

UNIVERSITY OF SCIENCE  
FACULTY OF INFORMATION TECHNOLOGY

Mai Duy Nam - Nguyễn Hữu Bình

THE DISAGREEMENT PROBLEM IN  
XAI ON IMAGE DATA VIA SALIENCY  
MAPS

BACHELOR THESIS  
HONORS PROGRAM

Ho Chi Minh City, June 2022

UNIVERSITY OF SCIENCE  
FACULTY OF INFORMATION TECHNOLOGY

Mai Duy Nam - 19120298  
Nguyễn Hữu Bình - 19120460

**THE DISAGREEMENT PROBLEM IN  
XAI ON IMAGE DATA VIA SALIENCY  
MAPS**

BACHELOR THESIS  
HONORS PROGRAM

**THESIS SUPERVISOR**  
Prof. Lê Hoài Bắc

Ho Chi Minh City, June 2022

# Comments of Thesis Supervisor

# Comments of Thesis Reviewer

# Acknowledgement

First and foremost, we would like to express our gratitude to our supervisor Prof. Lê Hoài Bắc, who introduced and motivated us to explore the field of Explainable Artificial Intelligence. We are grateful for his continuous encouragement, invaluable support and guidance throughout our research journey. We are grateful to the instructors at the Faculty of Information Technology for providing guidance throughout our four-year academic journey, which helped us acquire valuable base knowledge for this thesis. Last but not least, we would like to express our appreciation to our friends and family members for their constant support and encouragement, which served as a significant driving force for us to successfully finish this thesis.

# Proposal



**fit@hcmus**

UNIVERSITY OF SCIENCE - VNUHCM  
FACULTY OF INFORMATION TECHNOLOGY

BACHELOR THESIS PROPOSAL  
**THE DISAGREEMENT PROBLEM IN XAI  
ON IMAGE DATA VIA SALIENCY MAPS**

*Vấn đề bất đồng trong giải thích của XAI khi ứng dụng trên dữ liệu  
ảnh bằng saliency map*

## 1 GENERAL INFORMATION

**Thesis Supervisor:**

– Prof. Lê Hoài Bắc (Faculty of Information Technology)

**Students:**

1. Mai Duy Nam (ID: 19120298)
2. Nguyễn Hữu Bình (Id: 19120460)

**Project Type:** Research

**Timeline:** From *March 2023* to *July 2023*

## 2 CONTENTS

### 2.1 Introduction

In recent years, there has been a high rise in the use of AI models due to their high accuracy and usefulness. However, these models are often black boxes with very deep neural network structures and have an enormous number of parameters, meaning their internals are obscure and beyond human comprehension. In many real-world scenarios, AI systems are used to make critical decisions that affect human lives, such as medical diagnoses, autonomous vehicles, security, and surveillance. Interpreting these models' output is crucial in determining the appropriate response to these threats. Explainable AI (XAI) has emerged as a solution to prevent blind faith in model outputs and improve the interpretability of AI models. However, many AI practitioners who utilize visual AI models pay less attention to or are not aware of the disagreement problem between the XAI methods—the problem in which different XAI methods generate inconsistent explanations when evaluation on the same black box. Additionally, there is a lack of research examining the extent of disagreement between XAI methods in general [1], and methods that target images or speech input in particular.

Therefore, this work aims to investigate the extent of such disagreements and find a proper way to quantify their differences by conducting experiments with multiple visual XAI methods that explain a black-box model trained on different popular pre-trained image datasets. A variety of metrics will be used to quantify the differences. Moreover, we would propose solutions in case discrepancies arise by observing factors that made the explanations disagree.



## 2.2 Target

### 2.2.1 Rationale

As mentioned in 2.1, AI models are increasingly being used in various fields, but they are often black boxes that are difficult for humans to understand the decision-making process behind them.

Computer Vision (CV) is the most widely used among the different AI fields. Visual AI systems are now used to make critical decisions in various real-world scenarios that affect human lives, such as medical diagnoses, autonomous vehicles, security, and surveillance. Interpreting these models' output is crucial in determining the appropriate response to these threats. If these decisions are made solely by an AI system without any explanation, it can be difficult for humans to understand how they were made and trust the system's output. This can lead to legal, ethical, and social issues.

Numerous studies have been conducted to address this problem to improve the interpretability of these models. Many methods were created to explain black boxes using saliency map explainers, such as CAM [2], Grad-CAM [3], layer-wise relevance propagation [4], deep Taylor decomposition [5], etc.). While they all aim to explain the black-box models' decision-making processes, they use different methods to identify which features in the input image are important for the model's output. As a result, they may produce different saliency maps that highlight different regions of the image as having higher importance, leading to discrepancies in their explanations.

According to the paper [1] by Krishna et al., machine learning practitioners use multiple explanation models to gain insights into how a black-box works. Krishna et al. presented some experiments on the disagreement problems between various XAI methods, such as LIME, SHAP, and Gradient-based models. After that, they compared these methods using some metrics to quantify the extent

of disagreement. However, when applied to images, the aforementioned methods measure the disagreement using feature importance by regarding each pixel as one feature, which is not semantically meaningful. In this work, we explore the methods whose outputs are in the form of saliency maps by conducting experiments on many other suitable metrics to find whether the disagreement problem persists when applied to image data, what makes these disagreements arise, and how to quantify and resolve them.

### **2.2.2 Significance**

This comparative study has several potential contributions to the field of XAI.

Firstly, it can help improve the transparency and interpretability of ResNet family models by investigating how their behaviors vary when pre-trained on different datasets. This is especially important as the quality of the data used to train a model can significantly impact its performance and reliability. By comparing the interpretability of a pre-trained model, we can identify which datasets are most suitable for achieving a particular level of transparency and interpretability. This can help guide the development of more transparent and trustworthy AI systems.

Secondly, identifying the reasons for saliency map disagreement or inconsistency can help researchers develop new XAI methods and metrics that are more effective in explaining black-box models. This can contribute to the advancement of the field of XAI and improve the overall performance of visual AI systems.

Finally, conducting a comparative study of different visual XAI methods across different datasets can provide valuable insights into best practices and guide AI practitioners in selecting the most reliable and robust method for a given task.

## **2.3 Scope**

By utilizing multiple popular XAI methods, this study conducts experiments by explaining a pre-trained ResNet model on some common image datasets. The resulting explanations are then evaluated using some types of metrics to draw

practical conclusions about all the factors joining in the train model process and the disagreement issue between visual explanations.

## 2.4 Approach

In this study, we focus on factors that may affect the accuracy of saliency map explanations, such as datasets, the architecture of ResNet, and the mechanism of XAI methods.

With a pre-trained model on any given dataset, we will first use various XAI methods on ResNet to evaluate the interpretability level of this model family. Because different XAI methods may focus on different aspects of the data, leading to different explanations. So if a disagreement arises, we can evaluate the performance in terms of the accuracy of each algorithm to get insight into which types of methods fit with each dataset. Then, we will quantify the disagreement between the XAI methods by applying several metrics to quantify the disagreement between explanations to find a suitable metric for this task.

Next, we will use the most stable XAI algorithm to explain ResNet’s outputs on multiple datasets and observe the vary of interpretability through multiple datasets. Computer Vision datasets used to train computer vision models are vast in size, with a significant portion readily accessible online and might be combined from various sources. These sources may be unreliable and biased due to human error, which can lead to the models themselves inheriting such unfairness.

Finally, from these above results, plus the insights gained from analyzing the ResNet architecture & pre-trained datasets, we can draw conclusions about the disagreement between saliency maps.

## 2.5 Expectation

Expected outcomes of this research project include:

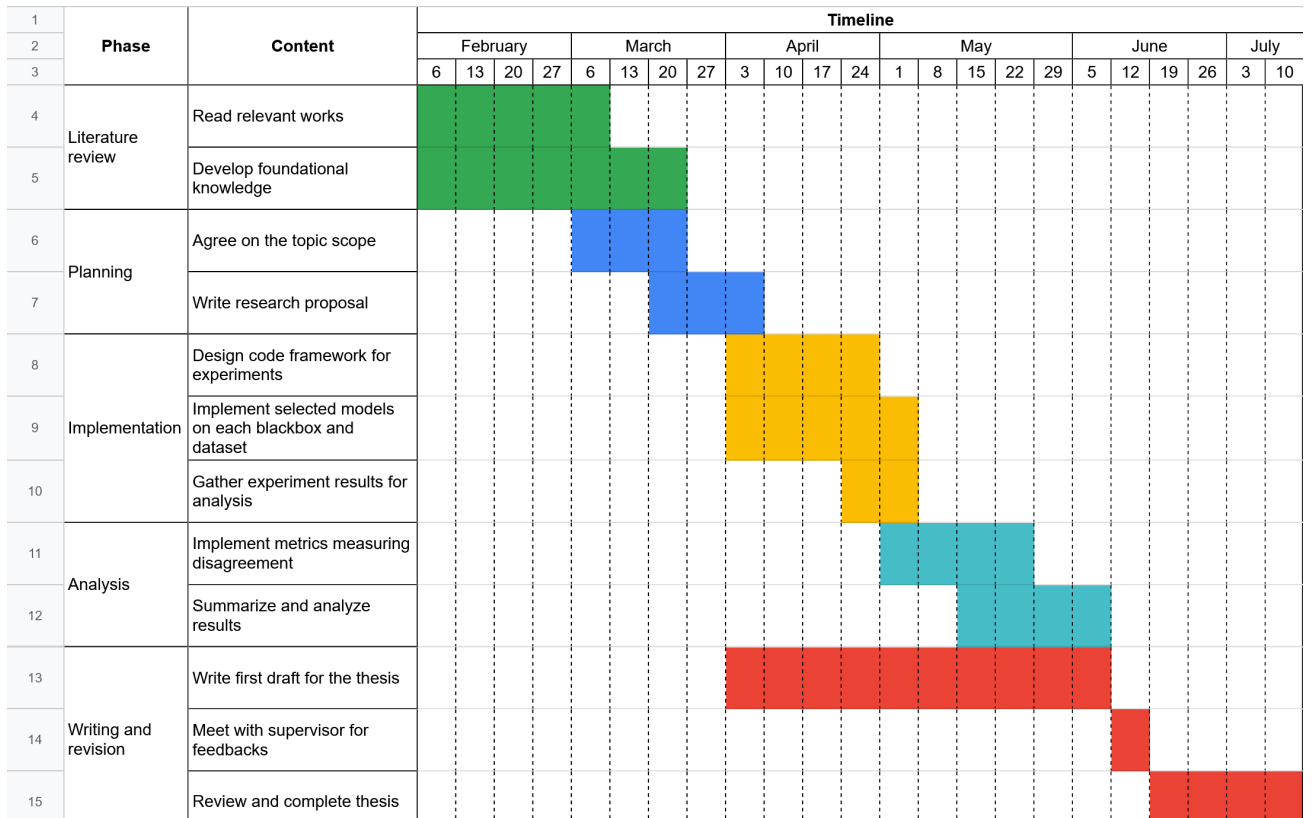
- A comparative study will show practical accuracy & performances of popular

visual XAI methods on ResNet.

- Identification of the reasons for inconsistency or disagreement between the models.
- Improvement in the transparency and interpretability of the ResNet model and understanding of its behaviors when pre-trained on different datasets.
- Practical way to choose which XAI methods to explain ResNet & metrics to measure disagreement degree.

## 2.6 Project Plan

The project plan is summarized in figure 1.



Hình 1: The project plan summarized in a Gantt chart, which includes the research phases, the activities in each phase, and the timeline for each activity.

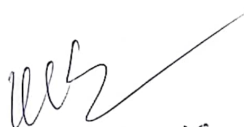
## References

- [1] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju, “The disagreement problem in explainable machine learning: A practitioner’s perspective,” *CoRR*, vol. abs/2202.01602, 2022.
- [2] B. Zhou, A. Khosla, À. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” *CoRR*, vol. abs/1512.04150, 2015.
- [3] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, “Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization,” *CoRR*, vol. abs/1610.02391, 2016.
- [4] A. Binder, G. Montavon, S. Bach, K. Müller, and W. Samek, “Layer-wise relevance propagation for neural networks with local renormalization layers,” *CoRR*, vol. abs/1604.00825, 2016.
- [5] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, “Explaining nonlinear classification decisions with deep taylor decomposition,” *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

*Ho Chi Minh City, April 2nd 2023*

**SUPERVISOR**

*(Signature)*



6/12/2023

**STUDENTS**

*(Signature)*



Mai Duy Nam Nguyễn Hữu Bình

# Contents

Comments of Thesis Supervisor	i
Comments of Thesis Reviewer	ii
Acknowledgement	iii
Proposal	iv
Table of contents	xii
Abstract	xvii
<b>1 Introduction</b>	<b>1</b>
1.1 Overview . . . . .	1
1.2 Problems . . . . .	3
1.3 Objectives . . . . .	4
1.4 Methodology . . . . .	4
1.5 Thesis Content . . . . .	5
<b>2 Background</b>	<b>8</b>
2.1 Overview of Explainable Artificial Intelligence . . . . .	8
2.2 Overview of Saliency-based XAI Methods . . . . .	11
2.2.1 Gradient-based methods . . . . .	11
2.2.1.1 Vanilla Gradient/Saliency . . . . .	12
2.2.1.2 Integrated Gradient . . . . .	13

2.2.1.3	Grad-CAM . . . . .	14
2.2.1.4	Guided Backpropagation . . . . .	16
2.2.1.5	Guided GradCAM . . . . .	17
2.2.1.6	Layer-wise Relevance Propagation . . . . .	18
2.2.1.7	Deep Label-Specific Feature Learning . . . . .	20
2.2.1.8	Gradient SHapley Additive exPlanation . . . . .	21
2.2.2	Perturbation-based methods . . . . .	23
2.2.2.1	Occlusion . . . . .	24
<b>3</b>	<b>Related Works</b>	<b>26</b>
3.1	The Disagreement Problem . . . . .	26
3.1.1	How is Disagreement Measured? . . . . .	28
3.1.2	Attempts in Resolving Disagreement . . . . .	29
3.2	Evaluating the Quality of Saliency Maps . . . . .	29
<b>4</b>	<b>Method</b>	<b>31</b>
4.1	Overview . . . . .	31
4.2	Mathematical Formulation . . . . .	32
4.2.1	Saliency Maps . . . . .	32
4.2.2	Metrics . . . . .	33
4.2.2.1	Structural Similarity Index Metric . . . . .	33
4.2.2.2	Feature Agreement . . . . .	34
4.2.2.3	Sign Agreement . . . . .	36
4.2.2.4	Rank Correlation . . . . .	36
4.3	Experiment Subjects . . . . .	37
4.3.1	Explanation Methods . . . . .	37
4.3.2	Black boxes . . . . .	37
4.3.2.1	Residual Network - 101 . . . . .	38
4.3.2.2	InceptionV3 . . . . .	39
4.3.3	Dataset . . . . .	41
4.3.3.1	Chest X-Ray Images with Pneumothorax Masks dataset . . . . .	41

4.3.3.2	Chest X-Ray Images (Pneumonia) . . . .	42
<b>5</b>	<b>Experiments &amp; Result</b>	<b>43</b>
5.1	Experiments . . . . .	43
5.1.1	Training Stage . . . . .	44
5.1.2	Analysis Stage . . . . .	45
5.2	Results . . . . .	45
5.2.1	Pneumothorax Dataset . . . . .	45
5.2.1.1	Structural Similarity Index Measure . . .	45
5.2.1.2	Feature Agreement . . . . .	46
5.2.2	Sign Agreement . . . . .	47
5.2.2.1	Rank Correlation . . . . .	50
5.2.3	Pneumonia Dataset . . . . .	50
5.2.3.1	Structural Similarity Index . . . . .	50
5.2.3.2	Feature Agreement & Sign Agreement . .	51
5.2.4	Overall Discussion . . . . .	51
<b>6</b>	<b>Conclusion</b>	<b>57</b>
6.1	Results . . . . .	57
6.2	Limitations and Future Works . . . . .	58
	<b>References</b>	<b>59</b>
<b>A</b>	<b>Supplementary Visualizations</b>	<b>65</b>
A.1	Disagreement Heatmaps Visualizations for the Pneumonia Dataset . . . . .	65



# List of Figures

1.1	Different saliency maps produced by different XAI methods. The bright areas indicate important areas. [19] . . . . .	3
2.1	Taxonomy of XAI methods [6] . . . . .	10
2.2	How Guided Grad-CAM heatmaps generated by combining Grad-CAM & Guided backpropagation method. Source [29]	19
2.3	Relevance scores propagating procedure. Source [7] . . . .	20
4.1	An overview of the steps we've taken to quantify the dis- agreement between the explanation methods. These steps are applied per black box. . . . .	32
4.2	Illustration of how feature agreement was measured for an example pair of saliency maps generated by Guided Back- propagation 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 $k$ . . . . .	35
4.3	Residual Block Diagram. Source [13] . . . . .	38
4.4	Resnet101 Architecture Diagram. Source [39] . . . . .	39
4.5	Naive Form of Inception Block. Source [13] . . . . .	39
4.6	Some major optimizations of InceptionV3 model. Source [37]	40
4.7	InceptionV3 Architecture. Source [37] . . . . .	40

5.1	Pairwise structural similarity indices (SSIM) between explanation methods generating for both black boxes (InceptionV3 and ResNet-101). The SSIM scores are computed using the average score over test set data points. Darker colors imply greater similarity in the structure of the two explanations. . . . .	47
5.2	Comparison of the saliency maps generated by the experimental methods with respect to InceptionV3 prediction. Lighter colors indicate more salient pixels. . . . .	48
5.3	Feature agreement between different methods with varying top- $k$ value. . . . .	49
5.4	Sign agreement between different methods. . . . .	53
5.5	The distributions of the saliency map values (within the top- $k$ magnitude) produced by the explanation methods, evaluating on ResNet model. . . . .	54
5.6	The distributions of the saliency map values (within the top- $k$ magnitude) produced by the explanation methods, evaluating on InceptionV3 model. . . . .	55
5.7	Rank correlation between different methods . . . . .	56
A.1	SSIM scores when evaluated on Pneumonia dataset. . . . .	66
A.2	Feature agreement scores when evaluated on Pneumonia dataset. . . . .	67
A.3	Sign agreement scores when evaluated on Pneumonia dataset. . . . .	68

# List of Tables

5.1	Experiment environments and requirements. . . . .	44
5.2	Comparison of the metric scores for every pair of the methods Guided Backpropagation, Guided Grad-CAM and Saliency. . . . .	52

# Abstract

Deep learning models have become increasingly complex and accurate, making them essential in various real-world applications like healthcare and security. Understanding the inner workings of these complex “black box” models is crucial for their appropriate use, creation, and trust. Explainable Artificial Intelligence (XAI) aims to make these complex systems interpretable through different methods, including saliency maps for convolutional neural networks. However, there is a growing concern about the “disagreement problem”, where explanation methods produce inconsistent and sometimes contradictory explanations. While this problem has been acknowledged for various explanation methods, disagreements within saliency methods have not been thoroughly explored. This thesis addresses this gap by measuring the extent of disagreement among popular saliency techniques using established metrics. We evaluated the disagreement for eight different saliency methods over two black boxes and found significant disagreement among saliency map explanations. We also observed complex patterns of disagreement among the experimented methods, and finally we indicated that the level of disagreement depends on the kind of black box being explained.

# Chapter 1

## Introduction

This chapter provides the context for our thesis and outlines the problem we aim to address. Firstly, we highlight the necessity for XAI in making complex models more interpretable to users and introduce saliency maps as a potential solution. We then present the disagreement problem, which arises when different XAI methods provide conflicting explanations for a complex black box. Ultimately, we state the goal of our thesis to explore the extent of disagreement among saliency methods, an area that has yet to receive extensive research attention.

### 1.1 Overview

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. Because in the end, deep learning systems can be unfair because they learn from data generated by humans, which inherently contain biases. Hence, comprehending the reasoning and decision-making processes of deep learning models is vital.

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. 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. This is crucial not only for transparency but also for establishing trust when employing artificial intelligence to address problems [10].

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 a newer, simpler model that can replicate 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

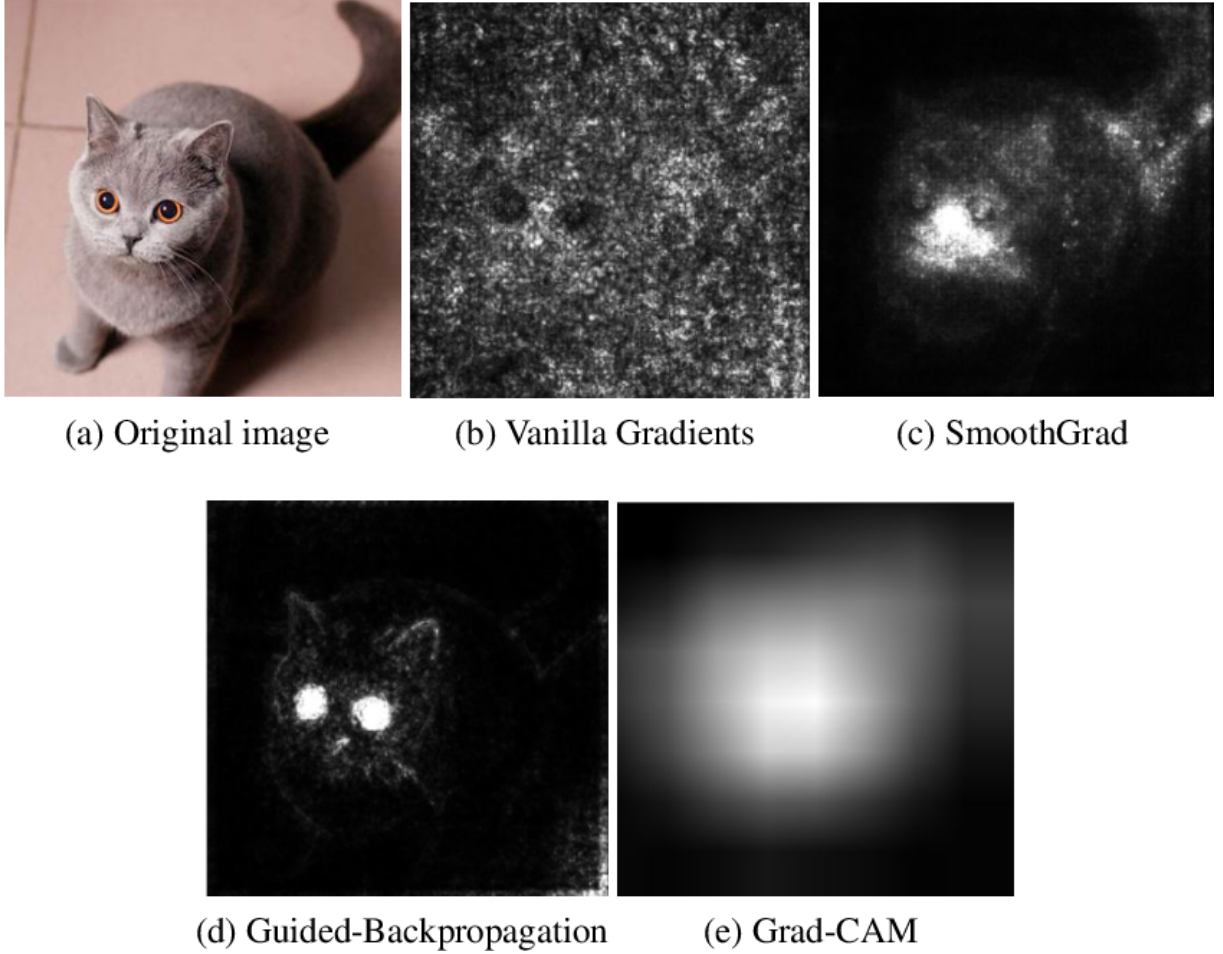


Figure 1.1: Different saliency maps produced by different XAI methods. The bright areas indicate important areas. [19]

or creating new versions. Figure 1.1 illustrates an example of using saliency map to explain a model prediction.

## 1.2 Problems

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.

The goal of XAI is to help users understand how complex AI systems work, thereby reducing incomprehensibility and increasing trust and confidence in the deployment of these systems. The disagreement problem poses a challenge to achieving this goal. If explanation methods do not agree on why the black box behaves in a certain way or provide conflicting explanations on what the black box considers important to its decision, it can cause confusion for users. Therefore, we believe this problem deserves the attention of XAI researchers and should be thoroughly studied.

## 1.3 Objectives

While the disagreement problem has been recognized for various explanation methods, saliency maps have not been extensively studied in this regard, as mentioned in 1.2. Saliency methods are widely used to explain convolutional neural network (CNN) black boxes. With numerous techniques available and no consensus evaluation framework to guide their development, it is highly likely that some level of disagreement exists among these methods. 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.

## 1.4 Methodology

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 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 such as structural similarity index measure (SSIM) [40], feature agreement, sign agreement, and rank correlation.

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:

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

## 1.5 Thesis Content

This thesis is organized into 6 chapters as followed:

- In chapter 1 we introduce the context for our thesis and the problem that we focus on. We present the need for XAI in making complex

models more interpretable to users and introduce saliency maps as a candidate solution. We then introduce the disagreement problem in which XAI methods differ from each other in explaining a complex black box. Finally, we state the goal of our thesis in attempting to study the disagreement between saliency methods, which currently lacks extensive research.

- In chapter 2 we provide an overview of XAI and some saliency-based methods. First, we cover XAI’s history, growth, and classification of approaches in XAI. Then, we introduce two types of saliency-based explanation methods: gradient-based and perturbation-based and provide an overview of their general ideas.
- In chapter 3 we discuss in detail the disagreement problem in XAI and introduce prior works that highlighted this problem. We also present several works on assessing the quality of saliency maps in practical use, since there is currently no work that highlights the disagreement between saliency-based methods specifically.
- In chapter 4 we present our approach to quantifying the disagreement between saliency-based methods. First, we describe the mathematical formulation for four metrics we used to measure disagreement, including feature agreement, sign agreement, rank correlation (inherited and adapted for saliency maps from [17]), and structural similarity index metric. Then, we select several methods (both gradient-based and perturbation-based) for generating explanations and two CNNs as black boxes (InceptionV3 and ResNet). We also provide an overview of the architecture of these two black boxes.
- In chapter 5 we provide a detailed description of our experiment in measuring disagreement, as well as a discussion about our findings. First, we describe the dataset, the environment configurations we used, the two CNNs that we used as black boxes for our experiment,

and how we measured disagreement. Then, we analyze the results and conclude some findings, which include a confirmation of disagreements among the explanation methods, a confirmation of complex disagreement patterns when evaluated on different black boxes, and an insight that the levels of disagreement are dependent on the kind of black box employed.

- In chapter 6 we conclude our thesis by providing a summary of our key findings, identifying some drawbacks in our work, and discussing future directions toward resolving the disagreement problem.

## Chapter 2

# Background

In this chapter, we set the foundation for our thesis by providing general knowledge about XAI, its history, and classify the field’s approaches. In addition, we also introduce some popular saliency-based methods, divided into gradient-based and perturbation-based approaches, and presents general overview about their workings.

### 2.1 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 [28, 36], when certain expert systems were designed to provide explanations for their outputs based on the rules they applied. Scientists have been engaged in continuous discussions about the significance of explanations in intelligent systems, specifically concerning decision-making, since the beginning of AI research.

In the context of modern deep learning, however, XAI has emerged as a novel research area in response to the increasing complexity and opacity of machine learning models. It aims to provide insights into AI systems’ decision-making processes and enable users to understand, trust, and effec-

tively interact with these systems. 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 [18]. 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. For instance, 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, 32]. However, it is important to note that these classification techniques are not fixed or absolute. They can differ significantly based on the specific characteristics of the methods, and methods may fall into multiple classes that overlap or do not overlap. In this context, different types of taxonomies and classification approaches will be briefly discussed, while a more comprehensive analysis of these taxonomies can be found in the referenced source [6].

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 with-

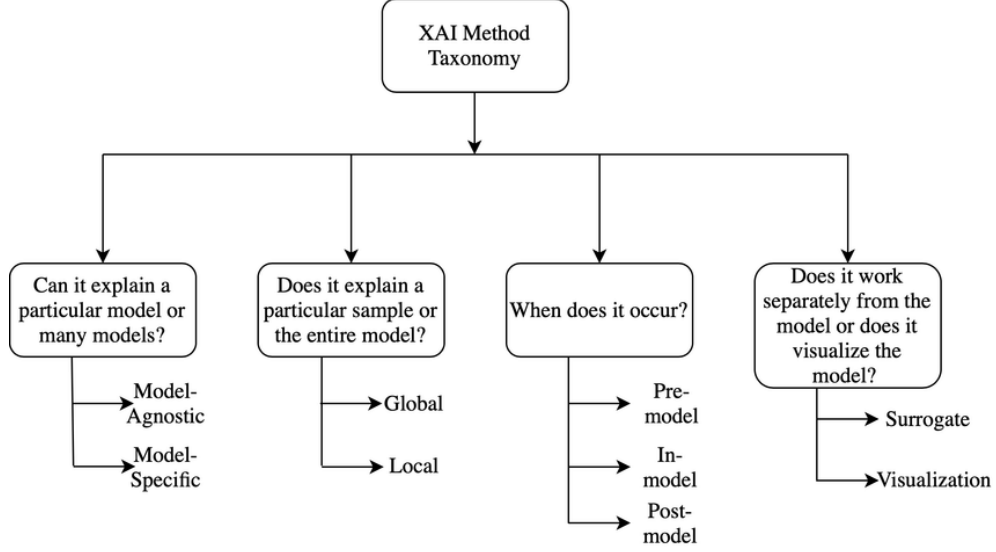


Figure 2.1: Taxonomy of XAI methods [6]

out modifying its internal workings. Post-hoc XAI methods provide an additional layer of interpretability by generating explanations that are independent of the underlying model’s architecture or learning algorithm.

Post-hoc XAI methods can be further categorized based on their approaches. A model-specific method is tailored to a particular machine learning model or architecture, leveraging its internal structure and characteristic to generate explanations. These methods include techniques like Layer-wise Relevance Propagation (LRP) [7], which rely on understanding the inner workings of deep neural networks. On the other hand, model-agnostic approaches, such as Local Interpretable Model-Agnostic Explanations (LIME) [24] and Shapley Additive Explanations (SHAP) [20], aim to provide explanations that are independent of the underlying model. These methods generate explanations by approximating the model’s behavior through sampling or perturbing the input data.

Another aspect of variation in XAI methods is the granularity of explanations. Some techniques focus on generating global explanations, providing insights into the overall behavior of the model across the entire dataset or feature space. Integrated Gradients [35] is an example of techniques that offer global explanations by assessing the impact of different features

on model predictions. In contrast, local explanation methods concentrate on explaining individual predictions or instances. These methods, such as LIME [24] and SHAP [20], highlight the specific features that influenced a particular prediction, providing more fine-grained explanations.

XAI methods also vary in terms of the types of explanations they generate. Some methods utilize textual or rule-based explanations to provide human-readable justifications for model decisions. Rule extraction techniques, for instance, aim to extract understandable rules from complex models, enabling users to comprehend the decision-making process. Other methods employ visualizations, such as saliency maps, to visually highlight important regions or features in the input data that contribute to model predictions. In this work, we will focus on this difference between explanations via saliency maps.

## **2.2 Overview of Saliency-based XAI Methods**

The concept of saliency maps [31] 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. This section outlines various saliency map explanation methods/algorithms that will be employed in our thesis. 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.

### **2.2.1 Gradient-based methods**

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]. Another advantage of gradient-based methods is that a single (or a few) forward pass and backpropagation allow us to evaluate the contributions of all features in the input (global contribution).

### 2.2.1.1 Vanilla Gradient/Saliency

Saliency method [31] was one of the initial approaches developed to visualize the input attribution of convolutional networks. As the term “saliency” is commonly associated with the overall approach of displaying input attribution known as the Saliency maps, this particular method is also referred to as Vanilla Gradient. The concept of the Saliency method originates from class visualization, which involves finding an image  $I$  that maximizes the score  $S_c$  for a particular class  $c$  while incorporating  $L2$  regularization. Formally, this can be expressed as follows:

$$\underset{I}{argmax}(S_c(I) - \lambda\|I\|_2^2) \quad (2.1)$$

In the case of deep CNN models, the class score  $S_c(I)$  is a highly non-linear function. However, we can approximate the class score  $S_c(I)$  with a linear function in the neighbourhood of an image  $I_0$  by utilizing the first-order Taylor expansion.

$$S_c(I) \approx w^T I + b \quad (2.2)$$

To compute the saliency map  $A$ , we just arrange the values in  $w$  to



match the shape of input image  $I$

$$A_{i,j} = \max_{ch} |w_{h(i,j,ch)}| \quad (2.3)$$

where  $ch$  is a color channel of the pixel  $(i, j)$  and  $h(i, j, ch)$  is an index of the  $w$  corresponding to the same pixel  $(i, j)$  and  $w$  is the derivative of  $S_c$  with respect to the image  $I$  at the image  $I_0$

$$w = \left. \frac{\partial S_c}{\partial I} \right|_{I_0} \quad (2.4)$$

### 2.2.1.2 Integrated Gradient

Integrated Gradients (IG) is an approach proposed by Sundararajan et al. [35] in 2017. This method is based on two axioms to which the authors assert that these principles should be upheld by all attribution methods: Sensitivity and Implementation Invariance. The specific definition of those two axioms is as follows:

- **Sensitivity:** An attribution method satisfies Sensitivity when it meets two conditions: The first condition is if we have two inputs and base-lines that differ in only one feature, but result in different predictions by the model, then the feature that distinguishes them should be assigned a non-zero attribution. In other words, when a single feature alteration leads to divergent model outputs, the attribution method should correctly identify and highlight the contribution of that differing feature.

The second condition is if the function implemented by the model does not depend (mathematically) on a particular variable or feature, then the attribution assigned to that variable should always be zero. In other words, if a feature has no influence on the model's predictions or decision-making process, it should not be attributed any significance or contribution.

- **Implementation Invariance:** The attribution values must be the same if two models are equivalent - Two models are considered functionally equivalent if their outputs are the same for all possible inputs, even if they have been implemented using different architectures or methodologies. The goal of Implementation Invariance is to ensure that the attributions produced by an attribution method remain consistent and identical when applied to functionally equivalent networks.

The main idea of this technique is to quantify the change in model output between an input  $x$  and a reference baseline  $x'$  (in the case of an image task, the baseline  $x'$  is often represented as a black image with all pixels set to 0). To compute the output variation, IG calculates the gradient of the model's output with respect to the input variables along the path from  $x'$  to  $x$ . The main idea is to integrate the gradients over this path to obtain a comprehensive measure of the input variable's impact on the model's output.

Let's assume our model is a mathematical function denoted as  $F : R^n \rightarrow [0, 1]$ , where  $n$  represents the number of input variables. The formula used to calculate the output variation attributed to each input variable,  $x_i$ , is as follows:

$$IG_i(x_i) = (x_i - x'_i) \times \int_{\alpha=0}^1 \frac{\partial F(x'_i + \alpha \times (x_i - x'_i))}{\partial x_i} d\alpha \quad (2.5)$$

By evaluating this integral for each input variable, this technique quantifies the contribution of each variable to the model's output variation between the input and the baseline. This provides valuable insights into the relative importance and influence of different input features on the model's decision-making process.

### 2.2.1.3 Grad-CAM

Grad-CAM (Gradient-weighted Class Activation Mapping) [29] is a more general version of CAM (Class Activation Mapping) [43]. With CAM,

the model architecture requires the use of Global Average Pooling (GAP) operation at the last convolutional layer followed by a fully-connected layer to make predictions. With this architecture, to explain the model’s decision, we utilize the results obtained after performing GAP on the feature map activations and weight them based on the weights of the final classification layer to generate a class activation map. Specifically, let  $F_k = \sum_{x,y} f_k(x,y)$  represent the result of applying GAP to the activation feature map  $k$  at the point  $(x,y)$ . The class activation map for class  $c$ , denoted as  $M_c$ , is calculated as follows:

$$M_c(x,y) = \sum_k F_k w_k^c \quad (2.6)$$

In this equation,  $w_k^c$  represents the weight associated with the  $k$ -th activation feature map for the class  $c$ . By summing up the weighted feature map activations, the Grad-CAM technique generates a class activation map that attributes the regions of the input image that contribute most significantly to the prediction for the target class.

By visualizing the class activation map, we can identify the spatial locations within the input image that the model focuses on when making predictions. These regions often correspond to the discriminative features relevant to the target class, providing insights into the model’s decision-making process.

Grad-CAM generalizes the CAM approach by eliminating the requirement of the model architecture. It leverages the gradient information flowing into the convolutional layers. This makes Grad-CAM applicable to a wide range of CNN architectures. The idea behind Grad-CAM remains using feature maps obtained from the chosen convolutional layer to calculate the contribution values from the input. The difference lies in weighting the feature maps using the gradient-derived value, denoted as  $\alpha$ . Specifically, the Grad-CAM algorithm consists of the following three steps:

Step 1: Calculate the variation of the output value for class  $c$ , denoted

as  $y_c$ , with respect to the activation feature map  $A^k$  in the desired convolutional layer that we want to explain.

Step 2: Compute the weight  $\alpha$  for the  $k$ -th activation feature map using the formula:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \quad (2.7)$$

Here,  $\frac{\partial y^c}{\partial A_{ij}^k}$  represents the partial derivative of the output class  $c$  with respect to the activation feature map  $A_{ij}^k$  at position  $(i, j)$ , and  $\frac{1}{Z} \sum_i \sum_j$  works as Global Average Pooling

Step 3: After obtaining the weights  $\alpha$ , we can calculate the Grad-CAM heatmap by summing the weighted feature map activations and applying the ReLU function. Specifically, the Grad-CAM heatmap for class  $c$  is computed as follows:

$$L^c = \text{ReLU} \left( \sum_k \alpha_k^c A^k \right) \quad (2.8)$$

The ReLU function is used here to consider only positive contribution values, while negative values are set to 0.

By following these steps, Grad-CAM generates a heatmap that highlights the important regions of the input image for the predicted class. These regions correspond to the areas where the model focuses its attention when making predictions. The higher the intensity in the Grad-CAM heatmap, the more influential the corresponding spatial locations are in determining the prediction for the target class.

#### 2.2.1.4 Guided Backpropagation

Guided Backpropagation (GBP) [34] is an explanation method based on the backpropagation mechanism. The idea behind GBP was developed by Alexey Dosovitskiy et al. building upon Deconvolution [42] and Saliency [31] techniques. The authors argued an issue with the Saliency

method proposed by Simonyan [31], which arises when negative gradient values are involved. These negative gradients could decrease the accuracy of deeper layers in the network being explained. To address this issue, the authors combined the two methods and introduced guidance into the Saliency method with the help of Deconvolution.

The term “guided” in this technique implies that we can selectively activate specific neurons to perform the backpropagation process. To achieve this, the method modifies the standard backpropagation step using the ReLU operation, allowing only positive gradients to be backpropagated. Negative gradient values are assigned a value of 0. Specifically, let  $f_i^l$  denote the feature map at layer  $l$ , and the backpropagation result at layer  $l$  is computed as follows:

$$\frac{\partial E}{\partial f_i^l} = \max \left( 0, \frac{\partial E}{\partial f_i^{l+1}} \right) \cdot \frac{\partial f_i^{l+1}}{\partial f_i^l} \quad (2.9)$$

Here,  $\frac{\partial E}{\partial f_i^l}$  represents the backpropagated error with respect to the feature map  $f_i^l$  at layer  $l$ . The ReLU operation ensures that only positive gradients are considered, filtering out negative gradients during the backpropagation process.

Guided Backpropagation is particularly useful when we want to focus on specific neurons or regions of the model during the explanation process. By selectively activating certain neurons and propagating positive gradients, GBP allows us to highlight the influential features and understand the model’s decision-making process at a more fine-grained level.

### 2.2.1.5 Guided GradCAM

Guided GradCAM is an XAI method that combines the principles of Grad-CAM 2.2.1.3 and Guided Backpropagation to provide more localized and fine-grained explanations of deep neural network predictions. Guided Grad-CAM inherits the strengths of Grad-CAM, which are class discrim-

ination and localization, with the ability to visualize by focusing only on the contributions in regions with positive gradients from the Guided Backpropagation method.

The main idea behind GuidedGradCAM is to generate a heatmap that represents the importance of each pixel or region in the input with respect to the predicted class. This heatmap is obtained by combining the gradient information from both the Grad-CAM and Guided Backpropagation methods. To combine these two methods, GuidedGradCAM utilizes the positive gradients obtained from Guided Backpropagation and multiplies them element-wise with the Grad-CAM weights. This multiplication enhances the relevance of positive features while suppressing the influence of negative features. The resulting guided Grad-CAM heatmap provides a more focused and accurate representation of the important regions in the input. Mathematically, let's denote the Grad-CAM weights for class  $c$  as  $W_c$  and the guided backpropagated gradients as  $L_{\text{guided}}$ . The GuidedGradCAM heatmap  $H_c$  for class  $c$  is obtained by element-wise multiplication of these two quantities:

$$H_c = \text{ReLU}(W_c \odot L_{\text{guided}}) \quad (2.10)$$

where  $\odot$  represents element-wise multiplication and ReLU ensures that only positive contributions are considered.

The resulting heatmap  $H_c$  highlights the regions in the input that have a strong positive influence on the prediction for class  $c$ . These regions indicate the areas that the model focuses on when making the prediction and provide insights into the important visual cues or features.

#### 2.2.1.6 Layer-wise Relevance Propagation

Layer-wise Relevance Propagation (LRP) was introduced by Bach et al. [7]. The fundamental idea behind LRP is to assign relevance scores to the input features or neurons based on their impact on the model predictions.

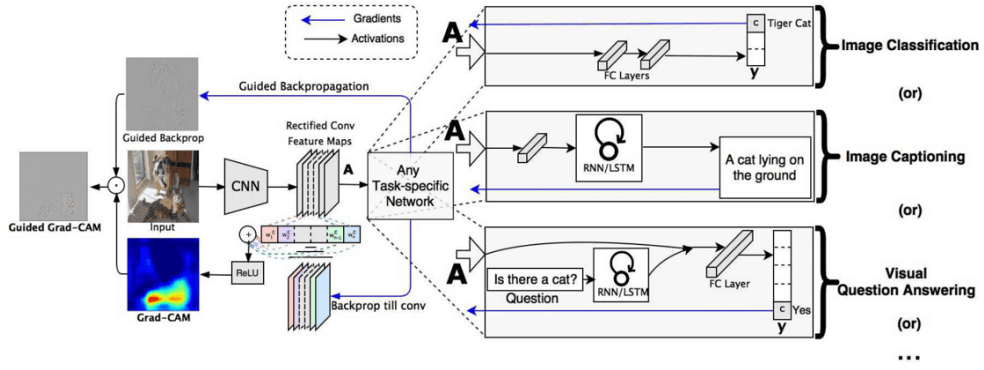


Figure 2.2: How Guided Grad-CAM heatmaps generated by combining Grad-CAM & Guided backpropagation method. Source [29]

LRP achieves this by redistributing the output relevance backward through the layers of the neural network.

LRP adheres to the conservation property, which states that the relevance received by a neuron must be evenly redistributed to the neurons in the lower layer. In simpler terms, whatever relevance a neuron receives, it must pass on an equal amount of relevance to the neurons in the next layer. This conservation property is one of the fundamental principles of LRP, although there are other properties that LRP also follows, this particular property is explicitly emphasized in its implementation. Let's denote  $j, k$  are two consecutive layers,  $z_{jk}$  quantifies the amount of contribution node  $j$  has to make node  $k$  relevant,  $z_{jk} = x_j^{(l)} w_{jk}^{(l, l+1)}$ , the guiding equation for propagating relevance scores is:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k \quad (2.11)$$

This relevance redistribution is guided by certain rules and principles that ensure the preservation of relevance throughout the network. One commonly used rule in LRP is the “epsilon rule” which solves the problem of gradient noise by introducing a small positive term,  $\epsilon$  to the denominator

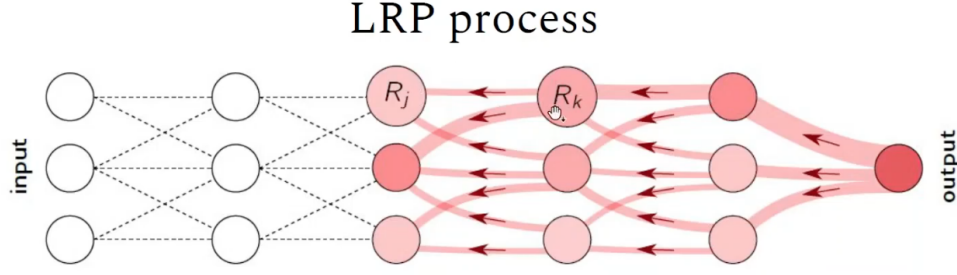


Figure 2.3: Relevance scores propagating procedure. Source [7]

The  $\epsilon$ -rule formula for LRP can be expressed as follows:

$$R_j = \sum_k \frac{z_{jk}}{\epsilon + \sum_i z_{ik}} R_k \quad (2.12)$$

An advantage of LRP is the consideration of both positive and negative contributions. This helps in understanding not only the positive factors that drive the model's prediction but also the inhibitory factors that may counteract certain features or neurons.

LRP is a broad family of methods, and different variants and modifications exist to adapt to various network architectures and specific application domains. The choice of propagation rules and modifications may vary depending on the specific requirements and interpretability objectives.

### 2.2.1.7 Deep Label-Specific Feature Learning

While LRP adhered to the conservation axiom, which ensures that the relevance received by a neuron is redistributed equally to the lower layer, there was still a challenge in determining how to allocate the net relevance among individual pixels within a layer. This ambiguity led to difficulties in accurately understanding the contribution of each pixel to the overall relevance. To address this issue, Deep Label-Specific Feature Learning (DeepLIFT) [30] introduces an additional axiom on how to propagate the relevance down.



DeepLIFT calculates the contributions of individual neurons by comparing their current activation with a reference activation. The reference activations for all neurons are determined by performing a forward pass, where the input parameters are propagated through the neural network. More details are presented in [30]. In simple terms, to explain a neural  $x$ , let's denote  $t$  as its current activations and  $t'$  as its reference activation. If the difference between these two values  $\delta t = t - t'$ , DeepLIFT assigns contribution scores  $C_{\delta x_i \delta t}$  to  $\delta x_i$  as follows:

$$\sum_{i=1}^n C_{\delta x_i \delta t} = \delta t \quad (2.13)$$

#### 2.2.1.8 Gradient SHapley Additive exPlanation

GradientSHAP (Gradient SHapley Additive exPlanation) [20] combines the gradient-based approach to attribution and the concepts of Shapley values from cooperative game theory.

In cooperative game theory, Shapley values are used to fairly distribute the contribution among players in a cooperative game. In the context of XAI, Shapley values measure the contribution of each feature to the prediction or output of the model.

To calculate the Shapley value of a feature, we consider all possible permutations of the features and measure their marginal contributions. In each permutation, we determine the difference in the prediction when the feature is included versus when it is not. By averaging these differences across all permutations, we obtain the Shapley value for that feature. Suppose  $F$  is the set of all features,  $S$  is all feature subsets of  $F$ . We compute the contribution value of a feature  $i$  as in the formula 2.14 represented in the work of Scott et al. [20].

$$\phi_i = \sum_{S \subset F \cup \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} (f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)) \quad (2.14)$$

where  $f_K$  is a model trained with the set of feature  $K$ .

With GradientSHAP, the Shapley values are computed using gradients. Gradients provide information about how changes in feature values affect the model's output. By considering the gradients of the model's predictions with respect to the input features, GradientSHAP determines the relative importance of each feature in the prediction process. Here's an overview of how the GradSHAP works:

1. Choose a reference dataset: Select a dataset that represents the background or baseline data distribution. This dataset serves as a reference point for calculating SHAP values.

2. Compute the expected model output: Calculate the average prediction of the model across the reference dataset. This provides the expected value of the model output, denoted as  $E[f(x)]$ .

3. Calculate the gradients: Determine the gradients of the model output with respect to the input features for a specific instance. Gradients indicate the rate of change of the model output for each feature.  $\nabla_x f(x)$ .

4. Combine the gradients with the expected values: To approximate the SHAP values, we linearize the model output around the expected values using the computed gradients. Therefore, the approximation of the SHAP value for each feature can be calculated as:

$$\phi_i = \sum_{j=1}^N (x_j - E[x_j]) \cdot \nabla_{x_j} f(x)$$

where  $\phi_i$  is the SHAP value for the  $i$ -th feature,  $N$  is the number of features,  $x_j$  is the value of the  $j$ -th feature for the input instance  $x$ , and  $E[x_j]$  is the expected value of the  $j^{th}$  feature calculated from the reference dataset.

5. Calculate the SHAP values: The above formula gives an approximation of the SHAP values for each feature. To improve the accuracy, we can use different reference instances for each feature and average the results. This can be done by sampling multiple times from the reference dataset and computing the SHAP values for each sample, then averaging the results. This process is called Monte Carlo sampling.

In summary, the SHAP Gradient Explainer uses the gradients of the model output with respect to input features to approximate SHAP values, which provide a fair distribution of the contribution of each feature towards the prediction for a specific instance.

### 2.2.2 Perturbation-based methods

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. By observing how the model’s predictions change in response to these perturbations, we can gain insights into the importance and impact of different features on the model’s decision-making process.

The process of perturbation-based analysis typically involves generating multiple perturbed versions of the input and evaluating the model’s predictions for each perturbed instance. This analysis provides a more comprehensive understanding of how variations in input features influence the model’s output. This characteristic 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” [33].

However, perturbation-based methods come with computational costs due to the need for generating and evaluating a large number of per-

turbed samples. The generated perturbations should ideally cover a diverse range of possible variations to ensure a robust analysis. Furthermore, perturbation-based methods require multiple forward passes which are computationally expensive compared to gradient-based methods since they need to experiment with a diverse set of generated perturbations. This computational burden limits the scalability and efficiency of perturbation-based methods, especially when dealing with complex models and high-dimensional input spaces.

It is worth noting that the results obtained from perturbation-based methods are specific to the generated sample and may not capture the model’s behavior under all possible samples in input space. Therefore, the interpretability provided by perturbation-based methods is constrained to the tested perturbations and may not fully represent the model’s general behavior.

### **2.2.2.1 Occlusion**

The main idea behind the Occlusion technique is to systematically occlude different regions of the input data and observe the resulting changes in the model’s predictions [42]. By comparing the predictions before and after occlusion, we can infer the importance or relevance of the occluded regions in influencing the model’s decision.

To apply the Occlusion method, we slide a predefined occlusion window across the input data, systematically covering different regions. The occlusion window can have various shapes and sizes depending on the nature of the input data. At each occlusion position, we replace the occluded region with a neutral value (e.g., zero or average value) and feed the modified input to the model to obtain a new prediction.

The difference between the original prediction and the prediction with the occlusion indicates the importance of the occluded region. Larger deviations in predictions suggest that the occluded region played a more influential role in the model’s decision-making process.

Mathematically, let's denote the original input as  $x$  and the occluded input as  $\tilde{x}$  obtained by occluding a specific region. The model's prediction on the original input is  $f(x)$ , and the prediction on the occluded input is  $f(\tilde{x})$ . The occlusion score, representing the importance of the occluded region, can be calculated using a similarity measure, such as the difference in predictions:

$$\text{Occlusion Score} = f(x) - f(\tilde{x}) \quad (2.15)$$

The advantage of Occlusion is that it does not rely on gradient information and can be applied to various types of models, including non-differentiable models. However, its effectiveness depends on the choice of the occlusion window size, shape, and stride, as well as the definition of the neutral value used for occlusion. Additionally, occluding small, local regions may not capture the global interactions between features in the input, potentially leading to limited interpretability.

## Chapter 3

# Related Works

In this chapter, we introduce the notion of the disagreement problem in XAI. Next, we explore some related research on this problem and show how there is a lack of extensive studies on the disagreement between general XAI methods, not to mention methods that utilize saliency maps. Then, we explore progress in the field regarding the assessment of the quality of saliency maps.

### 3.1 The Disagreement Problem

The problem of disagreement in the field of explainable artificial intelligence is a relatively recent issue. Brughmans et al. [9] characterized it as the occurrence of “conflicting or contradictory explanations” generated by different interpretation methods when assessing a given AI model. Roy et al. [25] 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. [22] is among the earliest works that have raised concerns about the issue of explanation disagreement. The authors utilized rank correlation, specifically Kendall’s  $\tau$  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.

Krishna et al. [17] went further by formularizing the concept of disagreement problem based on real-world scenarios. Their research involved surveying practitioners, including data scientists and machine learning engineers, to investigate their use of XAI methods in their daily work, their experiences with disagreements among different methods, and how they address such discrepancies. The findings revealed that practitioners often employ multiple explanation methods rather than relying on a single one to comprehend the functioning of black box models. Furthermore, the practitioners’ consideration of feature importance values varied, with factors such as mismatches in top features, feature order, and the sign of feature importance (whether positive or negative) taking precedence over the actual values generated by XAI methods. This observation is understandable given that XAI methods produce values with diverse meanings and properties. Using the above observations, the authors developed several metrics that capture the degree of disagreement between different XAI methods, which have laid the groundwork for further research on the problem. Our work is also influenced by this body of research.

Brughmans et al.’s study [9] investigates the issue of disagreement among counterfactual explanation algorithms, which describe a set of feature changes that can affect the predicted class of a black box model. In other words, such explanations explore the question of “what would the prediction be if a particular feature were to change in a certain way?”. This approach is different from saliency maps in which counterfactual explanations explain the prediction using *changes* in input features, while saliency maps interpret by *attributing* each pixel a feature importance value. The research confirmed the presence of disagreement among counterfactual ex-

planation methods and highlighted a notable discovery that the degree of disagreement is predominantly contingent on the dataset and methods utilized, rather than the type of classifiers employed.

Authors	Nb. of datasets	Data types	Nb. of xAI methods	Nb. of models	XAI
Neely et al. [1]	5	Text	6	2	LIME, Integrated Gradients (IG), DeepLift, Grad-SHAP, Deep-SHAP & attention-based explanations
Krishna et al. [2]	4	Tabular, Text & Image	6	8	LIME, Kernel SHAP, IG, Smooth-Grad, Input*Grad, Gradient
Roy et al. [3]	4	Tabular	2	1	LIME, SHAP
Brugmanns et al. [4]	40	Tabular	12	2	SHAP, Counterfactuals & Anchor

### 3.1.1 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. [17] 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). Brugmanns et al. [9] 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. [22] utilize Kendall’s  $\tau$  coefficient and Krishna et al. use Spearman’s  $\rho$  to measure the correlation between the relative ranking of the features.



In general, although differ in motivation, all the metrics above generally do not stress the importance of the values of the attribution, but instead, focus on the signs and the relative ordering of the features.

### **3.1.2 Attempts in Resolving Disagreement**

The attempts to address these disagreements have received scant attention in the literature. In practice, when faced with such disagreements, real-world practitioners often resort to selecting the method with which they are most familiar [17]. [25] 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.

## **3.2 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. Saliency maps, in general, are considered a valuable tool for facilitating comprehension of black box systems. 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. The authors recommend

using complementary global explanation methods in tandem with saliency maps to gain a more complete understanding of the complex black box system.

In medical imaging, saliency-based XAI methods play a predominant role [8] and are usually go-to techniques when explaining for clinicians [23]. Since its introduction in 2013 as a visualization method for explaining ConvNets architectures [31], the application of saliency maps to explain image-related tasks in the medical field has become increasingly popular. As mentioned in Section 1.3, several studies have been conducted to evaluate whether current image explanation algorithms meet the accuracy requirements in the medical field, such as [38, 5, 14]. Almost all of these studies share the same perspective as Cynthia Rudin’s research [23], 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 [26].

Concerns about the performance, robustness and fidelity of saliency methods have also been questioned. Nishanth et al. [5] found that many gradient-based methods demonstrate poor trustworthiness when used for high-stake domain of medical imaging. Eitel et al. [11] 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, 12, 16].

In summary, although saliency maps are very popular with multiple different techniques proposed, being an influential tools in many domains, there are still a lot of problems remain to be tackled regarding the usability of saliency maps.

# Chapter 4

## Method

In this chapter, we outline our approach for measuring the level of disagreement between various saliency-based methods. Firstly, we explain the mathematical formula behind four metrics that we used, which include feature agreement, sign agreement, rank correlation (inherited and adapted for images from [17]), and structural similarity index metric. Next, we select several saliency-based methods for generating explanations, including both gradient-based and perturbation-based. We also discuss the selection of two black box convolutional neural networks (CNNs) - InceptionV3 and ResNet - and provide an overview of their architecture.

### 4.1 Overview

Figure 4.1 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 chapter 5. 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

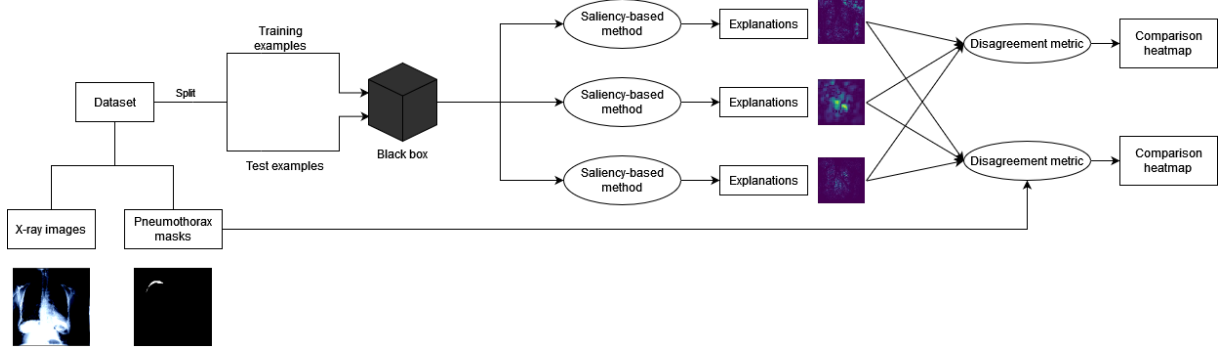


Figure 4.1: An overview of the steps we’ve taken to quantify the disagreement between the explanation methods. These steps are applied per black box.

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 in chapter 5, we applied this procedure to two different black boxes and analyzed the outcomes for both of them.

## 4.2 Mathematical Formulation

In this section, we establish certain mathematical notations that will be utilized to clarify the disagreement metrics.

### 4.2.1 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  $c$ :

$$e = c(b, x) \quad (4.1)$$

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  $c$ .

## 4.2.2 Metrics

### 4.2.2.1 Structural Similarity Index Metric

To capture the similarity in structure between two saliency maps, we use the Structural Similarity Index Metric (SSIM) [40]. 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:

$$\text{SSIM}(x, y) = [l(x, y)]^\alpha [c(x, y)]^\beta [s(x, y)]^\gamma \quad (4.2)$$

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;  $\alpha$ ,  $\beta$ , and  $\gamma$  are constants 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  $\mu_x$ ,  $\mu_y$  be the mean value of the saliency map  $x$  and  $y$ ;  $\sigma_x$ ,  $\sigma_y$  be the standard deviation of the  $x$  and  $y$ ;  $\sigma_{xy}$  be the covariance of  $x$  and  $y$ ;  $C_1$ ,  $C_2$  and  $C_3$  be small constants.

The luminance component is defined:

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (4.3)$$

Next, the contrast component is defined as:

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (4.4)$$

Finally, the structure component is defined as:

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (4.5)$$

The constants  $C_1$ ,  $C_2$  and  $C_3$  are typically set to small values in order to avoid zero division.  $\alpha$ ,  $\beta$  and  $\gamma$  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.

#### 4.2.2.2 Feature Agreement

We utilize the feature agreement introduced by [17], 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:

$$\text{FA}_k(e^{(1)}, e^{(2)}) = \frac{|\text{top}_k(e^{(1)}) \cap \text{top}_k(e^{(2)})|}{k} \quad (4.6)$$

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

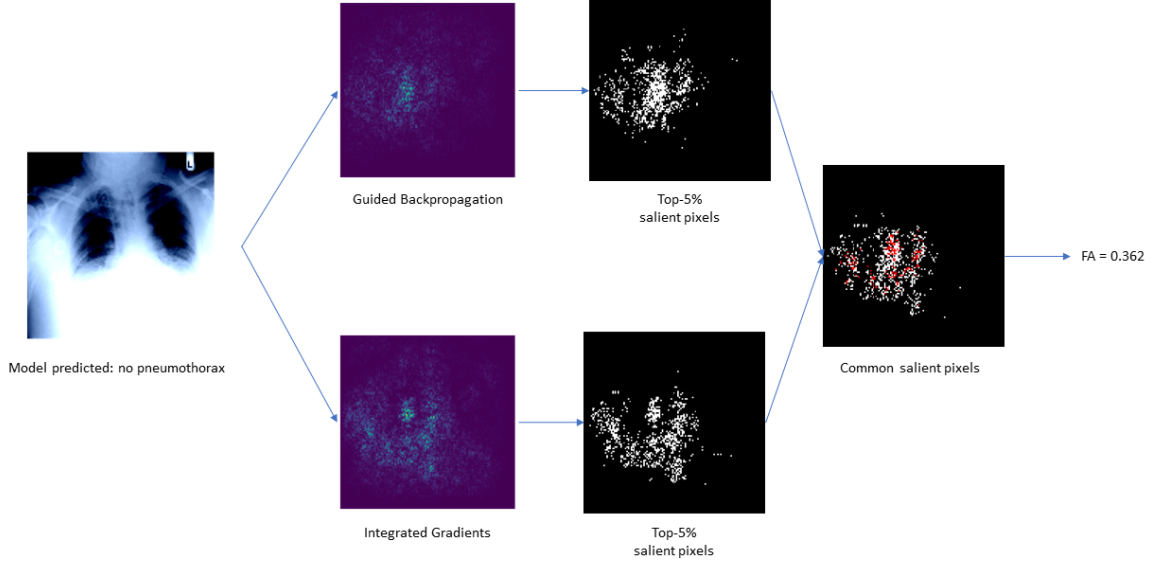


Figure 4.2: Illustration of how feature agreement was measured for an example pair of saliency maps generated by Guided Backpropagation and Integrated Gradients. We are interested in the top 5% of the most important pixels ( $k = 820$  in total, colored white). The feature agreement score is calculated by determining the number of common top salient pixels (highlighted in red) and dividing this by  $k$ .

- The 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 4.2 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  $128 \times 128$  image, this is equivalent to  $k = 820$ ).

#### 4.2.2.3 Sign Agreement

The sign agreement is quite similar to feature agreement, however, it is a stricter criterion [17]. The sign agreement between two saliency maps  $e^{(1)}$  and  $e^{(2)}$  with respect to the top- $k$  features is:

$$\text{SA}_k(e^{(1)}, e^{(2)}) = \frac{|\{(i, j) | (i, j) \in \text{top}_k(e^{(1)}) \cap \text{top}_k(e^{(2)}) \wedge e_{ij}^{(1)} \cdot e_{ij}^{(2)} > 0\}|}{k} \quad (4.7)$$

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).

#### 4.2.2.4 Rank Correlation

Another metric we adopted from the work of Krishna et al. [17] 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:

$$m(d) = \{d_{ij} | m_{ij} = 1, \forall i \in [1, m], \forall j \in [1, n]\} \quad (4.8)$$

Given a saliency map  $e$ , we define  $\text{rank}(e)$  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:

$$\text{RC}(e^{(1)}, e^{(2)}, m) = r_s \left( m(\text{rank}(e^{(1)})), m(\text{rank}(e^{(2)})) \right) \quad (4.9)$$



## 4.3 Experiment Subjects

### 4.3.1 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 GradCAM, Integrated Gradients, DeepLift, and LRP. Specific hyperparameters were required for some of these methods, which we specified as follows:

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

### 4.3.2 Black boxes

We opted to use XAI methods on the InceptionV3 [37] and Resnet101 [13] 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. These techniques leverage the inherent structure and characteristics of CNNs to produce visual explanations and highlight significant regions in an image. Additionally, we chose InceptionV3 and Resnet101 because of their demonstrated performance in previous work conducted by Narin, Ali et al. in their study on COVID-19 detection using X-ray images [21]. 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. By applying XAI methods to these architectures, we aim to gain a deeper understanding of their performance in medical imaging tasks.

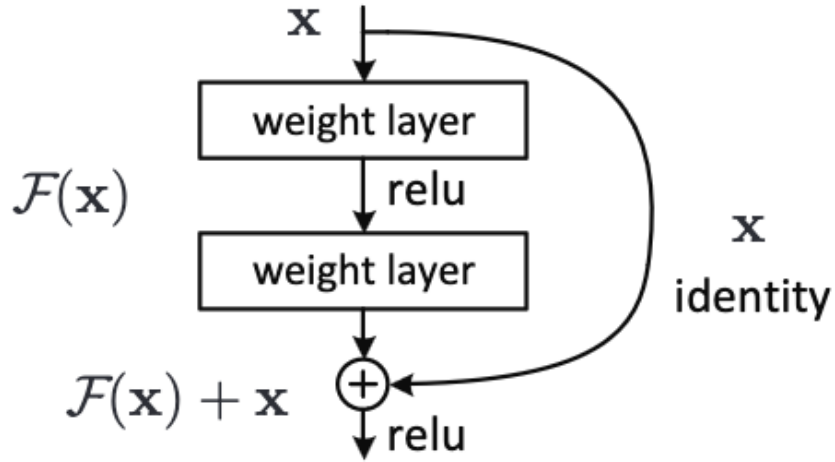


Figure 4.3: Residual Block Diagram. Source [13]

#### 4.3.2.1 Residual Network - 101

ResNet-101 is a convolutional neural network architecture that was introduced by He et al. in 2015 [13]. The residual network was developed to address the challenge of training very deep neural networks by utilizing residual connections.

The key idea of residual networks lies in their residual block design, which allows for the effective training of extremely deep models. In traditional deep networks, the increasing depth often leads to performance degradation due to the difficulty of learning mappings from input to output. A residual network solves this issue by introducing skip connections, or shortcut connections, that allow the network to learn residual mappings. By propagating gradients through these shortcuts, the model can effectively capture residual information and learn more efficiently.

The ResNet-101 architecture consists of 101 layers, including convolutional layers, pooling layers, and fully connected layers. The model employs a bottleneck structure in each residual block, reducing computational complexity while maintaining performance. The architecture has demonstrated impressive performance on various image recognition tasks, including the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) [27], where

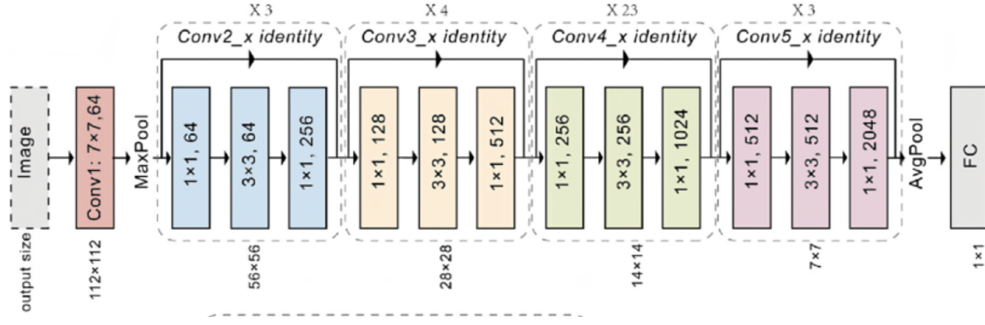


Figure 4.4: Resnet101 Architecture Diagram. Source [39]

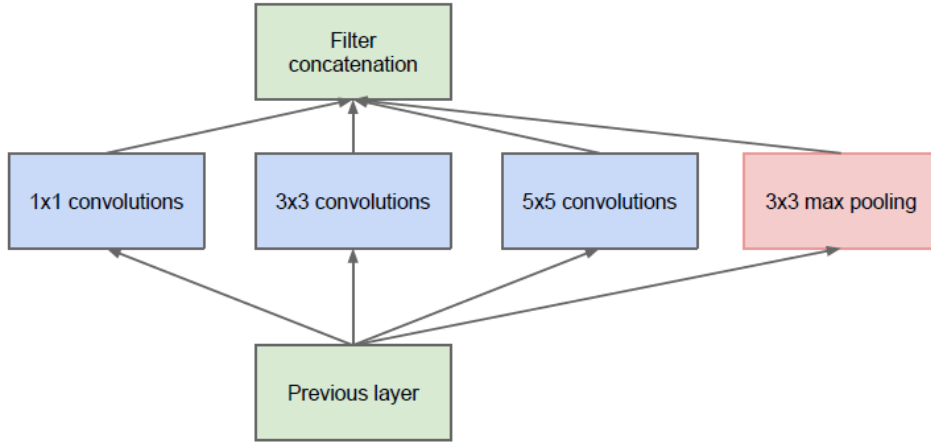


Figure 4.5: Naive Form of Inception Block. Source [13]

it achieved state-of-the-art results. With its ability to effectively train very deep models, ResNet-101 has contributed significantly to advancements in the field of deep learning and medical image analysis.

#### 4.3.2.2 InceptionV3

The InceptionV1 model addresses the problem of overfitting that arises when using deep layers of convolutions by employing parallel layers of multiple filter sizes at the same level. This wider architecture prevents the model from becoming excessively deep, mitigating overfitting risks.

InceptionV3 further enhances the original Inception design with several modifications. One key modification is the factorization of the  $5 \times 5$  convolutional layer into two  $3 \times 3$  convolutional layers, reducing computational

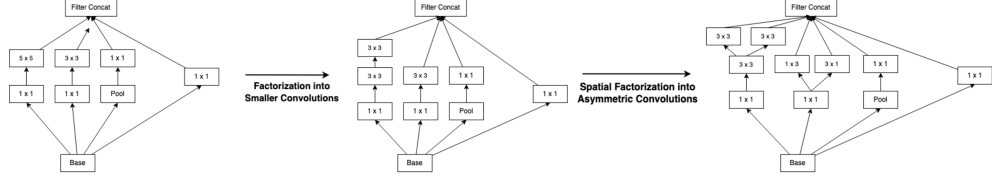


Figure 4.6: Some major optimizations of InceptionV3 model. Source [37]

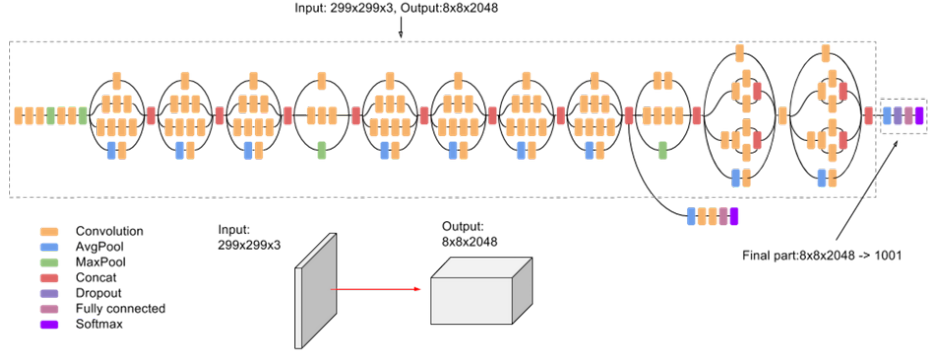


Figure 4.7: InceptionV3 Architecture. Source [37]

costs while maintaining expressive power. Another improvement involves spatial factorization, replacing  $n \times n$  convolutions with a  $1 \times n$  convolution followed by an  $n \times 1$  convolution. This two-layer solution reduces computational expenses by 33% when the number of input and output filters is equal.

The InceptionV3 model also introduces the use of auxiliary classifiers, which act as regularization techniques for the architecture. These auxiliary classifiers aid in training by providing additional supervision during the learning process. Additionally, efficient grid size reduction is achieved through the utilization of two parallel blocks of convolution and pooling, which are later concatenated.

Another modification is the reduction in the grid size of feature maps by employing two parallel blocks of convolution and pooling. This design allows for a more efficient reduction in grid size while maintaining the expressive power of the network. Specifically, if we start with a  $w \times w$  grid with  $k$  filters, after reduction, it results in a  $w/2 \times w/2$  grid with  $2k$  filters. By expanding the activation dimension in this manner, InceptionV3

ensures more efficient use of network capacity and computation. The two parallel blocks of convolution and pooling operate independently, capturing different types of information from the input. These blocks are then concatenated, combining the extracted features in a complementary way

These modifications in InceptionV3 enable improved performance and computational efficiency, making it a powerful model for various computer vision tasks. The Inception family of models has contributed significantly to the field of deep learning, demonstrating the effectiveness of wider architectures and factorization techniques in achieving state-of-the-art results.

### **4.3.3 Dataset**

In our study, we utilize two distinct datasets for the purpose of conducting a chest X-ray image classification task. By employing multiple datasets, we aim to enhance the generalizability of our findings.

#### **4.3.3.1 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 [41]. 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. This division facilitates the assessment of model performance on unseen data and enables the selection of optimal hyper-parameters.

#### 4.3.3.2 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 [15]. 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.

## Chapter 5

# Experiments & Result

In this chapter, we provide a detailed description of our experiment in measuring disagreement, as well as a discussion about our findings. First, we describe the dataset, the environment configurations we used, the two CNNs that we used as black boxes for our experiment, and how we measured disagreement. Then, we analyze the results and conclude that there are indeed disagreements among the explanation methods.

### 5.1 Experiments

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 4.3.3. 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.

The development environment in which the experiments were conducted are summarized in table 5.1.

Table 5.1: 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

### 5.1.1 Training Stage

In the training stage, we trained the InceptionV3 [37] and Resnet101 [13] models using popular tools such as pytorch, torchvision, scikit-learn, and pandas. Both models were initially pre-trained on the ImageNet dataset [27], which provided them with a strong foundation in recognizing and classifying various features. To adapt these models to our specific task of pneumothorax classification, we further trained them using the Chest X-ray Images with Pneumothorax Masks dataset [41]. 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  $\mu$  is (0.485, 0.456, 0.406) and its standard deviation  $\sigma$  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 ResNet101. This resizing ensured that the images were compatible with the input requirements of each model. The models were trained using 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 mon-



itored the models’ progress and selected the best-performing state based on their performance on the validation set.

### **5.1.2 Analysis Stage**

In the XAI analysis stage, we employed XAI methods provided by the Captum framework to generate explanations for the predictions made by our trained models. To measure the disagreement between these explanations, we implement some metrics introduced in 4.2.2 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.

## **5.2 Results**

In this section we discuss the results of our experiment for each of the disagreement metric. First, we analyze the disagreement scores of the explanation methods on both dataset separately. Then, we compare the levels of disagreement from both datasets and discuss some insights by looking at the overall view of the disagreement. Finally, we provide some remarks on the potential reason for the pattern of disagreement between the methods.

### **5.2.1 Pneumothorax Dataset**

#### **5.2.1.1 Structural Similarity Index Measure**

Figure 5.1 presents a comparison of SSIM scores for various explanation methods on the Pneumothorax dataset 4.3.3.1. These scores were obtained by averaging the results over the test dataset for each pair of saliency

map explanations. The computation of the SSIM score involves using the magnitude of the saliency maps, which is normalized to the range of  $[0, 1]$ .

The results show that the Guided Backpropagation and Guided GradCAM methods produce the most similar saliency maps among all method pairs. Both methods exhibit high similarity on both InceptionV3 and ResNet classifiers, with scores of 0.99 and 0.92, respectively.

Overall, the difference in SSIM scores between the two black boxes indicates that the inner architecture of the black box to some extent influences the appearance of the output explanations. Most of the scores are lower in ResNet, particularly the pairs Saliency–Guided Backpropagation and Saliency–Guided GradCAM, which show a stark difference in SSIM scores (0.71 and 0.72 reductions, respectively). Occlusion consistently shows lower similarity compared to other pairs for both black boxes because there is a difference in the granularity of the explanations produced by Occlusion and other methods. Figure 5.2 provides a comparison of the shapes of various saliency maps, where the saliency maps produced by Occlusion attribute in windows of pixels, while other saliency maps attribute down to the level of each pixel.

### 5.2.1.2 Feature Agreement

Figure 5.3 depicts the disagreement between different experiment explanation methods with varying top- $k$  values. The experiments reveal that there is a varying degree of disagreement, but the pair Guided Backpropagation–Guided GradCAM exhibits consistent agreement with each other. Additionally, as  $k$  increases, the degree of agreement tends to increase as well.

For InceptionV3, we generally observe that methods with structurally similar saliency maps produce agreeable explanations, while non-structurally similar methods produce higher disagreement. However, this pattern is not consistent when evaluating on ResNet. The agreement scores of LRP with Guided GradCAM and Guided Backpropagation are

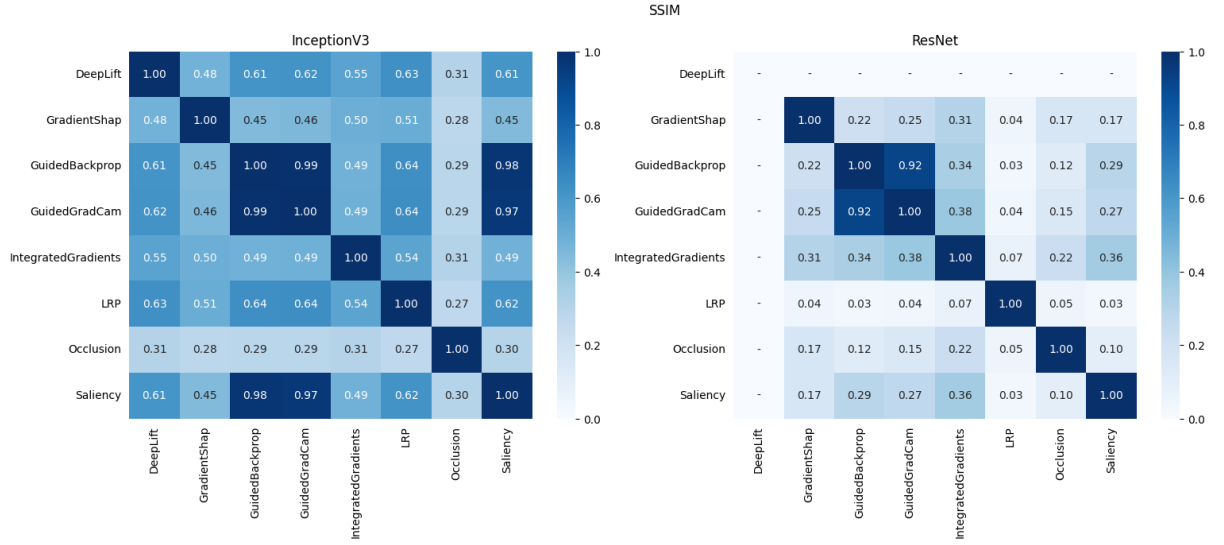


Figure 5.1: Pairwise structural similarity indices (SSIM) between explanation methods generating for both black boxes (InceptionV3 and ResNet-101). The SSIM scores are computed using the average score over test set data points. Darker colors imply greater similarity in the structure of the two explanations.

significantly lower, despite their similar structure. Moreover, the pairs between LRP and Saliency, respectively, with Guided Backpropagation and Guided GradCAM show the highest reduction in agreement scores. With  $k = 1638$ , for Saliency, the agreement scores with Guided Backpropagation and Guided GradCAM decrease from 0.92 and 0.91 to 0.36 and 0.38, respectively, with InceptionV3 and ResNet-101. In the case of LRP, the scores are both 0.55 on the two black boxes, down to 0.12 both, respectively.

### 5.2.2 Sign Agreement

As the conditions become more stringent, the amount of disagreement tends to increase. Figure 5.4 displays the average sign agreement score for pairs of explanation methods. Although the agreement scores for pairs that exhibit high agreement in feature agreement (such as Guided Backprop-

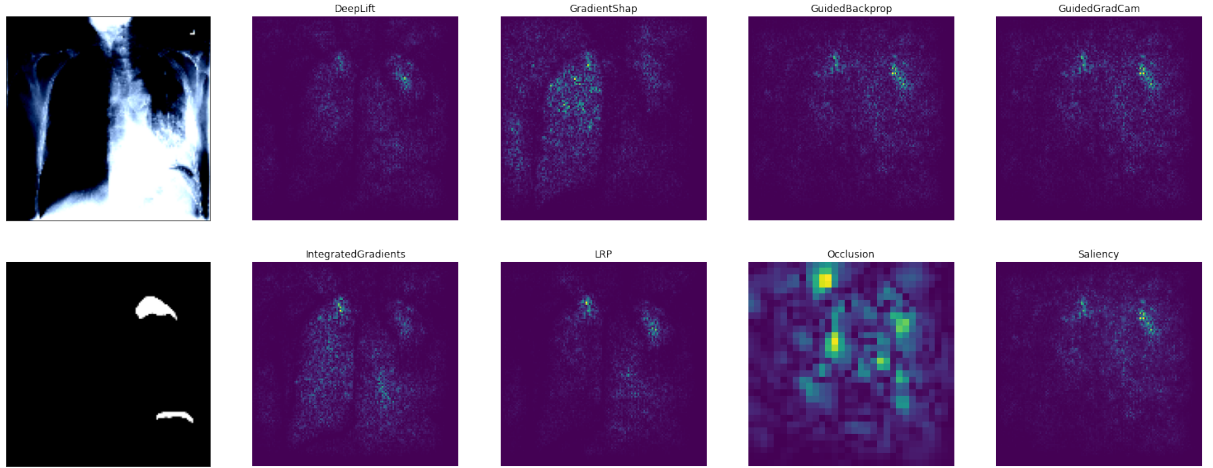


Figure 5.2: Comparison of the saliency maps generated by the experimental methods with respect to InceptionV3 prediction. Lighter colors indicate more salient pixels.

Guided GradCAM, Saliency–Integrated Gradients, etc.) decrease, they are still significant compared to other pairs. Many pairs demonstrate very low agreement, with scores lower than 0.1.

An interesting observation from this result is that most of the sign agreement scores for all pairs are approximately half of the corresponding feature agreement scores in both black boxes. This is because most explanation methods (with the exception of Occlusion and Saliency) generate saliency maps with the distribution of the attribution scores having a mean near 0. Figures 5.5 and 5.6 illustrate that all gradient-based methods distribute negative and positive values similarly among the most important pixels.

This suggests that the metric sign agreement may not provide useful information when used as a disagreement measure for saliency maps. However, our observations do not negate the potential use of this metric for other types of explanations. We urge further research on different types of explanation methods to confirm its usefulness.

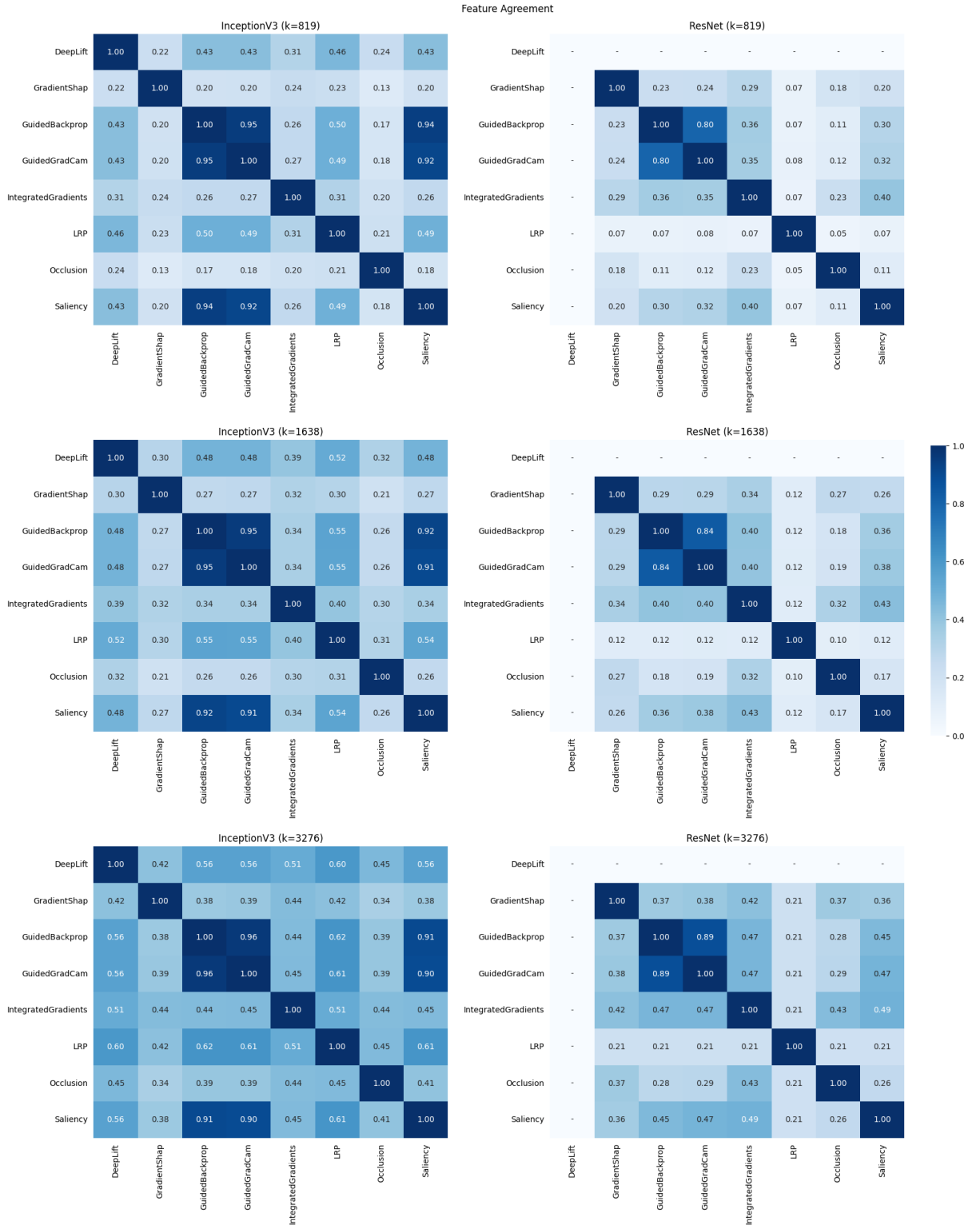


Figure 5.3: Feature agreement between different methods with varying top- $k$  value.

### 5.2.2.1 Rank Correlation

This section we present the disagreement scores between explanation methods when using the metric rank correlation. The scores are summarized in figure 5.7 and are computed on instances where pneumothorax masks are available (i.e., instances where the ground-truth label is has-pneumothorax). Computing the rank correlation score with respect to the ground-truth masks provides insight into how agreeable the explanation methods are in considering the relative importance of pixels in the ground-truth area. In simpler terms, the metric indicates whether two saliency maps consistently consider the importance ranking of pixels in the actual pneumothorax area for the black box’s decision.

The results show that highly agreeable pairs, as shown in previous experiments, remain consistent with each other in rank correlation, while pairs that strongly disagree with each other also remain that way.

### 5.2.3 Pneumonia Dataset

We refer to Appendix A for the disagreement heatmaps depicting the evaluation of explanation methods on the Pneumonia dataset. In this section, we analyze the insights derived from these heatmaps and draw parallels to the patterns of disagreement discussed in Section 5.2.1. It is important to note that the metric rank correlation is not applicable to this dataset due to the absence of segmentation masks.

#### 5.2.3.1 Structural Similarity Index

Figure A.1 displays a comparison of SSIM scores for different explanation methods applied to the Pneumonia dataset (see subsection 4.3.3.2). A notable observation is the significant difference in similarity scores between the explanation methods when applied to InceptionV3 and ResNet-101. For InceptionV3, most pairs of explanation methods yield high SSIM scores

(above 0.6), indicating a high level of similarity. In contrast, for ResNet-101, only four gradient-based methods, namely GradientSHAP, Guided Backpropagation, Guided Grad-CAM, and Integrated Gradients, exhibit moderate to high similarity. The remaining pairs demonstrate minimal to negligible similarity.

### 5.2.3.2 Feature Agreement & Sign Agreement

Figure A.2 and A.3 present the feature agreement and sign agreement scores for the methods on the Pneumonia dataset. Similar to the observations in Figure 5.3 and 5.4, the patterns of disagreement are repeated in the Pneumonia dataset. Pairs of methods that exhibit high agreement with each other, such as Guided Backpropagation and Guided Grad-CAM, consistently maintain their high agreement scores. The decrease in agreement scores for the Saliency pairs and the aforementioned methods, as discussed in Section 5.2.1.1 and illustrated in Figure 5.3, is also evident in the Pneumonia dataset.

### 5.2.4 Overall Discussion

One key observation from the analysis of the four metrics is that the agreement scores exhibit considerable variations when evaluating different black boxes. In particular, the disagreement exists more when explaining ResNet-101 than for InceptionV3. This observation also stands for all metrics and datasets.

However, when considering the same black box, the relative level of disagreement between pairs of explanation methods tends to remain consistent across all metrics and dataset (except for SSIM). This observation implies that the extent of disagreement is more influenced by the specific type of black box being explained rather than the source dataset used to train the black box. This finding contradicts the claim made in [9] that the magnitude of the disagreement issue is independent of the classifier.

It suggests 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. While there are variations in the levels of disagreement across different metrics, black boxes, and datasets, some pairs of explanation methods exhibit remarkable consistency. Particularly noteworthy are the three methods: Guided Grad-CAM, Guided Backpropagation, and Saliency. Table 5.2 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 InceptionV3, but consistently show disagreement when explaining ResNet-101, often by a significant margin.

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

Table 5.2: Comparison of the metric scores for every pair of the methods Guided Backpropagation, Guided Grad-CAM and Saliency.



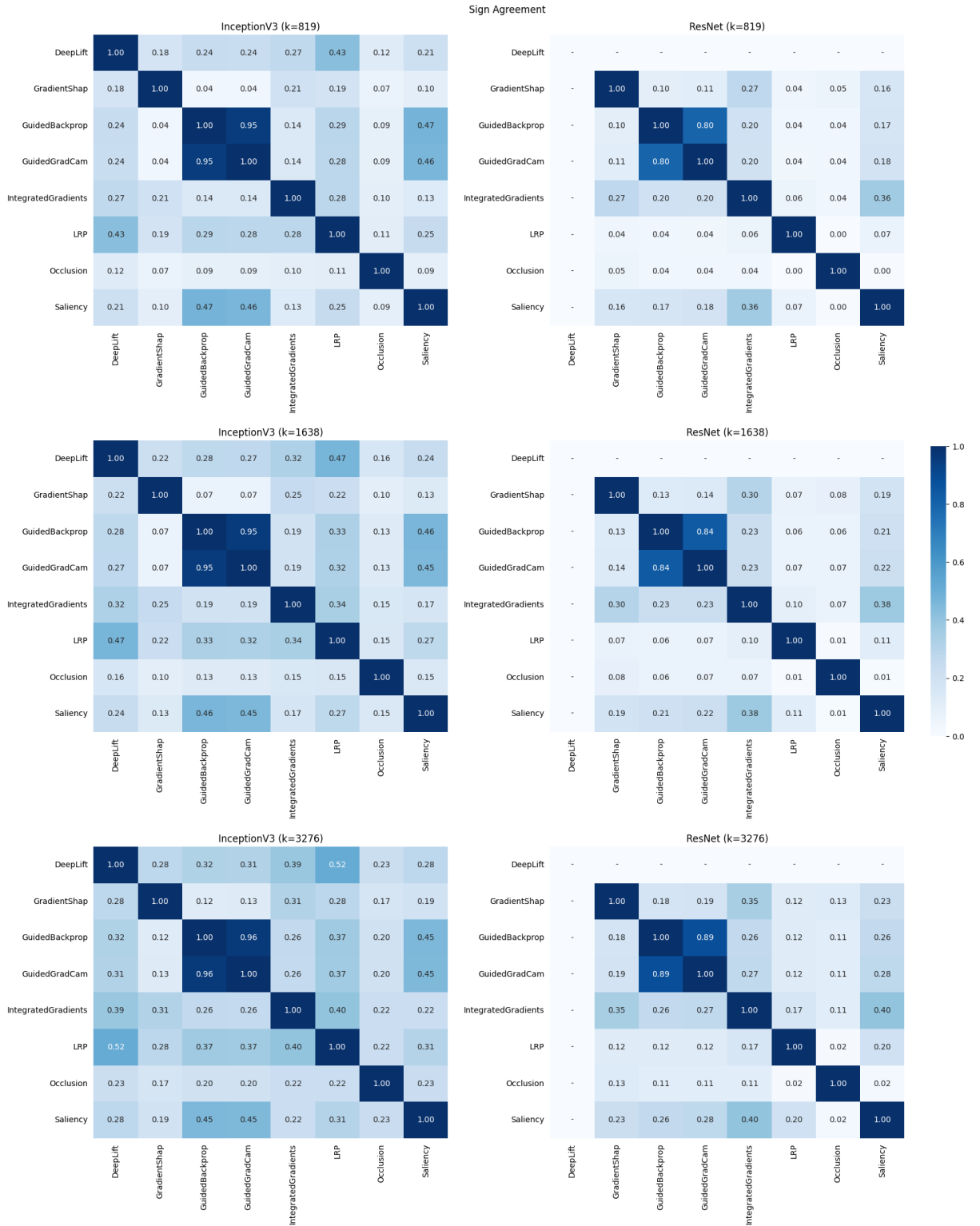


Figure 5.4: Sign agreement between different methods.

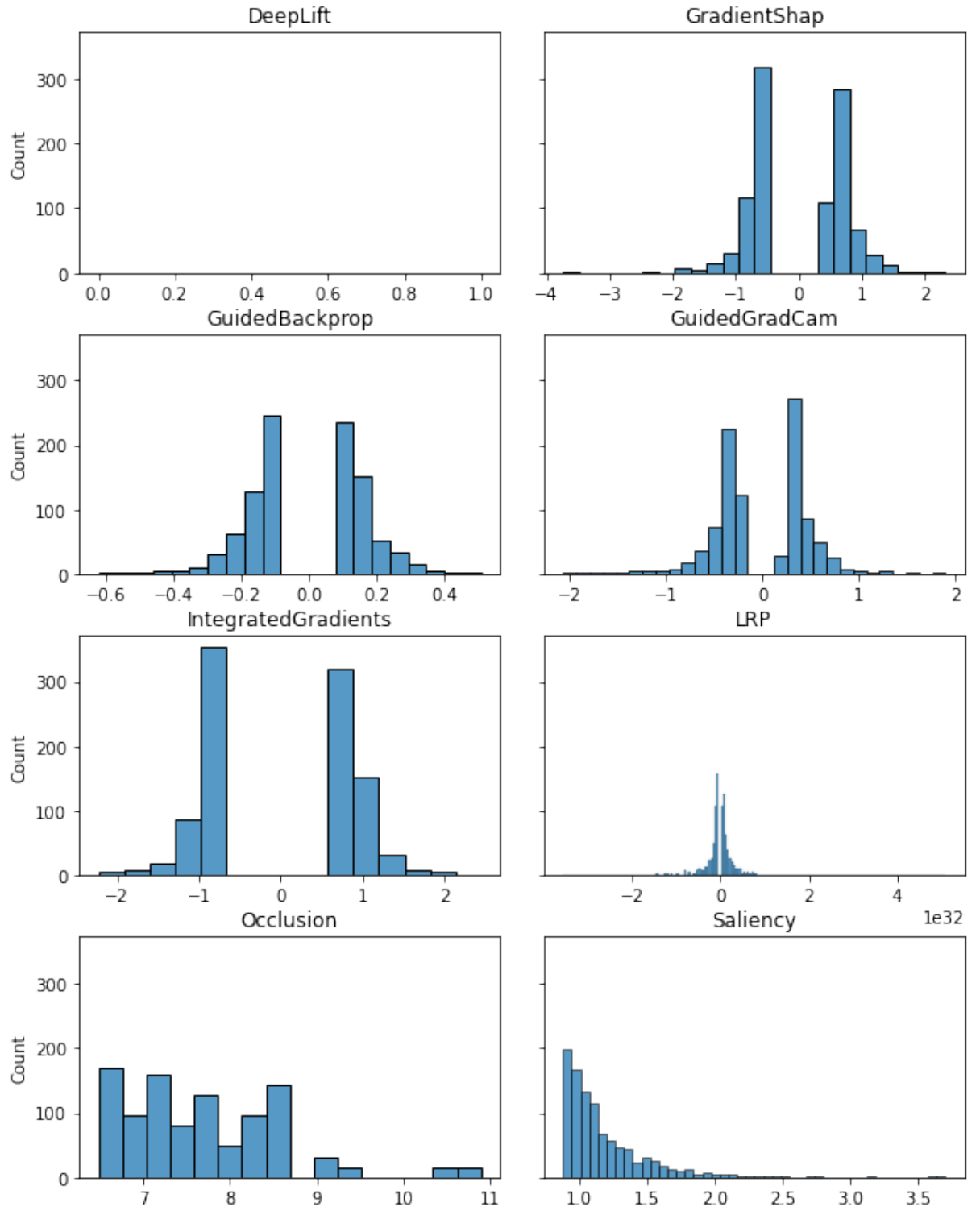


Figure 5.5: The distributions of the saliency map values (within the top- $k$  magnitude) produced by the explanation methods, evaluating on ResNet model.

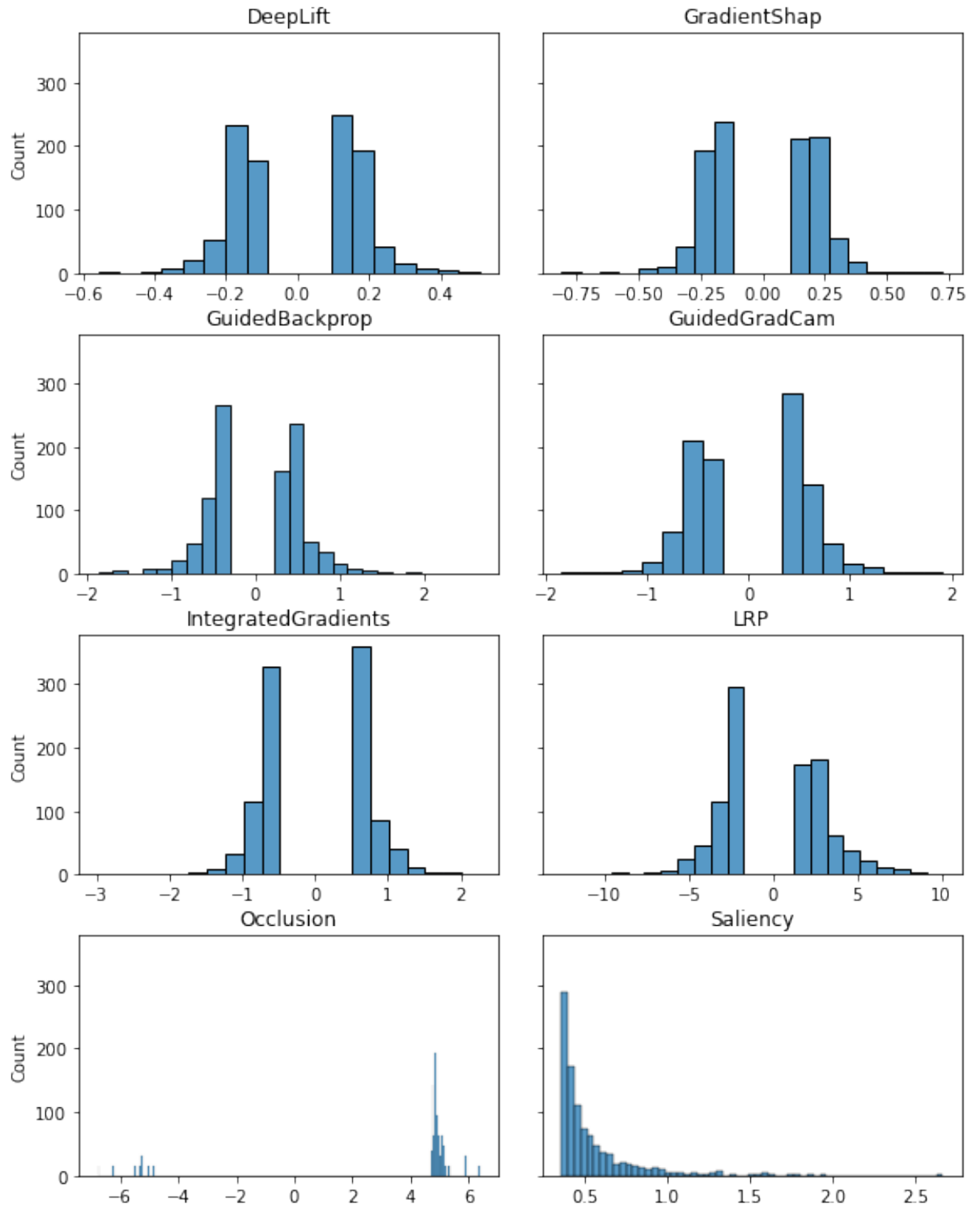


Figure 5.6: The distributions of the saliency map values (within the top- $k$  magnitude) produced by the explanation methods, evaluating on InceptionV3 model.

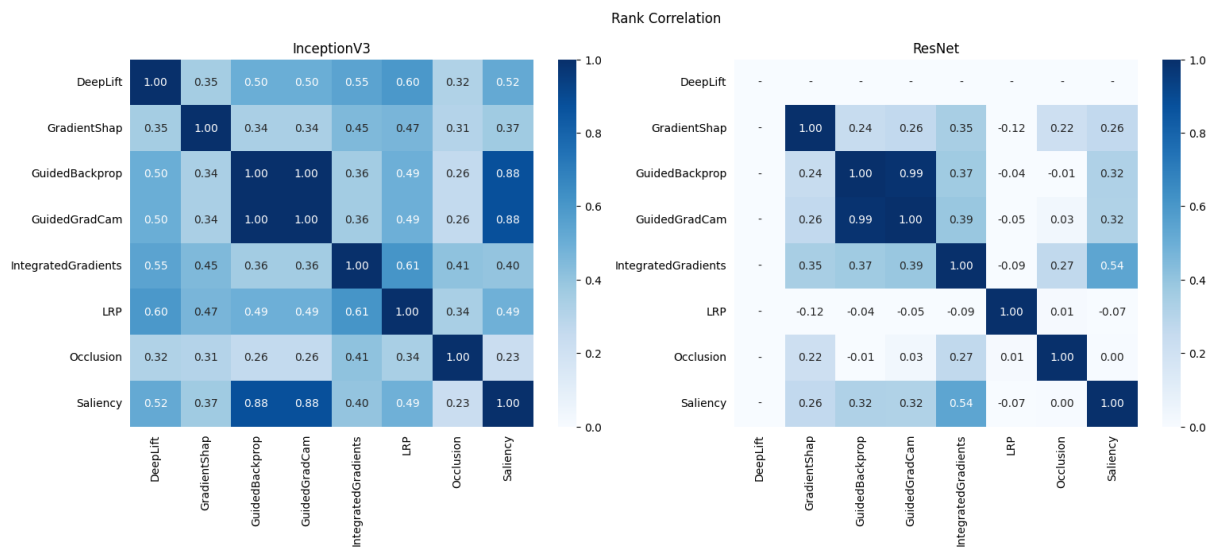


Figure 5.7: Rank correlation between different methods

## Chapter 6

# Conclusion

In this chapter we conclude our thesis by providing a summary of our key findings, identify some drawbacks in our work and discuss future directions towards resolving the disagreement problem.

### 6.1 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 the level of disagreement between eight 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.

## 6.2 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 explanations.

# References

## Tiếng Anh

- [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] Binder, Alexander et al. “Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers”. In: *CoRR* abs/1604.00825 (2016). arXiv: 1604.00825. URL: <http://arxiv.org/abs/1604.00825>.
- [8] 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>.
- [9] Brughmans, Dieter, Melis, Lissa, and Martens, David. *Disagreement amongst counterfactual explanations: How transparency can be deceptive*. 2023. arXiv: 2304.12667 [cs.AI].
- [10] Chetoui, Mohamed et al. “Explainable COVID-19 Detection on Chest X-rays Using an End-to-End Deep Convolutional Neural Network Architecture”. In: *Big Data and Cognitive Computing* 5.4 (2021). ISSN: 2504-2289. DOI: 10.3390/bdcc5040073. URL: <https://www.mdpi.com/2504-2289/5/4/73>.
- [11] 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.



- [12] 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.
- [13] He, Kaiming et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].
- [14] 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.
- [15] Kermany, Daniel. *Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification*. eng. 2018.
- [16] 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](https://doi.org/10.1007/978-3-030-28954-6_14).
- [17] 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>.
- [18] Lipton, Zachary Chase. “The Mythos of Model Interpretability”. In: *CoRR* abs/1606.03490 (2016). arXiv: 1606.03490. URL: <http://arxiv.org/abs/1606.03490>.
- [19] Lu, Xiaotian et al. “Crowdsourcing Evaluation of Saliency-based XAI Methods”. In: *CoRR* abs/2107.00456 (2021). arXiv: 2107.00456. URL: <https://arxiv.org/abs/2107.00456>.

- [20] Lundberg, Scott M. and Lee, Su-In. “A unified approach to interpreting model predictions”. In: *CoRR* abs/1705.07874 (2017). arXiv: 1705.07874. URL: <http://arxiv.org/abs/1705.07874>.
- [21] Narin, Ali, Kaya, Ceren, and Pamuk, Ziyne. “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/s10044-021-00984-y>.
- [22] Neely, Michael et al. “Order in the Court: Explainable AI Methods Prone to Disagreement”. In: (May 2021).
- [23] 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].
- [24] Ribeiro, Marco Tulio, Singh, Sameer, and Guestrin, Carlos. *“Why Should I Trust You?”: Explaining the Predictions of Any Classifier*. 2016. arXiv: 1602.04938 [cs.LG].
- [25] 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.
- [26] Rozario, Savio and Cevora, George. “Explainable AI does not provide the explanations end-users are asking for”. In: *ArXiv* abs/2302.11577 (2023).
- [27] 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.

- [28] 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>.
- [29] Selvaraju, Ramprasaath R. et al. “Grad-CAM: Why did you say that? Visual Explanations from Deep Networks via Gradient-based Localization”. In: *CoRR* abs/1610.02391 (2016). arXiv: 1610.02391. URL: <http://arxiv.org/abs/1610.02391>.
- [30] Shrikumar, Avanti, Greenside, Peyton, and Kundaje, Anshul. “Learning Important Features Through Propagating Activation Differences”. In: *CoRR* abs/1704.02685 (2017). arXiv: 1704.02685. URL: <http://arxiv.org/abs/1704.02685>.
- [31] Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. “Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps”. In: *CoRR* abs/1312.6034 (2013).
- [32] 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.
- [33] 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>.
- [34] Springenberg, Jost Tobias et al. *Striving for Simplicity: The All Convolutional Net*. 2015. arXiv: 1412.6806 [cs.LG].
- [35] Sundararajan, Mukund, Taly, Ankur, and Yan, Qiqi. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

- [36] 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.
- [37] 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>.
- [38] Taly, Ankur et al. “Using a deep learning algorithm and integrated gradient explanation to assist grading for diabetic retinopathy”. In: *Ophthalmology* (2019).
- [39] Tong, Yan et al. “Automated identification of retinopathy of prematurity by image-based deep learning”. In: *Eye and Vision* 7 (Aug. 2020), p. 40. DOI: 10.1186/s40662-020-00206-2.
- [40] 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.
- [41] Zawacki, Anna et al. *SIIM-ACR Pneumothorax Segmentation*. 2019. URL: <https://kaggle.com/competitions/siim-acr-pneumothorax-segmentation>.
- [42] Zeiler, Matthew D. and Fergus, Rob. “Visualizing and Understanding Convolutional Networks”. In: *CoRR* abs/1311.2901 (2013). arXiv: 1311.2901. URL: <http://arxiv.org/abs/1311.2901>.
- [43] Zhou, Bolei et al. “Learning Deep Features for Discriminative Localization”. In: *CoRR* abs/1512.04150 (2015). arXiv: 1512.04150. URL: <http://arxiv.org/abs/1512.04150>.

## Appendix A

# Supplementary Visualizations

### A.1 Disagreement Heatmaps Visualizations for the Pneumonia Dataset

In this supplementary chapter we present a collection of the disagreement heatmap visualizations generated for the pneumonia dataset.

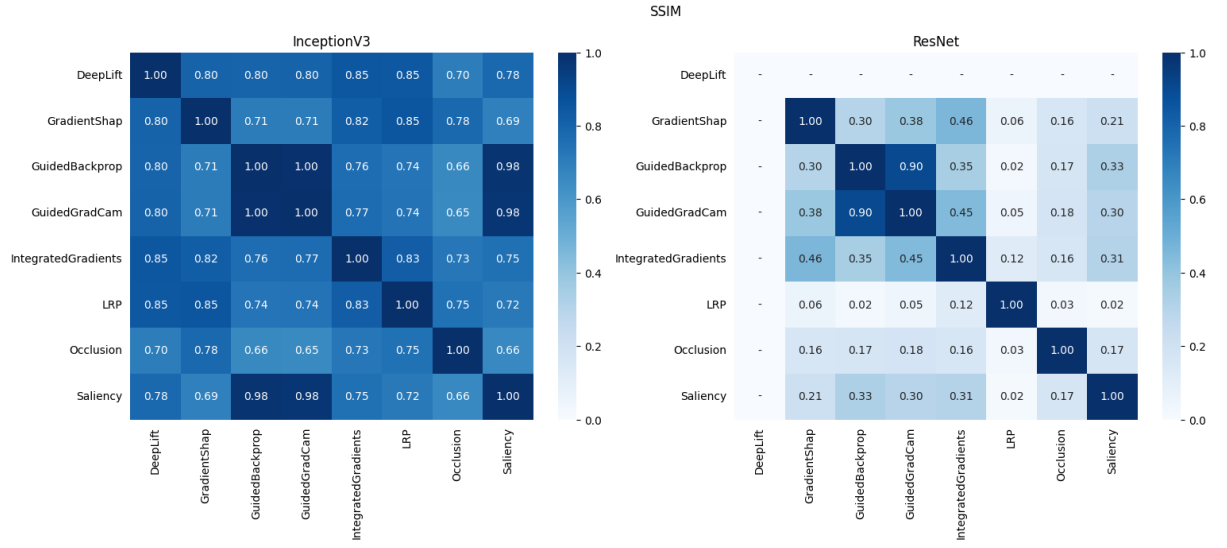


Figure A.1: SSIM scores when evaluated on Pneumonia dataset.

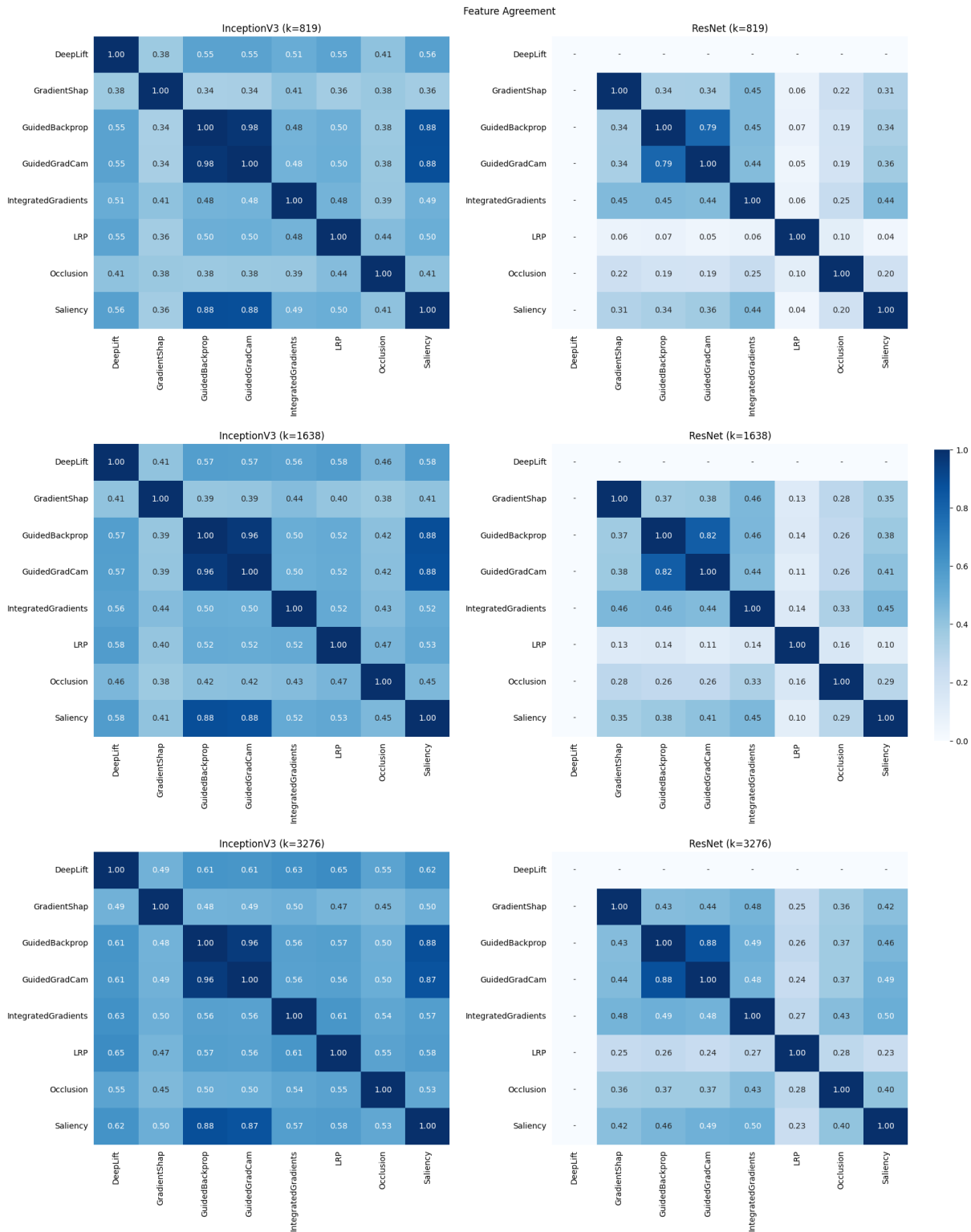


Figure A.2: Feature agreement scores when evaluated on Pneumonia dataset.

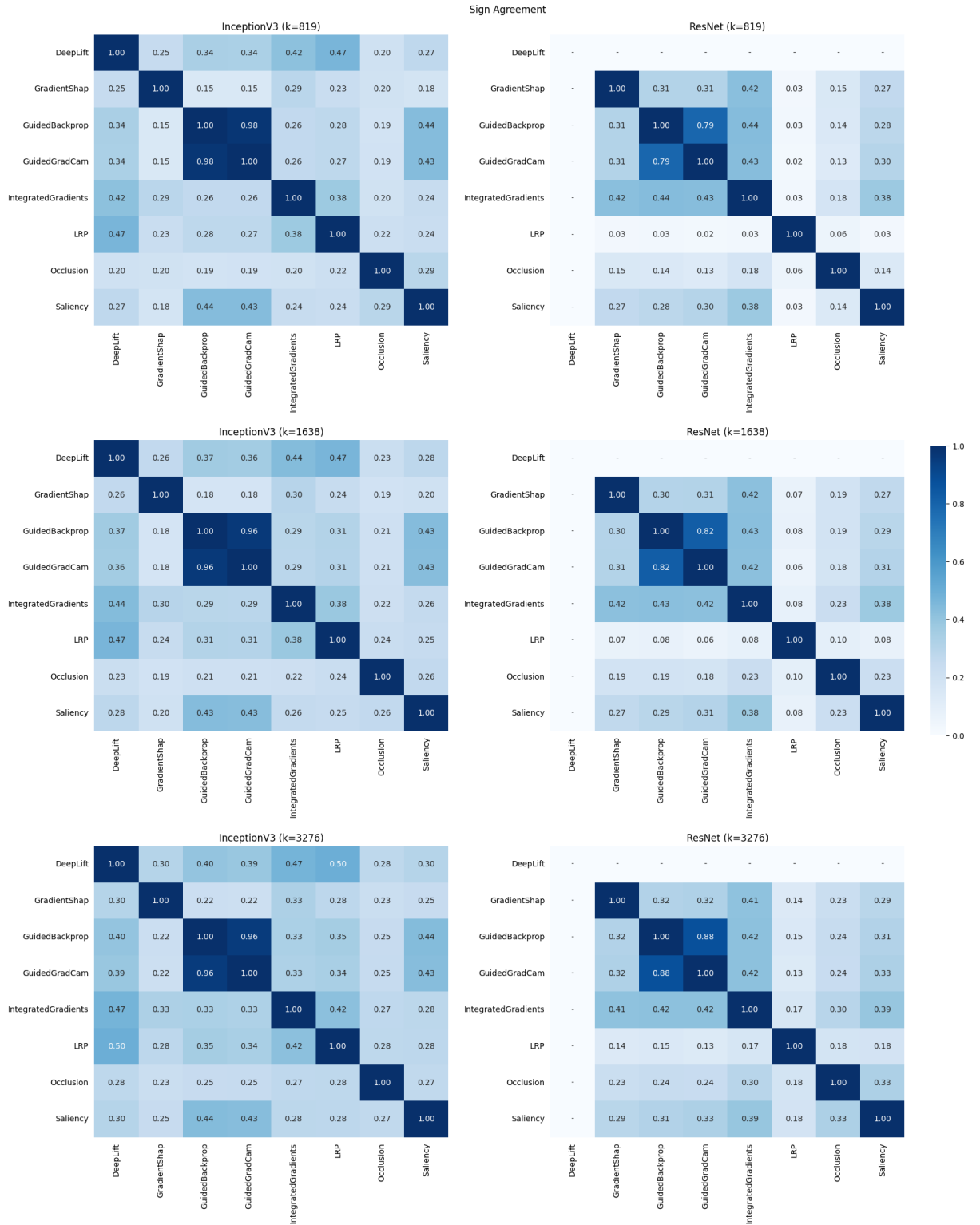


Figure A.3: Sign agreement scores when evaluated on Pneumonia dataset.