THE DISAGREEMENT PROBLEM IN XAI ON IMAGE DATA VIA SALIENCY MAPS

1st Nam D. Mai
Faculty of Information Technology
University of Science
Ho Chi Minh City, Vietnam
mdnam2410@gmail.com

2nd Binh H. Nguyen
Faculty of Information Technology
University of Science
Ho Chi Minh City, Vietnam
huubinh0409@gmail.com

3rd Bac H. Le
dept. name of organization (of Aff.)
name of organization (of Aff.)
Ho Chi Minh City, Vietnam
email address or ORCID

***Abstract*—Deep learning models have become increasingly crucial in real-world applications like healthcare and security due to their growing complexity and accuracy. Understanding these complex "black box" models is essential for their appropriate use, creation, and trust. Explainable Artificial Intelligence (XAI) aims to interpret these systems through methods like saliency maps for convolutional neural networks. However, a concern has arisen regarding the "disagreement problem", where explanation methods yield inconsistent and contradictory explanations. This thesis addresses this gap by measuring the extent of disagreement among popular saliency techniques using set-based metrics. Our evaluation of eight different saliency methods over two black boxes revealed significant disagreement among saliency map explanations, with complex patterns varying based on the explanation method and the specific model being explained. Furthermore, we identified limitations in current measurement metrics and encourage researchers to develop new, robust metrics that effectively capture various aspects of explanation disagreement.**

***Index Terms*—XAI, disagreement, saliency**

I. Introduction

In recent years, we have witnessed the rising popularity and effectiveness of deep learning models across diverse real-world applications. The power of these deep learning models has been amplified as the massive growth in computational resources allows models to become increasingly larger and more complex in dimensions, as evidence in the growing number of parameters and depth. Moreover, substantial investment in research has encouraged the development of more efficient architectures but has also made them more complex for users.

In some high-stake domains, artificial intelligence systems play a pivotal role in making consequential decisions that can significantly impact human lives, such as in healthcare, security, and automation. However, as deep learning models become more and more complex, we are at risk of being dependent on systems that we do not understand. Delegating the work to the models without seriously questioning the reasons for their decisions can be dangerous. For example, in the medical domain, the COVID-19 pandemic has motivated researchers to address its related challenges and contribute to disease prevention. Several studies have employed deep learning models to tackle computer vision problems such as anomaly detection in endoscopic, X-ray images, etc. With the development of XAI, researchers no longer solely focus on measuring the model performance by metrics, but also its interpretability

eXplainable Artificial Intelligence (XAI) is an active field that aims to tackle the issue of making complex models more interpretable to users. With the availability of interpretable models, trust and awareness can be gained, especially for high-stake applications.

In the literature, models that are complex and difficult to interpret are commonly known as "black boxes". There are various approaches to making a black box more understandable, such as creating an alternative simpler model replicating the behavior of the complex model. Another common method is to provide explanations for individual decisions made by the black box. In computer vision, where images are the input to the black boxes, these explanations are often provided in the form of "saliency maps", which assign scores to regions or pixels of the input images, indicating their influence on the model's output. Saliency maps help users to comprehend which parts of an image the model's prediction is based on, making it easier to identify weaknesses and limitations of the model, and thereby improving or creating new versions.

There are numerous methods available for providing explanations for the outcome of a black box, but there is a challenge in assessing the quality of these explanations. Without a standard evaluation framework, there is a risk of using explanations that are baseless, ambiguous, or even contradictory. This is known as the disagreement problem and has recently been identified as a concern. Studies have investigated the severity of this problem for various types of explanation methods, but to the best of our knowledge, there seems to be a scarcity of studies specifically examining saliency maps. Therefore, the purpose of this thesis is to investigate the disagreement problem among saliency methods. By doing so, we aim to shed light on various aspects of this problem and contribute to addressing it in the future.

In order to investigate the disagreement among saliency maps, we follow the current methodology utilized in existing literature. Initially, we train two distinct CNN models on the most two popular datasets to serve as our designated target black boxes. Next, we opt for eight various saliency techniques and assess them over the aforementioned black boxes in order to obtain explanations. Subsequently, we measure the level of

disagreement between the saliency methods for each black box by utilizing several metrics.

Our analysis findings suggest that there is a substantial level of inconsistency among the explanation methods. We have also confirmed that this phenomenon persists across different types of black boxes, although there are intricate and diverse patterns of disagreement. We have concluded that the degree of disagreement is influenced not only by the type of explanation methods but also by the specific type of black box being explained. Additionally, we have noted a limitation in the sign agreement metric, which is approximately half of the feature agreement score. This may make it superfluous in measuring disagreement, and we therefore encourage researchers to develop advanced metrics that can capture a broader range of aspects of inconsistency.

Overall, the contributions of our thesis include:

- Highlighting the existence of disagreement for saliency maps methods
- Quantifying the degree of disagreement for eight saliency methods over two black boxes
- Highlighting the variation of disagreement with respect to different kinds of black boxes

II. Background and Related Work

A. Overview of Explainable Artificial Intelligence

eXplainable Artificial Intelligence (XAI) is not a new concept that has emerged recently. The earliest research on XAI can be traced back to literature published four decades ago [23, 27], when certain expert systems were designed to provide explanations for their outputs based on the rules they applied. The advent of XAI can be traced back to the early 2000s when Lipton raised a renewed interest in the field, emphasizing the importance of explainability in machine learning algorithms by his influence paper in 2016 [16]. Since then, the number of publications in the XAI domain has witnessed exponential growth, demonstrating the increasing attention and significance given to this area of research. A bibliometric analysis conducted by Park and Lehman revealed a staggering 350% rise in the number of XAI-related papers published between 2017 and 2020 [2].

The significant growth of XAI leads to a notable diversity of approaches and techniques. This diversity is a result of researchers' varying perspectives and goals in their pursuit of improving the interpretability of AI systems. Various taxonomies have been suggested in the research literature to categorize different explainability methods [6, 25]. However, it is important to note that these classification techniques are not fixed or absolute. One notable variation is the distinction between interpretable-by-design and post-hoc XAI methods. Interpretable-by-design XAI methods involve incorporating interpretability during the model's development phase. These methods focus on designing inherently interpretable models, such as decision trees or rule-based systems, which provide explicit rules for decision-making. By building models with transparency in mind from the outset, interpretable-by-design methods offer direct interpretability without the need for additional post-hoc explanations. On the other hand, post-hoc XAI methods are applied after the model has been trained and are commonly used with black-box or complex models, including deep neural networks. These methods seek to explain the model's predictions without modifying its internal workings. In this work, we only focus on analysis post-hoc methods because of its popularity use.

B. Overview of Saliency-based XAI Methods

The concept of saliency maps [24] is based on the idea that certain regions or pixels in an image or input data play a crucial role in the model's prediction. By assigning importance scores to each pixel, saliency maps highlight the areas that have the highest influence on the model's output. The methods can be categorized into two groups, depending on the approach they utilize to calculate the contribution of each input pixel value: Gradient-based and Perturbation-based.

Based on the assumption that we know the parameters in the black box model, gradient-based methods first produce predictions through a forward pass and then utilize the predicted labels to propagate backward through each layer of the model, in sequential order, all the way to the input layer, estimating the contributions of the inputs during the backpropagation process. These methods, relying on derivatives, generate saliency maps that illustrate the contributions of each variable in the input space to the final prediction of the black box [4].

Unlike gradient-based methods, perturbation-based methods analyze how the model's predictions change when one or more input features are perturbed. These perturbations can be in the form of adding noise, modifying pixel values, or altering specific features in a controlled manner. This makes perturbation-based methods an optimal choice for analyzing the sensitivity of models [4]. Analyzing the sensitivity of models becomes particularly important when dealing with attacks where carefully crafted input modifications intentionally aim to alter the model's output, known as "adversarial attack" [26]. However, perturbation-based methods come with computational costs due to the need for generating and evaluating a large number of perturbed samples. The generated perturbations should ideally cover a diverse range of possible variations to ensure a robust analysis.

C. Evaluating the Quality of Saliency Maps

Despite limited research on the consistency of explanations produced by saliency methods, there exists a considerable body of literature evaluating the efficacy of saliency maps. Notably, Alqaraawi et al. [3] conducted an investigation into the potential application of saliency methods for non-expert users and demonstrated that saliency maps generated by the LRP algorithm can assist users in identifying image features that the black box model is sensitive to. However, the utility of saliency maps is not significant, and they may overlook critical features such as contrast, luminance, and others.

In medical imaging, saliency-based XAI methods play a predominant role [7] and are usually go-to techniques when

explaining for clinicians [19]. Since its introduction in 2013 as a visualization method for explaining ConvNets architectures [24], the application of saliency maps to explain image-related tasks in the medical field has become increasingly popular. Several studies have been conducted to evaluate whether current image explanation algorithms meet the accuracy requirements in the medical field, such as [29, 5, 12]. Almost all of these studies share the same perspective as Cynthia Rudin's research [19], which suggest that explanations generated by models should not be used in important tasks. According to experiment results, especially in image-related tasks in the medical field, saliency maps often exhibit similarities across different classification categories or are generally uninformative for end users [21]. Eitel et al. [9] showed that Gradient*Input, Guided Backpropagation, LRP and Occlusion vary in robustness and produce inconsistent explanations for classifying Alzheimer's disease when subjected to repeated black box retraining. Saliency methods are also found to be independent from the model, but are sensitive to the data [1, 10, 14].

III. The Disagreement Problem

The problem of disagreement in the field of explainable artificial intelligence is a relatively recent issue. Brughmans et al. [8] characterized it as the occurrence of "conflicting or contradictory explanations" generated by different interpretation methods when assessing a given AI model. Roy et al. [20] referred to the disagreement between explanation methods as "different (and even contradicting) explanations for the same model decisions". Nonetheless, the precise extent of disagreement between explanations and the specific aspects of disagreement remains to be conclusively defined.

The study conducted by Neely et al. [18] is among the earliest works that have raised concerns about the issue of explanation disagreement. The authors utilized rank correlation, specifically Kendall's τ coefficient, as a measure to assess the degree of disagreement among several explanation methods including LIME, Integrated Gradients, DeepLIFT, Grad-SHAP, Deep-SHAP, and attention-based explanations. The findings of the experiments indicated that there were notably low correlations observed between the results generated by the different methods, signifying that there is a high likelihood of disagreement among them.

A. How is Disagreement Measured?

Since there is a lack of common evaluation frameworks, the measurement of disagreement is yet to be concretely defined and generally depends on the nature of the problem and the explanation methods chosen. Currently, there are two approaches to measuring disagreement: using set-based metrics and using correlation.

The set-based metrics involve finding a set of feature attributions that share some common characteristics. The measure of agreement (or disagreement) then is based on the size of that set. Krishna et al. [15] developed some set-based metrics including feature agreement (using the set of shared features within the top-k feature of both explanations), signed agreement, rank agreement, and signed rank agreement (shared top-k features with the same signs, ranks, and both signs and ranks, respectively). Brughmanns et al. [8] adapted the set-based approach for counterfactual explanations with metrics such as relative feature exclusion, relative feature span, L0 distances, and feature disagreement (an improvement from Krishna et al.'s feature agreement that incorporates direction of disagreement). We refer to the original paper of the authors for a detailed description of these metrics. The correlation approach sees Neely et al. [18] utilize Kendall's τ coefficient and Krishna et al. use Spearman's ρ to measure the correlation between the relative ranking of the features.

B. Attempts in Resolving Disagreement

In practice, when faced with such disagreements, real-world practitioners often resort to selecting the method with which they are most familiar [15]. [20] seeks to mitigate disagreement by presenting the user only with the features on which LIME and SHAP agree while ignoring those on which they disagree. While this approach may create the appearance of consistency for the user, it does not address the underlying issue. If employed incorrectly, this approach may lead to even greater levels of misinterpretation. For instance, in cases where two approaches exhibit significant disagreement, the subset of features on which they both agree may be too small and insignificant to provide a meaningful explanation of the model's prediction.

IV. Method

A. Overview

Figure 1 summarizes our procedure for quantifying the disagreement between different saliency explanation methods. Initially, we train each targeted black box on the selected dataset that consists of X-ray images and matching masks. The specifics of this dataset will be discussed in depth in V-C. Subsequently, we apply the attribution algorithm of the chosen explanation methods to the test set, which generates an explanation for the black box per each test example. We then calculate the level of disagreement by using disagreement metrics for each pair of explanations per test example and aggregate the outcomes to create a comparison heatmap. Finally, we examine the heatmaps to determine whether there is any significant disagreement.

Note that this procedure is applied per black box. In our experiments, we applied this procedure to two different black boxes V-B and analyzed the outcomes for both of them.

B. Mathematical Formulation

1) Saliency Maps: For a given black box b and a test instance x, where x is an image of size $n \times n$, and an explanation method c, the saliency map explanation $e \in \mathbb{R}^{n \times n}$ for the black box b with respect to test instance x is obtained by running the attribution algorithm provided by c:

$$e = c(b, x) \tag{1}$$

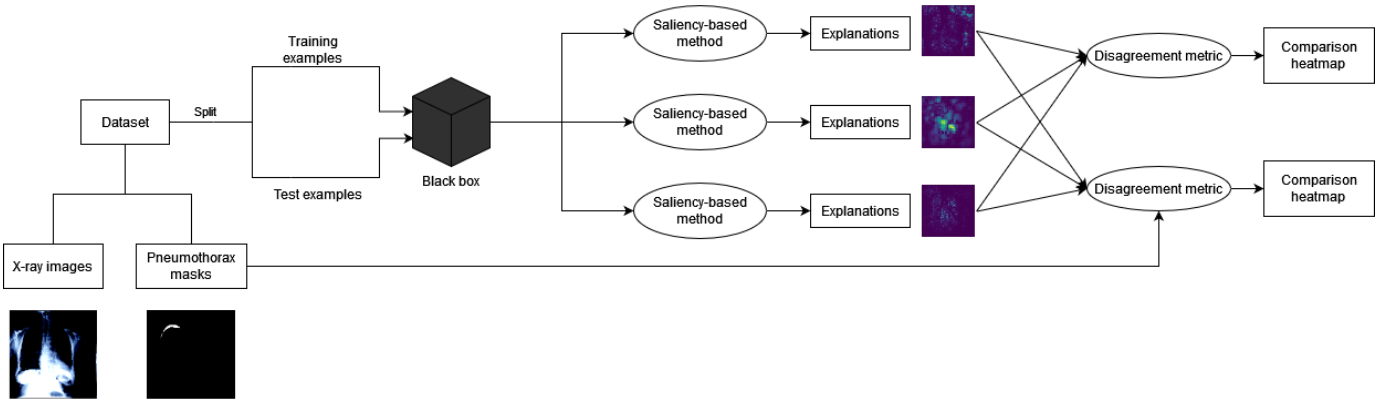

Fig. 1: An overview of the steps we've taken to quantify the disagreement between the explanation methods. These steps are applied per black box.

Each value e_{ij} represents the importance score that the pixel at (i, j) contribute to the prediction of the black box b at instance x. The range of values of e_{ij} depends on the nature of the explanation method c.

2) Metrics:

a) Structural Similarity Index Metric: To capture the similarity in structure between two saliency maps, we use the Structural Similarity Index Metric (SSIM) [30]. Despite its name, this metric captures more than just structural similarity between two images: it takes into account the contrast, luminance, and the structure of two images.

For the sake of simplicity, given two saliency maps $e^{(1)}$ and $e^{(2)}$, let $x = e^{(1)}$, $y = e^{(2)}$. Then, the formula for SSIM between two saliency maps x and y is:

$$\text{SSIM}(x, y) = [l(x, y)]^{\alpha}[c(x, y)]^{\beta}[s(x, y)]^{\gamma} \tag{2}$$

where $l(x, y)$, $c(x, y)$, and $s(x, y)$ is the comparison of the luminance, contrast, and structure of two images x and y, respectively; α, β, and γ are constansts that control the importance of the $l(x, y)$, $c(x, y)$, and $s(x, y)$, respectively.

Next, we present the definition of the three components of the SSIM. Let μ_x, μ_y be the mean value of the saliency map x and y; σ_x, σ_y be the standard deviation of the x and y; σ_{xy} be the covariance of x and y; C_1, C_2 and C_3 be small constants.

The luminance component is defined:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \tag{3}$$

Next, the contrast component is defined as:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \tag{4}$$

Finally, the structure component is defined as:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \tag{5}$$

The constants C_1, C_2 and C_3 are typically set to small values in order to avoid zero division. α, β and γ are usually set to 1, but they can be assigned with other values to emphasize or deemphasize their respective components.

In this work, because each saliency map differs in the range of its pixel values, we decided to measure SSIM using the magnitude of the saliency maps, normalized into the range $[0, 1]$ for fair comparison.

b) Feature Agreement: We utilize the feature agreement introduced by [15], adapted for images. For a saliency map e, a pixel (i, j) is within the top-k feature for e if $|e_{ij}|$ is among the set of top-k absolute values of all pixels of e. We denote the set of top-k pixels of e with $\text{top}_k(e)$. Then given two saliency map $e^{(1)}$ and $e^{(2)}$, the feature agreement between two maps is given by:

$$\text{FA}_k(e^{(1)}, e^{(2)}) = \frac{|\text{top}_k(e^{(1)}) \cap \text{top}_k(e^{(2)})|}{k} \tag{6}$$

Note that the above metric is equivalent to the following transformation and operation:

- Taking the absolute value of both saliency maps
- Convert both saliency maps into a binary map in which pixels within top-k are assigned the value 1, otherwise assigned the value 0
- The feature agreement of two saliency maps is computed by taking the size of the intersection of the 1s region and dividing by k

This metric captures how much the region that the two saliency maps consider most important overlaps with each other. Figure 2 presents the procedure to compute the feature agreement of a pair of explanations generated from Guided Backpropagation and Integrated Gradients. The feature agreement captures how much the two technique agrees for the top-5% most salient pixels (for a 128×128 image, this is equivalent to $k = 820$).

c) Sign Agreement: The sign agreement is quite similar to feature agreement, however, it is a stricter criterion [15]. The sign agreement between two saliency maps $e^{(1)}$ and $e^{(2)}$ with respect to the top-k features $\text{SA}_k(e^{(1)}, e^{(2)})$ is:

$$\frac{1}{k}|\{(i, j)|(i, j) \in \text{top}_k(e^{(1)}) \cap \text{top}_k(e^{(2)}) \wedge e_{ij}^{(1)} \cdot e_{ij}^{(2)} > 0\}| \tag{7}$$

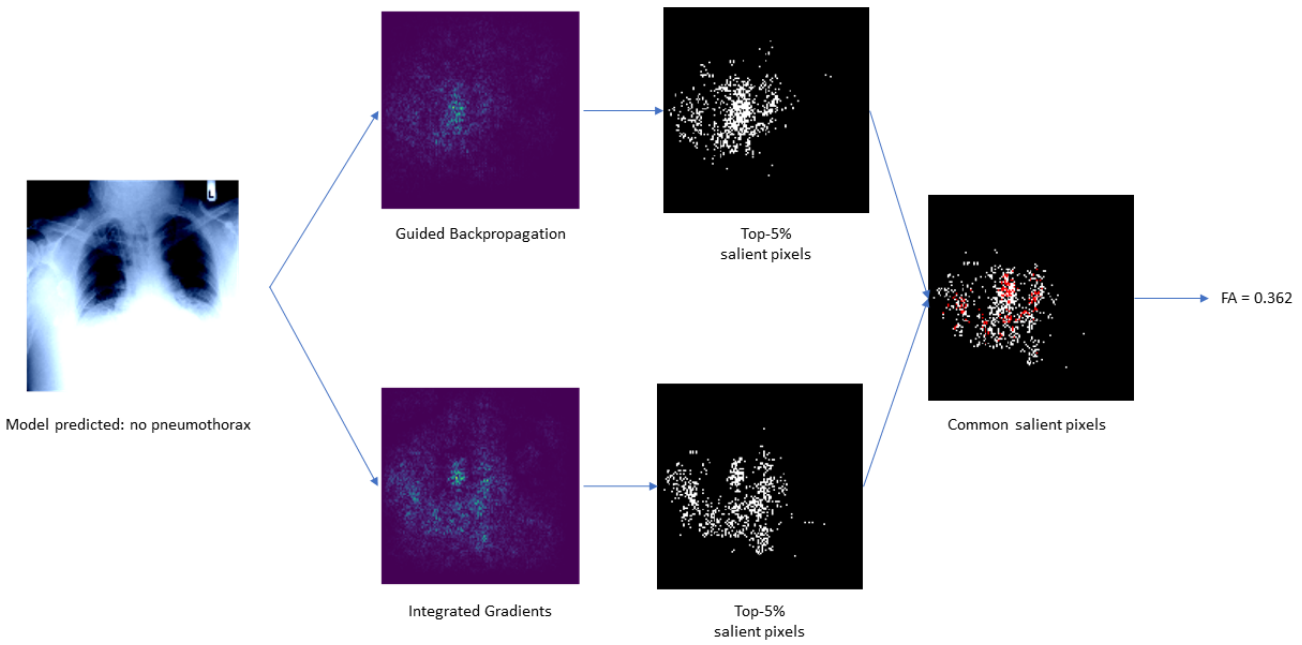

Fig. 2: Illustration of how feature agreement was measured for an example pair of saliency maps generated by Guided Backpropagation and Integrated Gradients. We are interested in the top 5% of the most important pixels ($k = 820$ in total, colored white). The feature agreement score is calculated by determining the number of common top salient pixels (highlighted in red) and dividing this by k.

The metric considers two saliency maps agree on a pixel if it is both significant in the two maps and has the same direction of contribution (negative or positive).

d) Rank Correlation: Another metric we adopted from the work of Krishna et al. [15] is the rank correlation metric. In the original work, the metric is computed using the rankings of a subset of the input features. This feature subset when adapted to saliency maps can be represented as a mask. For any given matrix real matrix d of size $m \times n$ (a saliency map can also be viewed as a matrix), a mask m is a binary matrix of the same size, and the operation $m(d)$ return the set of d_{ij} correspond to where $m_{ij} = 1$, that is:

$$m(d) = \{d_{ij} | m_{ij} = 1, \forall i \in [1, m], \forall j \in [1, n]\} \quad (8)$$

Given a saliency map e, we define rank(e) to be the matrix of the rank of the pixel values of e. Then the rank correlation for two saliency maps $e^{(1)}$ and $e^{(2)}$ with respect to a mask m is given by:

$$\text{RC}(e^{(1)}, e^{(2)}, m) = r_s \left(m(\text{rank}(e^{(1)})), m(\text{rank}(e^{(2)})) \right) \quad (9)$$

V. Experiments

A. Explanation Methods

We use the following saliency methods for our experiments: Occlusion as the single perturbation-based method, seven gradient-based methods including Vanilla Gradient or Saliency, GradientSHAP, Guided Backpropagation, Guided Grad-CAM, Integrated Gradients, DeepLift, and LRP. Specific hyperparameters were required for some of these methods, which we specified as follows:

- For Occlusion: we use an 8×8 occlusion window and a 4×4 stride.
- For Integrated Gradients and GradientSHAP: we use a baseline generated from a uniform distribution on the interval $[0, 1)$. $[0, 1)$.

B. Black boxes

We opted to use XAI methods on the InceptionV3 [28] and ResNet-101 [11] architectures for several reasons. One primary reason is that these models are based on Convolutional Neural Networks (CNNs), which are necessary for certain model-specific XAI techniques like Class Activation Mapping (CAM) and Grad-CAM. Additionally, we chose InceptionV3 and ResNet-101 because of their demonstrated performance in previous work conducted by Narin, Ali et al. on COVID-19 detection using X-ray images [17]. Although these models do not have sustainable G-ops (giga operations per second), they have shown good performance in specific applications, particularly in the medical domain.

C. Datasets

1) Chest X-Ray Images with Pneumothorax Masks dataset: This dataset contains the stage 1 train and test data extracted from the Kaggle SIIM-ACR Pneumothorax Segmentation competition [31]. This dataset encompasses a collection of medical images specifically curated for investigating and addressing the challenges associated with pneumothorax, a potentially life-threatening condition characterized by the presence of air in the pleural cavity, leading to lung collapse. This dataset comprises a total of 12,047 chest X-ray images, encompassing 10,675 samples designated for training and an additional 1,372 samples for testing. To ensure the reliability and generalizability of our models, we further partition the training samples into training and validation subsets using a ratio of 9:1, respectively.

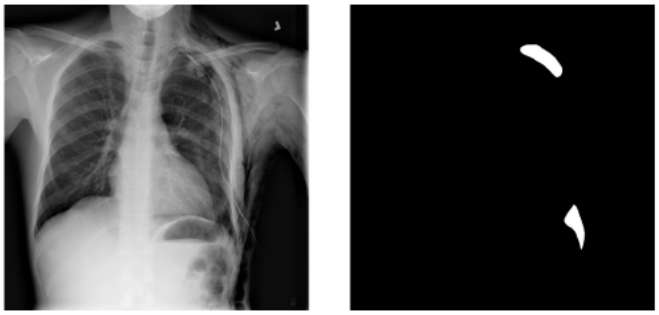

Fig. 3: A sample X-ray image of a patient with pneumothorax and a mask image marking the signs of pneumothorax extracted from the Chest X-Ray Images with Pneumothorax Masks dataset

2) Chest X-ray Images (Pneumonia): . This data includes a total of 5,863 JPEG images of X-Ray scans, encompassing 5216 samples for training and 624 samples for testing, which categorized into two classes: pneumonia and normal [13]. The chest X-ray images (anterior-posterior) were obtained from retrospective cohorts of pediatric patients aged one to five years old, specifically from Guangzhou Women and Children's Medical Center in Guangzhou. These images were part of the routine clinical care provided to the patients. To ensure the quality of the chest X-ray images used for analysis, a preliminary screening process was conducted. This involved removing scans that were of low quality or deemed unreadable. Subsequently, two experienced physicians reviewed and graded the diagnoses for the remaining images before they were considered suitable for training the AI system. To account for any potential grading errors, a third expert also examined the evaluation set.

D. Experiments Setup

To evaluate post-hoc XAI methods, our experiment consists of two stages. In the first stage, we train black boxes on a classification task using the designated Chest X-ray Images with Pneumothorax Masks dataset V-C. In the second stage, we apply XAI methods to the trained models to gain insights into their decision-making process and interpret the underlying factors that contribute to their predictions. By analyzing the explanations provided by the XAI methods, we aim to understand the model's behavior and assess its transparency and interpretability.

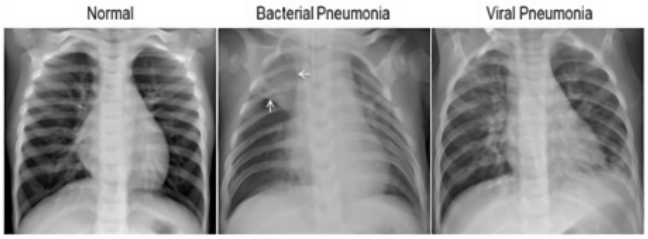

Fig. 4: X-ray images of normal lungs, viral or baterial pneumonia extracted from the Chest X-ray (Pneumonia) dataset

The development environment in which the experiments were conducted are summarized in table I.

TABLE I: Experiment environments and requirements.

CPU	Intel(R) Xeon(R) CPU @ 2.20GHz
RAM	1×13GB
GPU (number and type)	NVIDIA TESLA P100 GPUs 16GB
CUDA version	11.6
Programming language	Python 3.10

a) Training Stage: In the training stage, we trained the InceptionV3 [28] and ResNet-101 [11] models using popular tools such as pytorch, torchvision, scikit-learn, and pandas. Both models were initially pre-trained on the ImageNet dataset [22]. To adapt these models to our specific task of pneumothorax classification, we further trained them using the two datasets introduced in V-C. Before training the models, we performed preprocessing on the chest X-ray images. This involved normalization, where the image is transformed such that its mean μ is (0.485, 0.456, 0.406) and its standard deviation σ is (0.229, 0.224, 0.225). This normalization step helps to standardize the input data and improve the training process. Additionally, we resized the images to a specific dimension with length and width in pixels, respectively: (299, 299) for InceptionV3 and (244, 244) for ResNet-101. The models were trained on the classification task, aiming to categorize whether a given chest X-ray image exhibited signs of pneumothorax or not. For this binary classification task, we employed the cross-entropy loss function, which is commonly used for training classification models. The optimization of the models was carried out using the Adam optimizer, a popular choice known for its efficiency in handling complex models. The training process spanned 30 epochs, allowing the models to iteratively learn and improve their performance. Throughout the training phase, we monitored the models' progress and selected the best-performing state based on their performance on the validation set.

b) Analysis Stage: In the XAI analysis stage, we employed XAI methods provided by the Captum framework

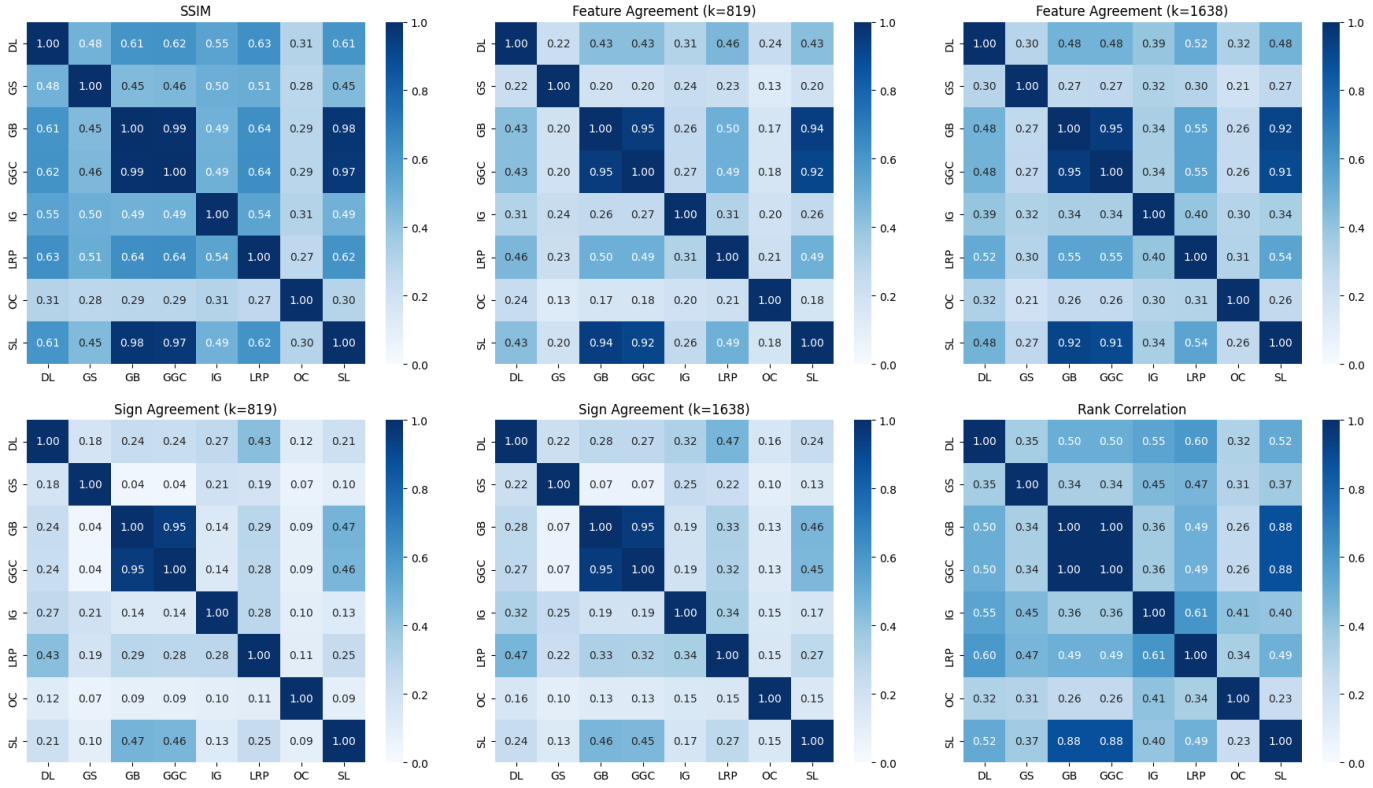

Fig. 5: Disagreement between explanation methods when explaining for InceptionV3, trained on the Pneumothorax dataset.

to generate explanations for the predictions made by our trained models. To measure the disagreement between these explanations, we implement some metrics introduced in IV-B2 by using popular python libraries such as numpy, torch, and skimage to perform the necessary calculations and computations. Once the metrics were computed, we visualized the results to gain a better understanding of the differences and similarities among the XAI methods by using seaborn library.

E. Results

In this section, we discuss some key findings of our thesis regarding the disagreement problem. We refer to our thesis document for the full list of the disagreement heatmap visualizations.

1) Varying level of disagreement: Figure 5 illustrates the diverse range of disagreement scores among different explanation methods when providing explanations for black box models trained on the Pneumothorax dataset. The figure demonstrates that there exists variability in the extent of disagreement exhibited by these methods. Overall, the tested explanation methods exhibit the highest level of agreement for the SSIM (Structural Similarity Index) and highest level of disagreement for sign agreement metric. This pattern is also observed consistently in the case of the Pneumonia dataset.

2) Different levels of disagreement when explaining for different black boxes: Our analysis reveals another finding, which is the notable variation in agreement scores when evaluating different black box models. Specifically, there is a higher level of disagreement observed when explaining the ResNet-101 model compared to the InceptionV3 model. This observation holds true for all metrics and datasets examined. Figure 6 presents a comparison of the SSIM scores for explanation methods when explaining the InceptionV3 and ResNet-101 models on the Pneumonia dataset. The figure highlights that most explanation methods generate highly similar saliency maps when explaining the InceptionV3 model. In contrast, for the ResNet-101 model, only a group of four metrics, namely GradientSHAP, Guided Backpropagation, Guided Grad-CAM, and Integrated Gradients, produce moderately to highly similar explanations. Other pairs exhibit significantly lower scores.

3) Independence of the source training data: However, the level of disagreement between pairs of explanation methods tends to remain consistent across all metrics and datasets (except for SSIM) when considering the same black box model. Figure 7 illustrates this by presenting feature agreement scores for the ResNet-101 model evaluated on both datasets. The heatmaps demonstrate a high degree of similarity, indicating that most pairs of explanation methods exhibit little change in feature agreement scores when explaining the same black box model trained on two different source training datasets. This observation suggests that the extent of disagreement is primarily influenced by the specific type of black box model being explained rather than the source dataset used to train the model. This finding contradicts the claim made in [8] that the magnitude of the disagreement issue is independent of the

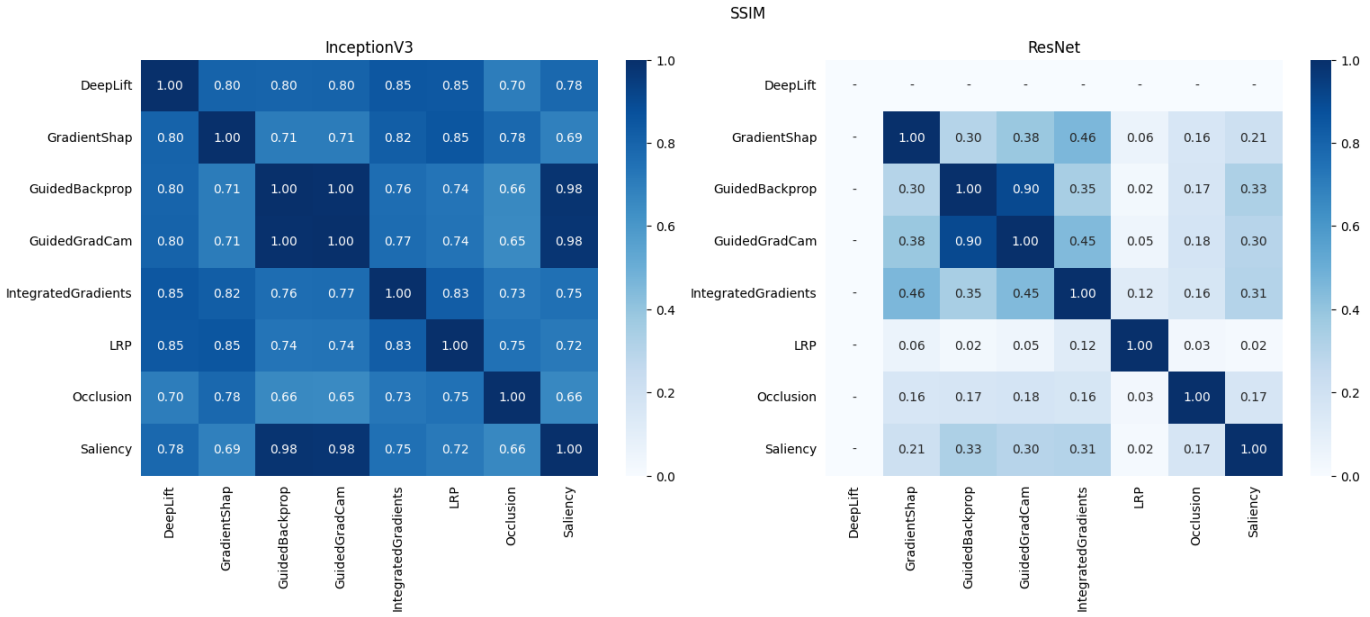

Fig. 6: Comparison of SSIM scores for InceptionV3 and ResNet-101 when evaluating on Pneumonia dataset.

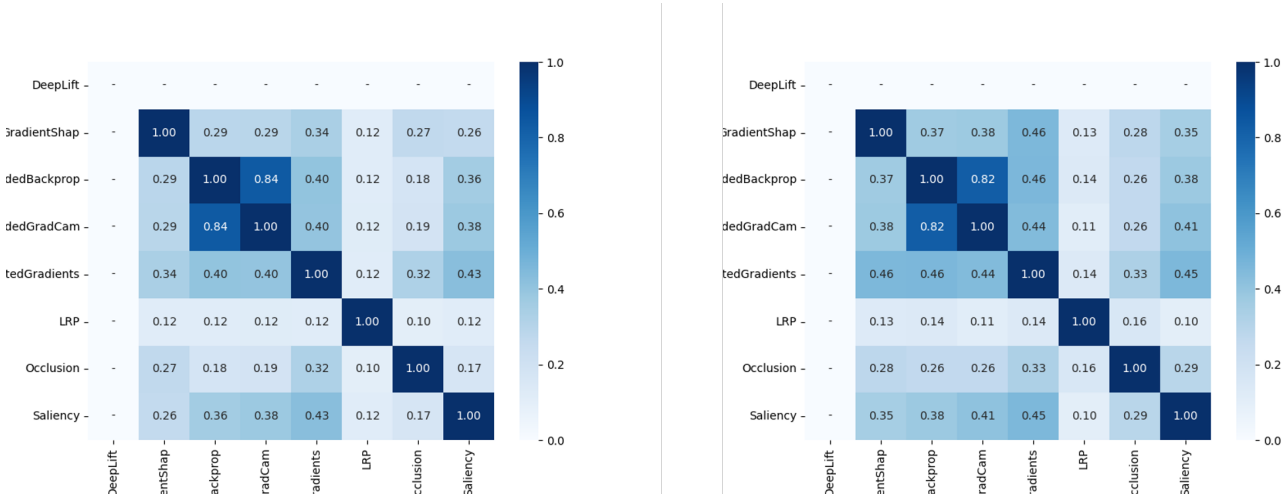

Fig. 7: Comparison of feature agreement scores of the explanation methods explaining for ResNet-101 when evaluating on two datasets. Figure on the left describes the feature agreement scores evaluated on the **Pneumothorax** dataset, while figure on the right describes said scores evaluated on the **Pneumonia** dataset.

classifier. It implies that the underlying problems may have deeper roots, potentially related to the characteristics of the explanation methods themselves. Therefore, we recommend further research to thoroughly investigate and provide a definitive explanation for this phenomenon.

4) Patterns of disagreement among the explanation methods: While the levels of disagreement vary across different metrics, black boxes, and datasets, certain pairs of explanation methods consistently display notable consistency. Particularly remarkable are the three methods: Guided Grad-CAM, Guided Backpropagation, and Saliency. Table II provides an overview of the agreement scores for these three pairs of methods. The Guided Grad-CAM–Guided Backpropagation pair consistently achieves high agreement scores across all scenarios. The pairs Saliency–Guided Grad-CAM and Saliency–Guided Backpropagation consistently exhibit agreement when explaining the InceptionV3 model, but consistently show disagreement when explaining the ResNet-101 model, often by a significant margin.

VI. Conclusion

A. Results

This thesis delves into the disagreement problem in XAI in general and specifically examines the disagreement between explanation methods that utilize saliency maps, an area that has not been previously researched. Our study involved measuring

Pair	Dataset	Feature Agreement (top 10%)		Sign Agreement (top 10%)		Rank Correlation		SSIM	
		InceptionV3	ResNet-101	InceptionV3	ResNet-101	InceptionV3	ResNet-101	InceptionV3	ResNet-101
GuidedBackprop–	Pneumothorax	**0.95**	0.84	**0.95**	0.80	**1.00**	0.99	**0.99**	0.92
GuidedGradCAM	Pneumonia	**0.96**	0.82	**0.96**	0.82	**-**	-	**1.0**	0.90
GuidedBackprop–	Pneumothorax	**0.92**	0.36	**0.47**	0.19	**0.88**	0.32	**0.98**	0.29
Saliency	Pneumonia	**0.88**	0.38	**0.43**	0.29	**-**	-	**0.98**	0.33
GuidedGradCAM	Pneumothorax	**0.91**	0.38	**0.45**	0.22	**0.88**	0.32	**0.97**	0.27
–Saliency	Pneumonia	**0.88**	0.41	**0.43**	0.31	**-**	-	**0.98**	0.30

TABLE II: Comparison of the metric scores for every pair of the methods Guided Backpropagation, Guided Grad-CAM and Saliency.

the level of disagreement between 8 saliency explanation methods over two black boxes, using four different metrics. Our findings can be summarized as follows:

- We confirmed that there is a considerable amount of disagreement in explanations between many saliency maps.
- We showed that the patterns of disagreement between the methods are complex.
- We indicated that the level of inconsistency varies based on the type of black box employed.

B. Limitations and Future Works

The disagreement problem is a novel issue that has yet to receive comprehensive research attention, resulting in a lack of diverse viewpoints and materials on the subject. However, we firmly believe that this is not a trivial issue and deserves greater attention. If not addressed adequately, the validity and transparency of XAI approaches, which are the key desired characteristics of XAI, may be seriously questioned. Our thesis is an attempt to explore a new aspect of this problem, and there is significant scope for further improvement.

We acknowledge some limitations of our work and suggest potential directions for future research. Firstly, we only tested the disagreement problem on two datasets with one classification task. We believe that further researches on a wider range of datasets and black boxes will reveal many useful insights. Secondly, the existing metrics are insufficient to capture the full extent of disagreement, and new metrics can be developed to address this limitation. Thirdly, while the existence of disagreement has been acknowledged, a thorough and systematic investigation into the reasons for disagreement has yet to be conducted. Future work should focus on uncovering the underlying reasons for this disagreement. Finally, we urge the XAI community to attach greater importance to this problem, as the goal of XAI is to assist humans in understanding machine learning and deep learning algorithms and to make them more interpretable, rather than introducing further confusion by providing inconsistent or contradictory explanation.

REFERENCES

References

[1] Adebayo, Julius et al. *Sanity Checks for Saliency Maps*. 2020. arXiv: 1810.03292 [cs.CV].

[2] Alonso, José Maria, Castiello, Ciro, and Mencar, Corrado. "A Bibliometric Analysis of the Explainable Artificial Intelligence Research Field". In: *International Conference on Information Processing and Management of Uncertainty*. 2018.

[3] Alqaraawi, Ahmed et al. "Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study". In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. IUI '20. Cagliari, Italy: Association for Computing Machinery, 2020, pp. 275–285. ISBN: 9781450371186. DOI: 10.1145/3377325.3377519. URL: https://doi.org/10.1145/3377325.3377519.

[4] Ancona, Marco et al. "Gradient-Based Attribution Methods". In: Sept. 2019, pp. 169–191. ISBN: 978-3-030-28953-9. DOI: 10.1007/978-3-030-28954-6_9.

[5] Arun, Nishanth et al. "Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging". In: *Radiology: Artificial Intelligence* 3.6 (2021), e200267. DOI: 10.1148/ryai.2021200267. eprint: https://doi.org/10.1148/ryai.2021200267. URL: https://doi.org/10.1148/ryai.2021200267.

[6] Arya, Vijay et al. *One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques*. 2019. arXiv: 1909.03012 [cs.AI].

[7] Borys, Katarzyna et al. "Explainable AI in medical imaging: An overview for clinical practitioners – Saliency-based XAI approaches". In: *European Journal of Radiology* 162 (2023), p. 110787. ISSN: 0720-048X. DOI: https://doi.org/10.1016/j.ejrad.2023.110787. URL: https://www.sciencedirect.com/science/article/pii/S0720048X23001018.

[8] Brughmans, Dieter, Melis, Lissa, and Martens, David. *Disagreement amongst counterfactual explanations: How transparency can be deceptive*. 2023. arXiv: 2304.12667 [cs.AI].

[9] Eitel, Fabian, Ritter, Kerstin, and (ADNI), Alzheimer's Disease Neuroimaging Initiative. "Testing the robustness of attribution methods for convolutional neural networks in MRI-based Alzheimer's disease classification". In: *Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI*

2019, Shenzhen, China, October 17, 2019, Proceedings 9. Springer. 2019, pp. 3–11.

[10] Ghorbani, Amirata, Abid, Abubakar, and Zou, James. "Interpretation of neural networks is fragile". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 3681–3688.

[11] He, Kaiming et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[12] Jin, Weina, Li, Xiaoxiao, and Hamarneh, Ghassan. "Evaluating explainable AI on a multi-modal medical imaging task: can existing algorithms fulfill clinical requirements?" In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 11. 2022, pp. 11945–11953.

[13] Kermany, Daniel. *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*. eng. 2018.

[14] Kindermans, Pieter-Jan et al. "The (Un)reliability of Saliency Methods". In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Samek, Wojciech et al. Cham: Springer International Publishing, 2019, pp. 267–280. ISBN: 978-3-030-28954-6. DOI: 10.1007/978-3-030-28954-6_14. URL: https://doi.org/10.1007/978-3-030-28954-6_14.

[15] Krishna, Satyapriya et al. "The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective". In: *CoRR* abs/2202.01602 (2022). arXiv: 2202.01602. URL: https://arxiv.org/abs/2202.01602.

[16] Lipton, Zachary Chase. "The Mythos of Model Interpretability". In: *CoRR* abs/1606.03490 (2016). arXiv: 1606.03490. URL: http://arxiv.org/abs/1606.03490.

[17] Narin, Ali, Kaya, Ceren, and Pamuk, Ziynet. "Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks". In: *Pattern Analysis and Applications* 24.3 (May 2021), pp. 1207–1220. DOI: 10.1007/s10044-021-00984-y. URL: https://doi.org/10.1007%5C%2Fs10044-021-00984-y.

[18] Neely, Michael et al. "Order in the Court: Explainable AI Methods Prone to Disagreement". In: (May 2021).

[19] Patrício, Cristiano, Neves, João C., and Teixeira, Luís F. *Explainable Deep Learning Methods in Medical Imaging Diagnosis: A Survey*. 2022. arXiv: 2205.04766 [eess.IV].

[20] Roy, Saumendu et al. "Why Don't XAI Techniques Agree? Characterizing the Disagreements Between Post-hoc Explanations of Defect Predictions". In: *2022 IEEE International Conference on Software Maintenance and Evolution (ICSME)*. 2022, pp. 444–448. DOI: 10.1109/ICSME55016.2022.00056.

[21] Rozario, Savio and Cevora, George. "Explainable AI does not provide the explanations end-users are asking for". In: *ArXiv* abs/2302.11577 (2023).

[22] Russakovsky, Olga et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[23] Scott, A. Carlisle et al. "Explanation Capabilities of Production-Based Consultation Systems". In: *American Journal of Computational Linguistics* (Feb. 1977). Microfiche 62, pp. 1–50. URL: https://aclanthology.org/J77-1006.

[24] Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps". In: *CoRR* abs/1312.6034 (2013).

[25] Singh, Amitojdeep, Sengupta, Sourya, and Lakshminarayanan, Vasudevan. "Explainable Deep Learning Models in Medical Image Analysis". In: *Journal of Imaging* 6 (June 2020), p. 52. DOI: 10.3390/jimaging6060052.

[26] Slack, Dylan et al. "Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods". In: *AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*. 2020. URL: https://arxiv.org/pdf/1911.02508.pdf.

[27] Swartout, William R. "Explaining and Justifying Expert Consulting Programs". In: *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'81. Vancouver, BC, Canada: Morgan Kaufmann Publishers Inc., 1981, pp. 815–823.

[28] Szegedy, Christian et al. "Rethinking the Inception Architecture for Computer Vision". In: *CoRR* abs/1512.00567 (2015). arXiv: 1512.00567. URL: http://arxiv.org/abs/1512.00567.

[29] Taly, Ankur et al. "Using a deep learning algorithm and integrated gradient explanation to assist grading for diabetic retinopathy". In: *Ophthalmology* (2019).

[30] Wang, Z. et al. "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4 (Apr. 2004), pp. 600–612. DOI: 10.1109/TIP.2003.819861.

[31] Zawacki, Anna et al. *SIIM-ACR Pneumothorax Segmentation*. 2019. URL: https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation.