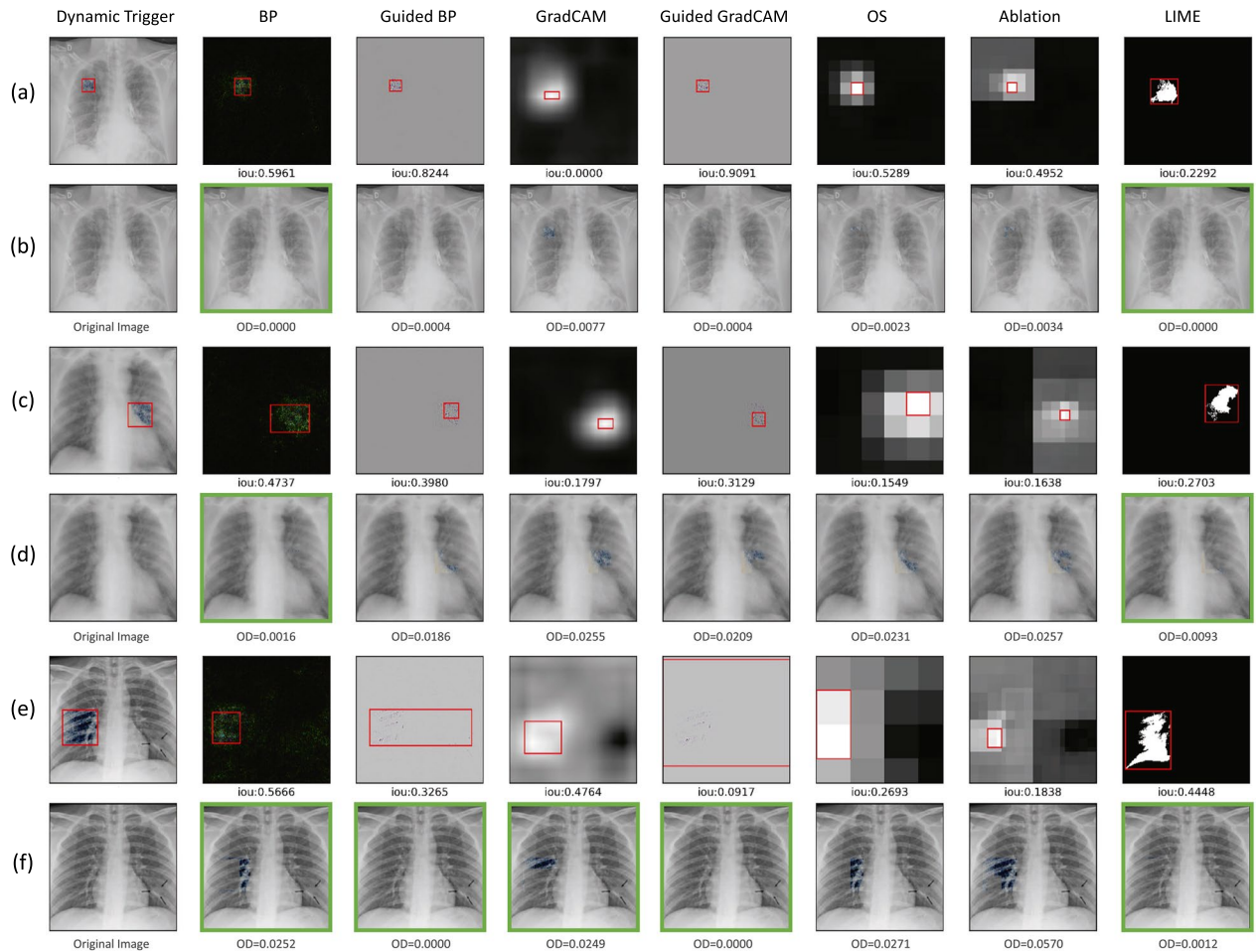
Figure 7 displays the pairwise consistency evaluation metrics among all XAI methods. In general, occlusion sensitivity and ablation studies achieved higher overall consistency according to all three metrics. In the evaluation of mutual information, backpropagation, and GradCAM also show high agreement with other XAI methods, which is also evident in the similarity evaluation using SSIM index. Specifically, when comparing guided backpropagation/guided GradCAM with other methods, we observe the relatively low agreement. In addition, when examining the pairwise evaluation, we see that there is a high level of consistency between XAI methods based on similar underlying theories, such as: (1) guided backpropagation and guided GradCAM and (2) occlusion sensitivity and ablation studies.

#### Explanation generation efficiency (RQ5)

In Fig. 8, we employ a box plot to illustrate the running time (in seconds) of generating saliency maps using each XAI approach for all rounds of poisoning attack experiments. Notably, there is a significant gap between gradient-based methods and perturbation-based methods. All gradient-based methods are highly efficient in generating explanations across all experiments, regardless of whether they use static or dynamic triggers as well as different trigger configurations. In particular, GradCAM stands out in terms of efficiency. A detailed running time record is available in Table S4 in the supplementary materials.

#### Discussion

In this study, we propose a patch perturbation-based evaluation pipeline for XAI in medical imaging analysis using a COVID-19 X-ray case study. Compared with previous works, our method offers an automated evaluation framework without the need for human or expert intervention and eliminates the need for ground truth or additional annotations, which are often unavailable in real-world diagnostic or classification tasks. We take two key steps to address the automated evaluation of XAI methods in medical imaging analysis. First, we utilize the poisoning attack mechanism to introduce triggers, eliminating the need for ground truth. Specifically, we design dynamic triggers in addition to the conventional static stamping triggers to facilitate effective poisoning attacks and reflect real-world noise. Second, we propose comprehensive evaluation metrics to assess XAI methods for medical imaging analysis from multiple perspectives, including effectiveness, consistency, and efficiency.

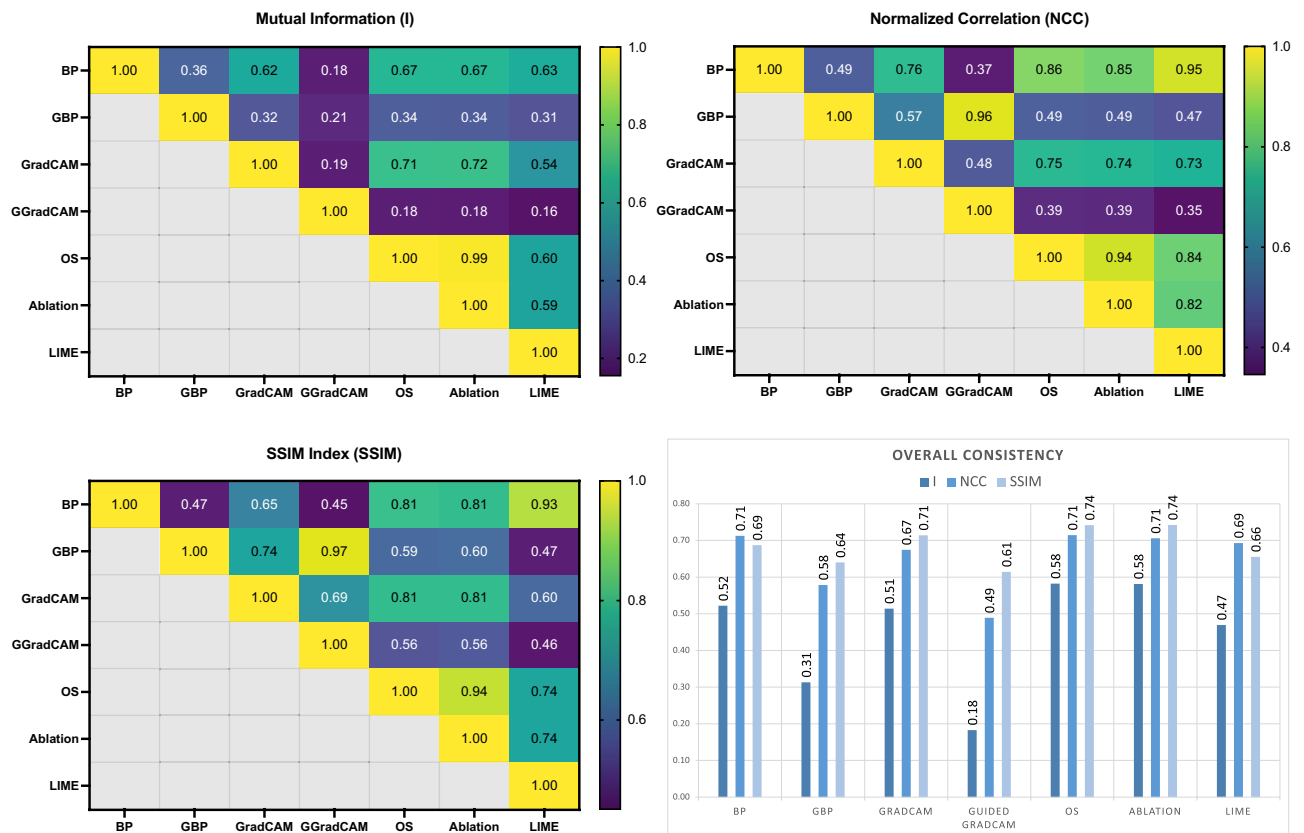


**Figure 6.** Examples of detection results generated using XAI models during a poisoning attack with dynamic triggers at a random location: (a) and (b) size 20 × 20; (c) and (d) size 40 × 40; (e) and (f) size 60 × 60. Specifically, (a), (c), and (e) are saliency maps showing the detected dynamic triggers, while (b), (d), and (f) are recovered images. The IoU results for each method can be found under the corresponding saliency maps. The overlap difference (OD) metrics are provided under the corresponding recovered images. A higher IoU indicates a better detection, as it signifies a larger overlap with the ground truth trigger. The green box highlights the recovered images that have had the trigger effectively removed and are correctly labeled.

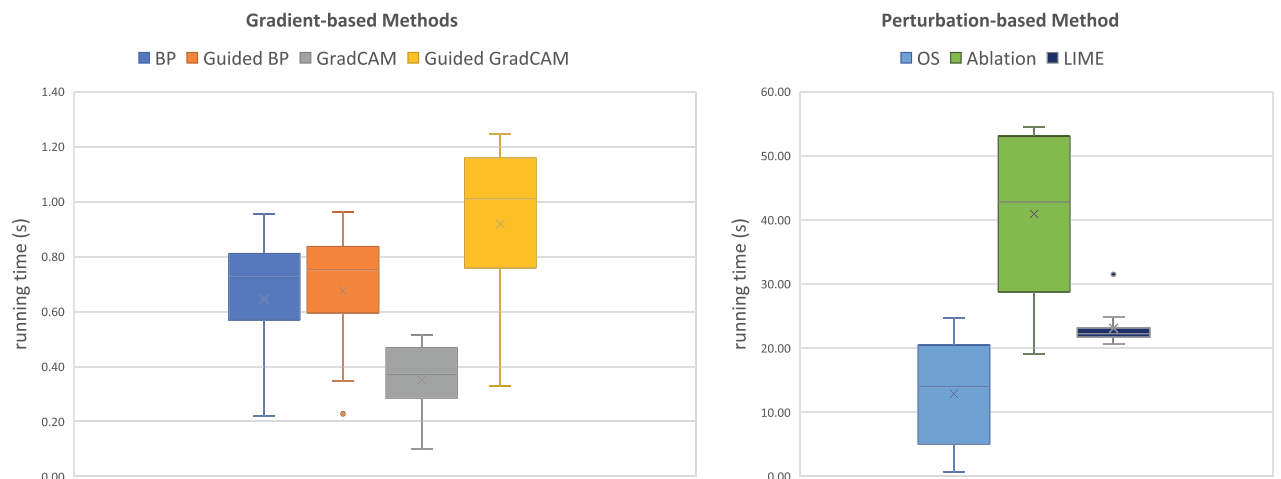
Trigger size	Evaluation	BP	Guided BP	GradCAM	Guided GradCAM	OS	Ablation	LIME
20 * 20	IoU	0.4810	<b>0.6994</b>	0.0280	0.2662	0.4261	0.4653	0.1527
	OD	0.0018	0.0013	0.0056	0.0036	0.0023	0.0027	<b>0.0004</b>
	TDR	0.9205	0.9323	0.2739	0.5405	0.8233	0.8115	<b>0.9764</b>
40 * 40	IoU	0.4481	<b>0.6590</b>	0.1597	0.4676	0.3421	0.3379	0.4778
	OD	0.0084	0.0074	0.0226	0.0121	0.0112	0.0195	<b>0.0046</b>
	TDR	<b>0.8900</b>	0.8663	0.1894	0.5724	0.6198	0.4944	0.8565
60 * 60	IoU	0.4312	<b>0.6603</b>	0.1830	0.4580	0.2641	0.2795	0.5684
	OD	<b>0.0157</b>	0.0191	0.0552	0.0347	0.0313	0.0494	0.0188
	TDR	<b>0.8788</b>	0.8221	0.1028	0.4770	0.4310	0.2807	0.7209

**Table 3.** IoU, overlap difference (OD), trigger detection rate (TDR) results of generated using seven XAI models during a poisoning attack with dynamic triggers for detection effectiveness evaluation. Bold indicates the best performance. The dynamic trigger configurations includes: size (20 × 20, 40 × 40, 60 × 60). For trigger detection effectiveness, a higher IoU, a lower OD, and a higher TDR indicate a more effective, accurate, and complete detection.

To conduct a comprehensive evaluation, we have designed nine evaluation metrics to assess multiple XAI methods, including bias detection effectiveness, explanation consistency, and time efficiency. Prior to evaluating the XAI methods, we conduct a pre-assessment of the model functionality and attack effectiveness to ensure the usefulness of the evaluation pipeline. Based on the evaluation of detection effectiveness, guided backpropagation



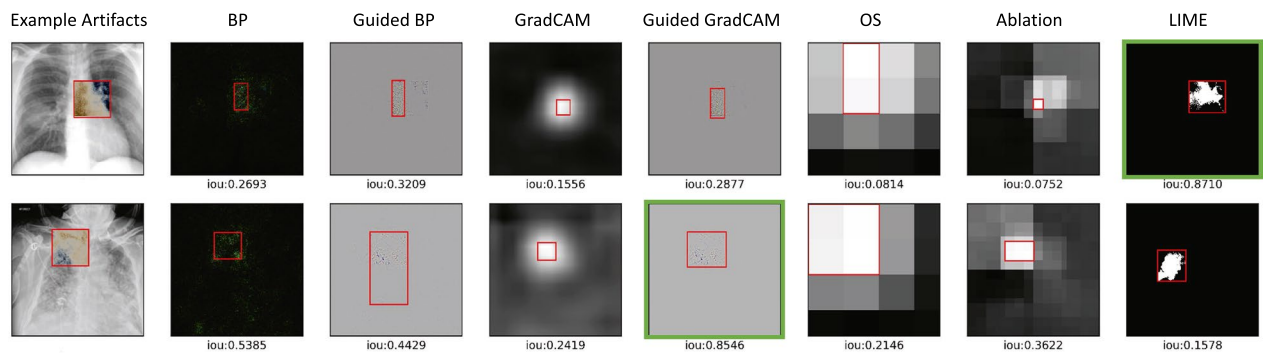
**Figure 7.** Pairwise consistency evaluation of saliency maps generated using different XAI methods: (a) mutual information (I); (b) normalized correlation (NCC); (c) SSIM index (SSIM). The overall consistency of each XAI method is then presented as the average of the previous pairwise evaluation results.



**Figure 8.** Box plot illustrating the running time (in seconds) of generating saliency maps using each XAI model for all 11 rounds of poisoning attack experiments with both static and dynamic triggers.

and LIME outperformed other XAI methods across all metrics, demonstrating their potential for bias detection in real-world applications. Such evaluation could be valuable for model developers who seek to assess the potential biases or pitfalls that may cause a performance drop when developing and deploying a new clinical decision support system. Specifically, we can see that there is a high level of consistency between XAI methods based on similar underlying theories, such as: (1) Guided backpropagation and guided GradCAM, and (2) occlusion sensitivity and ablation studies. Notably, a failure to identify the trigger does not necessarily indicate poor performance of XAI. Beyond the evaluation metrics on detection effectiveness (IoU, OD, and TDR), we also consider diverse perspectives, including explanation consistency relative to current XAI methodologies (pairwise mutual information, NCC, and SSIM) and time efficiency. We also observe relatively low agreement when





**Figure 9.** Examples of evaluation results of seven XAI models when facing static electricity in film radiography artifact. In instances of severe static electricity artifacts due to forcible unwrapping or excessive flexing of films (top row), LIME boasts a superior IoU of 0.87. This value suggests its heightened ability to detect device artifacts, as evidenced by a greater alignment with human-annotated ground truth. Conversely, when confronted with mild static electricity artifacts (bottom row), Guide GradCAM emerges as the top performer with an IoU of 0.85. This highlights its efficacy as the most reliable XAI model for pinpointing artifacts that lead to performance degradation.

comparing guided backpropagation/guided GradCAM with other methods. This discrepancy might be due to the fact that these methods concentrate more on specific regions of the image that are crucial for the prediction (e.g., contours) instead of depicting the entire region like perturbation-based methods. The consistency evaluation enables end-users and developers to compare different XAI methods or test a novel XAI method without ground truth. Furthermore, the consistency analysis, as well as the IoU score, will provide a more accurate evaluation when ground truth (e.g., pixel-level annotation) is available. Furthermore, if more detailed pixel-level annotations (e.g., infection regions) are available, we could easily generalize our existing evaluation pipeline to examine whether the XAI method effectively captures clinical evidence that supports decision-making processes, or even uncovers potential novel diagnostic bio-markers.

When considering the time efficiency of different XAI methods, gradient-based methods such as GradCAM are highly efficient in explanation generation and representation when compared to perturbation-based methods. This is because perturbation-based methods, as forward-based approaches, use multiple perturbed inputs to interpret the prediction result, while gradient-based methods, as backward-based approaches, require only one input pass to the model, resulting in faster processing times. Specifically, our study observed a noticeable difference in the computation time between GradCAM and Guided GradCAM, with the latter proving to be more time-consuming. GradCAM involves two primary steps: (1) a forward pass computes the output scores, and (2) a backward pass calculates the gradients of the class score with respect to the feature maps of a convolution layer. Assuming  $n$  to represent the number of such feature maps, the time complexity of these operations is approximately  $O(m * n)$ , where  $m$  corresponds to the size of the model parameters. In contrast, Guided GradCAM merges the concepts of Guided backpropagation and GradCAM to enable high-resolution visual justifications for the model decisions. Besides the forward and backward pass identical to GradCAM ( $O(m * n)$ ), Guided GradCAM necessitates an additional step of guided backpropagation. This additional computation effectively doubles the workload as gradients are calculated for each individual neuron, as opposed to each feature map in GradCAM. Consequently, the time complexity for Guided GradCAM escalates to approximately  $O(2m * n)$ . In our study, we further examined the temporal efficiency associated with various evaluation metrics to provide users with valuable insights for optimal selections. Our observations revealed that except for the metric of mutual information, which demands a considerably elongated execution period (approximately 250 seconds for an image sized  $128 \times 128$ ), the remaining metrics prove to be highly time-efficient (within a range of 0.01–0.02 s). This efficiency enables these metrics for potential integration into real-time clinical decision-support scenarios to improve model transparency.

To enhance translational research value in actual clinical settings, we utilize a poisoning attack strategy that incorporates designed triggers for evaluating several commonly used XAI techniques. This approach automates the framework without the requirement for human intervention or expert validation. Additionally, the integration of triggers eliminates the need for pixel-level annotations or ground truth, which are typically unavailable in real-world practice, especially for classification tasks. Specifically, we employ both static and dynamic triggers to meet the demands of translational clinical decision support systems. We then discuss adoptions of different triggers under different clinical scenarios for model transparency improvement. Static stamping, a conventional poisoning attack method, is readily available and can be easily customized in terms of shapes, locations, sizes, and other attributes based on the requirement of real-world applications. Static stamping can be employed to replicate hardware artifacts, such as ring artifacts, tube arcing, out-of-field artifacts, and air bubble artifacts<sup>69</sup>, which result from real-world imaging challenges. Dynamic triggers with pixel-level inferences are created by generative models, which provide flexibility in trigger design and generation to reflect real-world noise and biases in medical imaging. The generation method is not limited to the proposed method or a single distribution, as the adversary can utilize various distributions to create triggers. Employing different distributions allows the adversary to modify the appearance of the triggers used to mimic real-world biases, such as patient-based artifacts due to patient movement or the presence of metallic materials (e.g., motion artifact<sup>70</sup>, transient interruption of

contrast<sup>71</sup>, clothing and jewelry artifact), physics-based artifacts when acquiring imaging (e.g., beam hardening<sup>72</sup>, partial volume averaging, aliasing artifact, photon starvation<sup>73</sup>, quantum mottle, truncation artifact<sup>74</sup>), and helical and multichannel artifacts (e.g., windmill artifact, cone beam effect<sup>75</sup>, multiplanar reconstruction artifact<sup>72</sup>, zebra stripes<sup>76</sup>, stair-step artifact) during medical imaging reconstruction processes. Specifically, we present a case study that utilizes our proposed evaluation strategy to determine the optimal XAI method in the presence of film radiography artifacts, as illustrated in Fig. 9. By leveraging static and dynamic trigger generation, the proposed XAI method evaluation pipeline can be generalized to specific real-world settings, facilitating translational research.

Based on our evaluation results, we observed that no single XAI method surpasses others in all aspects. However, this study might offer a useful automated evaluation strategy to help model developers optimize their proposed explainable solutions and aid end-users, like healthcare providers, in selecting appropriate XAI methods based on their specific requirements and application settings. There are several limitations to the current study. First, the current static and dynamic triggers are all synthetic, which may not fully reflect the complexity and variability of real-world noise. Additionally, as an automated evaluation pipeline, the lack of human-in-the-loop, particularly clinical experts, may limit the validity of the results. Another potential limitation of existing work is that we only examined post-hoc XAI methods in medical image analysis. In future work, we will include more inherent approaches (e.g., attention mechanisms<sup>77</sup>) in clinical decision support systems with state-of-the-art model architecture (e.g., ViT). Meanwhile, we will specifically investigate model reliability, the relationship between model accuracy and predictive probability (e.g., Expected Calibration Error (ECE)<sup>78</sup>), to reveal how different XAI methods calibrate model confidence as an additional evaluating dimension. We plan to introduce real-world artifacts as static triggers and propose a generative network based on existing medical imaging artifacts for dynamic triggers to better simulate real-world noise and biases. Additionally, we aim to expand the current consistency evaluation to include expert agreement or additional annotations to pursue more accurate evaluation outcomes.

## Conclusion

In this study, we proposed an automated evaluation pipeline for XAI methods of explanation generation and representation using patch perturbation in medical imaging analysis. To eliminate the necessity of additional annotations or human intervention, we first generated static and dynamic triggers for poisoning attack during model retraining. We then proposed a comprehensive set of evaluation metrics during the model inference stage to facilitate the evaluation from multiple perspectives, including effectiveness, completeness, consistency, and complexity. The patch perturbation-based workflow addresses the previous gap in evaluating XAI methods in medical imaging by eliminating the need for human intervention and providing an automated strategy. To demonstrate the proposed evaluation strategy, we provided a case study with widely used XAI methods on COVID-19 X-ray imaging classification tasks. Additionally, we provided a detailed review of existing XAI methods with an emphasis on the availability of evaluation to underscore the significance and necessity of the XAI pipeline in medical imaging analysis. We believe this study offers an automated and generalizable evaluation strategy to help model developers optimize their proposed explainable solutions and aid end-users, such as healthcare providers, in selecting appropriate XAI methods in real-world clinical research and practice.

## Data availability

The datasets analysed during the current study are publicly available in the COVID-19 Radiography Database repository<sup>65,66</sup>.

## Appendix: Notation

Table 4 summarizes the notation frequently used in this study.

Notations	Description
$\mathbf{x}, \mathbf{x}_i$	Original input images
$\tilde{\mathbf{x}}, \tilde{\mathbf{x}}_i$	Patch-perturbed images
$\hat{\mathbf{x}}, \hat{\mathbf{x}}_i$	Recovered images
$y, y_i$	Original labels
$\tilde{y}, \tilde{y}_i$	Target labels for patched images
$f(\cdot)$	Baseline classification model
$f'(\cdot)$	Poisoned classification model
$\mathbf{p}, \mathbf{p}_s, \mathbf{p}_d$	Attached patches for perturbation
$\mathbf{m}$	Binary masks indicating the patch locations
$\mathbf{s}, \mathbf{s}_i$	Saliency maps
$M$	Number of cleaned samples in test set
$N$	Number of poisoned samples in test set
$\varepsilon$	Pixel-wise perturbation amount
$\alpha$	Poisoning ratio

**Table 4.** Summary of the notations.

Received: 21 March 2023; Accepted: 1 November 2023

Published online: 09 November 2023

## References

- Shen, D., Wu, G. & Suk, H.-I. Deep learning in medical image analysis. *Annu. Rev. Biomed. Eng.* **19**, 221–248 (2017).
- Shi, F. *et al.* Review of artificial intelligence techniques in imaging data acquisition, segmentation, and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* **14**, 4–15 (2020).
- Giuste, F. *et al.* Explainable artificial intelligence methods in combating pandemics: A systematic review. In *IEEE Rev. Biomed. Eng.* **16**, 5–21. <https://doi.org/10.1109/RBME.2022.3185953> (2022).
- Alvarez Melis, D. & Jaakkola, T. Towards robust interpretability with self-explaining neural networks. *Adv. Neural Inf. Process. Syst.* **31**, 1–10 (2018).
- Rivera, S. C. *et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: The spirit-AI extension. *Lancet Digit. Health* **2**, e549–e560 (2020).
- Liu, X. *et al.* Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The consort-AI extension. *Lancet Digit. Health* **2**, e537–e548 (2020).
- Antoniadi, A. M. *et al.* Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Appl. Sci.* **11**, 5088 (2021).
- Arrieta, A. B. *et al.* Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf. Fusion* **58**, 82–115 (2020).
- Zhou, J., Gandomi, A. H., Chen, F. & Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics* **10**, 593 (2021).
- Brunese, L., Mercaldo, F., Reginelli, A. & Santone, A. Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from x-rays. *Comput. Methods Prog. Biomed.* **196**, 105608 (2020).
- Mondal, A. K., Bhattacharjee, A., Singla, P. & Prathosh, A. xViTCOS: Explainable vision transformer based COVID-19 screening using radiography. *IEEE J. Transl. Eng. Health Med.* **10**, 1–10 (2021).
- Signoroni, A. *et al.* BS-Net: Learning COVID-19 pneumonia severity on a large chest X-ray dataset. *Med. Image Anal.* **71**, 102046 (2021).
- Duran-Lopez, L., Dominguez-Morales, J. P., Corral-Jaime, J., Vicente-Diaz, S. & Linares-Barranco, A. COVID-XNet: A custom deep learning system to diagnose and locate COVID-19 in chest x-ray images. *Appl. Sci.* **10**, 5683 (2020).
- Wang, Y. *et al.* Semantic-powered explainable model-free few-shot learning scheme of diagnosing COVID-19 on chest x-ray. *IEEE J. Biomed. Health Inform.* **26**, 5870–5882 (2022).
- Mahmud, T., Rahman, M. A. & Fattah, S. A. CovXNet: A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest x-ray images with transferable multi-receptive feature optimization. *Comput. Biol. Med.* **122**, 103869 (2020).
- Ozturk, T. *et al.* Automated detection of COVID-19 cases using deep neural networks with x-ray images. *Comput. Biol. Med.* **121**, 103792 (2020).
- Haghanifar, A., Majdabadi, M. M., Choi, Y., Deivalakshmi, S. & Ko, S. COVID-cxnet: Detecting COVID-19 in frontal chest x-ray images using deep learning. *Multimed. Tools Appl.* **81**, 30615–30645 (2022).
- Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint [arXiv:1312.6034](https://arxiv.org/abs/1312.6034) (2013).
- Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I* **13**, 818–833 (Springer, 2014).
- Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. arXiv preprint [arXiv:1412.6806](https://arxiv.org/abs/1412.6806) (2014).
- Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, 618–626 (2017).
- Karim, M. R. *et al.* Deepcovidexplainer: Explainable COVID-19 diagnosis from chest x-ray images. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 1034–1037 (IEEE, 2020).
- Chatterjee, S. *et al.* Exploration of interpretability techniques for deep COVID-19 classification using chest x-ray images. arXiv preprint [arXiv:2006.02570](https://arxiv.org/abs/2006.02570) (2020).
- Karthik, R., Menaka, R. & Hariharan, M. Learning distinctive filters for COVID-19 detection from chest x-ray using shuffled residual CNN. *Appl. Soft Comput.* **99**, 106744 (2021).
- Ghorbani, A., Abid, A. & Zou, J. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. **33**, 3681–3688 (2019).
- Lin, T.-C. & Lee, H.-C. COVID-19 chest radiography images analysis based on integration of image preprocess, guided grad-cam, machine learning and risk management. In *Proceedings of the 4th International Conference on Medical and Health Informatics*, 281–288 (2020).
- Meyes, R., Lu, M., de Puiseau, C. W. & Meisen, T. Ablation studies in artificial neural networks. arXiv preprint [arXiv:1901.08644](https://arxiv.org/abs/1901.08644) (2019).
- Ribeiro, M. T., Singh, S. & Guestrin, C. “why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144 (2016).
- Ahsan, M. M. *et al.* Study of different deep learning approach with explainable ai for screening patients with COVID-19 symptoms: Using ct scan and chest x-ray image dataset. arXiv preprint [arXiv:2007.12525](https://arxiv.org/abs/2007.12525) (2020).
- Teixeira, L. O. *et al.* Impact of lung segmentation on the diagnosis and explanation of COVID-19 in chest x-ray images. *Sensors* **21**, 7116 (2021).
- Punn, N. S. & Agarwal, S. Automated diagnosis of COVID-19 with limited posteroanterior chest x-ray images using fine-tuned deep neural networks. *Appl. Intell.* **51**, 2689–2702 (2021).
- Zhang, A. *et al.* A-fni: Learning attributions from deep networks via feature map importance. arXiv preprint [arXiv:2104.05527](https://arxiv.org/abs/2104.05527) (2021).
- Adebayo, J. *et al.* Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* **31**, 1–11 (2018).
- Tsiknakis, N. *et al.* Interpretable artificial intelligence framework for COVID-19 screening on chest x-rays. *Exp. Ther. Med.* **20**, 727–735 (2020).
- Petsiuk, V., Das, A. & Saenko, K. Rise: Randomized input sampling for explanation of black-box models. arXiv preprint [arXiv:1806.07421](https://arxiv.org/abs/1806.07421) (2018).
- Yang, M. & Kim, B. Bim: Towards quantitative evaluation of interpretability methods with ground truth. arXiv preprint [arXiv:1907.09701](https://arxiv.org/abs/1907.09701) (2019).
- Nguyen, H. T. T., Cao, H. Q., Nguyen, K. V. T. & Pham, N. D. K. Evaluation of explainable artificial intelligence: Shap, lime, and cam. In *Proceedings of the FPT AI Conference*, 1–6 (2021).
- Rieger, L. & Hansen, L. K. A simple defense against adversarial attacks on heatmap explanations. arXiv preprint [arXiv:2007.06381](https://arxiv.org/abs/2007.06381) (2020).

39. Heo, J., Joo, S. & Moon, T. Fooling neural network interpretations via adversarial model manipulation. *Adv. Neural Inf. Process. Syst.* **32**, 1–12 (2019).
40. Lin, Y.-S., Lee, W.-C. & Celik, Z. B. What do you see? evaluation of explainable artificial intelligence (XAI) interpretability through neural backdoors. arXiv preprint [arXiv:2009.10639](https://arxiv.org/abs/2009.10639) (2020).
41. Liu, Y. *et al.* Trojaning attack on neural networks. In *Network and Distributed System Security Symposium* (2018).
42. Saha, A., Subramanya, A. & Pirsivash, H. Hidden trigger backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 34, 11957–11965 (2020).
43. Chen, X., Liu, C., Li, B., Lu, K. & Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. arXiv preprint [arXiv:1712.05526](https://arxiv.org/abs/1712.05526) (2017).
44. Gu, T., Dolan-Gavitt, B. & Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint [arXiv:1708.06733](https://arxiv.org/abs/1708.06733) (2017).
45. Salem, A., Wen, R., Backes, M., Ma, S. & Zhang, Y. Dynamic backdoor attacks against machine learning models. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS & P)*, 703–718 (IEEE, 2022).
46. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint [arXiv:1412.6572](https://arxiv.org/abs/1412.6572) (2014).
47. Fu, Y., Zhang, S., Wu, S., Wan, C. & Lin, Y. Patch-fool: Are vision transformers always robust against adversarial perturbations? arXiv preprint [arXiv:2203.08392](https://arxiv.org/abs/2203.08392) (2022).
48. Gu, J., Tresp, V. & Qin, Y. Are vision transformers robust to patch perturbations? In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII*, 404–421 (Springer, 2022).
49. Dong, X. *et al.* Greedyfool: Distortion-aware sparse adversarial attack. *Adv. Neural Inf. Process. Syst.* **33**, 11226–11236 (2020).
50. Samek, W., Binder, A., Montavon, G., Lapuschkin, S. & Müller, K.-R. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans. Neural Netw. Learn. Syst.* **28**, 2660–2673 (2016).
51. Zhang, J. *et al.* Top-down neural attention by excitation backprop. *Int. J. Comput. Vis.* **126**, 1084–1102 (2018).
52. Li, X.-H. *et al.* Quantitative evaluations on saliency methods: An experimental study. arXiv preprint [arXiv:2012.15616](https://arxiv.org/abs/2012.15616) (2020).
53. Rebuffi, S.-A., Fong, R., Ji, X. & Vedaldi, A. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8839–8848 (2020).
54. Dombrowski, A.-K. *et al.* Explanations can be manipulated and geometry is to blame. *Adv. Neural Inf. Process. Syst.* **32**, 1–12 (2019).
55. Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I. & Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Adv. Neural Inf. Process. Syst.* **32**, 1–12 (2019).
56. Wehbe, R. M. *et al.* DeepCOVID-XR: An artificial intelligence algorithm to detect COVID-19 on chest radiographs trained and tested on a large US clinical data set. *Radiology* **299**, E167–E176 (2021).
57. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint [arXiv:1409.1556](https://arxiv.org/abs/1409.1556) (2014).
58. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **22**, 1345–1359 (2010).
59. Goodfellow, I. *et al.* Generative adversarial networks. *Commun. ACM* **63**, 139–144 (2020).
60. Rieger, L. & Hansen, L. K. Aggregating explanation methods for stable and robust explainability. arXiv preprint [arXiv:1903.00519](https://arxiv.org/abs/1903.00519) (2019).
61. Ning, R., Li, J., Xin, C., Wu, H. & Wang, C. Hibernated backdoor: A mutual information empowered backdoor attack to deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence* vol. 36, 10309–10318 (2022).
62. Cheng, S., Liu, Y., Ma, S. & Zhang, X. Deep feature space trojan attack of neural networks by controlled detoxification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 1148–1156 (2021).
63. Jia, S., Song, Y., Ma, C. & Yang, X. IOU attack: Towards temporally coherent black-box adversarial attack for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6709–6718 (2021).
64. Chen, X. *et al.* One-shot adversarial attacks on visual tracking with dual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10176–10185 (2020).
65. Chowdhury, M. E. *et al.* Can AI help in screening viral and COVID-19 pneumonia?. *IEEE Access* **8**, 132665–132676 (2020).
66. Rahman, T. *et al.* Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Comput. Biol. Med.* **132**, 104319 (2021).
67. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 1–12 (2019).
68. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. arXiv preprint [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
69. Wong, S. S. M., Cheung, J. S. W., Lee, T. W. K., Tsai, C. S. C. & Ng, A. W. H. Ct brain artefact due to air bubbles in the oil cooling system: Characteristic band-like configuration on sagittal reformatted image. *Jpn. J. Radiol.* **36**, 90–95 (2018).
70. Boas, F. E. *et al.* CT artifacts: Causes and reduction techniques. *Imaging Med.* **4**, 229–240 (2012).
71. Magaya, E. Case report: Transient interruption of contrast. *S. Afr. Radiograph.* **54**, 27–29 (2016).
72. Barrett, J. F. & Keat, N. Artifacts in CT: Recognition and avoidance. *Radiographics* **24**, 1679–1691 (2004).
73. Mori, I., Machida, Y., Osanai, M. & Inuma, K. Photon starvation artifacts of X-ray CT: Their true cause and a solution. *Radiol. Phys. Technol.* **6**, 130–141 (2013).
74. Chen, Y., Li, K., Li, Y., Hsieh, J. & Chen, G.-H. Reduction of truncation artifacts in ct images via a discriminative dictionary representation method. In *Medical Imaging 2016: Physics of Medical Imaging*, vol. 9783, 341–347 (SPIE, 2016).
75. Flohr, T. & Ohnesorge, B. Multi-slice ct technology. *Multislice CT* 41–69 (2007).
76. Stadler, A., Schima, W., Ba-Ssalamah, A., Kettenbach, J. & Eisenhuber, E. Artifacts in body MR imaging: Their appearance and how to eliminate them. *Eur. Radiol.* **17**, 1242–1255 (2007).
77. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7794–7803 (2018).
78. Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural networks. In *International Conference on Machine Learning*, 1321–1330 (PMLR, 2017).

## Acknowledgements

This work was supported by a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research to Professor May D. Wang.

## Author contributions

J.S. and W.S. contributed to the study design, data pre-processing, statistical analysis, model development, result analysis, and writing of the manuscript, including figures and tables. F.G., Y.V., and L.T. contributed to the data collection, data annotation, study design, result evaluation, clinical validation, and manuscript revision. M.W. contributed to the study design, result evaluation, and extensive refining of the manuscript. All authors reviewed the manuscript.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-46493-2>.

**Correspondence** and requests for materials should be addressed to M.D.W.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023