An Evolutionary Algorithm for Black-Box Adversarial Attack Against Explainable Methods

Anonymous authors
Paper under double-blind review

Abstract

The challenge of deep neural network (DNN) explainability continues to be a significant hurdle in developing trustworthy AI, particularly in essential fields like medical imaging. Despite progress in explainable AI (XAI), these methods remain susceptible to adversarial images, emphasizing the urgent need for robustness evaluation. While many current adversarial attack techniques focus on specific explanation strategies, emerging research has introduced black-box methods capable of targeting multiple approaches. However, such methods often necessitate a large number of queries due to the complexity of pixel-level modifications. In response, we propose an innovative attack method that employs semi-transparent, RGB-valued circles to create perturbations, optimizing their features via an evolutionary strategy, drastically reducing the number of tunable optimization parameters required. Through experiments on medical image datasets, our method demonstrates superior performance compared to current leading techniques. This study further underscores the vulnerabilities of XAI methods in critical sectors such as medical imaging, advocating for more robust solutions.

1 Introduction

Deep neural networks (DNNs) have revolutionized the field of computer vision, driving significant advancements across a variety of tasks Lin et al. (2014); Simonyan & Zisserman (2015); Springenberg et al. (2015). In healthcare, artificial intelligence is becoming a transformative force, offering groundbreaking solutions for diagnosis, treatment, and patient care Chaddad et al. (2023). Yet, the black-box nature of many DNNs raises concerns regarding their explainability, accountability, and trustworthiness Quinn et al. (2021); Rane et al. (2023); Rosenbacke et al. (2024b). To address these issues and bolster trust, explainable artificial intelligence (XAI) has emerged as a pivotal area of research. By understanding the the decision-making processes of complex DNNs, XAI fosters confidence among healthcare providers and patients Dosilovic et al. (2018). In the realm of computer vision, explanation methods frequently generate attribution maps that visualize feature importance, illustrating how different elements contribute to a DNN's predictions Simonyan et al. (2014); Shrikumar et al. (2017); Selvaraju et al. (2017); Lundberg & Lee (2017); Böhle et al. (2024).

Despite the progress in existing explainability methods, recent studies reveal their vulnerability to adversarial inputs Tamam et al. (2023); Huang et al. (2023); Baniecki & Biecek (2024). These inputs, which are subtly altered by imperceptible perturbations (as illustrated in Figure 1), have shown to potentially impact both the attribution maps and classification of a DNN simultaneously. The occurrence of such adversarial examples in real-world scenarios Dong et al. (2025); Wang et al. (2023) is particularly troubling for areas where DNN explainability is crucial or legally required, such as autonomous driving Omeiza et al. (2022) and healthcare Chaddad et al. (2023); Hao et al. (2024); van der Velden et al. (2022). Therefore, the development of adversarial attack methods has emerged as a critical research avenue for assessing the robustness of explainability methods Tamam et al. (2023); Huang et al. (2023).

Initially, research efforts primarily focused on crafting adversarial images that target XAI methods by utilizing insights into the underlying DNN architecture, known as white-box attacks Wang et al. (2023); Moosavi-Dezfooli et al. (2016); Zhang et al. (2020); Ghorbani et al. (2019). However, these strategies often

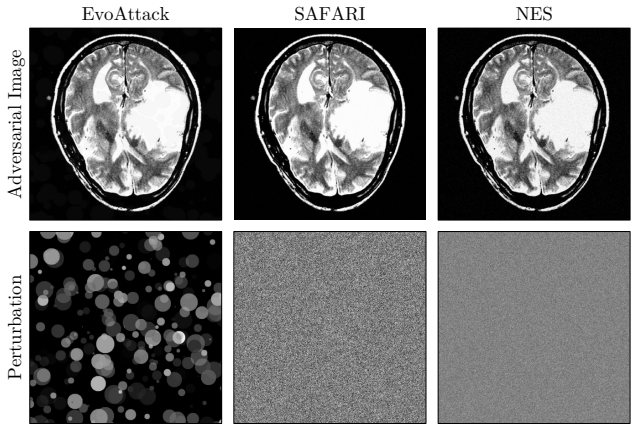

Figure 1: Adversarial images and perturbations generated by the NES Tamam et al. (2023), SAFARI Huang et al. (2023) and the proposed EvoAttack algorithm when attacking an image from the Br35h datasets. We observe that the adversarial perturbation generated by NES and SAFARI perturbs every pixel of the image whereas the perturbation generated by the proposed method is constructed using a set of 300 *circle* shapes

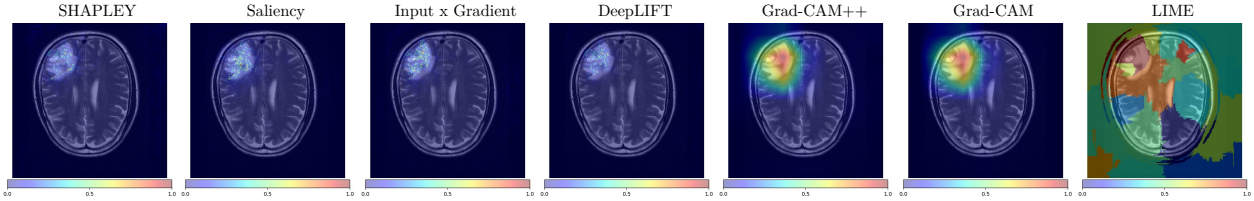

Figure 2: Attribution maps generated by XAI methods. These methods are applied to explain a 'true' decision for tumor classification on the Br35h dataset made by the trained VGG-16 classifier. We observe that DeepLIFT, SHAPLEY, Saliency and Input x Gradient methods produce high-granularity attribution maps, emphasizing important pixels. In contrast, Grad-Cam, Grad-Cam++ and LIME provide attributions that capture more global features, highlighting broader regions of the image.

lack the ability to generalize across different explanation techniques. Consequently, recent investigations have shifted toward the black-box scenario, where only the input-output pairs of the DNN and the XAI method are available Tamam et al. (2023); Huang et al. (2023). To achieve this, existing attacks largely rely on meta-heuristic approaches Tamam et al. (2023); Huang et al. (2023), inspired by evolutionary algorithms Li et al. (2024).

While existing methods have successfully generated adversarial images against XAI techniques, they face key limitations. Firstly, these methods often require extensive querying of both the DNN and the XAI method to achieve meaningful distortions in attribution maps. This dependency poses substantial challenges in environments where query budgets are limited or expensive, whether due to financial constraints Ilyas et al. (2018); Dhabliya et al. (2024) or time restrictions Keddous et al. (2023). As a result, conducting robustness evaluations that involve adversarial attacks with high query budgets becomes costly, impacting both financial resources and development time. This issue stems from the use of population-based approaches combined with the inherently high-dimensional nature of the search space—for instance, attacking an image from the HAM10000 dataset Tschandl et al. (2018) with dimensions $(450 \times 600 \times 3)$ results in searching through a space of $810,000$ dimensions. Secondly, existing attacks often overlook the varying granularity of XAI methods' explanations when designing perturbations, as shown in Figure 2. Current approaches tend to modify all pixels independently, which is effective when targeting XAI methods that produce detailed, pixel-level explanation maps, such as SHAPLEY or Saliency methods. However, these approaches struggle against XAI methods that emphasize broader, global regions, leading to increased robustness in methods like

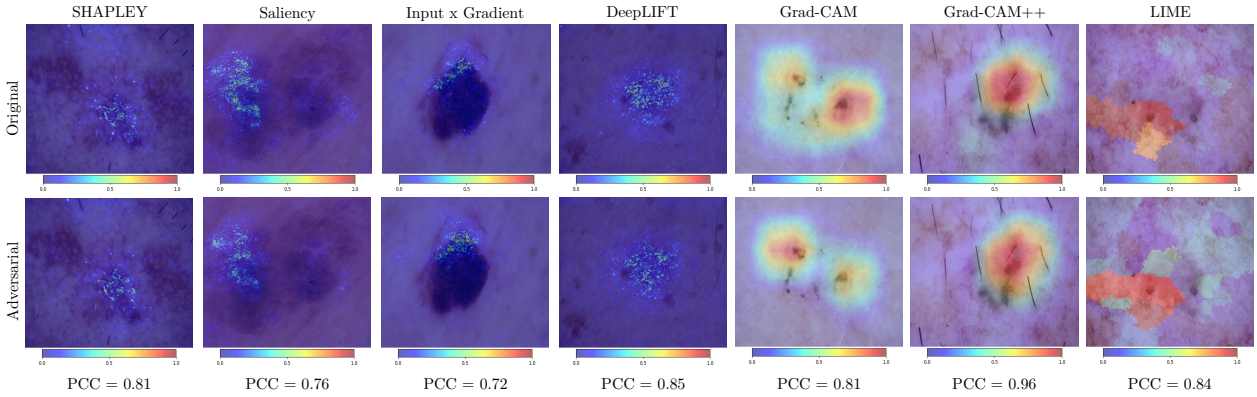

Figure 3: Task 1 (misclassification with preserved explanation) adversarial images produced by the EvoAttack method, along with the respective generated attribution maps. Both original and adversarial explanations on the HAM10000 images are visually similar with PCC values of 0.7 and above. Both sets of explanations highlight seemingly relevant regions of the image, however, all adversarial images cause the underlying VGG-16 DNN to misclassify.

Grad-CAM and Grad-CAM++ Huang et al. (2023). This oversight highlights the need for more adaptive attack strategies that consider the explanatory granularity of different XAI techniques.

To address these limitations, we propose a novel attack method inspired by image approximation techniques from the computational-art community Lambert et al. (2013); Garbaruk et al. (2022); Tian & Ha (2022). Our approach involves creating adversarial perturbations using a set of RGB-valued, semi-transparent shapes. These shapes are optimized via an evolutionary strategy to maximize distortion in the attribution maps generated by XAI methods. By focusing on the intrinsic characteristics of each shape, this approach substantially reduces the search space and remains consistent across different image sizes.

The rest of this paper is organized as follows: In Section 2, we provide an overview of related works, highlighting their contributions and limitations. Section 3 outlines our proposed attack scenario and offers an in-depth explanation of our method's implementation. The proposed evaluation metrics, alongside empirical results, are presented and analysed in Section 4. Finally, Section 5 concludes the paper and suggests potential directions for future research.

2 Related Works

Adversarial attacks on DNNs have become one of the most active research areas within the machine learning community, predominantly focusing on targeting DNN classifiers Williams & Li (2023b); Dong et al. (2025); Ilyas et al. (2018); Williams & Li (2023a); Madry et al. (2017); Andriushchenko et al. (2020). More recently, there has been increasing interest in exploring the effects of adversarial perturbations on attribution maps produced by XAI methods, addressing both white-box Heo et al. (2019); Moosavi-Dezfooli et al. (2016); Zhang et al. (2020); Ghorbani et al. (2019); Kindermans et al. (2019); Subramanya et al. (2019); Dombrowski et al. (2019); Kuppa & Le-Khac (2020) and black-box Tamam et al. (2023); Huang & Zhang (2020) settings. Our work situates itself within this domain by focusing on the black-box scenario. Unlike existing black-box approaches that demand extensive query budgets, our method is designed to achieve substantial distortion with a notably limited query budget, addressing key constraints faced by current strategies.

As DNNs are increasingly deployed in real-world applications, ensuring that their decision-making processes are interpretable becomes essential Doshi-Velez et al. (2017). This underscores the importance of developing robust and accurate explanations for DNNs. Adversarial attacks against XAI methods are designed to assess the sensitivity of explanations to minor changes in input images Alvarez-Melis & Jaakkola (2018). While traditional adversarial attacks on DNNs focus on inducing misclassification, attacks against XAI methods aim to challenge the reliability of provided explanations by either 1) causing minimal distortion to an attribution

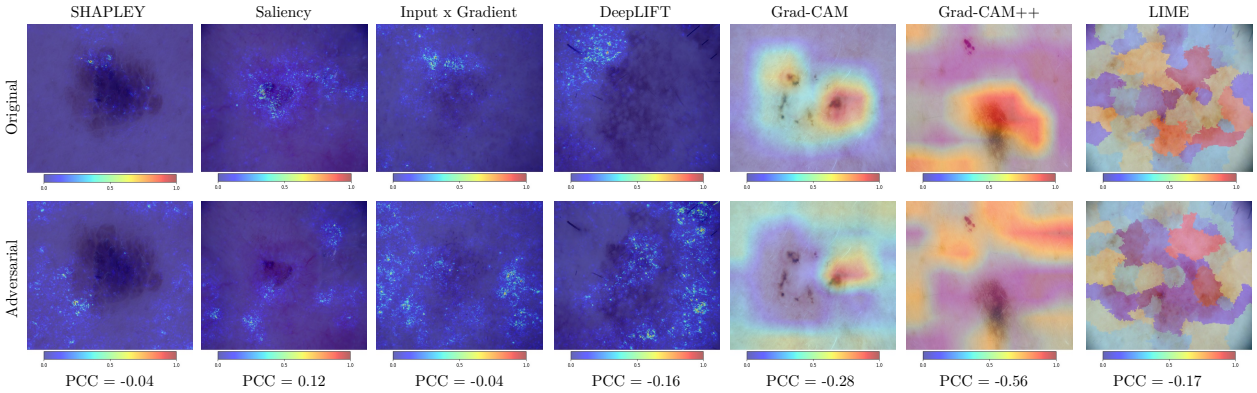

Figure 4: Task 2 (correct classification with distorted explanation) adversarial images produced by the EvoAttack method, along with the respective generated attribution maps. The original and adversarial explanations on the HAM10000 images are visually dissimilar with PCC values of -0.04 and below. Dispite the distortion, all adversarial images are correctly classified by the VGG-16 DNN.

map while causing misclassification Huang et al. (2023), or 2) maximizing the distortion of an attribution map whilst maintaining the correct classification Tamam et al. (2023) (shown in Figures 3 and 4).

In their pioneering work, Ghorbani et al.Ghorbani et al. (2019) introduced the concept of adversarial attacks on explanations by targeting gradient-based feature attributions in convolutional DNNs. The authors iteratively perturbed inputs in the direction that altered the explanation's gradient. Concurrently, research by Kindermans et al.Kindermans et al. (2019) highlighted the sensitivity of explanations to slight input transformations, although they did not directly propose methods for constructing such attacks. Subsequent works exploiting the gradient information have constructed adversarial images by altering all pixels in the image Zhang et al. (2020) in addition to localised regions (patches) Selvaraju et al. (2017).

While most existing methods focus on the white-box setting, recent efforts have shifted toward developing black-box approaches. In the black-box scenario, only the input-output pairs of the DNN and XAI methods are accessible, allowing these techniques to be applied to any explanation method that produces an attribution map Huang et al. (2023). Existing black-box techniques often employ heuristic optimization methods to craft adversarial images by solving a constructed loss function. Tamam et al.Tamam et al. (2023) utilize the Natural Evolutionary Strategy method Wierstra et al. (2014) to minimize the loss function previously proposed by Dombrowski et al. Dombrowski et al. (2019). Their attack iteratively updates a single solution by sampling a set of solutions from a normal distribution at each iteration to estimate the gradient of the loss function. Conversely, the attack method by Huang et al.Huang et al. (2023) does not rely on any gradient information. Instead, they adapt the POBA-GA genetic algorithm developed by Chen et al.Chen et al. (2019) to evolve a population of solutions through the genetic operators crossover and mutation. The authors demonstrate the superior performance of their method by targeting both gradient-based and perturbation-based explanation methods.

Despite recent advancements in black-box adversarial attacks against XAI methods, the high number of queries required for both the DNN and the XAI method raises concerns about their practical applicability to financial Dhabliya et al. (2024) or time Keddous et al. (2023) restricted scenarios. Moreover, the excessive effort to construct adversarial images questions the realistic assessment of explanation robustness against adversarial images Wu et al. (2021). Furthermore, existing methods do not consider the granularity of the XAI methods' explanation and perturb input images by modifying individual pixels. In doing so, these methods have shown to perform well on granular XAI methods such as saliency maps and DeepLIFT but struggle against

For an in-depth survey on adversarial attacks against explainable AI methods, we refer readers to Baniecki & Biecek (2024).

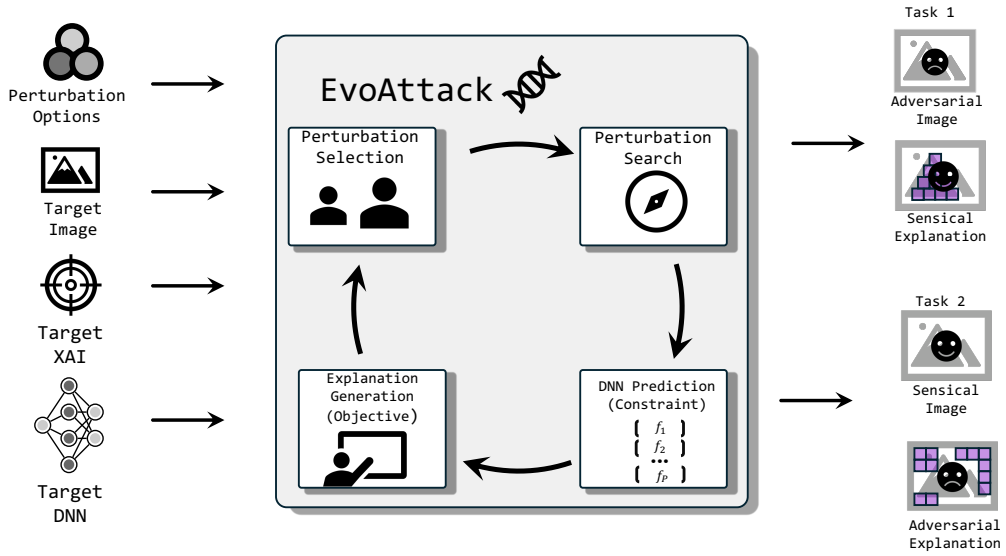

Figure 5: Flowchart demonstrates the process of the proposed EvoAttack method. The proposed method follows the $(1+1)$-ES structure and iteratively generates a single child solution by applying random changes to its parent. If the parent solution is dominated (defined in Section 3.4) by its child, it is replaced for the next generation, otherwise the parent remains.

3 Proposed Method

The aim of our method is to create adversarial images that either: 1) minimize distortion in the attribution map while causing an incorrect classification or 2) maximize the distortion in the attribution map while keeping the DNN classification unchanged Huang et al. (2023), all while operating within a constrained query budget. We design the perturbation as a set of semi-transparent RGB-valued circles, which reduce the search space dimension to their properties. Similar to existing approaches, we generate adversarial perturbations by optimizing a distance-based objective function. We begin this section by formulating the problem, followed by a detailed description of each component of our proposed method. The overall structure of our approach is summarized in Figure 5.

3.1 Problem Formulation

Consider a trained DNN classifier $f : \mathcal{X} \subseteq [0,1]^{h\times w\times 3} \rightarrow \mathbb{R}^P$ which takes a benign RGB image $\mathbf{x} \in \mathcal{X}$ of height h and width w and outputs a label $y = \underset{p\in\{1,\cdots,P\}}{\operatorname{argmax}} f_p(\mathbf{x})$, with P representing the total number of class labels. Further, let $g(\cdot,\cdot)$ be an explanation function, where both the trained DNN f and benign image $\mathbf{x}$ are inputs. In this work, we attack XAI methods g that output an attribution map $g(\cdot,\cdot) \rightarrow \mathbb{R}^{h\times w}$ where the height and width match that of the input image $\mathbf{x}$.

To preserve the semantic integrity of the image, we adhere to existing attack methods by constraining the perturbation size using the l_∞ norm Huang et al. (2023); Tamam et al. (2023); Dombrowski et al. (2019). Consequently, we aim to generate a perturbation $\boldsymbol{\delta}$ that solves the following optimization problems:

Task 1:

$$\begin{aligned} \underset{\boldsymbol{\delta}}{\text{minimize}} \quad & \mathcal{D}(g(\mathbf{x}), g(\mathbf{x}+\delta)) \\ \text{subject to} \quad & ||\boldsymbol{\delta}||_\infty \leq \epsilon, \\ & \mathcal{L}(f; \mathbf{x}+\boldsymbol{\delta}, y_q) < 0, \\ & 0 \leq \mathbf{x}+\boldsymbol{\delta} \leq 1, \end{aligned} \tag{1}$$

,

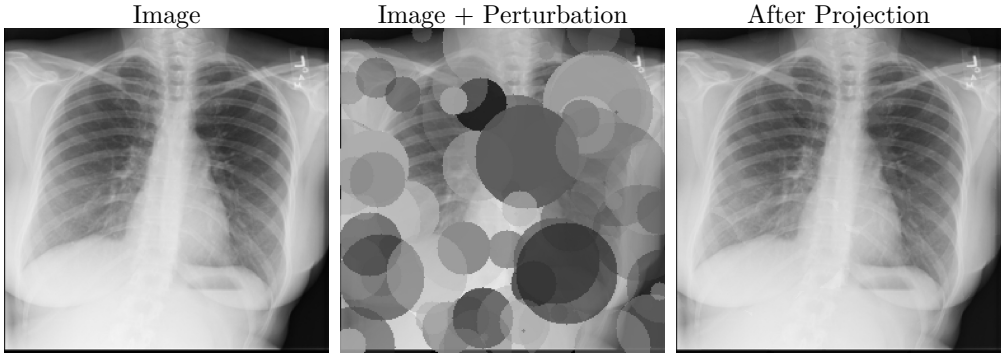

Figure 6: The process of repairing an adversarial image $\mathbf{x} + \delta$ (described in equation (5)) to ensure it satisfies the constraints of (1) and (2), where $\epsilon = 6/255$

Task 2:

$$\begin{aligned} \underset{\boldsymbol{\delta}}{\text{maximize}} \quad & \mathcal{D}(g(\mathbf{x}), g(\mathbf{x}+\delta)) \\ \text{subject to} \quad & ||\boldsymbol{\delta}||_\infty \leq \epsilon, \\ & \mathcal{L}(f; \mathbf{x}+\boldsymbol{\delta}, y_q) > 0, \\ & 0 \leq \mathbf{x}+\boldsymbol{\delta} \leq 1, \end{aligned} \tag{2}$$

where $\mathcal{D}$ represents the distance measure between the attribution maps of the benign and adversarial images and ϵ controls the extent of the perturbation's impact on the benign image. We follow the setup of Huang et al. by defining $\mathcal{D}$ as the $1/PCC$ where PCC is the Pearson Correlation Coefficient. We ensure that the value of the loss function $\mathcal{L}(\cdot)$ is negative when $\mathbf{x} + \boldsymbol{\delta}$ results in misclassification. This is achieved by defining the loss in the constraint as the margin loss:

$$\mathcal{L}(f; \mathbf{x}+\boldsymbol{\delta}, y) = f_y - f_{y_q}, \tag{3}$$

where y corresponds to the true labels of $\mathbf{x}$ and $y_q = \underset{q \neq y}{\text{argmax}}\ f_p(\mathbf{x})$ is a label corresponding to a class other than the true class y.

3.2 Perturbation Initialization

To alleviate the issues associated with the large number of tunable values of an image perturbation, we construct a perturbation by overlaying N RGB-valued semi-transparent circles, drawing inspiration from techniques prevalent in the computational art field Lambert et al. (2013); Garbaruk et al. (2022); Tian & Ha (2022).

In this work, we construct the adversarial perturbation $\boldsymbol{\delta}$ through the concatenation of N shapes:

$$\boldsymbol{\delta} = \boldsymbol{\delta}^1 \oplus \boldsymbol{\delta}^2 \cdots \oplus \boldsymbol{\delta}^N \tag{4}$$

where $\oplus$ denotes the concatenation operator and $\boldsymbol{\delta}^a$ the a-th shape applied to a blank array. In the proposed attack, each shape $\boldsymbol{\delta}^a$, for $a \in 1, \cdots, N$, is represented by a vector consisting of seven elements: the centre's coordinates (c_1^a, c_2^a), the radius $r^a \times$ (Max Diameter) $\times (w \cdot h)$, where 'Max Diameter' is a tunable parameter controlling the size of the circles, and w, h are the width and height of the attacked image. Finally, the vector includes the RGB values R^a, G^a, B^a, and the transparency T^a. These elements are normalized to continuous values between 0 and 1 and are initially sampled randomly from a uniform distribution, $\boldsymbol{\delta}^a \sim \mathcal{U}(0, 1)$. Constructing the perturbation in this manner reduces the number of optimization variables significantly—from $h \times w \times 3$ (which totals $150,528$ for a typical $(224, 224, 3)$ ImageNet image) to $N \times 7$, where N is the number of circle shapes used to construct the perturbation defined by the user. Importantly, this method of perturbation construction is also invariant to the image size.

3.3 Adversarial Image Construction

Given an adversarial perturbation $\boldsymbol{\delta}$ and a benign image $\mathbf{x}$, where both $\mathbf{x}$ and $\boldsymbol{\delta}$ belong to $\mathbb{R}^{h \times w \times 3}$, the corresponding adversarial image $\mathbf{x}^*$ is generated by equation (4) which overlaps all shapes from $\boldsymbol{\delta}$ onto $\mathbf{x}$. To ensure $\mathbf{x}^*$ complies with the constraints of (1) and (2), we project the pixels of the constructed adversarial image as follows:

$$\mathbf{x}_i^* = \begin{cases} \mathbf{x}_i + \epsilon & \text{if } \mathbf{x}_i^* > \mathbf{x}_i + \epsilon \\ \mathbf{x}_i - \epsilon & \text{if } \mathbf{x}_i^* < \mathbf{x} - \epsilon \\ \mathbf{x}_i^* & \text{otherwise} \end{cases}, \tag{5}$$

where i are the pixel indices. For greyscale images, we convert the RGB values of each circle to their greyscale equivalents before placing them onto the benign image. We visualize the projection process in Figure 6 which ensures the constraints of (1) and (2) are satisfied.

3.4 Perturbation Optimization

To optimize the properties of each shape $\boldsymbol{\delta}^a$, we utilize a single-solution evolutionary strategy known as the $(1+1)$-ES. In each iteration, a child solution δ^{**} is generated by perturbing its parent δ^* using values sampled from a normal distribution $\sigma \cdot \mathcal{N}(0, I)$. The parameter σ is adjustable, allowing a balance between exploration—searching unexplored regions of the solution space—and exploitation—fine-tuning the current solution. A larger σ facilitates exploration, while a smaller value encourages exploitation. The child solution replaces the parent if it demonstrates superior performance on the task at hand.

Traditionally, the selection process is based solely on comparing the objective values of the parent and child solutions. However, due to the constraints outlined in equations (1) and (2), a more nuanced selection process is warranted. Huang et al. addressed this issue for (1) by implementing a heuristic method related to population dynamics and resolved (2) by multiplying the objective value of solutions not satisfying the constraint with -1. Alternatively, Tamam et al. proposed a methodology that combines $\mathcal{D}$ and $\mathcal{L}$ using a weighted sum of its estimated gradients. Nonetheless, the single-solution architecture of our proposed method limits its ability to incorporate these existing strategies.

Instead, we draw inspiration from previous research in evolutionary computation that tackles objectives constrained by complex conditions using dominance functions Coello Coello & Mezura-Montes (2002); Williams & Li (2023b); Williams et al. (2023), similar to techniques in the multi-objective domain Deb (2001). Specifically, for each attack task, given the parent and child solutions δ^* and δ^{**}, respectively, δ^{**} replaces its parent δ^* if one of the following conditions is satisfied:

Definition 1 (Task 1 Domination) *:*

- $\mathcal{L}(\delta^*) \geq 0$ *and* $\mathcal{L}(\delta^{**}) < 0$.
- *Both* $\mathcal{L}(\delta^*) < 0$ *and* $\mathcal{L}(\delta^{**}) < 0$ *and* $\mathcal{D}(g(\mathbf{x} + \delta^{**})) < \mathcal{D}(g(\mathbf{x} + \delta^*))$
- *Both* $\mathcal{L}(\delta^*) > 0$ *and* $\mathcal{L}(\delta^*) > 0$ *and* $\mathcal{L}(\delta^{**}) < \mathcal{L}(\delta^*)$

Definition 2 (Task 2 Domination) *:*

- $\mathcal{L}(\delta^*) \leq 0$ *and* $\mathcal{L}(\delta^*) > 0$.
- *Both* $\mathcal{L}(\delta^*) > 0$ *and* $\mathcal{L}(\delta^*) > 0$ *and* $\mathcal{D}(g(\mathbf{x} + \delta^{**}) > \mathcal{D}(g(\mathbf{x} + \delta^*)$
- *Both* $\mathcal{L}(\delta^*) < 0$ *and* $\mathcal{L}(\delta^*) < 0$ *and* $\mathcal{L}(\delta^{**}) < \mathcal{L}(\delta^*)$.

These domination definitions are designed to guide our attack method in generating perturbation values that either minimize (for task 1) or maximize (for task 2) the distortion to the explanation, while simultaneously satisfying the respective classification constraints.

4 Experiments

Most existing studies assess their methods by attacking explanation techniques applied to DNN classifiers trained on ImageNet Deng et al.. In contrast, our work concentrates on classification tasks within the medical imaging domain, where the explainability of a DNN's decisions is critically important Chaddad et al. (2023); Hao et al. (2024); van der Velden et al. (2022). Therefore, we target explanation methods used with DNN classifiers trained on three distinct medical image datasets. The experimental setup is outlined in Section 4.1, which is followed by a comparative analysis of leading XAI attack methodologies, including SAFARI Huang et al. (2023) and NES Tamam et al. (2023), detailed in Section 4.2. Subsequently, we employ our proposed method to rank the robustness of various XAI techniques, as presented in Section 4.3. In Section 4.4, we evaluate XAI methods applied to adversarially trained models that incorporate adversarial images generated by our attack method within their training processes. Lastly, Section 4.5 presents an ablation study that examines the significance of different components and parameters of our proposed approach.

4.1 Experimental Setup

Datasets: Our experiments focus on evaluating XAI methods applied to DNN classifiers trained on three distinct medical image datasets. First, we utilize the HAM10000 dataset Tschandl et al. (2018), which comprises $10,000$ images of common pigmented skin lesions categorized into seven types of cancerous lesions. Second, we employ the Br35h dataset Hamada (2020), containing $3,000$ brain MRI scans that feature cases with and without tumors. Lastly, we examine the COVID-QU-Ex dataset Tahir et al. (2021), which includes $33,920$ chest X-rays classified into three categories: COVID-19, Pneumonia, and Normal. Consistent with the setup used in existing literature, we perform attacks on 100 randomly selected, correctly classified images from the test set of each dataset Huang et al. (2023); Tamam et al. (2023). All images are resized to dimensions $(224 \times 224 \times 3)$ before being processed by the DNN.

Explanation and Classifier Settings: In this study, we explore seven XAI methods to assess their robustness and efficiency. Specifically, we evaluate Grad-CAM Selvaraju et al. (2017), Grad-CAM++ Chattopadhyay et al. (2017), Saliency Map Simonyan et al. (2014); Selvaraju et al. (2017), DeepLIFT Shrikumar et al. (2017), GradientxInput Shrikumar et al. (2017), LIME Peng & Menzies (2021), and SHAPLEY Lundberg & Lee (2017). For the classifiers, we employ the architectures of MobileNet Howard et al. (2019), AlexNet Krizhevsky et al. (2012), and VGG-16 Simonyan & Zisserman (2015). Given the relatively small size of our medical datasets, we fine-tune models pre-trained on ImageNet using the PyTorch library Paszke et al. (2019). Each model undergoes fine-tuning over 10 epochs, with a batch size of 32 and a learning rate of 1×10^{-4}, utilizing the ADAM optimizer Kingma & Ba (2015) and Cross Entropy Loss. The datasets are divided into training, validation, and testing subsets with a ratio of 70%/10%/20%. Detailed performance metrics are provided in Section A.1 in the Appendix. All experiments were executed on an NVIDIA RTX A6000 GPU system.

Parameter Settings: To ensure that perturbations result in minimal semantic alterations to the images, we adopt settings from previous adversarial attack research targeting image classification DNNs Rusu et al. (2022). For RGB images from the ImageNet and HAM10000 datasets, we set $\epsilon = 8/255$, whereas for greyscale images from the Br35h and COVID-QU-Ex datasets, $\epsilon = 6/225$ Dong et al. (2025). Consistent with prior studies focused on efficient adversarial attacks on DNN classifiers, we set $K = 5000$ Williams & Li (2023a;b) for all attacks. As elaborated in Section 3, our approach involves three adjustable parameters: σ, N, and Max Diameter. Here, σ governs the exploration of the method, N signifies the number of shapes used to create the perturbation, and Max Diameter determines the largest possible size of the perturbation circles. The specific values for these parameters are listed in Table 8, with justification provided in Section 4.5.

Performance Metrics: For evaluating the effectiveness of the proposed method, we adopt the Pearson Correlation Coefficient (PCC) as a measure of the distortion caused to the attribution maps, following previous research Huang et al. (2023). PCC values near 1 indicate strong positive correlation, values close to 0 imply no correlation, and values approaching -1 suggest strong negative correlation. Additionally, we report the percentage of generated adversarial images that meet the constraints specified in (1) and (2). To ensure fair comparison, distortion measurements are only conducted on images that satisfy their respective

Table 1: Table presents the Pearson Correlation Coefficient (PCC) along with the percentage of images that satisfy the respective task constraint when attacking images from the HAM10000, Br35h and COVID-QU-Ex datasets. We provide the mean and variance of each metric over 10 runs.

DeepLIFT	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**80.4%(15.528)**†	**0.7(0.034)**†	**75.81%(4.88)**†	**0.04(0.138)**†
NES	32.44%(8.265)‡	0.19(0.054)‡	55.9%(16.028)‡	0.34(0.186)‡
SAFARI	40.11%(12.184)‡	0.38(0.141)‡	64.37%(8.04)‡	0.49(0.207)‡
Saliency	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**81.07%(14.756)**†	**0.6(0.096)**†	**83.1%(3.319)**†	**−0.0(0.11)**†
NES	34.11%(8.108)‡	0.16(0.128)‡	54.9%(17.283)‡	0.27(0.227)‡
SAFARI	36.8%(14.534)‡	0.31(0.173)‡	63.83%(6.72)‡	0.35(0.25)‡
Grad-CAM	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**82.4%(13.258)**†	**0.9(0.021)**†	**87.05%(8.31)**†	**−0.55(0.338)**†
NES	35.11%(8.338)‡	0.19(0.033)‡	54.23%(18.132)‡	0.39(0.217)‡
SAFARI	39.13%(16.982)‡	0.31(0.087)‡	65.7%(6.236)‡	0.49(0.228)‡
Shapley	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**85.07%(10.771)**†	**0.71(0.033)**†	**77.14%(3.037)**†	**0.08(0.134)**†
NES	35.11%(8.338)‡	0.2(0.106)‡	56.23%(15.616)‡	0.37(0.216)‡
SAFARI	37.47%(15.201)‡	0.41(0.126)‡	60.17%(8.287)‡	0.49(0.214)‡
Input x Gradient	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**81.08%(13.627)**†	**0.58(0.074)**†	**75.47%(5.346)**†	**0.08(0.133)**†
NES	35.44%(8.466)‡	0.14(0.082)‡	56.56%(15.207)‡	0.27(0.2)‡
SAFARI	39.77%(11.836)‡	0.38(0.118)‡	63.2%(5.467)‡	0.4(0.221)‡
Grad-CAM++	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**83.07%(11.757)**†	**0.88(0.111)**†	**83.05%(13.764)**†	**−0.44(0.436)**†
NES	35.77%(8.618)‡	0.19(0.18)‡	55.56%(16.444)‡	0.38(0.248)‡
SAFARI	40.77%(12.631)‡	0.56(0.209)‡	63.76%(4.64)‡	0.53(0.283)‡
LIME	Task 1		Task 2	
Method	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
EvoAttack	**72.69%(12.176)**†	**0.89(0.039)**†	**82.61%(10.283)**†	**−0.54(0.318)**†
NES	39.33%(8.225)‡	0.17(0.026)‡	60.84%(17.84)‡	0.36(0.212)‡
SAFARI	39.66%(16.296)‡	0.29(0.096)‡	66.12%(12.988)‡	0.51(0.245)‡

† denotes the performance of the method significantly outperforms the compared methods according to the Wilcoxon signed-rank test Wilcoxon (1992) at the 5% significance level; ‡ denotes the corresponding method is significantly outperformed by the best performing method (shaded).

task constraints. For ranking the robustness of XAI methods, we utilize the dominance relations described in Definitions 1 and 2 to evaluate and rank the considered XAI methods.

Given the stochastic nature of our method, each experiment is repeated over 10 different random seeds. For each metric, we aggregate its value across all model architectures. For each metric, we combine the results across model architectures and report its mean and variance. To statistically verify whether the improvements achieved by our method are significant relative to other algorithms, we employ the Wilcoxon signed-rank test Wilcoxon (1992) at a 5% significance level, as is standard practice within the Evolutionary Optimization field Williams & Li (2023a); Storn & Price (1997); Deb (2001).

4.2 Result Analysis

To ensure a fair comparison between the proposed EvoAttack, SAFARI, and NES methods, we set the population size for both SAFARI and NES to 50. This configuration allows for 100 iterations of their respective optimization cycles.

	HAM10000		Br35h		COVID-QU-Ex	
XAI Method	Task 1	Task 2	Task 1	Task 2	Task 1	Task 2
DeepLIFT	4.19 (0.982)	4.15 (1.269)	4.02 (1.535)	4.9 (1.368)	4.04 (1.043)	**2.71(0.825)**
Grad-CAM	5.48 (1.095)	5.88 (1.649)	4.14 (1.529)	5.84 (1.839)	6.2 (0.965)	5.71 (1.285)
SHAPLEY	4.87 (0.875)	3.12 (0.968)	4.69 (1.467)	5.0 (1.535)	4.31 (1.012)	3.04 (0.943)
Input x Gradient	**2.43(0.86)**	**2.82(0.948)**	**3.93(1.437)**	**4.15(1.637)**	3.34 (0.947)	2.9 (1.02)
Grad-CAM++	6.17 (1.516)	5.49 (1.669)	4.42 (1.373)	5.16 (2.013)	6.52 (0.726)	5.89 (1.489)
Saliency	3.38 (0.963)	2.93 (1.037)	4.51 (1.34)	4.95 (1.367)	**2.54(0.79)**	4.35 (1.262)

Table 2: Table presents the robustness ranking of XAI methods across each task and dataset using the proposed EvoAttack as the attack method. We provide the mean and variance of each metric over 10 runs.

Constraint satisfaction: The statistical outcomes for the proposed method and comparison attacks are presented in Table 1, with arrows indicating whether better metric values are larger or smaller. The constraint in Task 1, is defined as the misclassification of the image, whereas the constraint in Task 2 is defined as the correct classification of the image. For the EvoAttack and SAFARI methods, the constraint satisfaction rates of below 100% indicate that even random perturbations cause the underlying DNN to misclassify, with the attacks unable to generate perturbations leading to correct predictions. For NES, the weights used for summing the to objectives also impacts its ability to satisfy the constraint.

The proposed EvoAttack method significantly outperforms both SAFARI Huang et al. (2023) and NES Tamam et al. (2023) across all XAI methods and datasets. As discussed in Section 1, existing methods often rely on the assumption of ample computational query budgets; constraining the number of queries leads to considerable performance degradation. Specifically, both NES and SAFARI struggle to fulfill the task 1 constraint of inducing adversarial classifications across all datasets. This aligns with previous findings that highlight the difficulty in generating adversarial images, specifically when the number of queries is limited. Conversely, the EvoAttack method shows greater success in generating adversarial images. When evaluating performance related to the task 2 condition, the differences are less pronounced, particularly when compared to SAFARI. With the task 2 constraint being satisfied when added perturbations do not cause misclassification, this result can be attributed to the robustness of the underlying DNN classifiers against pixel-level perturbations.

When analysing the constraint satisfaction rates across datasets (see Tables 10, 12 and 11 within the Appendix) all attack methods exhibit performance degradation on the Br35h dataset. Despite clear regions of interest in images from all datasets, the diagnostic information in Br35h images is more explicit (presence of a tumor within the image). Consequently, the classifiers may have learned robust features such as tumor presence rather than non-robust spurious correlations, also known as shortcut learning Wang et al. (2024).

Finally, we observe only minor discrepancies in constraint satisfaction rates across XAI methods. This uniformity is expected for EvoAttack and SAFARI as both prioritize task constraints before focusing on explanation distortion.

Explanation distortion: In assessing the distortions in explanations caused by the attack methods, EvoAttack consistently outperforms both SAFARI and NES across all experimental scenarios, demonstrating significant superiority in the majority of cases (see Figure 7 for visuali comparison). Notably, there is a pronounced performance disparity when targeting XAI methods that generate less granular, more region-focused explanations, such as Grad-CAM, Grad-CAM++, and LIME. Conversely, when attacking more granular XAI methods like DeepLIFT, Saliency maps, SHAPLEY, and Input x Gradient, the performance gap between EvoAttack and the other attack methods narrows. This outcome is anticipated due to the less granular perturbation strategy employed by EvoAttack, emphasizing the importance of accounting for the granularity of XAI explanations when designing perturbations. However, it also underscores the adaptability of the proposed method to effectively target both granular and non-granular XAI methods.

Comparing task performances, the attack methods generally achieve greater success in maximizing the distortion of explanations for correctly classified images rather than minimizing distortion for misclassified ones. This indicates that XAI-generated explanations are more susceptible to task 2 attacks, where the classification remains accurate, but the explanation is distorted. Although this scenario poses less risk to

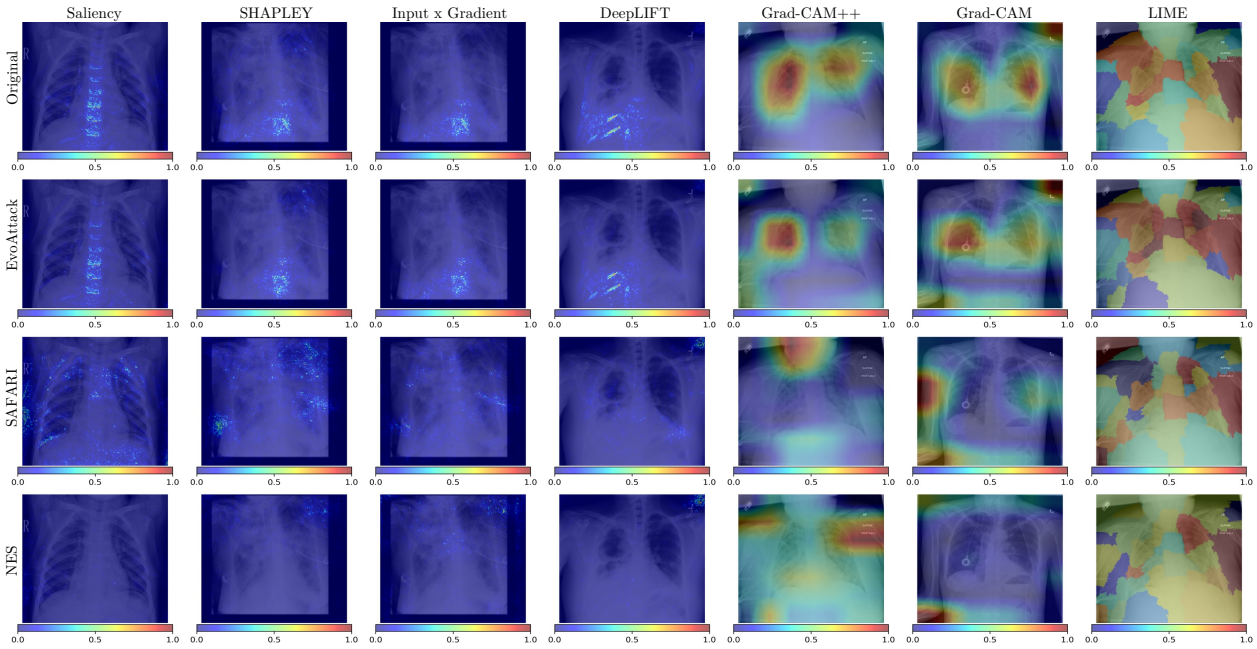

Figure 7: Original COVID-QU-Ex and adversarial images constructed by the proposed EvoAttack, SAFARI and NES attacks, along with attribution maps generated by the respective XAI method. Adversarial images are generated by attacks deployed within the Task 1 scenario. For the majority of the images, the proposed method is able cause larger distortions to the original explanation, compared to explanations on SAFARI and NES generated adversarial images.

patient safety due to correct disease classification, it could undermine trust in AI systems among medical practitioners, potentially reducing their willingness to rely on AI for assistance Rosenbacke et al. (2024a).

Additionally, performances vary across different XAI methods among the attack strategies. EvoAttack notably manipulates less granular explanations, such as Grad-CAM, Grad-CAM++, and LIME, more effectively in all experimental setups compared to the granular methods, underscoring their specific vulnerabilities to structured perturbations. Among granular XAI methods, the greatest challenge appears in attacking Saliency maps and Input x Gradient, particularly for task 1. For example, when targeting images from the HAM10000 and COVID-QU-Ex datasets, no attack achieves an average PCC value above 0.6, which is a benchmark for consistency between explanations. Conversely, for task 2, EvoAttack is able to reduce the average PCC of those XAI values below 0.4, indicating inconsistent explanations Huang et al. (2023), highlighting the greater vulnerability of XAI methods when distorting correctly classified images.

More visual comparisons of adversarial images and explanations is provided in Section A.5 in the Appendix.

4.3 XAI Robustness Comparison

The findings in Section 4.2 highlight the superior performance of the proposed method over existing attack techniques, showcasing its utility for robustness evaluations. However, the combination of the task's objective and constraint (as detailed in Section 3) makes it difficult to use a single metric value for ranking. Therefore, to rank the robustness of the targeted XAI methods, we employ the task-specific domination relations defined in 1 and 2.

For each attacked image, we utilize non-dominated sorting Deb et al. (2002) to rank each XAI method based on the performance of EvoAttack in targeting the image, where lower ranks correspond to poorer attack results, indicating greater robustness of the XAI method. This procedure is repeated across all 100 images for each model architecture. We repeat this procedure over the 10 different random seeds, with the average rank serving as a measure of the XAI method's overall performance.

Table 3: Table presents the Pearson Correlation Coefficient (PCC) along with the percentage of images that satisfy the respective task constraint when attacking XAI method applied to adversarial trained classifiers. We provide the mean and variance of each metric over 10 runs.

DeepLIFT	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	98.06 (2.04)	0.71 (0.118)	97.01 (2.544)	0.08 (0.162)
Br35h	56.00 (2.376)	0.66 (0.148)	56.99 (2.182)	0.56 (0.158)
COVID-QU-Ex	98.01 (2.577)	0.57 (0.154)	61.08 (2.508)	-0.0 (0.235)
Saliency	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	98.07 (2.214)	0.54 (0.112)	98.91 (2.782)	0.17 (0.16)
Br35h	59.29 (2.816)	0.65 (0.124)	59.0 (2.218)	0.51 (0.146)
COVID-QU-Ex	98.95 (2.46)	0.46 (0.11)	67.01 (2.86)	-0.1 (0.083)
Grad-CAM	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	98.07 (2.034)	0.77 (0.238)	99.94 (2.531)	-0.36 (0.391)
Br35h	40.0 (2.46)	0.60 (0.209)	48.0 (2.334)	0.68 (0.365)
COVID-QU-Ex	98.92 (2.31)	0.78 (0.268)	75.06 (2.367)	-0.73 (0.481)
Shapley	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	97.9 (2.185)	0.76 (0.11)	96.91 (2.139)	0.13 (0.176)
Br35h	41.0 (2.687)	0.63 (0.154)	42.0 (2.385)	0.56 (0.139)
COVID-QU-Ex	98.96 (2.866)	0.59 (0.151)	60.93 (2.174)	-0.0 (0.223)
Input x Gradient	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	97.94 (2.476)	0.48 (0.123)	96.99 (2.005)	0.12 (0.147)
Br35h	52.03 (2.82)	0.67 (0.131)	41.0 (2.913)	0.56 (0.118)
COVID-QU-Ex	97.95 (2.708)	0.52 (0.1)	60.9 (2.6)	0.01 (0.087)
Grad-CAM++	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	98.02 (2.541)	0.86 (0.154)	99.9 (2.416)	-0.33 (0.401)
Br35h	46.0 (2.105)	0.49 (0.259)	46.0 (2.344)	0.69 (0.34)
COVID-QU-Ex	97.92 (2.327)	0.87 (0.154)	75.06 (2.933)	-0.73 (0.482)
LIME	Task 1		Task 2	
Dataset	Constraint Satisfied (↑)	PCC (↑)	Constraint Satisfied (↑)	PCC (↓)
HAM10000	99.31 (1.409)	0.77 (0.101)	99.4 (0.18)	-0.41 (0.392)
Br35h	38.11 (1.201)	0.49 (0.259)	46.0 (2.344)	0.69 (0.34)
COVID-QU-Ex	97.92 (2.327)	0.87 (0.154)	75.06 (2.933)	-0.73 (0.482)

The final robustness ranking of XAI methods, as presented in Table 2, indicates that Input x Gradient frequently achieves better average rankings across most attack instances. Despite the Saliency method attaining better average PCC values, it achieves higher overall rankings on 2 out of 3 datasets. This discrepancy arises because the metric values in Section 4.2 focused solely on attack instances that satisfied the constraints, whereas this ranking accounts for all evaluated images. For example, during Task 1 attacks on images from the HAM10000 dataset (see Table 11 within the Appendix), the Saliency method shows greater resistance to attacks, resulting in a higher average PCC value. By employing the domination relation, the ranking considers the attack's ability to satisfy the constraint, demonstrating that EvoAttack achieves lower constraint satisfaction when targeting Input x Gradient compared to Saliency, leading to Input x Gradient's superior average rank. Finally, in Task 1 attacks on COVID-QU-Ex images, the Saliency method demonstrates superior robustness, while DeepLIFT proves most resilient against Task 2 attacks.

4.4 Evaluation of Adversarial Training

The results in Section 4.2 demonstrated the effectiveness of the proposed attack method in successfully compromising XAI methods, outperforming existing attacks. Given that adversarial training has been recognized as a promising approach to mitigating the vulnerabilities posed by adversarial attacks, this section evaluates its potential for enhancing the robustness of XAI methods against the EvoAttack.

Adversarial Training Setup: In this study, we implement an adversarial training procedure similar to those used in prior research, which incorporates image perturbations during the training phase Madry et al. (2017); Chernyak et al. (2021). Specifically, each training iteration involves augmenting all images with random EvoAttack perturbations, resulting in batches composed of both benign and adversarial images, thereby doubling the batch size. To ensure the DNN is exposed to various forms of the EvoAttack perturbation structures, we randomly sample parameters (N) and Maximum Diameter for each perturbation from the grid used within our ablation study. A detailed description of the employed adversarial training process along with the performance metrics of the adversarially trained classifiers is provided in Section A.3 in the Appendix.

Task 1: The results presented in Table 3 illustrate the impact of adversarial training on different XAI methods. First, we see that adversarially trained HAM10000 DNN classifiers become more vulnerable to adversarial attacks. This phenomenon is likely attributed to the large class imbalance in the HAM10000 dataset, which has shown to be an issue for adversarial training Wang et al. (2022). Comparing with Br35h and COVID-QU-Ex classifiers, we see the use of adversarial training improved their robustness against the proposed EvoAttack.

Comparing the performance of the different XAI methods' we see the overall use of adversarial training has improved the robustness across all XAI methods. In particular, we see greater jumps in robustness for the less granular XAI methods, Grad-CAM, Grad-CAM++ and LIME. Despite these results demonstrating the potential of adversarial training, these methods are still the least vulnerable across the different XAI methods. Comparing the the more granular XAI methods, the improved robustness of Saliency maps and Input x Gradient XAI methods brings their respective PCC values near or below 0.6 across all three datasets. This indicates that, although adversarial explanations are not perfectly inconsistent with their benign counterparts, they remain close to or below the threshold of consistency. As a result, the distortions in the attribution maps could become noticeable.

Task 2: Similar to Task 1, we witness an enhancement in robustness across all XAI methods which can be described by the increased average PCC values. In the case of Br35h images, EvoAttack's ability to meet the constraint is diminished, reflecting improved robustness across XAI methods, with all average PCC values exceeding 0.5. This suggests that EvoAttack faces challenges in altering the original explanation to be inconsistent while ensuring the classifier predicts accurately.

Similar observations arise with the COVID-QU-Ex dataset, where EvoAttack's success in meeting the constraint is reduced, while average PCC values for most XAI methods increase. This indicates that adversarial training has degraded EvoAttack's ability to distort explanations while retaining accurate classifier predictions.

Despite the impact of adversarial training in improving the robustness across XAI methods, EvoAttack is still able to distort attribution maps to PCC values of below 0.4, which indicates that inconsistency was achieved.

4.5 Ablation Study

The proposed method incorporates three tunable parameters: (N), (σ), and the maximum circle diameter (Max Diameter) expressed as a percentage of the image. We employ a grid search over the parameter space to determine their optimal values. Specifically, we explore $(N \in 100, 300)$, $(\sigma \in 0.1, 0.2, 0.3)$, and maximum circle diameters ranging from $(20\%, 30\%, 40\%, 50\%)$ of the original image size. These parameter ranges are based on commonly set values used in the evolutionary Skiscim & Golden (1983) and computational art Tian & Ha (2022) fields. To evaluate the performance of each parameter configuration, we conduct attacks on each explanation for each task using a VGG-16 ImageNet classifier with 100 correctly classified images from the validation set. To compare the performance of different parameter configurations, we employ the same methodology as described in Section 4.3.

Configuration Performance Analysis: As illustrated in Table 8, the optimal configuration varies significantly across XAI methods and tasks. Similar to previous studies that highlight the lack of consensus among XAI methods, this suggests that the vulnerabilities between them may differ. Nonetheless, some patterns emerge across different XAI methods. Firstly, for Task 1, which aims to induce misclassification

	Task 1			Task 2		
XAI Method	N	Max Diameter (%)	σ	N	Max Diameter (%)	σ
DeepLIFT	100	80.0	0.1	300	60.0	0.3
Saliency	100	80.0	0.3	200	70.0	0.2
Grad-CAM	100	80.0	0.3	100	80.0	0.2
LIME	100	80.0	0.3	100	80.0	0.2
Shapley	100	80.0	0.3	300	60.0	0.3
Input x Gradient	100	80.0	0.1	300	60.0	0.3
Grad-CAM++	100	80.0	0.1	100	80.0	0.3

Figure 8: Chosen EvoAttack parameters for attacking the considered XAI methods.

while preserving the explanation, the best-performing configurations consistently feature a maximum circle diameter of 80% of the image size. Conversely, configurations with smaller circle diameters perform worse. This indicates that utilizing larger local perturbations (i.e., larger circles) is more effective in influencing the DNN classifier while minimizing distortion in the attribution map.

Another observed pattern is that all optimal configurations use ($N = 100$) circles. A likely explanation for this is that adding more shapes increases the number of variables involved, which might demand a larger computational budget for effective optimization Williams et al. (2021); Eltaeib & Mahmood (2018). Keeping a constant budget might lead to prematurely halting the optimization process, thus affecting performance negatively.

For Task 2, which aims to distort the explanation while maintaining correct classification, we observe greater variation in the best performing configurations. However, similar configuration performance are seen across different XAI methods. Specifically, the results indicate that XAI methods that produce similar granularity in attribution maps are effectively attacked with comparable parameter setups. For example, when targeting highly granular attribution maps from DeepLIFT, Input x Gradient, and SHAPLEY (see Figure 2), the proposed method achieves superior performance using smaller diameter circles. Conversely, XAI methods like Grad-CAM, Grad-CAM++, and SHAPLEY, which highlight broader regions rather than individual pixels, are more susceptible to larger circular perturbations. This behaviour stems from altering the perturbations' granularity by changing circle size, with smaller circles constructing perturbations more closely resembling pixel-level perturbations.

A surprising result from the ablation study was the performance of parameter configurations when targeting the Saliency XAI method. Unlike other granular methods, the Saliency method proved more vulnerable to medium to large circle perturbations, while remaining robust against smaller circles. This difference might be attributed to its level of granularity. Whereas SHAPLEY and DeepLIFT produce sparse maps emphasizing specific pixels, Saliency maps highlight broader regions. Although the Input x Gradient method also emphasizes broader areas, its multiplication with the input image may also have an impact, requiring smaller circle diameters.

These results underscore the advantages of the proposed attack method, EvoAttack. By adjusting the size of the circular shapes, EvoAttack effectively manages the trade-off related to granularity when targeting explanation methods—an aspect that existing strategies lack. In conclusion, we recommend adopting the optimal configurations detailed in Table 8 for the XAI methods considered in this study. We provide the performance across all configurations in Figure 9 within the Appendix. .

5 Conclusion, Limitations and Future Work

Conclusion: This research introduces a novel adversarial attack specifically designed to target XAI methods in computer vision. Unlike most existing approaches that modify each pixel of the benign image, our method constructs adversarial perturbations by concatenating RGB-valued circular shapes. We optimize the parameters of these shapes using a (1+1)-evolutionary strategy, a widely used optimization heuristic in evolutionary computation. To enhance the attack's efficacy, we conducted an ablation study assessing the influence of various parameters on performance across several XAI methods. The results demonstrate that larger circles effectively manipulate less granular XAI methods like Grad-CAM, Grad-CAM++, and Lime, while smaller circles yield better results against granular XAI methods such as DeepLIFT and Input x

Gradient. Compared to state-of-the-art attack techniques, the proposed method consistently outperforms them in all attack setups, showcasing its efficiency and effectiveness.

We leveraged the EvoAttack method to evaluate and rank the robustness of XAI methods. Given the complexity of assessing XAI methods using both PCC and constraint metrics, we employed the EvoAttack domination relation for each task to rank the resistance of XAI methods with respect to each attacked image. By averaging the rank of each XAI method across all attacked images and underlying classifiers, we formulated a comprehensive ranking. This approach allowed us to incorporate all data regarding XAI distortion alongside constraint satisfaction, culminating in a unified ranking table for each dataset.

To counter the proposed attack, we developed an adversarial training procedure that incorporates random EvoAttack-style perturbations into the training process. Attacking adversarially trained models revealed enhanced robustness in most XAI methods; however, we noted a decrease in classifier robustness on the HAM10000 dataset, accompanied by improved XAI robustness. In the Task 2 attack scenario, adversarially trained classifiers decreased EvoAttack's effectiveness in redirecting the DNN classifier towards correct predictions, suggesting that while the classifier became more susceptible to random EvoAttack perturbations, it also developed a stronger resistance to subsequent manipulations. This study underscores the potential of adversarial training as a crucial strategy for defending against attacks, in addition to the utility of considering both classifier and XAI performance when evaluating the robustness of human-in-the-loop systems. Nevertheless, our study highlights the need for further exploration, marking this as a vital area for future research.

Limitations and Future Work: This study focuses on enhancing the robustness of XAI techniques, with an emphasis on medical imaging datasets. While the proposed method demonstrates promising results, there are limitations and avenues for future research. Firstly, the hyper-parameters were optimized using a basic grid-search approach, aimed at analysing the impact of varying parameter values. In future work, more advanced hyper-parameter optimization frameworks, such as Bayesian Optimization or methods assisted by large-language models, should be explored. Snoek et al. (2012); Zhang et al. (2023).

Additionally, this study employed a predefined l_∞ constraint ϵ based on existing recommendations. Exploring minimum-norm attacks, which identify the minimal ϵ value necessary to compromise an AI system, could provide deeper insights into the XAI and classifier vulnerabilities Williams & Li (2023a). Future work should investigate these types of attacks to better understand the weaknesses of current approaches. Moreover, while this research examined perturbations that modify all pixels, future studies should consider alternative perturbations, such as sparse attacks, where only a limited number of pixels are altered.

Our experiments highlight the potential of adversarial training in reducing the impact of EvoAttack on classifier and XAI robustness. However, they also reveal EvoAttack's capability to manipulate explanations to remain consistent or inconsistent, depending on the task, while satisfying respective constraints. Future research should focus on developing more sophisticated adversarial training methodologies to bolster the robustness of both DNN classifiers and XAI methods. For example, rather than using random perturbations during adversarial training, executing the proposed attack with a limited number of queries could pinpoint more vulnerable regions of images, thereby exposing the model to more harmful perturbations and potentially improving its robustness. Alternatively, defence strategies like adding random noise to output probabilities have been suggested to reduce the efficacy of malicious attacks. We anticipate that methods like these will be employed as overall DNN robustness is enhanced.

To foster continued research in this domain, we plan to publicly release our implementation, datasets, and evaluation scripts upon acceptance of the paper.

Broader Impact Statement

In this work, we introduce a novel adversarial attack method against explainable AI (XAI) techniques that can account of the varying granularities in explanations, as well as reducing the dimension of the search space. We apply our proposed attack to the robustness ranking of various XAI methods across three different medical image datasets. This study underscores the necessity of evaluating both classifier and XAI system robustness.

With the healthcare domain, our research demonstrates two significant risks: first, the possibility of medical professionals trusting incorrect AI diagnoses due to seemingly plausible explanations, potentially endangering patient safety; and second, the risk of decreasing trust between healthcare practitioners and AI systems due to distorted explanations of accurate diagnoses, which could slow down the diagnostic process by necessitating additional human evaluations. To address these risks, we explore different mitigation strategies, such as adversarial training, which showed promise in enhancing the resilience of XAI systems against adversarial threats.

We hope our work will lead to further research into adversarial training strategies and encourage practitioners to rigorously test the robustness of XAI systems before deployment. Our ultimate goal is to advance the safe and effective integration of AI in critical domains like healthcare.

References

David Alvarez-Melis and Tommi S. Jaakkola. On the robustness of interpretability methods. *CoRR*, 2018.

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: A query-efficient black-box adversarial attack via random search. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*. Springer, 2020.

Hubert Baniecki and Przemyslaw Biecek. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Inf. Fusion*, 2024.

Moritz Böhle, Navdeeppal Singh, Mario Fritz, and Bernt Schiele. B-cos alignment for inherently interpretable cnns and vision transformers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable AI techniques in healthcare. *Sensors*, 23(2):634, 2023.

Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, 2017.

Jinyin Chen, Mengmeng Su, Shijing Shen, Hui Xiong, and Haibin Zheng. POBA-GA: perturbation optimized black-box adversarial attacks via genetic algorithm. *Comput. Secur.*, 2019.

Bronya Roni Chernyak, Bhiksha Raj, Tamir Hazan, and Joseph Keshet. Constant random perturbations provide adversarial robustness with minimal effect on accuracy. *CoRR*, 2021.

Carlos A. Coello Coello and Efrén Mezura-Montes. Handling constraints in genetic algorithms using dominance-based tournaments. In *Adaptive Computing in Design and Manufacture V*, 2002.

Kalyanmoy Deb. *Multi-objective optimization using evolutionary algorithms*. Wiley-Interscience series in systems and optimization. Wiley, 2001.

Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.*, 2002.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*.

Dharmesh Dhabliya, Swati Saxena, Jambi Ratna Raja Kumar, Dinesh Kumar Pandey, NV Balaji, and X Mercilin Raajini. Exposing the financial impact of ai-driven data analytics: A cost-benefit analysis. In *World Conference on Communication & Computing (WCONF)*, 2024.

Ann-Kathrin Dombrowski, Maximilian Alber, Christopher J. Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems 32*, 2019.

Junhao Dong, Junxi Chen, Xiaohua Xie, Jianhuang Lai, and Hao Chen. Survey on adversarial attack and defense for medical image analysis: Methods and challenges methods and challenges. *ACM Comput. Surv.*, 2025.

Finale Doshi-Velez, Mason Kortz, Ryan Budish, Chris Bavitz, Sam Gershman, David O'Brien, Stuart Schieber, James Waldo, David Weinberger, and Alexandra Wood. Accountability of AI under the law: The role of explanation. *CoRR*, 2017.

Filip Karlo Dosilovic, Mario Brcic, and Nikica Hlupic. Explainable artificial intelligence: A survey. In *41st International Convention on Information and Communication Technology, Electronics and Microelectronics, MIPRO 2018, Opatija, Croatia, May 21-25, 2018*. IEEE, 2018.

Tarik Eltaeib and Ausif Mahmood. Large-scale evolutionary optimization using multi-layer differential evolution. In *2018 9th IEEE Annual Ubiquitous Computing, Electronics Mobile Communication Conference (UEMCON)*, 2018.

Julia Garbaruk, Doina Logofatu, Costin Badica, and Florin Leon. Digital image evolution of artwork without human evaluation using the example of the evolving mona lisa problem. *Vietnam Journal of Computer Science*, 2022.

Amirata Ghorbani, Abubakar Abid, and James Y. Zou. Interpretation of neural networks is fragile. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI*, 2019.

Ahmed Hamada. Br35H :: Brain Tumor Detection 2020. `https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection`, 2020. Accessed: 2024-10-15.

Jinkui Hao, William R Kwapong, Ting Shen, Huazhu Fu, Yanwu Xu, Qinkang Lu, Shouyue Liu, Jiong Zhang, Yonghuai Liu, Yifan Zhao, et al. Early detection of dementia through retinal imaging and trustworthy ai. *npj Digital Medicine*, 2024.

Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. In *Advances in Neural Information Processing Systems 32*, 2019.

Andrew Howard, Ruoming Pang, Hartwig Adam, Quoc V. Le, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, and Yukun Zhu. Searching for mobilenetv3. In *International Conference on Computer Vision, ICCV*, 2019.

Wei Huang, Xingyu Zhao, Gaojie Jin, and Xiaowei Huang. SAFARI: versatile and efficient evaluations for robustness of interpretability. In *IEEE/CVF International Conference on Computer Vision, ICCV*, 2023.

Zhichao Huang and Tong Zhang. Black-box adversarial attack with transferable model-based embedding. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Proceedings of Machine Learning Research. PMLR, 2018.

Fekhr Eddine Keddous, Nadiya Shvai, Arcadi Llanza, and Amir Nakib. Inference acceleration of deep learning classifiers based on RNN. In *IEEE International Conference on Image Processing, ICIP*. IEEE, 2023.

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer, 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

Aditya Kuppa and Nhien-An Le-Khac. Black box attacks on explainable artificial intelligence(xai) methods in cyber security. In *nternational Joint Conference on Neural Networks, IJCNN*. IEEE, 2020.

Nicholas Lambert, William H. Latham, and Frederic Fol Leymarie. The emergence and growth of evolutionary art: 1980-1993. In *International Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2013, Anaheim, CA, USA, July 21-25, 2013, Art Gallery*, 2013.

Nan Li, Lianbo Ma, Guo Yu, Bing Xue, Mengjie Zhang, and Yaochu Jin. Survey on evolutionary deep learning: Principles, algorithms, applications, and open issues. *ACM Comput. Surv.*, 2024.

Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hemanth Nadipineni. Method to classify skin lesions using dermoscopic images. *CoRR*, 2020.

Daniel Omeiza, Helena Webb, Marina Jirotka, and Lars Kunze. Explanations in autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.*, 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. 2019.

Kewen Peng and Tim Menzies. Documenting evidence of a reuse of '"why should I trust you?": explaining the predictions of any classifier'. In *ESEC/FSE: 29th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, 2021.

Thomas P. Quinn, Manisha Senadeera, Stephan Jacobs, Simon Coghlan, and Vuong Le. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *J. Am. Medical Informatics Assoc.*, 28, 2021.

Nitin Rane, Saurabh Choudhary, and Jayesh Rane. Explainable artificial intelligence (xai) in healthcare: Interpretable models for clinical decision support. *SSRN*, 2023.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review. *JMIR AI*, 2024a.

Rikard Rosenbacke, Åsa Melhus, Martin McKee, and David Stuckler. How explainable artificial intelligence can increase or decrease clinicians' trust in ai applications in health care: Systematic review. *JMIR AI*, 2024b.

Andrei A. Rusu, Dan Andrei Calian, Sven Gowal, and Raia Hadsell. Hindering adversarial attacks with implicit neural representations. In *International Conference on Machine Learning, ICML*, 2022.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning, ICML*. PMLR, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations, ICLR*, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *2nd International Conference on Learning Representations*, 2014.

Christopher C. Skiscim and Bruce L. Golden. Optimization by simulated annealing: A preliminary computational study for the TSP. In *Proceedings of the 15th conference on Winter simulation, WSC 1983, Arlington, VA, USA, December 12-14, 1983*, 1983.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, 2012.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

Rainer Storn and Kenneth V. Price. Differential evolution - A simple and efficient heuristic for global optimization over continuous spaces. 1997.

Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV*, 2019.

Anas M. Tahir, Muhammad Enamul Hoque Chowdhury, Amith Khandakar, Tawsifur Rahman, Yazan Qiblawey, Uzair Khurshid, Serkan Kiranyaz, Nabil Ibtehaz, M. Sohel Rahman, Somaya Al-Máadeed, Sakib Mahmud, Maymouna Ezeddin, Khaled Hameed, and Tahir Hamid. COVID-19 infection localization and severity grading from chest x-ray images. *Comput. Biol. Medicine*, 2021.

Snir Vitrack Tamam, Raz Lapid, and Moshe Sipper. Foiling explanations in deep neural networks. *Trans. Mach. Learn. Res.*, 2023.

Yingtao Tian and David Ha. Modern evolution strategies for creativity: Fitting concrete images and abstract concepts. In *Artificial Intelligence in Music, Sound, Art and Design - 11th International Conference, EvoMUSART 2022, Held as Part of EvoStar 2022, Madrid, Spain, April 20-22, 2022, Proceedings*, Lecture Notes in Computer Science. Springer, 2022.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *CoRR*, 2018.

Bas H. M. van der Velden, Hugo J. Kuijf, Kenneth G. A. Gilhuijs, and Max A. Viergever. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Anal.*, 2022.

Jiakai Wang, Donghua Wang, Jin Hu, Siyang Wu, Tingsong Jiang, Wen Yao, Aishan Liu, and Xianglong Liu. Adversarial examples in the physical world: A survey. *CoRR*, 2023.

Shunxin Wang, Raymond Veldhuis, Christoph Brune, and Nicola Strisciuglio. A survey on the robustness of computer vision models against common corruptions, 2024.

Wentao Wang, Han Xu, Xiaorui Liu, Yaxin Li, Bhavani Thuraisingham, and Jiliang Tang. Imbalanced adversarial training with reweighting. In *IEEE International Conference on Data Mining, ICDM*. IEEE, 2022.

Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *J. Mach. Learn. Res.*, 2014.

Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*. Springer, 1992.

Phoenix Williams and Ke Li. Camopatch: An evolutionary strategy for generating camoflauged adversarial patches. In *Advances in Neural Information Processing Systems*, 2023a.

Phoenix Neale Williams and Ke Li. Black-box sparse adversarial attack via multi-objective optimisation CVPR proceedings. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, 2023b.

Phoenix Neale Williams, Ke Li, and Geyong Min. Large-scale evolutionary optimization via multi-task random grouping. In *2021 IEEE International Conference on Systems, Man, and Cybernetics, SMC*, 2021.

Phoenix Neale Williams, Ke Li, and Geyong Min. Sparse adversarial attack via bi-objective optimization. In *Evolutionary Multi-Criterion Optimization - 12th International Conference, EMO*, 2023.

Jing Wu, Mingyi Zhou, Ce Zhu, Yipeng Liu, Mehrtash Harandi, and Li Li. Performance evaluation of adversarial attacks: Discrepancies and solutions. *CoRR*, 2021.

Michael R. Zhang, Nishkrit Desai, Juhan Bae, Jonathan Lorraine, and Jimmy Ba. Using large language models for hyperparameter optimization. *CoRR*, 2023.

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