Technische Universität Berlin

Faculty of Electrical Engineering and Computer Science
Dept. of Computer Engineering and Microelectronics
Remote Sensing Image Analysis Group

Explainable Artificial Intelligence for Multi-Label Remote Sensing Image Classification

---

Master of Science in Computer Science

June 4, 2024

Jonas Klotz

Matriculation Number: 383652

Supervisor: Prof. Dr. Begüm Demir

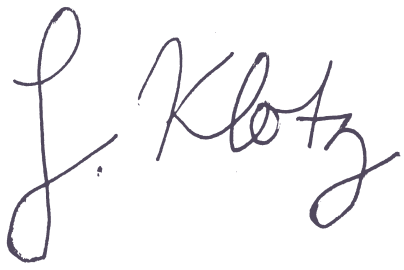
Advisor: Tom Oswald Burgert

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe angefertigt habe. Sämtliche benutzten Informationsquellen sowie das Gedankengut Dritter wurden im Text als solche kenntlich gemacht und im Literaturverzeichnis angeführt. Die Arbeit wurde bisher nicht veröffentlicht und keiner Prüfungsbehörde vorgelegt.

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

Berlin, June 4, 2024



.....

Name Surname

Abstract

Deep neural networks have demonstrated remarkable success in remote sensing (RS) tasks; however, their inherent "black-box" nature raises concerns about interpretability and robustness. These models often rely on spurious correlations (SC) between the annotated class label and biased features of the inputs, such as background or secondary objects. Explainable artificial intelligence (xAI) is increasingly becoming an essential tool for debugging such models. Although xAI methods have been extensively studied in classical computer vision (CV) tasks, their application in RS, particularly for multi-label classification (MLC), remains underexplored

In this thesis, seven xAI methods and 34 xAI metrics across six categories were theoretically studied in this regard. The analysis identified issues with backpropagation-based methods and perturbation-based methods. Localisation and Faithfulness metrics were considered particularly important for RS, while Complexity metrics were deemed less useful. Moreover, the xAI methods and 16 selected xAI metrics were empirically assessed on the two RS MLC datasets DeepGlobe and BigEarthNet and further contrasted with experiments on the single-label CV dataset Caltech101. A qualitative analysis showed that Grad-CAM and Guided Grad-CAM were best for the CV dataset, while non-backpropagation methods (Grad-CAM, LIME, Occlusion) were preferred for RS datasets. The quantitative analysis found Occlusion best for the CV dataset, and Grad-CAM and Guided Grad-CAM best for RS datasets.

Furthermore, two xAI-guided training methods that utilize xAI to enhance training performance and decision-making are evaluated: (i) the Right for the Right Reasons (RRR) loss alongside a proposed extension; and (ii) the xAI-based label propagation (LP) strategy for the appropriate application of the augmentation strategy CutMix in MLC task. RRR-guided training modestly improved performance while enhancing the model's reasoning capabilities, demonstrated by a reduced reliance on spurious correlations, particularly with non-backpropagation methods. Additionally, the usage of xAI methods for the LP strategy for CutMix also shows performance enhancements. Notably, there was an average increase in mean Average Precision of approximately 2% for ResNET and 0.5% for VGG compared to the baseline. Finally, a correlation analysis contextualizes the improvement of xAI-guided training methods with the theoretical and empirical results of xAI methods and metrics. The analysis showed that success in Randomisation and Localisation metrics can serve as predictor for xAI-guided training performance.

Zusammenfassung

Tiefe neuronale Netze haben bemerkenswerte Erfolge bei Fernerkundungsaufgaben erzielt, aber ihr inhärenter „BlackBox“Charakter wirft Bedenken hinsichtlich ihrer Interpretierbarkeit und Robustheit auf. Diese Modelle stützen sich häufig auf Scheinkorrelationen zwischen dem annotierten Klassenlabel und sekundären Eingabemerkmalen wie dem Hintergrund. Erklärbare Künstliche Intelligenz wird zunehmend zu einem wichtigen Werkzeug, um Fehler in solchen Modellen zu korrigieren. Obwohl Erklärungsmethoden im klassischen Computersehen Aufgaben ausführlich untersucht wurden, ist ihre Anwendung in der Fernerkundung, insbesondere in der Multi Label Klassifikation, noch wenig erforscht.

In dieser Arbeit wurden diesbezüglich sieben Erklärungsmethoden und 34 Erklärungsmetriken in sechs Kategorien theoretisch untersucht. Die Analyse identifizierte Probleme mit rückpropagierenden Methoden und perturbationbasierten Methoden. Lokalisierungs und Treue Metriken wurden als besonders wichtig für Fernerkundung erachtet, während Komplexitätsmetriken als weniger nützlich eingestuft wurden. Zusätzlich wurden die Erklärungsmethoden und 16 ausgewählte Erklärungsmetriken empirisch auf den beiden Multi Label Fernerkundungsdatensätzen DeepGlobe und BigEarthNet evaluiert und mit Experimenten auf dem Single Label Computersehensdatensatz Caltech101 verglichen. Eine qualitative Analyse zeigte, dass GradCAM und Guided GradCAM am besten für Caltech101 geeignet waren, während nicht rückpropagierende Methoden (GradCAM, LIME, Occlusion) für die Fernerkundungsdatensätze bevorzugt wurden. Die quantitative Analyse ergab, dass Occlusion am besten für Caltech101 geeignet war, während GradCAM und Guided GradCAM am besten für Fernerkundungsdatensätze geeignet waren.

Darüber hinaus wurden zwei erklärungsbasierte Trainingsmethoden evaluiert, die erklärbare Künstliche Intelligenz zur Verbesserung der Trainingsleistung und der Schlussfolgerung nutzen: (i) der Right for the Right Reasons Verlust zusammen mit einer vorgeschlagenen Erweiterung; und (ii) die erklärungsbasierte Label Propagations Strategie zur adäquaten Anwendung der Augmentierungsstrategie CutMix in der Multi Label Klassifikationsaufgabe. Das auf Right for the Right Reasons basierende Training führte zu einer leichten Leistungssteigerung, während es die Schlussfolgerungsfähigkeit des Modells verbesserte, was sich in einer verringerten Abhängigkeit von falschen Korrelationen zeigte. Darüber hinaus zeigte die Verwendung von Erklärungsmethoden für die Label Propagationsstrategie für CutMix ebenfalls Leistungsverbesserungen. Bemerkenswert war eine durchschnittliche Erhöhung der mittleren Genauigkeit von etwa 2% für ResNET und 0,5% für VGG im Vergleich zu einem Referenzmodell. Abschließend wurde eine Korrelationsanalyse durchgeführt, um die Verbesserung der erklärungsbasierten Trainingsmethoden mit den theoretischen und empirischen Ergebnissen der Erklärungsmethoden und Metriken in Beziehung zu setzen. Die Analyse zeigt, dass die Ergebnisse der Randomisierungs und Lokalisierungsmetriken die Leistung des erklärungsbasierten Trainings vorhersagen können.

Contents

List of Tables

List of Figures

1 Introduction

Over the past decade, Deep Learning (DL) has emerged as the dominant approach for solving complex tasks, such as image classification [3], object detection [4] and semantic segmentation [5], across various domains, achieving remarkable state-of-the-art performance. Given the considerable significance of these tasks within the domain of Earth Observation (EO) and Remote Sensing (RS), DL has emerged as a prevailing paradigm in RS research [6, 7, 8]. However, despite their impressive results, deep learning models often remain enigmatic black boxes, with almost no human understanding of their decision-making processes [9].

This lack of transparency in Artificial Intelligence (AI) models is becoming increasingly problematic as more decisions rely on these black-box systems, or are even automatically decided by them [10]. For critical applications, new regulatory approaches emerged, especially in the European Union, like the General Data Protection Regulation [11] or the Artificial Intelligence Act [12]. But there are additional important issues to consider. For example, there is significant potential to gain scientific knowledge from these models, which can lead to wider societal benefits [10]. In addition, understanding these models is critical to overcome poor model performance caused by several factors, including limited or biased training data, the presence of outliers, adversarial data, and model overfitting [10].

To open this black box Explainable Artificial Intelligence (xAI) has emerged as a research focus to reveal the contributing factors behind DL model predictions, enabling a better understanding of their internal mechanisms. The research interest in xAI has grown strongly in the recent past, leading to the invention of many sophisticated explanation methods ([13], [14], [15], [16], [2], [17], [18], [19]).

The general need for xAI arises from various concerns regarding the trustworthiness and transparency of AI systems. Several factors contribute to the lack of trust in AI models, including incomplete problem formulation and the complexity of real-world tasks. Real-world tasks often involve numerous variables, leading to challenges in providing comprehensive descriptions.

To build trust and ensure ethical practices in AI systems, Doshi-Velez et al. [20] highlight the importance of several fundamental factors. **Fairness** is critical, requiring that AI systems make unbiased decisions, especially in sensitive contexts such as credit scoring, where decisions must be free of discrimination based on gender, race, or other protected characteristics. **Privacy** is another key element, requiring that AI models comply with privacy laws to protect sensitive information such as medical records. **Reliability** ensures that AI models produce consistent outputs even with small variations in input, which is essential for safety-critical applications such as self-driving cars to prevent unexpected behaviour. **Acceptance** focuses on fostering human trust in AI systems, which is enhanced when systems can provide understandable explanations for their decisions, e.g. by helping healthcare professionals and patients to understand medical diagnoses and treatments. Finally, **Causality** is necessary to ensure that AI

models capture only causal relationships that are present in the data. This is essential to avoid so-called Spurious Correlations (SC), shortcuts or "Clever-Hans" predictors [21, 22, 23], which are the misinterpretation of correlations as causal relationships and can lead to inaccurate predictions and potentially harmful outcomes.

SC in AI models often arise from uncontrolled confounding biases present during data collection [21]. For example, in the PascalVOC challenge, the class 'horse' is spuriously correlated with a watermark on the images [22]. Such correlations also occur in medical datasets, where the SC can manifest as hospital tags in COVID-19 radiographs [24], or as skin marks in skin lesion detection datasets [25]. xAI can serve as valuable tool to identify and eliminate these SCs, and therefore, enhance the model [26]. This process, known as xAI-guided training or xAI-augmentation, can improve several model properties including performance, convergence speed, robustness against adversarial attacks, efficiency, fairness, and reasoning [27]. Integrating these capabilities into RS applications is crucial, as emphasised by Tuia et al. [28], who identified xAI as one of the six main research directions in the field of EO.

However, xAI has received comparatively little attention in RS [28]. Gevaert [29] reports that the most common application of xAI in EO is to identify and debug incorrect model behaviour. Furthermore, she highlights a critical gap in xAI practice: the failure to evaluate the effectiveness of explanation methods within their intended contexts. Often explanations are provided without verifying their usefulness to the target audience, leading to a disconnect between the generation of explanations and their practical usefulness [29].

This argument can be extended to not only highlight the mismatch in usefulness to the audience but also to question the overall applicability of current explanatory methods to typical RS image data. One of the main reasons for this mismatch is that much of the xAI research has focused on Single-label Classification (SLC) tasks. A SLC task is to assign exactly one label from a set of discrete classes to each instance in the dataset. However, the RS domain often involves more complex challenges such as Multi-label Classification (MLC) or pixel-level classifications (i.e. semantic segmentation). This complexity arises from the nature of satellite imagery data, which often contains multiple equally relevant classes or objects within a single scene, adding layers of complexity that xAI methods designed for simpler tasks may not adequately address. Such complexity, coupled with the unique characteristics of RS data - such as repetitive textures, nadir views and multispectral bands - presents significant difficulties for xAI methods originally designed for simpler SLC tasks in conventional Computer Vision (CV) contexts.

Against this background, the specific objectives of this study are outlined below, focusing on the evaluation and improvement of the applicability of xAI methods to the challenges of MLC. RS image data:

Objective A: *Assessment of xAI methods and metrics on MLC RS image data.*

- Conduct a theoretical analysis to assess the compatibility of xAI methods with the unique characteristics of RS images.
- Evaluate the utilisation of xAI metrics to measure the performance of explanation methods for RS image data.
- Implement and adapt these methods and metrics for MLC.
- Empirically compare these methods and metrics for CV and RS image data.

Seven well-known explanation methods (Deep Learning Important FeaTures (DeepLIFT),

Gradient-weighted Class Activation Mapping (GradCAM), Guided Gradient-weighted Class Activation Mapping (Guided GradCAM), Integrated Gradients (IG), Local Interpretable Model-agnostic Explanations (LIME), Layer Relevance Propagation (LRP), and Occlusion Sensitivity (Occlusion)) will first be assessed theoretically. They will then be evaluated through experiments using carefully selected xAI metrics from six categories: Axiomatic, Complexity, Faithfulness, Localisation, Randomisation, and Robustness. Additionally, a qualitative assessment is conducted to evaluate the explanations. The experiments will be conducted using three datasets: a RGB CV SLC dataset (Caltech101), a RGB MLC RS dataset (DeepGlobe), and a multispectral MLC RS dataset (BigEarthNet-S2 (BEN)). The results will be compared across datasets, specifically focusing on differences in task and data characteristics.

As outlined previously, the primary application of xAI in RS is to understand the black-box models, however, it can be also used to enhance them. Weber et al. [27] state that xAI-guided training can improve several properties of AI models: performance, convergence, robustness, efficiency, and reasoning. The second objective of this paper is to investigate whether the training performance and reasoning of Machine Learning (ML) models on RS image data can be enhanced through the utilisation of explanatory methods.

Objective B: *Assessment of xAI-guided training approaches for MLC RS image data.*

- Evaluate xAI-guided model training to improve performance and reasoning.
- Link these results with the previous investigation to determine which explanation methods are effective for xAI-guided MLC RS image data.
- Assess whether success in explanation metrics can predict better performance gains using xAI-guided methods.

Here, two methods will be applied: Right for the Right Reason (RRR) loss [30], a loss-augmentation method incorporating explanations into the loss function to improve the reasoning of models, and CutMix with xAI Label Propagation for Multi-Label Classification (CutMix with xAI LP), an improvement of the CutMix augmentation method for MLC that uses explanations to propagate labels correctly for CutMix-generated augmentations, aiming to enhance the training performance of the models.

In general, this thesis is divided into eight chapters, each dealing with a specific aspect of the research into the applicability of xAI methods and metrics to MLC-based RS image data.

Chapter 1 introduces the motivation for the study, outlines the research objectives and provides an overview of the structure of the thesis.

Chapter 2 presents the background to the study, discusses related work, and defines key terms and concepts that are essential for understanding the following chapters.

Chapter 3 discusses the various explanatory methods used in this thesis, including Occlusion, IG, LRP, DeepLIFT, GradCAM, Guided GradCAM and LIME. The challenges of applying these methods to MLC RS images are also discussed.

Chapter 4 focuses on evaluating the quality of explanations. It details the different types of metrics, followed by a discussion of desiderata for explanation metrics in RS MLC image classification. Furthermore, the most appropriate set of metrics is selected based on their suitability for MLC RS images.

Chapter 5 explores the integration of explanations into the training process, investigating the

use of RRR loss and CutMix with xAI LP to improve model performance and reasoning. Both of these xAI-guided training methods will be further evaluated in the experiments.

Chapter 6 describes the datasets and experimental setup used in this study. This includes an overview of the characteristics of the datasets and an explanation of the methodology used in the experiments.

The experimental results and discussion are presented in Chapter 7. This chapter presents a qualitative assessment of the explanation methods, a quantitative analysis using selected explanation metrics, and an evaluation of xAI-guided training, focusing on the effectiveness of RRR loss and CutMix with xAI LP. Also the results of the xAI-guided training are set in context with the quantitative analysis, by evaluating if any of the explanation metrics or categories can be used as estimators for better xAI-guided training success.

Finally, Chapter 8 concludes the thesis by summarising the key findings, discussing the limitations of the study and suggesting future research opportunities in the area of xAI for MLC RS image data.

2 Background

This chapter outlines the framework of this thesis, including a comprehensive review of the relevant literature, definitions of relevant terminology and the notation used throughout.

2.1 Related Work

Adadi and Berrada [31] outline four primary motivations for xAI, which are critical in framing its role within ML: *explain to justify*, focusing on understanding the model's reasoning to justify its decisions; *explain to control*, which uses insights into the decision process to identify and correct errors; *explain to improve*, leveraging explanations to enhance model performance; and *explain to discover*, helping researchers gain a deeper understanding of the task. Most xAI research in RS has been focused on *explain to control* [29, 32], indicating a predominant interest in diagnosing and mitigating errors in the models rather than expanding into other motivations which might offer broader benefits, such as improving model performance or enhancing scientific understanding of complex EO tasks. This aligns with findings from Leluschko and Tholen, who conducted a review on the stakeholders and goals within human-centred xAI applied in RS [33]. Their research highlights an underrepresentation of non-developer stakeholders in this area, suggesting a narrow focus that may overlook broader opportunities to leverage other goals. Although the number of publications has increased recently, there remains a significant gap in the publication frequency of ML research versus xAI research in the field of EO, with ML publications outnumbering xAI publications by approximately a factor of 70 [32].

Recently, there has been a surge in adapting xAI methods to suit RS properties, with half of the relevant publications emerging in the last year and none before 2021 [32]. The modifications primarily enhance the Class Activation Mapping (CAM) and Gradient-weighted Class Activation Mapping (GradCAM) methods for various RS applications and domains [32]. For instance, Feng et al. [34] developed a new CAM method, Self-Matching CAM (SMCAM), tailored for Synthetic Aperture Radar (SAR) images, where target objects typically occupy a small portion of the image. Instead of the conventional approach of upsampling the feature map of the convolutional layer to the input image size, their method downsamples the input image to match the feature map of the last convolutional layer. This technique produces saliency maps that localize targets in SAR images with greater precision compared to the GradCAM method. To evaluate their SMCAM, two different occlusion tests were conducted to compare various CAM approaches. These tests revealed that while most methods showed similar performance when the most influential pixels were perturbed, the SMCAM uniquely exhibited only a minor drop in prediction difference when the most salient pixels were not occluded. This result indicates that the SMCAM's explanations more accurately focus on the actual target object within the image Huang et al. [35] present the Encoder-Classifier-Reconstruction Class Activation Mapping (ECR-CAM) Neural Networks (NN) specifically designed for RS images

with multiple objects. The network architecture includes four integral modules: encoder, classifier, reconstruction, and a CAM module. Their approach demonstrates improved classification accuracy and superior capability in localizing target objects using CAM. De Lucia et al. [36] have developed variants of CAM specifically for hyperspectral images, where the traditional 2D saliency map is extended into a 3D volume. This adaptation allows each voxel to attribute importance across different channels in-depth, offering detailed pixel-wise and spectral cumulative attributions. Additionally, their study includes an evaluation of faithfulness, comparing the performance of different CAM methods. They also use a class sensitivity metric to measure the correlation between the attributions of different target classes, enhancing the understanding of model decisions across various spectral dimensions.

Höhl et al. [32] identify challenges in understanding and validating explanations due to the complex properties of EO data, which can obscure the semantics of objects or individual pixels within a RS scene. They emphasize two main areas of focus: firstly, the creation of interpretable input spaces to simplify the data attributes, and secondly, the use of domain knowledge to aid in the interpretation of explanations.

One extensive study for explanation methods for MLC in RS has been conducted by Kakogeorgiou et al. [37], where 10 methods, namely: Saliency [38], Input $\times$ Gradient [39], Integrated Gradients [40], Guided Backpropagation [41], GradCAM and Guided GradCAM [19], LIME [42], Occlusion [18], DeepLIFT [39] as well as their combination with the SmoothGrad approach [43] were evaluated. The authors conducted a qualitative assessment and quantitative evaluation regarding the following metrics: Max-Sensitivity [44], Area Under the MoRF perturbation curve (Selectivity) [45], File Size [46] and Computational Time. Furthermore, they examined and discussed the differences in the explanations for multiple labels that co-exist in the same image. According to the max-sensitivity metric, Occlusion, GradCAM, and LIME proved to be the most reliable methods. Among these, GradCAM was identified as the most computationally efficient option.

Kucklick [47] explored the reliability of GradCAM for real estate appraisal tasks, highlighting its sensitivity to modifications in network weights and label randomisation. This study includes model and data randomisation tests to underscore the instability of GradCAM under varying conditions, suggesting a need for caution when deploying this technique in critical applications such as property valuation. Abdollahi and Pradhan [48] utilized SHapley Additive exPlanations (SHAP) [49] to rank the input parameters and select appropriate features for classification to extract vegetation covers. Wolanin et al. [50] proposed Regression Activation Maps (RAM) to explain their model to estimate the crop yield in the Indian Wheat Belt, gaining valuable insights into the factors contributing to variations in yield. Furthermore, an investigation was conducted by [51], using LRP [52] to highlight the distinctive attributes of various loss functions. Su et al. [53] used several CAM approaches to evaluate ResNet models, focusing on addressing large variance challenges and assessing localisation capabilities. They generated segmentation maps from CAM, GradCAM, GradCAM++ [54], Smooth GradCAM++ [55] and Score-weighted Class Activation Mapping [15] by thresholding their attributions. Of these, GradCAM demonstrated superior localisation accuracy and an enhanced ability to identify complex features in images characterised by significant variance.

Most recently in 2023, Mohan and Peeples [56], conducted a study on explanation methods for RS. They implemented 3 state-of-the-art model architectures: ConvNeXt [57], Vision transformers (ViT) [58], and FocalNets [59] and examined them with different explanation meth-

ods, namely High-Resolution Class Activation Mapping (HiResCAM) [60], LIME [42], Gradient SHapley Additive exPlanations (GradSHAP) [49], Saliency maps [18], and Occlusion [18]. They found that HiResCAM was the most versatile explanation method regarding their quantitative evaluation with the Quantus framework [61].

An additional difficulty is that the lack of standardised and objective evaluation frameworks for xAI methods poses significant challenges for practitioners outside the xAI domain. While various metrics and tools exist for evaluation, effective incorporation of xAI into an ML pipeline requires domain-specific knowledge of xAI to ensure accurate evaluation of the methods employed by a RS practitioner [32]. To the best of my knowledge, no studies have assessed the usability of these metrics specifically for RS images, highlighting a critical gap in the field that needs to be addressed.

There are relatively few applications of xAI-guided training in RS. One application of loss-augmentation was conducted by Cheng et al. [62], where the authors utilized GradCAM to enhance model training through feature unlearning (FUL). They employed the RRR method to refine forestry classification, demonstrating how targeted adjustments in training can lead to improved model accuracy and relevance for leaf classification. A notable example is provided by Beker et al. who used data augmentation to generate synthetic data for training a Convolutional Neural Network (CNN) model aimed at detecting volcanic deformation [63]. Through an explanatory analysis using GradCAM on real data, the researchers discovered that the model incorrectly identified volcanic deformation in locations such as salt lakes and slope-induced signals - patterns not present in the synthetic training set. These findings led to improvements in prediction performance by fine-tuning the final layer of the CNN model using a hybrid dataset containing both synthetic and real-world imagery to account for the previously missed patterns. Another example of data augmentation is described in the work of Xiong et al. [64], where a mask generated from the GradCAM output is used during training to occlude regions of highest model activation in each image. This technique aims to encourage the network to explore other features in the image, thus facilitating a more comprehensive understanding and use of the available data. More recently, Burgert et al. [65] proposed a data-augmentation strategy specifically designed for MLC RS images, called CutMix with xAI LP. This method addresses a limitation found in the popular CutMix technique [66], which struggles with MLC contexts. CutMix with xAI LP extends the traditional approach by propagating pixel-level label information into the augmented images. If no ground truth is available for the dataset, explanations are used to derive this pixel-level label information, ensuring that the augmented images retain accurate annotations for training.

2.2 Definitions and Terminology

Defining the meaning of the word explanation is a challenging task, heavily depending on the context. From philosophy to psychology there have been many approaches, but finding one universal answer is difficult. In the scope of this thesis, explanations are seen in the context of xAI. Here, Nauta et al. [1] define an explanation as *"a presentation of (aspects of) the reasoning, functioning and/or behaviour of a machine learning model in human-understandable terms"* [1]. Where they use the phrasing reasoning, functioning and/or behaviour, which they adopt from Gilpin et al [67]. Gilpin et al. define "reasoning" as explaining how specific inputs produce specific outputs. "Functioning" pertains to the inner workings and data structures of a model and "Be-

havior" describes the model's overall operation by examining inputs and outputs, simplifying interpretation [67]. Furthermore, Nauta et al. [1] adopt the phrase "human-understandable" from Doshi-Velez and Kim [20]. Understanding is defined as the capacity of humans to identify connections, alongside grasping the context of an issue[68]. The notion of understanding can be categorized into mechanistic ("How does something operate?") and functional understanding ("What is its intent?")[69]. An explanator or explanation method is the system or method that explains [70]. These methods can be categorized by the *Task Type* (e.g. regression, classification, segmentation), the Input Data Type (e.g. image, text) and the *Level of interpretability* [70]. Different levels of interpretability are, for instance, *Intrinsic explanations*, also called (*ante-hoc*) [71], which are inherently interpretable models. *Post-hoc explanations* are used to explain a model without changing it [70]. their portability (whether the explanator is *Model-specific* or *Model-agnostic*) and the *explanation locality*. Here *Model-agnostic* or black box methods, have only access to the model's input and output, while *Model-specific* or white box methods, have full access to the model's internals [70]. *Global Explanations* are explaining the whole model or *Local Explanations* explaining a model's specific output [70]. Additionally, the output of the explanation method can be divided into the object of explanation, the actual output type and the output's presentation [70].

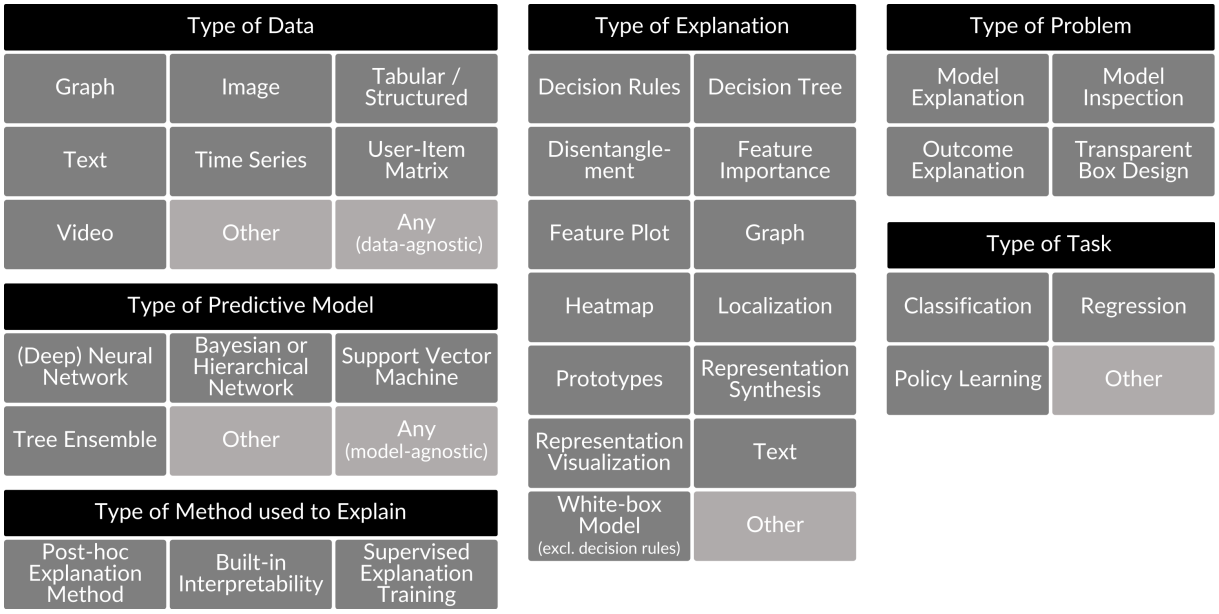


Figure 2.1: Categorisation of xAI methods along 6 dimensions (taken from [1])

A more detailed categorisation is given by Nauta et al. [1]. They organise xAI along six dimensions, as shown in Figure 2.1. The first dimension is the type of data, which includes categories such as text, graphs, images and videos; this paper focuses specifically on image data, which is prevalent in RS applications. The second dimension concerns the type of predictive models, ranging from NNs, support vector machines, to model-agnostic methods; the focus here is on NN due to its widespread use in RS image classification. The third dimension concerns the general type of explanation, with a particular focus in this study on post-hoc explanation methods. This approach allows the application of explanation methods across all classifiers. The fourth dimension concerns the type of explanation used, specifically heatmaps,

which are defined as maps with at least two dimensions that visually highlight non-binary features such as attributions, activations, sensitivities, attentions, or saliencies. In this thesis, terms such as 'heatmap', 'saliency map' and 'feature attribution map' are used interchangeably. The fifth dimension, the nature of the problem, focuses on outcome explanation, which is concerned with explaining the outcome or prediction of a model on a particular input instance, thus providing local interpretability. Finally, the study focuses on classification tasks, specifically addressing the challenges of land cover classification.

Notation

The formalisation in this thesis is adopted from Hedström [72]. Let $\boldsymbol{f}$ be a black-box model in a supervised classification framework, parameterized by θ. This model, trained on the dataset $\boldsymbol{X}_{\text{tr}} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}$, maps an input $\boldsymbol{x} \in \mathbb{R}^D$ to an output class. For a SLC task let $y \in \{1, \ldots, C\}$ for a MLC tasks, the output is a vector $\mathbf{y}_i \in \{0,1\}^L$, where $L \in \mathbb{N}$ denotes the number of classes, with each entry $(\mathbf{y}_i)_l$ indicating the presence (1) or absence (0) of class l.

The model is defined by the function:

$$\boldsymbol{f}(\boldsymbol{x}; \theta) = \hat{y},$$

where $f : \mathbb{X} \mapsto \mathbb{Y}$ denotes the mapping from the input space $\mathbb{X}$ to the label space $\mathbb{Y}$. For RS images, the inputs $\mathbf{x}_i$ fall within $\mathbb{R}^{C \times H \times W} \subset \mathbb{X}$, with C, H, and W representing the number of image bands, height, and width respectively. To quantitatively estimate the performance of model f, the prediction error is computed on a given test dataset $\boldsymbol{X}_{\text{te}}$, where each prediction $\hat{y}$ is associated with a label y. For clarity and when detail is not necessary, the notation simplifies the model function as $f(\boldsymbol{x})$, omitting the parameter θ.

Furthermore, an explanation method is formally defined as a function designed to visualize the reasoning behind a specific prediction $\hat{y}$ of the model f:

$$\Phi(\boldsymbol{x}, f, \hat{y}; \lambda) = \boldsymbol{e},$$

where $\Phi : \mathbb{R}^D \times \mathbb{F} \times \mathbb{Y} \mapsto \mathbb{R}^D$ is an explanation function parameterised by λ. This function attributes importance to each feature in $\boldsymbol{x}$, typically visualized in an explanation map $\boldsymbol{e} \in \mathbb{R}^D$. For clarity, the notation is abbreviated as $\Phi(\boldsymbol{x})$ when the additional parameters are implicit.

Since there is no ground truth explanation available, it is not possible to compute the prediction error for an explanation function Φ. Instead, a generalized notation for the quality estimation of attribution-based explanation methods is provided. Define $\Psi_\tau : \mathbb{E} \times \mathbb{R}^D \times \mathbb{F} \times \mathbb{V} \mapsto \mathbb{R}$ as a quality estimator that is parameterised by τ. This estimator takes an explanation and returns a scalar value—termed the "quality estimate"—that indicates the quality of the explanation. The evaluation of an explanation, or explanation metric can be formalized as:

$$\Psi(\Phi, \boldsymbol{x}, f, \hat{y}; \tau) = \hat{q},$$

where Ψ represents the quality estimator. For simplicity in notation, $\Psi(\Phi, \boldsymbol{x}, f, \hat{y}; \tau)$ is abbreviated to $\Psi(\Phi, \boldsymbol{x})$.

Explanation methods or metrics often compare an input $\boldsymbol{x}$ to a reference input or baseline $\bar{\boldsymbol{x}}$.

Furthermore, a perturbation in the image space is defined as:

$$\tilde{\boldsymbol{x}} = \mathcal{P}_{\mathbb{X}}(\boldsymbol{x}, M),$$

with $\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}, M)$ being the perturbation function, where M is the number of input features that are perturbed. One example of a common perturbation is the setting of indices to a baseline. Let $\boldsymbol{x}_{[\boldsymbol{x}_s = \overline{\boldsymbol{x}}_s]}$ denote that the subset of indices $s \in |S| \subseteq d$ of sample $\boldsymbol{x}$ is set to the baseline value $\overline{\boldsymbol{x}}$. Conversely, $\boldsymbol{x}_{[\boldsymbol{x}_k = \overline{\boldsymbol{x}}_k]}$ defined the perturbation where the top-k features attributed by explanation method Φ are perturbed to the baseline.

3 Explanation Methods

This section defines and investigates the explanation methods that will be analysed in this thesis, providing a comprehensive analysis of their principles, implementations, and possible applications.

As stated in Section 2.2, this thesis focuses on posthoc explanation methods for the classification output of a NN for image data, producing feature attribution maps. Much of the research in xAI for CV relates to this approach, which is classified as attribution-based explanations [73]. Bai et al. [73] also provide an additional categorisation for methods that generate attribution-based explanation, dividing them into propagation-based methods and perturbation-based methods. Perturbation-based approaches perturb the model input and evaluate changes in the output. Depending on these changes, relevance is attributed to the input. Propagation-based methods can be further categorised into response-based, CAM-based and backpropagation-based (BP-based) approaches. Response-based methods "use feature maps in the inference process to interpret the decision making" [73] of a model, CAM-based methods weight the CAM of convolutional layers of a model, while BP-based methods backpropagate some relevancy measure to explain the model.

In the following sections, the explanation methods considered in this thesis are explained in more detail. Further explanatory methods that are not considered, are listed in the Appendix 1.

3.1 Occlusion Sensitivity

One of the first explanation methods is the **Occlusion Sensitivity (Occlusion)** method by [18]. It is a perturbation-based approach that involves occluding parts of the input image x with a back patch P of size $ps \times ps$. The patch is systematically shifted across the image with a horizontal and vertical stride of ps, ensuring that every pixel in the image is covered by the patch.

For each position $\lambda \in \Lambda$, where Λ is the set of all possible top-left corner positions of the black patch, the perturbed image x' is created by setting the pixel values in the patch to 0. This process generates $\left\lceil \frac{H}{ps} \right\rceil \times \left\lceil \frac{W}{ps} \right\rceil$ occluded variations of the original image, each with a different region covered by the patch. These occluded images are fed into the network, resulting in a new prediction for each perturbed variation. The classification probability for the target class, $f_c(x')$, is recorded for each occluded image. The saliency map $S : \Lambda \rightarrow \mathbb{R}$ is then computed, where $\Lambda = \{1, \ldots, \left\lceil \frac{H}{ps} \right\rceil\} \times \{1, \ldots, \left\lceil \frac{W}{ps} \right\rceil\}$. Each entry of $S(\lambda)$ represents the importance of the region covered by the patch at position λ and is calculated as:

$$S(\lambda) = 1 - f_c(x'). \tag{3.1}$$

The intuition behind this formula is that if an important region is occluded, the confidence for the target classification should significantly drop, resulting in a higher saliency score. The saliency map S is then upsampled to the size of the original image. The biggest problem is its inefficiency: the generation of $\left\lceil \frac{H}{ps} \right\rceil \times \left\lceil \frac{W}{ps} \right\rceil$ occlusions, each of which requires a model prediction. This is highly impractical for large images such as those found in RS data.

3.2 Integrated Gradients

Integrated Gradients (IG) [74] is a BP-based explanation method that attributes relevance by backpropagating the difference to a reference.

Consider a function $f : \mathbb{R}^n \to [0,1]$ representing a NN, an input $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, and a baseline input $\overline{x}$. The baseline input is typically chosen to represent the absence of features (e.g., an all-zero vector for images). The integrated gradient along the i-th dimension is defined as:

$$\Phi_{\mathrm{IG}_i}(\mathbf{x}) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial f(\overline{x} + \alpha(\mathbf{x} - \overline{x}))}{\partial x_i} \, d\alpha, \tag{3.2}$$

where α is a scalar that interpolates between the baseline $\overline{x}$ and the input $\mathbf{x}$. The idea is that if the function f changes with respect to an input feature x_i between $\overline{x}$ and $\mathbf{x}$, the integrated gradient for x_i should be non-zero. This ensures that if a feature contributes to the change in the output, it will be reflected in the attributions.

Sundararajan et al. [74] define two axioms that the method is supposed to fulfil. First, the **Completeness** or summation-to-delta property, which states that the sum of the attributions across all input features should be equal to the difference between the function's output at the input and the baseline. Mathematically:

$$\sum_{i=1}^{n} \Phi_{\mathrm{IG}_i}(\mathbf{x}) = f(\mathbf{x}) - f(\overline{x}). \tag{3.3}$$

This property ensures that the attributions account for the entire prediction difference between the input and the baseline. This criterion is also fulfilled byDeepLIFT and LRP. The second axiom is **Implementation Invariance**, meaning that if two models are functionally equivalent (i.e., they produce the same output for all inputs), the attributions should be identical. This property ensures that the attributions are independent of the specific implementation details of the model. This axiom does not hold for DeepLIFT and LRP.

Integrated gradients can be generalized to path methods, where the straight-line path between $\overline{x}$ and $\mathbf{x}$ is replaced by any smooth path $\gamma(t)$ such that $\gamma(0) = \overline{x}$ and $\gamma(1) = \mathbf{x}$. The path integrated gradient along the i-th dimension is defined as:

$$\Phi_{\mathrm{PathIG}_{\gamma,i}}(\mathbf{x}) = \int_0^1 \frac{\partial f(\gamma(t))}{\partial \gamma_i(t)} \frac{d\gamma_i(t)}{dt} \, dt. \tag{3.4}$$

Integrated gradients is a special case of this where $\gamma(t) = \overline{x} + t(\mathbf{x} - \overline{x})$.

In practice, the integral in the definition of IG is approximated using a summation:

$$\Phi_{\mathrm{IG}_i}^{\mathrm{approx}}(\mathbf{x}) = (x_i - x_i') \times \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(\bar{x} + \frac{k}{m}(\mathbf{x} - \bar{x}))}{\partial x_i}, \tag{3.5}$$

where m is the number of steps used in the approximation. Typically, m ranges between 20 and 300 to achieve a good approximation.

For the IG method specifically, Holzinger et al. [75] state that the main drawbacks are: the need for a baseline, that the model must be differentiable, and that it is integrated along the shortest path between the pixel to be attributed and the corresponding baseline pixel, which is not appropriate for all data. Compared to the Occlusion method, IG is much more efficient, as the attribution is calculated in a single backward pass.

3.3 Layer Relevance Propagation

Layer Relevance Propagation (LRP) [52, 2] is another BP-based attribution method. It assigns a relevance score R to each input feature of the network, indicating how much the feature contributed to the final output. R is determined by propagating the output decision back through the network to the input layer. LRP interprets the network as a flow graph (see Figure 3.1) and the redistribution of prediction relevance information backwards through the network using specific propagation rules. The theoretical justification of LRP is based on the Deep Taylor Decomposition [76].

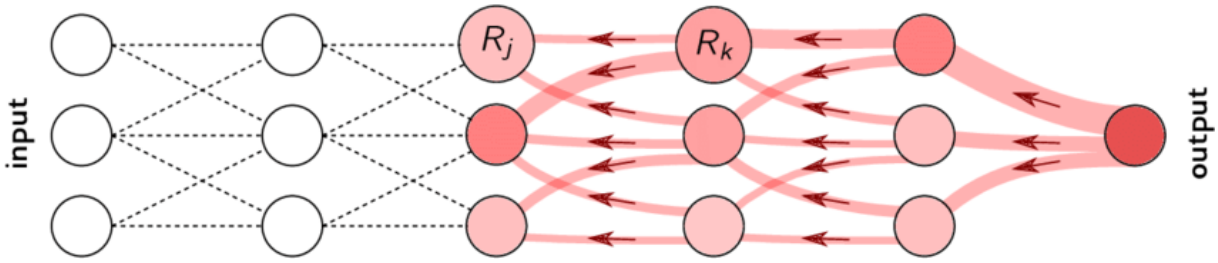


Figure 3.1: Visualisation of LRP (taken from [2])

The R's for a given layer l are transmitted to the neurons of the preceding layer by implementing the following rule, where j and k represent neurons in two successive layers of the NN:

$$R_j = \sum_k \frac{z_{jk}}{\sum_j z_{jk}} R_k. \tag{3.6}$$

This rule ensures that the relevance is conserved through the layers, ultimately distributing the output relevance back to the input features. Here, z_{jk} quantifies how much j has contributed to k's relevance.

The **Basic Rule (LRP-0)** [77] redistributes the relevance scores in alignment with the individual contributions of inputs to neuron activation, relevant to deep NNs utilizing rectifier (ReLU) non-linearities:

$$R_j = \sum_k \frac{a_j w_{jk}}{\sum_{0,j} a_j w_{jk}} R_k. \tag{3.7}$$

When applied across the entire network, it is equivalent to the Gradient $\times$ Input method, [39]. However, the basic rule is prone to gradient noise.

The **Epsilon Rule (LRP- ϵ)** [77] improves the basic rule byadding a small positive term ϵ in the denominator:

$$R_j = \sum_k \frac{a_j w_{jk}}{\epsilon + \sum_{0,j} a_j w_{jk}} R_k. \tag{3.8}$$

Finally, the **Gamma Rule** $(\mathbf{LRP} - \gamma)$ [77] is proposed by the authors. This rule favours the effect of positive contributions over negative contributions.

$$R_j = \sum_k \frac{a_j \cdot \left(w_{jk} + \gamma w_{jk}^{+}\right)}{\sum_{0,j} a_j \cdot \left(w_{jk} + \gamma w_{jk}.^{+}\right)} R_k \tag{3.9}$$

The parameter γ controls how many positive contributions are considered. As γ increases, the negative contributions start to disappear. The prevalence of positive contributions has a limiting effect on how large positive and negative relevance can grow in the propagation phase. This helps to deliver more stable explanations.

One question is how to distribute these rules across layers. Montavon et al. [2] state that upper layers, with approximately 4,000 neurons mix different concepts of classes. Thus, a rule like LRP-0 tends to be better, as it closely mirrors the function and its gradient, omitting these mix-ups. Middle layers offer clearer concept separation but face issues with layer stacking and convolution weight sharing, leading to unwanted variations. Here, LRP-ϵ helps by filtering out these variations. Lower layers, while similar to middle layers, benefit more from LRP-γ, as this rule distributes the relevance more evenly across features rather than focusing on individual pixels. [2] LRP can be used to explain a model's prediction in a single backward pass, making it an efficient method.

3.4 Deep Learning Important FeaTures

Deep Learning Important FeaTures (DeepLIFT) [39] is a BP-based algorithm designed to interpret the predictions of deep NNs by attributing the network's output to its input features. Unlike gradient-based methods, which often struggle with vanishing gradients and discontinuities, DeepLIFT explains the output of a NN in terms of the difference between the actual input and a reference input (baseline). This method allows for the attribution of output changes to specific input changes, even when the gradient is zero.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ be the input vector to a NN, f be the NN function, and t be the target output neuron. The reference input, denoted as $\bar{x}^0$, is chosen to represent a baseline state, and the corresponding reference activation of t is $t^0 = f(\bar{x}^0)$. The difference-from-reference for the target neuron t is defined as:

$$\Delta t = t - t^0 \tag{3.10}$$

DeepLIFT assigns contribution scores $C_{\Delta x_i \Delta t}$ to each input feature x_i, which represent the amount of the difference-from-reference in t that can be attributed to the difference-from-reference in x_i. These scores satisfy the summation-to-delta property:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t. \tag{3.11}$$

The contribution score $C_{\Delta x_i \Delta t}$ can be non-zero even when the gradient $\frac{\partial t}{\partial x_i}$ is zero, addressing a fundamental limitation of other gradient-based methods, like Guided Backpropagation (see Subsection 3.5). To propagate contributions through the network, DeepLIFT introduces multipliers. For a given input feature x_i and target neuron t, the multiplier $m_{\Delta x_i \Delta t}$ is defined as:

$$m_{\Delta x_i \Delta t} = \frac{C_{\Delta x_i \Delta t}}{\Delta x_i}. \tag{3.12}$$

These multipliers are analogous to partial derivatives but are calculated over finite differences rather than infinitesimal ones. Using these multipliers, DeepLIFT applies a chain rule for backpropagation similar to that used in standard NN training. For intermediate neurons y_j, the chain rule for multipliers is given by:

$$m_{\Delta x_i \Delta t} = \sum_j m_{\Delta x_i \Delta y_j}. \tag{3.13}$$

Similar to LRP, DeepLIFT utilizes different rules for the backpropagation for different layers. The **Linear Rule** handles linear transformations, such as those in dense or convolutional layers, let y be a linear function of its inputs x_i:

$$\begin{aligned} y &= b + \sum_i w_i x_i \\ \Delta y &= \sum_i w_i \Delta x_i \\ C_{\Delta x_i \Delta y} &= w_i \Delta x_i \\ m_{\Delta x_i \Delta y} &= w_i, \end{aligned} \tag{3.14}$$

with y being the difference-from-reference, $C_{\Delta x_i \Delta y}$ the contribution score for each Δx_i and $m_{\Delta x_i \Delta y}$ the corresponding multiplier. For nonlinear activations, such as ReLU, tanh, or sigmoid the **Rescale Rule** is defined. Let $y = f(x)$. The difference-from-reference for y is:

$$\begin{aligned} \Delta y &= (x) - f(x^0) \\ C_{\Delta x \Delta y} &= \Delta y \\ m_{\Delta x \Delta y} &= \frac{\Delta y}{\Delta x}. \end{aligned} \tag{3.15}$$

Since y has a single input x, the contribution score is given directly by the summation-to-delta property. This rule ensures that contributions are scaled appropriately based on the ratio of the

change in output to the change in input. The **RevealCancel Rule** can be utilised when positive and negative contributions have to be treated differently. For a neuron y with inputs x_i, let Δy^+ and Δy^- represent the positive and negative components of Δy, respectively:

$$\Delta y = \Delta y^+ + \Delta y^-. \tag{3.16}$$

The contributions $C_{\Delta x_i^+ \Delta y^+}$ and $C_{\Delta x_i^- \Delta y^-}$ are computed by considering the average impact of positive and negative differences, alleviating issues where positive and negative contributions might cancel each other out. Similar to the other BP-based methods DeepLIFT can also be calculated in a single backward pass.

3.5 Gradient-weighted Class Activation Mapping

Class Activation Mapping (CAM) [78] provide a visual representation of where specific classes are activated within the images in a deep learning model. This technique involves the aggregation of activation maps from the final convolutional layer before a global pooling layer using linear weights, highlighting regions of interest that significantly impact the output class.

Mathematically, let f denote the model, with l representing the global pooling layer that follows the final convolutional layer $l-1$ and precedes the fully connected layer $l+1$. For a given class c, the CAM for that class is calculated as:

$$L^c_{\text{CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A_k^{l-1}\right), \tag{3.17}$$

where $\alpha_k^c = w^c_{l,l+1}[k]$ is the weight associated with the k-th channel at the fully connected layer following the pooling layer. However, CAM is limited to network architectures that incorporate global pooling layers, and cannot applied to models that include additional fully connected layers before the classification layer, such as VGG [79].

Gradient-weighted Class Activation Mapping (GradCAM) [19] extends CAM by incorporating gradient information flowing into any convolutional layer, typically the last one due to its high-level semantics and spatial information retention. The class discriminative localisation map, L^c_{GradCAM}, for a specific class c is computed by first capturing the gradients of the class score y^c with respect to the activations A_k of convolutional layer l, which produces K feature maps each of dimension $u \times v$. The importance weights α_k^c for each feature map are determined by globally averaging the gradients over all spatial locations (i, j):

$$\alpha_k^c = \frac{1}{Z}\sum_{i,j}\frac{\partial y^c}{\partial A^k_{ij}}. \tag{3.18}$$

These weights represent a partial linearisation of the network and denote the importance of each feature map k for the target class c. A weighted combination of these maps followed by a ReLU ensures that only positive influences on the class prediction are visualized:

$$\Phi^c_{\text{GradCAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right). \tag{3.19}$$

Similar to the Occlusion method, the resulting attribution map is upscaled to the image size.

Guided Backpropagation (GBP), introduced by Springenberg et al. [41], combines ideas from Deconvolution [18] and Saliency [38] techniques to improve the visualisation of CNNs. The authors identified a problem with the flow of negative gradients in the saliency method, which reduces the accuracy of visualisations in higher layers. GBP addresses this by focusing on the ReLU activation function and masking non-positive values during backpropagation to remove noise from negative gradients. This method effectively guides the backpropagation process, producing clearer and more interpretable explanations by retaining only positive contributions to the output, thereby producing sharper visualisations of network features.

Guided Gradient-weighted Class Activation Mapping [19] combines GradCAM with the GBP approach. Selvaraju et al. [19] aimed to enhance the blurry GradCAM attribution map with the sharp GBP [41] attribution map to create a clearer final explanation. Formally:

$$\Phi_{\text{GuidedGradCAM}}(\boldsymbol{x}) = \Phi_{\text{GradCAM}}(\boldsymbol{x}) \odot \Phi_{\text{GDP}}(\boldsymbol{x}), \tag{3.20}$$

where $\odot$ denotes the Hadamard product.

Although GradCAM and Guided GradCAM are model-agnostic, calculated in a single backward pass and applicable to various CNN architectures, they still suffers from limitations in scenarios where gradients do not provide informative insights into the model's decision process:

The saturation gradient in deep NNs can be noisy and tend to disappear due to saturation in the sigmoid or flat zero gradient region in ReLU activations. This leads to visually noisy gradients in the output with respect to input or internal layer activations [15]. False confidence arises because Φ^c_{GradCAM} uses a linear combination of activation maps. Given the activation maps A_{il} and A_{jl} along with the corresponding weights α_{ci} and α_{cj}, where $\alpha_{ci} \geq \alpha_{cj}$, it is assumed that the input region responsible for producing A_{il} is at least as significant as the region producing A_{jl} with respect to the target class 'c'. However, GradCAM can present instances of deceptive certainty, where maps with higher weights contribute less to the output of the network compared to a zero baseline [15].

3.6 Local Interpretable Model-agnostic Explanations

Local Interpretable Model-agnostic Explanations (LIME) [42] is a framework to create locally faithful and interpretable surrogate models that explain the predictions of a black-box machine learning model on an individual instance basis. These explanations are generated by perturbing the input data and observing the variations in model predictions. For an input $\boldsymbol{x} \in \mathbb{R}^d$, its interpretable representation is $\boldsymbol{x}' \in \{0,1\}^{d'}$. In the context of images, $\boldsymbol{x}'$ may represent the presence or absence of superpixels, where a superpixel is defined as a contiguous patch of similar pixels, simplifying the input features into more manageable segments.

G denotes the class of interpretable models, such as linear models or decision trees, used within the LIME framework. In LIME, an interpretable model $g \in G$ is defined with domain $\{0,1\}^{d'}$ and serves as the explanation model. For instance, g can be a linear model, which is inherently interpretable. Consider a linear classifier $g(x) = w^T x$. A feature x_i is relevant to the prediction if $w_i x_i > 0$. If $w_i = 0$, the feature is effectively ignored by the model; if $x_i = 0$,

it indicates that the feature is absent in the input. The relevance of each feature x_i can be quantified by $R_i = w_i x_i$, a method often referred to as explanation by decomposition.

$$\Phi(x) \approx \sum_{i=1}^{d} R_i. \tag{3.21}$$

To maintain simplicity, the complexity of g is controlled by a regularisation term $\Omega(g)$, which might be the number of nonzero weights in a linear model or the depth of a decision tree. The black-box model being explained is denoted as $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}$. Additionally, $\Pi_x(z)$ is defined as a proximity measure that quantifies the closeness of an instance z to x, thereby defining a locality around x. The faithfulness of g in approximating $\boldsymbol{f}$ within this locality is quantified by $\mathcal{L}(\boldsymbol{f}, g, \Pi_x)$. The optimisation objective to find the best explanation model is then formulated as:

$$\Phi_{\text{LIME}}(x) = \text{argmin}_{g \in G} \mathcal{L}\left(\boldsymbol{f}, g, \Pi_x\right) + \Omega(g). \tag{3.22}$$

Sparse Linear Explanations: One application for this framework could be to use LIME to generate sparse linear explanations, where $g(z') = w_g \cdot z'$. This form of explanation prioritizes simplicity by focusing only on the most influential features. The proximity measure used in this context is an exponential kernel $\Pi_x(z) = \exp\left(-D(x,z)^2/\sigma^2\right)$, where D is the L2 distance, suitable for comparing image data.

The regularisation term $\Omega(g) = \infty \mathbb{I}\left[\left\|w_g\right\|_0 > K\right]$ imposes a hard constraint on the model complexity by limiting the number of non-zero weights in g to the number of superpixels that can be active in g to K. If the number of non-zero weights exceeds K, the regularisation term renders the model infinitely complex.

$$\Phi_{\text{LIME}}\left(\boldsymbol{f}, g, \Pi_x\right) = \sum_{z, z' \in \mathcal{Z}} \Pi_x(z) \left(\boldsymbol{f}(z) - g\left(z'\right)\right)^2. \tag{3.23}$$

A limitation of LIME is that it only indirectly solves the explanation problem by relying on a surrogate model. Consequently, the quality of the explanation depends largely on the quality of the surrogate fit, which may require dense sampling and result in high computational costs. In addition, sampling introduces uncertainty, leading to non-deterministic behaviour and variable explanations for the same input sample [75]. This is also reflected in the complexity of LIME, as training the surrogate is very time-consuming.

3.7 Challenges of Explanation Methods for Multi-Label Remote Sensing Images

Interpreting ML models, particularly in the context of multi-label RS images, poses significant challenges due to the complex patterns and repetitive textures that span multiple spectral bands, making it difficult to explain model predictions. Explanation methods often struggle with RS data due to unique image properties such as different sources, scales, geographic relationships and temporal dependencies [80, 32]. The resolution and extent of RS define the granularity and boundaries of observed features, with some, such as land cover or mountains, lacking clear boundaries and exhibiting continuous, irregular shapes. This spatial resolution

has implications for semantic interpretation, as the ability to distinguish and interpret features depends heavily on the presence or absence of information, a primary challenge in RS and xAI [80, 32].

Moreover, the spectral resolution, influenced by each sensor's capacity to capture data across different wavelengths, plays a crucial role in the information extraction process, distinguishing multispectral and hyperspectral images from RGB images. Current literature and xAI evaluations, however, often overlook the diverse spectral properties and resolutions among platforms, compromising the effectiveness of explanations [32]. For instance, De Lucia et al. [36] showed that the traditional GradCAM is ineffective for hyperspectral images processed with 3D CNNs. They propose 3D GradCAM, an extension specifically designed for 3D CNNs. Huang et al. [35] demonstrated that CAM-related methods struggle with RS images containing multiple objects of interest. In an example where a RS image contained five aeroplanes, GradCAM managed to locate only one. The authors attribute this limitation to the inherent complexity of RS images, which typically encompass more, smaller, and intricately arranged objects compared to natural images. However, the line of argumentation can also be applied to multi-label images.

Adebayo et al. [81] state that BP-based explanation methods tend to produce edge-like visual artefacts. This is because edges in the image cause a distinct change in pixel values, creating unique activation patterns in the ReLU units. This results in varying gradients around the edges, making them stand out in the attribution map. In contrast, uniform regions of the image where pixel values are similar will have the same activation pattern and therefore similar gradient values. As RS image data often contains repetitive local textures, this focus on edges could lead to misleading saliency maps that highlight irrelevant features. As a result, important patterns in RS data may be missed.

However, BP-based methods also suffer from general problems. Nie et al. revealed that GBP and DeconvNet, rather than emphasizing class-relevant pixels or visualizing the learned weights, effectively perform partial image recovery. This indicates that these methods do not directly relate to the decision-making processes of neural networks [82]. In addition, Sixt and Landgraf [83] show that methods like LRP tend to converge to rank-1 matrices, ignoring the contributions of later layers and leading to class-insensitive explanations. These methods produce similar saliency maps for different classes and fail to highlight class-specific features. They find that onlyDeepLIFT maintains class sensitivity by considering both positive and negative relevance scores, preventing convergence to rank-1 matrices and accurately representing the contributions of the final layers.

One challenge, particularly for the Occlusion method for RS images, is that these images often contain classes that occupy large contiguous areas, such as forests, water bodies, and urban areas. When using a fixed patch size $ps \times ps$ for occlusion, there is a risk that the occluded patch may not cover enough of a particular class to significantly affect its classification probability. This is particularly problematic for classes that are spread over large areas because even when a patch is occluded, sufficient class information remains in the visible parts of the image.

Formally, let c be a class in the MLC, and let $\text{Area}(c)$ be the total number of pixels covered by class c in the image x. Let P be the patch of size $ps \times ps$ used for occlusion, and let T be a threshold representing the minimum number of class c pixels that must be occluded to significantly affect the classification probability for that class. If the total area occupied by class

c in the image is such that:

$$\text{Area}(c) > ps^2 + T, \tag{3.24}$$

then the classification probability $f_c(x')$ for c may not decrease significantly. This is because the occlusion patch does not remove enough class-specific information from the image x to affect the classifier's confidence. To address this issue, one possible solution is to use adaptive patch sizes that vary according to the characteristics of the class and the spatial resolution of the image. Adaptive patches can be dynamically sized based on the extent of the class in the image, ensuring that a sufficient amount of class information is occluded to significantly affect the classification probability. However, pixel-wise ground truth is not always available.

Furthermore, there is an additional limitation to the Occlusion method and its sliding window approach. Classic CV images typically contain unique, complex features that are representative for their class, such as a cat's ears. If these features are occluded, the prediction certainty should decrease significantly. However, for a land cover class such as forest, there are more repetitive, simple features that occupy the entire area of the class, making texture the more important feature. For example, a nadir view of trees in a forest class demonstrates this simplicity and repetition, with each tree having the same relevance to the final prediction. If one or more of these repetitive features are masked, the final result should not change significantly. This problem also applies to other perturbation-based methods.

As the name implies, perturbation-based methods perturb the input image and evaluate changes in the model's prediction. Depending on whether the pixels or patches of the input image are perturbed, the explanation has a different granularity [84]. Usually, this perturbation is done by occluding [18], blurring [85], introducing noise [44] or replacing parts of the image [86]. However, they often produce samples that do not lie in the original distribution of the data, so-called Out-of-Distribution (OoD) samples [87]. This can be a problem, as noisy samples can produce a prediction with high confidence, e.g. an MNIST classifier producing a prediction with 91% confidence on a Gaussian noise input [88, 87]. For example, research by [89] has shown that when there are discrepancies in the distribution between perturbation samples and input data, explanatory methods such as LIME [42] and SHAP [49] can be misleading, resulting in unbiased explanations for biased classifiers. To address these issues, Qiu et al. [87] propose to approximate the probability of a sample coming from the original distribution and use it to weight the attributions. They apply it to LIME, RISE and Occlusion [18], improving the explanations visually.

In the context of multispectral RS images, the concept of 'absence of features' is even more difficult to construct than for classical RGB images. Defining a baseline to perturb areas of an image is more difficult because different spectral bands (e.g. RGB, near-infrared, thermal) capture different information. Therefore, common baselines, such as all-zero vectors, may not be appropriate for all bands. For example, an all-zero baseline in thermal imagery is not meaningful and can misrepresent the importance of features. Another example is that an all-zero baseline may not be appropriate for a near-infrared band because some infrared radiation is always present in real-world scenarios. Inappropriately chosen baselines can lead to inaccurate IG that do not accurately reflect the importance of features. This is the case for four of the explanatory methods considered: LIME, Occlusion, as they are both perturbation-based techniques, Additionally, it holds for IG and DeepLIFT, as they both use a 'difference-to-reference' approach.

Moreover, Hooker et al. [90] show that the perturbation introduces artefacts and leads to a distribution shift. They propose the RemOve And Debias (ROAD) framework with an expensive model retraining step to adapt to the distribution shift and evaluate explanations. Rong et al. [91] show that simple occlusion can leak class information that may not be present in the feature values. To address these issues, Brock and Chung [92] propose the fidelity estimation of explanation methods by examining model accuracy curves when input features are perturbed according to the Most Relevant First (MoRF) and Least Relevant First (LeRF) orders. The method involves calculating accuracy curves by first perturbing images using a blurring technique based on the importance scores of input pixels. The areas above these curves, up to a certain fraction of the perturbed pixels, are quantified to assess the impact of artefacts introduced by the perturbation. By comparing the changes in accuracy with random masks and corresponding masks, the influence of the artefact can be isolated, allowing a more accurate estimate of the fidelity of an explanation method. However, this problem could be mitigated by considering the multi-label nature of RS data. Inpainting with an image of a different class should not be a problem. If the reference map is known, the same image can be used; otherwise, a different image with no overlapping multi-labels could be used. Due to the repetitive local textures, inserting a different image may not result in an OoD perturbation. For future work, this approach could be implemented and tested, possibly using the inlier score of [87]. However, this is beyond the scope of the current study.

Nonetheless, in order to address these issues, this thesis adopts a multi-faceted evaluation approach using different explanatory metrics, which are discussed in the following chapter.

4 Evaluating the Quality of an Explanation

The evaluation of xAI method, together with xAI itself, poses an entirely new research field compared to the evaluation of traditional AI classifiers. The main reason behind this is, that an explanation, as defined in section 2.2 is a "presentation [...] in human-understandable terms" [1], meaning that it is naturally difficult to measure. It cannot be compared to a ground truth and measured with metrics like a classification problem. Furthermore, there are more variables the explanation quality depends on, than just the method. For instance, even the best explanation method will provide an implausible-looking explanation when the model that is explained has an error in its reasoning [67]. However, the explanation would still faithfully reflect the model's behaviour. Zhang et al. [93] identify this as one of the main problems when explanations are evaluated visually. Because of this lack of ground truth, the evaluation function is unverifiable, leading to an also unverifiable explanation evaluation, which Hedström et al. refer to as the *Challenge of Unverifiability* [72]: The lack of ground truth labels, makes it impossible to verify the accuracy for an explanation.

Doshi-Velez and Kim [20] categorize the evaluation of explanation methods into three levels. The highest level is called *application-grounded evaluation*, which includes experiments with domain experts in a real-world setting, allowing the method to be assessed by its intended users for specific tasks [20]. The intermediate level, *human-grounded evaluation*, consists of studies with non-expert individuals performing simplified tasks that capture the core of the actual application [20]. This level is appropriate when the aim is to assess broader aspects of explanation quality rather than a specific objective, or when accessing the target audience is impractical due to factors like high costs or limited availability of domain experts. These controlled human experiments can yield either subjective outcomes through user-reported perception of quality, or objective outcomes by measuring task performance. The third level, known as *functionally-grounded evaluation*, does not require human testing but instead relies on computational metrics to evaluate interpretability [20]. In their opinion, human evaluation is typically the "gold standard" to evaluate explanations [20].

However, others argue for more quantifiable evaluations for xAI: Ancona et al. [94] suggest that user studies inherently favour simpler, more expected explanations at the expense of methods that may more accurately represent network behaviour. Furthermore, findings from user studies are often neither replicable nor comparable [95] and conducting these studies is both time-consuming and costly [96]. Additionally, involving users in the evaluation process might unintentionally mix assessing the accuracy of explanations with the correctness of the predictive model itself [1]. A quantitative approach enables formal comparisons among different explanation techniques, providing a more objective method of evaluation [96]

Often, researchers present individual examples that appear sensible and pass the initial test

of having "face-validity" [20], but these examples can be selectively chosen and potentially misleading [97]. Many experts contend that relying solely on such anecdotal evidence is inadequate and can be deceptive [81]. Leavitt and Morcos [98] note the frequent assumption that an explanation method and its outcomes are reliable, cautioning that explanations that look intuitively right can lead to wrong outcomes. They argue that the absence of quantitative evaluation hinders interpretability research because anecdotal assessments do not provide a solid foundation for verification. Furthermore, various studies highlight that evaluating the plausibility and persuasiveness of an explanation to humans is distinct from assessing its accuracy, and these criteria should not be merged [1].

To address the challenge of proper quantification, numerous metrics have been developed to assess the quality of an explanation. Hedström et al. [61] classify these metrics into six primary categories: a) *Faithfulness*, b) *Robustness*, c) *Complexity*, d) *Randomisation*, e) *Axiomatic* and f) *Localisation*. However, as the research field is still evolving, there is no established consensus on which specific metric should be used within these broad categories. This lack of standardisation makes it difficult to compare results across different studies, even if they are evaluating the same fundamental aspect of their xAI method.

A different categorisation is given by Nauta et al. [1]: the CO12 criteria, which involves the 12 criteria: *Correctness, Completeness, Consistency, Continuity, Contrastivity, Covariate Complexity, Compactness, Composition, Confidence, Context, Coherence, Controllability*. Correctness pertains to the accuracy of an explanation in reflecting the true operation of the model. Completeness involves the coverage of the entire model behaviour within the explanation. Consistency ensures that the explanation remains uniform across different implementations or when given identical inputs. Continuity demands that small variations in input should not lead to disproportionate changes in the explanation, promoting stability. Contrastivity enhances the explanation by clearly distinguishing between different possible outcomes and addressing "why not?" or "what if?" scenarios[1]. Further, Covariate Complexity relates to the explanation's complexity and the extent to which it uses human-understandable concepts. Compactness emphasizes the importance of conciseness in explanations, advocating for maximizing information while minimizing detail. Composition refers to the structural and presentational aspects of an explanation, affecting how it is perceived and understood. Confidence integrates probabilistic information to quantify the certainty of explanations, providing a measure of reliability [1]. Context assesses the relevance of the explanation to the user's specific needs and practical scenarios, ensuring the explanation's utility. Coherence guarantees that the explanation aligns with existing knowledge and beliefs, making it plausible and reasonable to users. Finally, Controllability offers users the ability to interact with and modify the explanation, tailoring it to their specific requirements and enhancing personal engagement with the AI system. Together, these criteria form a comprehensive framework for evaluating and designing explanations in AI systems, ensuring they are both effective and user-centric. [1]

Le et al. [99] recommend Quantus [61] as the most apparent choice for evaluating explanations due to its maturity, thorough documentation, and comprehensive coverage of different explanation types and Co-12 criteria. Moreover, they discourage using different toolkits, as the implementation of a metric has a significant impact on the output value [99]. Therefore, in this thesis, Quantus is used for the evaluation. As the Quantus framework utilizes the metric categories as proposed by Hedström et al. [61], this thesis adopts Hedström's perspective. Initially, the various categories will be described more thoroughly; subsequently, each metric will

be theoretically described and analyzed for usability with RS MLC images. The notation for the evaluation metrics follows the guidelines defined in Section 2.2 and is also adopted from Hedström [72].

The first of the six metrics categories, as described by Hedström et al. [61], is **Faithfulness**. This metric category tries to measure how closely aligned an explanation is to the true inner decision-making process of the corresponding model. A good explanation is expected to truly explain the decision of the model. This is a fairly important category, as the internal behaviour of the model cannot be evaluated by humans. The second category is **Robustness** [61], which is measuring how invariant the explanation is to small perturbations of the input data. A key point is that this metric might lead to explanations that are more resistant to adversarial perturbations. In this context, an adversarial perturbation refers to an image that is perceptually similar yet, despite having the same predicted label by the neural network, receives a significantly different interpretation [100]. The **Complexity** [61] of an explanation regards its overall size. For images, a high-quality explanation is expected to produce a saliency map that concentrates on specific, crucial areas of the image, rather than indiscriminately highlighting everything as significant to the explanation [101]. **Randomisation** metrics [61] assesses how significantly an explanation changes when the internal layers of the model, and thus the final prediction output, are randomized. This metric is closely associated with the category of faithfulness. Effective explanations under this metric should closely correlate with the prediction. A failure to do so could lead to generic explanations that reflect only the internal structure of the model rather than the specific prediction it makes. **Axiomatic** [61] metrics test whether explanations fulfil certain axiomatic properties. The final category, **Localisation** [61] assesses the extent to which an explanation concentrates on the actual regions of interest. Reliance on ground truth data is critical in this category, as it provides the precise locations of classified objects through bounding boxes or reference maps.

An additional difficulty in the quantitative evaluation of an explanation is the parameterisation of the explanation method [91]. An example of this is the choice of perturbation strategy, which is discussed in more detail in the section 3.7. These difficulties can be transferred to the parameterisation of the evaluation metric, which often also requires a perturbation strategy. These difficulties make the quantitative evaluation of an explanation "easy to get [...] wrong" [72].

4.1 Faithfulness Metrics

The first category to examine closer is **Faithfulness**. Metrics in this category quantify the alignment of the model internal and the explanation. Usually, these metrics evaluate if features that are deemed more relevant affect the prediction of a model strongly. To determine that the input $\boldsymbol{x}$ is perturbed. Hedström [72] defines this category as:

$$\Psi_{\mathrm{F}}(\Phi, f, \boldsymbol{x}, \mathcal{P}, M) = \mid (f(x) - f(\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}, M)) \mid \tag{4.1}$$

with $\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}, M)$ being the perturbation function where M is the number of input features that are perturbed. The selection of $\mathcal{P}$ is crucial for the reliability of such metrics [102], and discussed more thoroughly in Sections 3.7 and 4.7. In the following, metrics such as Faithfulness Correlation, Faithfulness Estimate, Pixel-Flipping, Region Perturbation, Selectivity, Sensitivity-

n, Iterative Removal Of Features, RemOve And Debias, Infidelity, and Sufficiency are defined and analyzed.

Faithfulness Correlation (FC) [101] estimates the faithfulness of an explanation method Φ iteratively. The core idea is that when a subset of indices $s \in |S| \subseteq d$ of sample $\boldsymbol{x}$ is set to a baseline value $\overline{\boldsymbol{x}}$, defined as $\boldsymbol{x}_{[\boldsymbol{x}_s=\overline{\boldsymbol{x}}_s]}$, the change in the model's output should be proportional to the sum of attribution scores of these features in $\boldsymbol{x}_s$.

$$\Psi_{\text{FC}} = \text{corr}_{s \in |S| \subseteq d} \left(\sum_{i \in S} \Phi(\boldsymbol{x})_i, f(\boldsymbol{x}) - f\left(\boldsymbol{x}_{[\boldsymbol{x}_s=\overline{\boldsymbol{x}}_s]}\right) \right), \tag{4.2}$$

Higher values indicate a stronger correlation between the explanation method's attribution and the model's behaviour, thus being preferred. The Quantus implementation [61] iteratively replaces a random subset of given attributions with a baseline value, then measures the correlation between the sum of this attribution subset and the difference in function output.

When applied to RS data, this approach may encounter issues. Depending on the amount and locality of the indices to be replaced, random replacement of pixels may not be effective. This method works well for CV images containing unique, complex features that significantly decrease prediction confidence when removed. However, for RS images, a random replacement strategy may not sufficiently disrupt the repetitive texture to influence the prediction, potentially resulting in a lack of correlation. Thus, the metric was not included in the further analysis.

Faithfulness Estimate (FE) [103] is a simplification of FC. It evaluates the similarity between the predictions for a masked input and the attribution of this masked input. Mathematically, it is expressed as:

$$\Psi_{\text{FE}} = SIM_{k \in |S| \subseteq d} \left(\Phi(\boldsymbol{x}_s), \left(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}_{[\boldsymbol{x}_k=\overline{\boldsymbol{x}}_k]}) \right) \right), \tag{4.3}$$

with $\boldsymbol{x}_{[\boldsymbol{x}_k=\overline{\boldsymbol{x}}_k]}$ denoting the vector after the top-k most relevant features have been replaced. For efficient calculation, Quantus [61] implements this by sorting the indices to be perturbed by their relevance according to the explanation method, in descending order. Perturbing these indices should show a positive correlation with the model's prediction confidence. In the implementation, the Pearson correlation coefficient ρ is used as the similarity measure SIM, although other measures are possible.

The main difference to FC is that the masking is not random but ordered by the attributed relevance. For a RS use case, this approach offers some improvement but remains inadequate as the relevant pixels do not necessarily need to be nearby. Consequently, this method may still fail to disrupt the texture characteristics sufficiently to reduce prediction confidence. Therefore, this metric is not selected for further evaluation.

Monotonicity by Arya et al. [104] evaluates whether incrementally adding positive evidence increases the classification probability for a given class. The process starts with a baseline image $\overline{\boldsymbol{x}}$ where all features are set to a neutral value. The features in $\overline{\boldsymbol{x}}$ are then iteratively replaced with their actual values from $\boldsymbol{x}$ in ascending order of their relevance scores according to $\Phi(\boldsymbol{x})$. Let $\boldsymbol{x}_{[\boldsymbol{x}_k=\overline{\boldsymbol{x}}_k]}$ represent the image after replacing the top k most relevant features with $k \in |S| \subseteq d$. For each $\boldsymbol{x}_{[\boldsymbol{x}_k=\overline{\boldsymbol{x}}_k]}$, the classification probability $\boldsymbol{f}\left(\boldsymbol{x}_{[\boldsymbol{x}_k=\overline{\boldsymbol{x}}_k]}\right)$ for the specified class is calculated. The Monotonicity metric is satisfied if the sequence of classification probabilities

$\{p_k\}$ is monotonically increasing, i.e. $p_1 \leq p_2 \leq \ldots \leq p_n$.

$$\Psi_{\text{MO}} = \begin{cases} 1 & p_1 \leq p_k \leq \ldots \leq p_n, \\ 0 & \text{otherwise} \end{cases}. \tag{4.4}$$

Because this metric starts from a baseline and iteratively adds features sorted by relevance, it is well suited to evaluating explanations for RS images. The reasoning is that even if the explanation assigns high relevance to scattered pixels, starting from a baseline, adding these pixels should only increase prediction accuracy. This is particularly relevant for MLC cases, where the presence of highly relevant pixels should be sufficient to increase prediction confidence. Therefore, it will be used in the experiments.

Pixel-Flipping (PF) [77] originally flipped pixels from 0 to 1 sorted by their assigned relevance for greyscale images. However, to adapt this method for images with more channels, the flipping process has been changed to a perturbation approach. Here, Hedström's notation is followed [72], where PF is defined as the Area Under the Curve (AUC) of the prediction confidence over a set of masked pixels:

$$\Psi_{\text{PF}} = \sum_{i=1}^{n} \left(f(\boldsymbol{x}'_i) + f(\boldsymbol{x}'_{i+1}) \right) \cdot \frac{\boldsymbol{x}'_{i+1} - \boldsymbol{x}'_i}{2}, \tag{4.5}$$

where n is the number of discrete perturbation steps, $\boldsymbol{x}'_i$ and $\boldsymbol{x}'_{i+1}$ are the x-values for the i^{th} and $(i+1)^{\text{th}}$ perturbation steps. For faithful explanations, a steep degradation of prediction scores is expected when attributions are iteratively replaced in descending order. Thus, a lower AUC value indicates better performance.

The reasoning here is similar to the FC metric, which can be seen as an improvement over PF. Flipping individual pixels does not change the prediction score as drastically for RS images. It is therefore not adopted.

Region Perturbation (RP) [45] is an extension of PF that involves flipping an area rather than a single pixel. Similarly, a lower value of AUC is indicative of better performance.

$$\Psi_{\text{RP}} = \sum_{i=1}^{n} \left(f(\boldsymbol{x}'_{R_i}) + f(\boldsymbol{x}'_{R_{i+1}}) \right) \cdot \frac{\boldsymbol{x}'_{R_{i+1}} - \boldsymbol{x}'_{R_i}}{2}, \tag{4.6}$$

where again n is the number of discrete perturbation steps, however, $\boldsymbol{x}'_{R_i}$ and $\boldsymbol{x}'_{R_{i+1}}$ are the $\boldsymbol{x}$-values after perturbing regions R_i and R_{i+1}, respectively.

For the implementation, a parameter for the region size is specified to determine the size of the area to be flipped. An additional parameter is the order of removal or the removal strategy. Samek et al. [45] define two orders: Most Relevant First (MoRF) and Least Relevant First (LeRF). The former removes the most important pixels first, while the latter removes the least important pixels. These strategies are described in more detail in the Section 4.7. To analyse the influence of removal orders on RS data, this method will only be examined on the RS datasets.

Selectivity (SEL) [105] is closely related to PF and RP. It is calculated by AUC of the curve of the prediction value when the most relevant features are removed. Let $\boldsymbol{x}_{[x_k = \bar{x}_k]}$ with $k \in |S| \subseteq d$

be the perturbation after the k most important features attributed by Φ are removed.

$$\Psi_{\text{SEL}} = \text{AUC}_{k \in |S| \subseteq d}(f(\boldsymbol{x}_{[\boldsymbol{x}_k = \bar{\boldsymbol{x}}_k]})). \tag{4.7}$$

Similar to the metrics before, a sharp drop in the function's value, summarized by a low AUC score, indicates that the correct features have been identified as relevant.

Although this metric shares some of the disadvantages of FE, namely that individual pixels must have a strong influence on the prediction confidence, it is still considered for the quantitative analysis. This consideration is purely from a scientific point of view, as it will be compared with the IROF metric described next.

Iterative Removal Of Features (IROF) [106] combines elements of RP and SEL. Like SEL it measures how the prediction changes when the most relevant input features are removed, but like RP it considers segments rather than individual pixels. Each image $\boldsymbol{x}$ is partitioned into a set of segments $\left\{S^l\right\}_{l=1}^{L}$, where $s_{i,j}^l = 1$ indicates that pixel $x_{i,j}$ belongs to segment l. Let $\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}^l$ be the masked image for segment l. The segments are then sorted in descending order by their average relevance. Let $\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}^k$ be the image with the k most relevant features removed. The prediction $f(\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}^k)$ is computed iteratively for all $k \in 0, \ldots, L$, resulting in a prediction curve. To normalise the curve, the predictor value is divided by $f(\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}^0)$, resulting in

$$\Psi_{\text{IROF}} = \frac{1}{N} \sum_{n=1}^{N} AOC \left(\frac{f(\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}^k)}{f(\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}^0)} \right)_{k=0}^{L}. \tag{4.8}$$

A faithful explanation method should have a steep decrease in this function. However, unlike SEL or RP, Area Over the Curve (AOC) is measured, so a higher AOC value is desired.

This method, tailored to the characteristics of RS image data, considers segments rather than individual pixels in its removal strategy. While the use of the mean attribution value may not perform as well on natural images, where complex features consisting of edges are most relevant, it is particularly effective at capturing the repetitive textures typical of RS data. Therefore, this method is used in the experiments.

RemOve And Debias (ROAD) [102] follows a similar approach to SEL. It measures the accuracy of f in an iterative process of removing the k most important pixels in MoRF order. The main novelty in this method is the perturbation strategy. The authors demonstrate that simple perturbation strategies can leak class information through the shape of the perturbed mask [102]. They propose *Noisy Linear Amputations*, where each perturbed pixel is approximated by the weighted mean of its neighbours.

However, due to the nature of RS images, class leakage may not be possible. For example, in an image that contains a *jaguar*, if this particular object is removed, the perturbed shape may still resemble a "*jaguar* shape". Conversely, if a segment containing forest is occluded, it is not possible to tell whether it is forest or grassland. Neighbouring pixels may slightly increase the probability of a contiguous forest patch, but it may not be a discriminative feature. In addition, this method involves a computationally expensive retraining step. It is therefore not considered.

Sensitivity-n (SENS-N) [94] is based on the principle that segments of input features have a

more significant impact than individual pixels. This metric asserts that the decrease in output when multiple inputs are occluded should correspond to the sum of their relevances. An attribution method satisfies SENS-N if, for every subset of features of cardinality n, the sum of the attributions equals the change in the output caused by occluding these features. Formally, for all subsets of features $x_S = [x_1, \ldots, x_n] \subseteq x$, the following condition should hold:

$$\Psi_{\text{SENS}-\text{N}} = \begin{cases} 1 & \sum_{i=0}^{N} R_i^c(x) = \boldsymbol{f}(x) - \boldsymbol{f}(\overline{\boldsymbol{x}}), \\ 0 & \text{otherwise} \end{cases}, \tag{4.9}$$

where $\overline{\boldsymbol{x}}$ is defined as a baseline input from which all features have been removed. This formulation essentially represents the definition of **Completeness** or Summation to Delta, making SENS-N a generalisation of these concepts.

However, SENS-N can be seen as a characteristic of an explanation method. The authors state that Occlusion-1 satisfies Sensitivity-1 by construction, and IG, DeepLIFT, and ϵ-LRP satisfy SENS-N under specific conditions [94]. Therefore, it is not considered in the analysis.

Infidelity [44] is a broader concept than SENS-N. It also relates to the completeness axiom, which asserts that the total of the attributions should equal the difference between $\boldsymbol{f}$'s output at its input and its baseline. Mathematically, it is defined as follows:

$$\Psi_{\text{INFD}} = \mathbb{E}_{\mathbf{I} \sim \mu_{\mathbf{I}}} \left[\left(\mathbf{I}^T \Phi(\boldsymbol{f}, \boldsymbol{x}) - (\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x} - \mathbf{I})) \right)^2 \right], \tag{4.10}$$

where $\mathbf{I} \in \mathbb{R}^d$ is a random variable with probability measure $\mu_{\mathbf{I}}$ representing a meaningful perturbation. $\mathbf{I}$ signifies significant perturbations around $\boldsymbol{x}$ and can be specified in various ways. Similar to SENS-N and with the same reasons, Infidelity is not considered for further experiments.

Sufficiency [107] tries to quantify whether similar explanations have the same prediction label. A key requirement for Φ is that if a particular property π justifies the classification of an instance $\boldsymbol{x}$, then any other instance $\boldsymbol{x}'$ possessing property π should be classified similarly, regardless of whether the explanation $\boldsymbol{e}(\boldsymbol{x}')$ provided differs from π. For instance, if the presence of a cat's ear in an image justifies classifying the image as a cat, then any other image that also shows a cat's ears should be classified as a cat, even if a different attribute is highlighted in the explanation.

Explanations $\boldsymbol{e}$ are considered sufficient if, for any instance $\boldsymbol{x} \in \mathcal{X}$ and property $\pi \in \boldsymbol{e}$, it can be determined whether π applies to $\boldsymbol{x}$, represented as the relation $A(\boldsymbol{x}, \pi)$. This relation, independent of the actual or predicted label, must be both well-defined and verifiable by humans. The set of instances sharing the same property as $\boldsymbol{x}$'s explanation is defined by:

$$C_{\boldsymbol{x}} = \{\boldsymbol{x}' \in \mathcal{X} : A(\boldsymbol{x}', \boldsymbol{e}(\boldsymbol{x}))\}. \tag{4.11}$$

Consistency of explanations varies across instances, reflecting a spectrum of sufficiency rather than a binary value. Sufficiency is quantified using a probability distribution over $C_{\boldsymbol{x}}$, assessing the uniformity of predictions within this set:

$$\Psi_{\text{SUFF}} = \Pr_{\boldsymbol{x}' \in_{\mu} C_{\boldsymbol{x}}} \left(\boldsymbol{f}(\boldsymbol{x}') = \boldsymbol{f}(\boldsymbol{x}) \right), \tag{4.12}$$

with $x' \in_\mu C_x$ stating that x' is drawn from the distribution μ, restricted to the set C_x, meaning that they share the same property. The global sufficiency metric quantifies the expected value of local sufficiency. Local sufficiency evaluates the likelihood that the prediction label of a specific sample matches the labels of other samples to which the same explanation is applicable. For instance, if an image is explained as "contains cat's ears", the local sufficiency metric calculates the probability that another sample described as "contains cat's ears", receives an identical prediction label.

This metric is effective for images with unique and complex features, but may not be appropriate for RS images. For example, if the property π is defined to indicate the presence of trees in an image, this does not necessarily indicate a forest, as trees can be found in urban areas or along agricultural boundaries. A forest is characterised by the repeated occurrence of multiple trees. This complexity makes the metric difficult to evaluate and somewhat ambiguous, and it is therefore excluded from consideration.

4.2 Robustness Metrics

The second category to evaluate quantifies the **Robustness** of explanation methods. Similar, to adversarial attacks on NN [108], explanations can also be fooled [100, 89]. Thus, ensuring that the explanation is invariant to minimal perturbations in the input space is crucial for critical applications. For instance, Slack et al. fooled LIME and SHAP to evaluate a strongly biased racist classifier as non-biased [89]. Hedström [72] defines this invariance as $f(\boldsymbol{x}) \approx f\left(\mathcal{P}_{\mathbb{X}}(\boldsymbol{x})\right)$ for small perturbations of the input: $\left\|\mathcal{P}_{\mathbb{X}}(\boldsymbol{x}) - \boldsymbol{x}\right\|_p < \varepsilon$. In the following, robustness metrics such as the Maximum Sensitivity, Average Sensitivity, Local Lipschitz Estimate, Continuity, Consistency, Relative Input Stability, Relative Output Stability, and Relative Representation Stability are defined and analyzed.

Average Sensitivity (AS) and **Maximum Sensitivity (MS)** [44] evaluate the sensitivity of the explanation method by perturbing samples using a Monte-Carlo-based approximation.

$$\begin{aligned} \Psi_{\text{MS}} &= \max_{\boldsymbol{x}+\delta \in \mathcal{N}_\epsilon(\boldsymbol{x}) \leq \varepsilon} \left[\frac{\|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}+\delta)\|}{\|\boldsymbol{x}\|} \right], \\ \Psi_{\text{AS}} &= \text{avg}_{\boldsymbol{x}+\delta \in \mathcal{N}_\epsilon(\boldsymbol{x}) \leq \varepsilon} \left[\frac{\|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}+\delta)\|}{\|\boldsymbol{x}\|} \right], \end{aligned} \tag{4.13}$$

where ε defines the radius of a discrete, finite-sample neighbourhood around each input sample $\boldsymbol{x}$. This neighbourhood denoted as $\mathcal{N}_\epsilon(\boldsymbol{x})$, includes all samples in the set X that are within a distance of ε from $\boldsymbol{x}$. A lower score is indicative of more robustness.

One can assume that a classifier trained on MLC RS image data is more robust, than for classical CV data because there are no unique features. Thus, similar to the prediction, the explanation of a prediction should remain relatively consistent for minor perturbations (unless adversarial) of the input. Therefore, measuring the average sensitivity should not be as meaningful as measuring the maximum sensitivity. Thus, the MS metric is considered.

The **Local Lipschitz Estimate (LLE)** [103] works similar to the MS method. However, it estimates the Lipschitz constant of the explanation, which measures how much the explanation changes with respect to the input under slight perturbation. The LLE method is defined as

follows:

$$\Psi_{\text{LLE}} = \max_{x+\delta \in \mathcal{N}_\epsilon(x) \leq \epsilon} \left[\frac{|\Phi(x) - \Phi(x+\delta)|_2}{|x - (x+\delta)|_2} \right], \tag{4.14}$$

where lower values indicate less change concerning the change in input, which is desirable.

However, LLE does not have a sophisticated sampling strategy like MS or AS, thus it is not considered.

Continuity [105] , refers to the property of an explanation method being a continuous function. Montavon et al. [105] introduce this metric, assuming the prediction function $f(x)$ is also continuous. This metric ensures the following behaviour: if two data points are nearly equivalent, then their prediction explanations should also be nearly equivalent. Explanation continuity is quantified by identifying the strongest variation of the explanation $\Phi(x)$ in the input domain:

$$\Psi_{\text{CONT}} = \max_{x,\tilde{x}} \left[\frac{\|\Phi(x) - \Phi(\tilde{x})\|_1}{\|x - \tilde{x}\|_2} \right]. \tag{4.15}$$

This metric, being more axiomatic, defines a characteristic of the explanation function and is less dependent on the data; therefore, it is not suitable for the experiments.

Consistency [107] is defined as the property that samples receiving the same explanation should also have the same predicted label. This is formalized through the measure of homogeneity in predictions within the set of instances that share the same explanation.

$$\Psi_{\text{L}-\text{CONS}} = \Pr_{x' \in \mu_{C_\pi}} (f(x') = f(x)), \tag{4.16}$$

$$\Psi_{\text{G}-\text{CONS}} = \mathbb{E}x \in \mu_X[\Psi\text{L} - \text{CONS}(x)], \tag{4.17}$$

where $C_\pi = x' \in X : \Phi(x') = \pi$ denotes the set of instances that are assigned the explanation π by the explanation function Φ and μ represents the distribution over the instances. A higher consistency score indicates that the explanation method reliably assigns the same label to instances with the same explanations, reflecting the internal coherence of the explanation method.

A potential criticism of the consistency metric is that it assumes that the explanation function Φ perfectly captures the relevant features of the prediction function f. This assumption may be unrealistic in practical scenarios where explanations may not fully encapsulate the model's decision boundaries or complex interactions within the data. Furthermore, the metric's reliance on a homogeneous distribution of labels within the set C_π could lead to misleading evaluations. For example, if the explanation function Φ provides overly broad explanations, different samples within C_π could have the same explanation but different underlying reasons for their predictions, artificially inflating the consistency score without truly reflecting the robustness of the model. It is therefore not considered for further evaluation.

The **Relative Input Stability (RIS)** [86] assesses how the explanation changes relative to alterations in the input data. Formally, for an original input x and its perturbed version x',

with their respective explanations $\Phi(\boldsymbol{x})$ and $\Phi(\boldsymbol{x}')$, it is defined as:

$$\Psi_{\text{RIS}} = \max \left(\frac{\left\| \frac{\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')}{\Phi(\boldsymbol{x})} \right\|_p}{\max \left(\left\| \frac{x - x'}{x} \right\|_p \right)} \right), \tag{4.18}$$

where $\|\cdot\|_p$ denotes the L_p norm. For numerical reasons a small constant ϵ_{min} is added to the denominator to prevent division by zero. Higher RIS values indicate greater instability of the explanation.

The **Relative Representation Stability (RRS)** [86] measures the stability of explanations in relation to changes in the model's internal representations for an input $\boldsymbol{x}$ and its perturbed version. Defined as:

$$\Psi_{\text{RRS}} = \max \left(\frac{\left\| \frac{\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')}{\Phi(\boldsymbol{x})} \right\|_p}{\max \left(\left\| \frac{L_x - L_{x'}}{L_x} \right\|_p \right)} \right), \tag{4.19}$$

where L_x and $L_{x'}$ represent the internal representations of x and x' of a chosen layer, respectively. Like RIS, higher values of RRS suggest more unstable explanations.

The **Relative Output Stability (ROS)** [86] quantifies stability based on changes in the model's output predictions, defined for inputs $\boldsymbol{x}$ and $\boldsymbol{x}'$, and their explanations $\Phi(\boldsymbol{x})$ and $\Phi(\boldsymbol{x}')$, as:

$$\Psi_{\text{ROS}} = \max \left(\frac{\left\| \frac{\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')}{\Phi(\boldsymbol{x})} \right\|_p}{\max \left(\left\| f_{\text{LOGIT}}(\boldsymbol{x}) - f_{\text{LOGIT}}((\boldsymbol{x}') \right\|_p \right)} \right), \tag{4.20}$$

where $f_{\text{LOGIT}}(\boldsymbol{x})$ are the output logits for $\boldsymbol{x}$. Higher ROS values denote explanations that are less stable concerning output changes.

All relativity metrics (RIS, RRS, ROS) are approximated iteratively. As the metrics provide a broad overview of models' robustness, RIS and ROS were considered in the examination.

4.3 Localisation Metrics

Localisation metrics evaluate the match between predicted and actual object positions using ground truth reference maps or bounding boxes, focusing on spatial accuracy. However, these metrics often assume that predictions are determined solely by the object or its parts, a notion challenged by [109]. Label dependencies are not considered in this metric category, e.g. a patch of sand next to a patch of ocean is usually a beach and not a desert. However, Arias-Duart et al. [109] suggest that a reliable method will assign higher relevance to pixels that support the identified class. Thus, both the object and its context are essential in assessing the explanatory power and accuracy of a model. If the ground truth is known, this category is the most important. Thus, all Localisation metrics (Pointing-Game, Top-K Intersection, Relevance Mass Accuracy, Relevance Rank Accuracy, Attribution Localisation), except Focus were considered.

Pointing-Game (PG) [110] captures whether the feature of maximal attribution lies on the

ground truth mask, a binary mask indicating the true features contributing to the model's output. It is defined as follows:

$$\Psi_{\text{PG}} = \begin{cases} 1 & \text{if } \arg\max_i \Phi_i(\boldsymbol{x}, \boldsymbol{f}, \hat{y}; \lambda) \in \boldsymbol{s}^{\text{gt}} \\ 0 & \text{otherwise} \end{cases} \tag{4.21}$$

where $\Phi_i(\boldsymbol{x})$ represents the i^{th} input feature of highest attribution, and $\boldsymbol{s}^{gt} \in \mathbb{R}^D$ denotes the binary ground truth mask.

However, the authors note that the Pointing Game is trivial for images with large dominant objects [110]. In RS images where only one class is present, the highest attribution will inevitably lie within the target class, making the metric less informative. Therefore, this metric is particularly relevant for classes with multiple labels, where the distinction in attribution becomes more significant.

Attribution Localisation (AL) [111] measures the ratio of positive attributions within the targeted object relative to the total positive attributions across the image. To quantify this, the inside-total relevance ratio, Ψ_{AL}, and a weighted variant, $\Psi_{\text{AL},w}$, are introduced, taking into account the object size. The formulas are given as:

$$\Psi_{\text{AL}} = \frac{R_{\text{in}}}{R_{\text{tot}}} \tag{4.22}$$

$$\Psi_{\text{AL},w} = \Psi_{\text{AL}} \cdot \frac{S_{\text{tot}}}{S_{\text{in}}} \tag{4.23}$$

Here, Ψ_{AL} represents the inside-total relevance ratio without considering object size, and $\Psi_{\text{AL},w}$ is a weighted variant that accounts for the object size. R_{in} is the sum of positive relevance within the bounding box, R_{tot} is the total sum of positive relevance in the image, and S_{in} and S_{tot} represent the size of the bounding box and the image, respectively, in pixels. The subscript w indicates the addition of a normalisation factor in $\Psi_{\text{AL},w}$, considering the size of the image and object.

Top-K Intersection (TKI), as proposed by Theiner et al. [112], extends the PG method by computing the intersection between a ground truth mask and the binarized explanation at the top k feature locations. This metric measures the pixel-wise intersection of the k most important features from $e(x, \hat{c})$ to evaluate the difference between two explanation maps (top-k intersection) for assessing manipulated explanation maps. Specifically, it defines the influence of a concept s visible in image x concerning a prediction $\hat{c}$ of the model. The pixel-wise intersection tki between the binary reference map $m(x, s)$ and the binary mask of the top-k features $e_k(x, \hat{c})$ is given by:

$$\Psi_{\text{TKI}} = \frac{1}{k} \sum_{i=1}^{w} \sum_{j=1}^{h} m(x, s)_{i,j} \wedge e_k(x, \hat{c})_{i,j}, \tag{4.24}$$

where $m(x, s)$ and $e_k(x, \hat{c})$ are binary masks in $\{0, 1\}^{w \times h}$, and $\wedge$ denotes the pixel-wise boolean AND operation. If all top-k pixels are within the shape of the concept s, then $\Psi_{\text{TKI}} = 1$.

Relevance Mass Accuracy (RMA) and **Relevance Rank Accuracy (RRA)**, proposed by [113],

evaluate the distribution of relevance within the ground truth mask. These metrics assess whether the majority of relevance is correctly placed within the ground truth mask, considering both relevance mass and relevance ranking. This non-binary evaluation of the heatmap's alignment with the ground truth is suitable for contexts beyond weak object localisation. Both metrics yield values in the range $[0, 1]$, with higher values indicating a more accurate relevance saliency map.

The RMA is computed as the ratio of the sum of the relevance values within the ground truth mask to the sum of all relevance values in the entire image. It measures how much "mass" the explanation method assigns to pixels within the ground truth. It can be expressed as:

$$\Psi_{\text{RMA}} = \frac{\boldsymbol{e}_{\text{in}}}{\boldsymbol{e}_{\text{tot}}},$$
$$\text{with } \boldsymbol{e}_{\text{in}} = \sum_{\substack{k=1 \\ \text{s.t. } k \in S_{\text{tot}}}}^{|S_{\text{tot}}|} \boldsymbol{e}^k \text{ and } \boldsymbol{e}_{\text{tot}} = \sum_{k=1}^{N} \boldsymbol{e}^{k'} \tag{4.25}$$

where $\boldsymbol{e}^k$ is the relevance value at pixel $k \in \boldsymbol{x}$, S_{tot} is the set of pixel locations within the ground truth mask, $|S_{\text{tot}}|$ is the number of pixels in this mask, and N is the total number of pixels in the image.

The RRA measures the concentration of high-intensity relevance values within the ground truth. It is calculated as follows: Let K be the size of the ground truth mask. Select the K highest relevance values, count how many of these values lie within the ground truth pixel locations, and divide by the size of the ground truth. This can be expressed as:

$$P_{\text{top } K} = \left\{ p_1, p_2, \ldots, p_K \mid R_{p_1} > R_{p_2} > \ldots > R_{p_K} \right\}, \tag{4.26}$$

where $P_{\text{top } K}$ is the set of pixels with relevance values $R_{p_1}, R_{p_2}, \ldots, R_{p_K}$ sorted in decreasing order until the K-th pixel. The rank accuracy is then computed as:

$$\Psi_{\text{RRA}} = \frac{\left| P_{\text{top } K} \cap S_{\text{tot}} \right|}{|S_{\text{tot}}|}. \tag{4.27}$$

The **Focus** [109] score quantifies the coherence of the explanations generated by these methods, focusing on their ability to highlight relevant features in a class-specific context. Using mosaic images composed of instances from different classes, the Focus score assesses the concentration of explanation relevance on target class regions, thus introducing a visual pseudo-precision metric. Formally, the focus score for a given mosaic m and target class $c(m)$ is defined as

$$\Psi_{\text{FOCUS}} = \frac{\Phi_{c(m)}(\boldsymbol{x}_1) + \Phi_{c(m)}(\boldsymbol{x}_2)}{\Phi_{c(m)}(m)}, \tag{4.28}$$

where $\Phi_{c(m)}(\boldsymbol{x}_1) + \Phi_{c(m)}(\boldsymbol{x}_2)$ is the sum of positive relevance within the target class images in the mosaic, and $\Phi_{c(m)}(m)$ is the total positive relevance assigned across the entire mosaic. The higher the Focus score, the more accurately the method identifies and highlights the relevant features for the target class, facilitating an objective comparison of feature attribution methods

and their alignment with model behaviour.

The Focus score may not be suitable for MLC tasks because it assumes a single target class for evaluating the relevance distribution. In multi-label scenarios, where an instance may belong to multiple classes, this metric does not take into account the relevance of features across all classes.

4.4 Complexity Metrics

Explanations for machine learning models should be simple and understandable to ensure they are useful for human interpretation. Complexity metrics aim to quantify the simplicity and interpretability of these explanations by assessing factors such as sparseness and entropy. In the following, complexity metrics such as Sparseness, Complexity and Effective Complexity are defined and analyzed.

Sparseness (SP) [114] is a method for evaluating the sparsity of explanations and is defined as the Gini index of the explanation. It is calculated by summing the product of the ranks of the input features and their attributions and dividing by the sum of the attributions as follows:

$$\Psi_{\text{SP}} = \frac{\sum_{i=1}^{D}(2i - D - 1) \cdot \hat{\boldsymbol{e}}^i}{D(D-1)\sum_{i=1}^{D}\hat{\boldsymbol{e}}^i}, \tag{4.29}$$

where $\hat{\boldsymbol{e}}^i$ is the attribution of the i^{th} input feature, and D is the total number of input features. A higher Sparseness score indicates a more sparse and concise explanation.

Complexity (CO) [101] is defined using the Shannon entropy calculation, which measures the amount of uncertainty or randomness in the explanation map. It is calculated by summing the product of the probabilities of the attributions and the logarithm of the probabilities of the attributions:

$$\begin{aligned} \Psi_{\text{CO}} &= \mathbb{E}_i\left[-\ln\left(\mathbb{P}_\Phi\right)\right] = -\sum_{i=1}^{D}\mathbb{P}_\Phi(i)\ln\left(\mathbb{P}_\Phi(i)\right), \\ \text{with} \quad \mathbb{P}_\Phi(i) &= \frac{|\Phi_i(\boldsymbol{x})|}{\sum_{j\in[D]}\left|\Phi_j(\boldsymbol{x})\right|}; \mathbb{P}_\Phi = \left\{\mathbb{P}_\Phi(1), \ldots, \mathbb{P}_\Phi(D)\right\}, \end{aligned} \tag{4.30}$$

where $|\cdot|$ denotes the absolute value, and $\mathbb{P}_\Phi(i)$ denotes the fractional contribution of feature $\boldsymbol{x}_i$ to the total quantity of the attribution. A higher entropy indicates a higher level of uncertainty or randomness, i.e., a higher complexity. A uniformly distributed attribution would have the highest possible complexity score.

However, as RS images typically have uniformly distributed pixels and features, this metric might not be suitable for evaluating the explanations in such contexts.

Effective Complexity [115] measures how many attributions in absolute values exceed a certain threshold.

$$\Psi_{\text{ECO}} = \#|\boldsymbol{e}^k| < \epsilon \text{ for } k \in \{1, \ldots, N\}, \tag{4.31}$$

with $\#|\boldsymbol{e}^k| < \epsilon$ being the total number of pixels that are attributed a higher value than ϵ.

A low effective complexity indicates that many features can be ignored with minimal impact

on the prediction, reducing cognitive load while maintaining simplicity and broad applicability. However, the problems with Complexity metrics are discussed more thoroughly in Section 4.7. However, to include a complexity metric in the evaluation this one is included.

4.5 Randomisation Metrics

Randomisation metrics are designed to evaluate the robustness and reliability of explanation methods by introducing randomness into the model parameters or the prediction targets. These metrics help to ensure that the generated explanations are not merely artefacts of specific parameter settings or model configurations but rather reflect meaningful and consistent insights about the model's behaviour. In this section, two randomisation metrics are discussed: the Model Parameter Randomization Test (MPRT) and the Random Logit (RL) method Both assess the stability and validity of explanations under randomized conditions.

Model Parameter Randomization Test (MPRT) [81] measures the correlation between an explanation from a randomly parameterised model $f\left(\boldsymbol{x}; \mathcal{P}_{\mathbb{F}}(\theta; v)\right) = \hat{f}$ and the original model f for each separate layer v of the network. To generate one quality estimate per sample, then the average of the correlation coefficients over all the layers in the network is calculated, denoted as V. Formally:

$$\Psi_{\text{MPR}} = \frac{1}{V} \sum_{v=1}^{V} \text{corr}\left(\Phi^{v}(\boldsymbol{x}, \boldsymbol{f}), \Phi^{v}(\boldsymbol{x}, \hat{\boldsymbol{f}})\right). \tag{4.32}$$

The limitations of model randomisation-based sanity checks for evaluating explanations are outlined by Binder et al. [116]. They show that such checks often yield high scores with uninformative attribution maps and fail to significantly alter explanations after randomisation due to preserved activation scales, thus revealing their inadequacy as a criterion for ranking attribution methods.

Random Logit (RL) method proposed by [83] is defined using the Structural Similarity Index Measure (SSIM) over the explanation of the ground truth label and an explanation of non-target class y'.

$$\Psi_{\text{RL}} = \text{SSIM}\left(\Phi(\boldsymbol{x}, \boldsymbol{f}, \hat{y}; \lambda), \Phi\left(\boldsymbol{x}, \boldsymbol{f}, y'; \lambda\right)\right), \tag{4.33}$$

where $\Phi(\boldsymbol{x}, f, \hat{y}; \lambda)$ is the explanation generated for the prediction $\hat{y}$ and $\Phi\left(\boldsymbol{x}, f, y'; \lambda\right)$ is the explanation generated for a non-target class y'. Lower values indicate that the explanations are not correlated which is desirable. To provide a robust measure of the model's behaviour, this metric was incorporated into the experiments.

4.6 Axiomatic Metrics

Axiomatic metrics evaluate explanation methods based on fundamental principles or axioms. These metrics ensure explanations adhere to theoretical properties, enhancing their robustness and reliability. Notable axiomatic metrics include Completeness, Non-Sensitivity, and Input Invariance. However, a deeper analysis or use of these metrics in experiments is beyond the scope of this thesis, as the data domain should not influence the axiomatic properties of explanation methods.

Completeness, as defined by [74], evaluates whether the sum of attributions equals the dif-

ference between the function values at the input x and baseline x'. This property is also known as Summation to Delta [39] or SENS-N [94].

The **Non-Sensitivity** metric [115] assesses whether the total attribution is proportional to the explainable evidence at the model output. It ensures that a method assigns zero importance only to features on which the model is not functionally dependent. This aligns with the Sensitivity(b) axiom proposed by [74]. The complexity of the explanation can be gauged by the number of non-zero attributions.

Input Invariance [117] ensures that attributions remain unchanged when the input is shifted, provided that the model's prediction does not change in response to the shift.

4.7 Desiderata for Explanation Metrics in Remote Sensing Multi-Label Image Classification

Just as the effectiveness of explanation methods is strongly related to the characteristics of data, the same holds for their evaluation metrics. In the following, each metric category is analysed theoretically.

Complexity: The inherent texture of RS image data introduces challenges for Complexity metrics. Because this texture serves as a discriminative feature, a larger number of input features tends to influence the prediction, typically leading to more complex explanations. Here, 'complex explanation' refers to assigning significant values to a large number of input features. The complexity metrics assess the conciseness of explanations, specifically how few features are needed to visualise a model's prediction. For example, Effective Complexity (ECO) [115] quantifies the number of attributions that exceed a predetermined threshold. Consequently, explanations for multi-label images in RS are expected to perform poorly on these metrics due to the greater presence of relevant pixels due to the texture and possibly multiple Object of Interest (OoI) associated with the same class. The importance of performance within these metrics could arguably be considered less significant.

Localisation: Localisation metrics are valuable for RS because they test the alignment between the OoI and a prediction's explanation. However, in MLC, where each instance may have multiple labels, label dependencies can arise. A label dependency means that the presence (or absence) of one label can significantly influence the likelihood of another class being present in the image. This relationship is particularly common in scenarios such as land cover classification, where certain features or objects (e.g. ships) typically appear in certain environments (e.g. oceans). These dependencies are overlooked by Localisation metrics. This could be solved by designing a Localisation metric that considers these label dependencies. Furthermore, the reliability of Localisation metrics depends heavily on accurate and up-to-date ground truth data. Obtaining accurate ground truth annotations is costly and can be error-prone, especially in remote or inaccessible areas where verification is difficult. Inaccuracies or obsolescence in ground truth data can significantly affect the reliability of the metrics, potentially leading to incorrect assessments of model focus.

Randomisation and Axiomatic: The Randomisation and Axiomatic categories seem to be unaffected by the data domain. Randomisation tests typically modify parts of the model and assess the impact on prediction outcomes, focusing solely on the model rather than the data

type. Axiomatic metrics evaluate whether explanation methods adhere to certain axiomatic properties, which depend on the method and not on the data.

Faithfulness: The category with the biggest challenges for MLC RS images is Faithfulness. As already stated, RS images tend to have a lower amount of discriminating shapes than natural images. Especially in low-resolution data, the most discriminative feature often is the texture of the ground, which is local and repetitive.

This unique characteristic can pose problems with many of the common Faithfulness metrics. Often, metrics iteratively swap pixels or regions in order of importance and evaluate the change in confidence in the prediction, e.g. SEL [105], RP [45] use this strategy. They measure how quickly the prediction confidence drops when the most important features are removed. These metrics may underestimate the explanation methods because the texture is more evenly distributed in the image. To destroy the discriminative power of the texture, more features would have to be removed, leading to a seemingly worse performance in these metrics. If the predicted class occupies a large part of the image, this large amount needs to be reduced until there are no continuous patches of the class left.

To evaluate xAI methods for their Faithfulness, this has to be taken into consideration. One possible inference from this could be that metrics add iteratively add important features to a reference baseline e.g. the Monotonicity Metric by [104], might be working better than metrics that remove the most important features, e.g. the Faithfulness Correlation [101]. However, Monotonicity only looks if the prediction curve is monotonically increasing, and does not consider a sudden sharp increase in the function curve. Conceptually, a metric dealing with this could be designed similarly to IROF [106], calculating the means of the attributed relevance for each superpixel, but instead of iteratively removing the segments sorted by relevance, segments could be added to a baseline. To adapt current Faithfulness metrics to this problem, the curve drop could be seen in relation to the class coverage in the ground truth map, or if there is no ground truth the incline of the prediction function could be considered in a sliding window. However, this would be out of the scope of this thesis and could be evaluated in the future.

Faithfulness under Removal Orders: As already stated, almost all Faithfulness metrics iteratively remove input features based on their attributed relevance (e.g. FC, FE, PF, RP, SEL, IROF, SENS-N, see Section 4.1). Each of these metrics needs a strategy defining the order of removal. Following established literature [118, 45, 102], two predominant strategies, Least Relevant First (LeRF) and Most Relevant First (MoRF), are employed. The MoRF approach prioritises feature removal based on descending relevance as determined by the explanation method. This process is performed iteratively over different k values, removing the k most relevant features from each sample. The faithfulness of the method is then quantified by the change of the prediction function, e.g. the Area Under the Curve (AUC) of the prediction curve, with a lower score indicating higher faithfulness due to a rapid decrease in prediction confidence following the removal of critical features. This principle is utilized for instance by the SEL metric [2].

The removal strategy is illustrated in Figure 4.1, which shows an image from the DeepGlobe dataset. This example includes the classes urban land and agricultural land, with the urban land segment occupying a smaller area of the image. The first column shows the original sample and its reference map. Each other column shows the image with the top-k-percent of the image masked with a black baseline, for $k \in [0.05, 0.1, 0.2, 0.5, 0.9]$. The top row of the Figure shows the urban land class, the last column shows the GradCAM attributions for both classes.

The subplot titles show the prediction logit for the corresponding image and class. The prediction curve is also visualised in Figure 4.3. As mentioned previously, when the explanation has a high faithfulness the prediction curve is supposed to drop fast when more and more relevant pixels are removed. For the urban land class, the strategy works as expected. The explanation is a coherent patch and is located in the class area. For an increasing masked percentage of the image, the model's prediction logit drops fast. For instance, when $k = 0.2$, meaning 20% of the image is covered, most of the urban land class is not visible anymore. This leads to the prediction logit dropping to $p = 0.05$. However, it can be seen that for the agricultural class, this is not the case. Even for 50% of the most relevant features masked, the prediction logit remains at $p = 0.64$. This is because, as the class occupies a large area of the image, it is still not occluded completely.

In contrast, this approach is more effective for a traditional CV image. Here, even if a class covers a large area, its distinguishing features (e.g. a cat's ears) are typically unique and complex. These features are considered the most relevant and are masked first. Removing them quickly reduces prediction confidence, as the model relies heavily on such features for its predictions. This contrasts with RS images, where the repetitive nature means that even after removing significant portions of a class, the remaining pixels can still maintain some model confidence. Due to the nature of the MLC task, only a minimal number of remaining pixels are required to maintain confidence, requiring the removal of a large number of pixels before a noticeable drop in prediction confidence occurs in the MoRF strategy.

The LeRF strategy, as depicted in Figure 4.2, operates oppositely to the MoRF strategy. This approach involves the iterative removal of the least important features, as identified by the explanation method. In contrast to the MoRF strategy, a higher AUC score is indicative of better performance for the LeRF strategy, reflecting the model's stability when non-critical features are removed. The structure of this plot is similar to the one presented in Figure 4.1. Here, it can be seen that even for the urban land class, which occupies only a small part of the input the masking works. Even for $k = 0.9$, the class-relevant features are still contained in the image and the prediction logit remains stable. The prediction curve is also visualised in Figure 4.4.

Figures 4.3 and 4.4 show the prediction curves for the respective strategies applied to the same sample. Here, the MoRF strategy shows a strong discrepancy between the two classes. Specifically, the AUC for urban land is 0.06, illustrating a steep decrease in the prediction certainty as relevant features are removed. Conversely, the AUC for the agricultural class is 0.58, indicating less impact from feature removal. This discrepancy between classes is a huge flaw in the removal strategy and thus in many faithfulness metrics. For both classes the predictor remains faithful, however, due to the difference in class coverage the metric assigns the predictor a low faithfulness.

In contrast to the MoRF strategy, the LeRF strategy shows a different behaviour, as shown in Figure 4.4. The AUC remains consistent at 0.79 for both classes, indicating robustness to the area each class occupies within the image. This consistency is supported by the visualisation in Figure 4.2, which shows that significant class-relevant information persists over a range of k values. The LeRF strategy, which removes the least relevant pixels first and considers a higher AUC score as an indication of better performance, is also effective for MLC images. RS images. In this context, the prediction confidence remains relatively stable until almost all the pixels representing a class are removed, regardless of the class size. In general, this aligns with the findings from Rong et al. [91]: Explanation metrics are highly sensitive to their parametrisation,

Figure 4.1: GradCAM: Visualisation of the MoRF strategy, using a sample from the DeepGlobe dataset containing urban and agricultural land, with urban areas occupying a smaller portion. Illustrates prediction curves for varying k values in the set $[0.05, 0.1, 0.2, 0.5, 0.9]$. The p values for Urban Land are: $p = 0.77$, $p = 0.34$, $p = 0.05$, $p = 0.01$, $p = 0.03$ and for p for Agricultural Land: $p = 0.98$, $p = 0.98$, $p = 0.96$, $p = 0.64$, $p = 0.31$. The first column shows the original image and its reference map. The top row presents the urban land class, and the bottom row the agricultural land class, each image is masked. The last column shows the attribution map. The subtitle of each plot shows the k value and the prediction logit for the corresponding class.

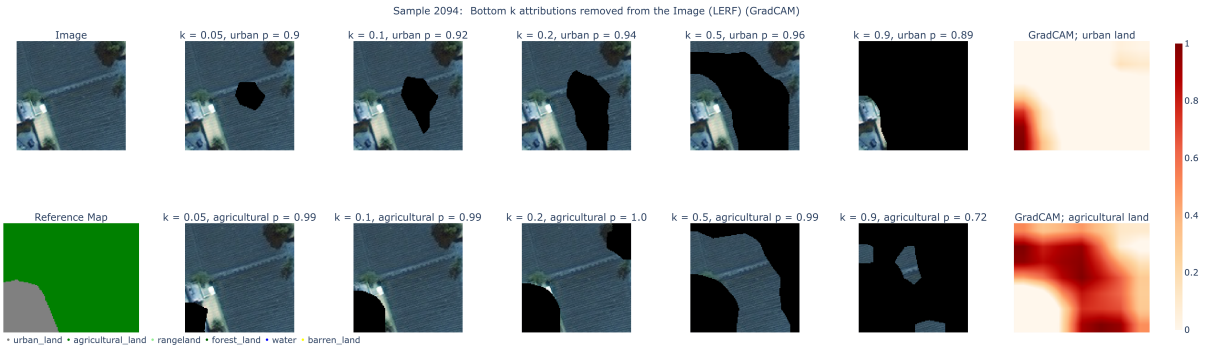


Figure 4.2: GradCAM: Visualisation of the LeRF strategy, using a sample from the DeepGlobe dataset containing urban and agricultural land, with urban areas occupying a smaller portion. Illustrates prediction curves for varying k values in the set $[0.05, 0.1, 0.2, 0.5, 0.9]$. The p values for Urban Land are: $p = 0.9$, $p = 0.92$, $p = 0.94$, $p = 0.96$, $p = 0.89$ and for Agricultural Land: $p = 0.99$, $p = 0.99$, $p = 1$, $p = 0.99$, $p = 0.72$. The first column shows the original image and its reference map. The top row presents the urban land class, and the bottom row the agricultural land class, each image is masked. The last column shows the attribution map. The subtitle of each plot shows the k value and the prediction logit for the corresponding class.

especially removal order and perturbation strategy. Different orders of pixel removal often give opposite results. For example, local attribution methods may perform well under one sequence while performing poorly under another, as noted by Hooker et al. [90]. This poses a significant challenge to the objective comparison of different attribution methods [91].

Figure 4.5 presents the application of the MoRF removal strategy using the DeepLIFT method on the same sample, with k values ranging over $[0.05, 0.1, 0.2, 0.5, 0.9, 1]$. The plot is structured the same as the previous ones. For urban land, the prediction probability p starts at 0.71 and progressively decreases as k increases, finally reaching $p = 0.03$ when all original features are eliminated. In contrast, the agricultural class maintains high prediction probabil-

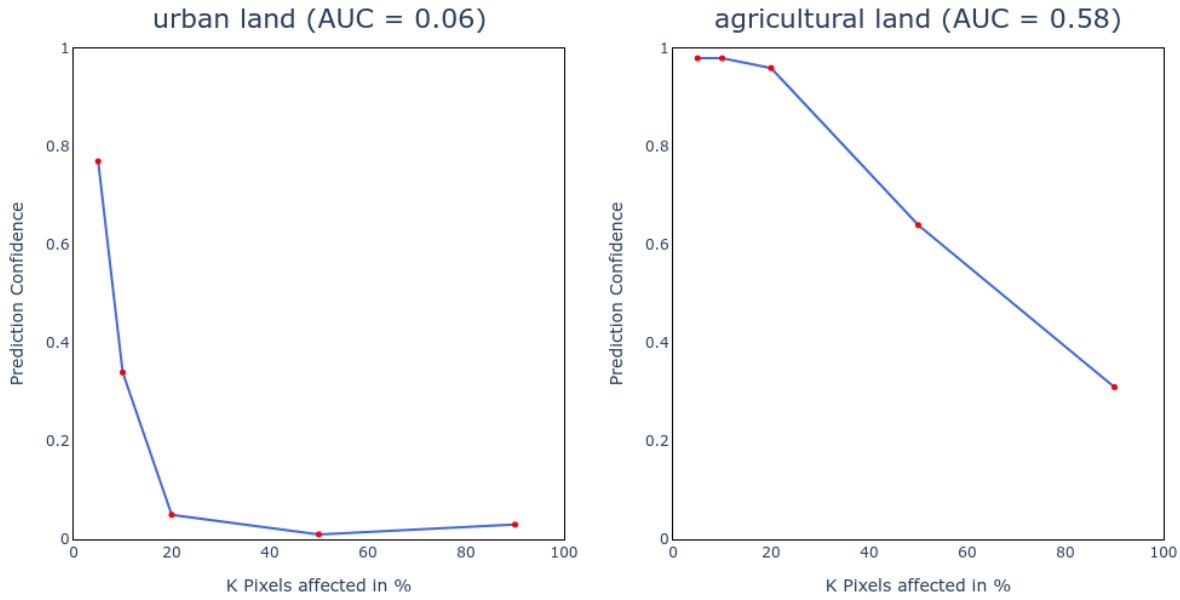


Figure 4.3: GradCAM: Prediction curve for the MoRF strategy applied to a DeepGlobe dataset sample, illustrating the variance in model prediction between urban and agricultural land classes, for images masked with different k values.

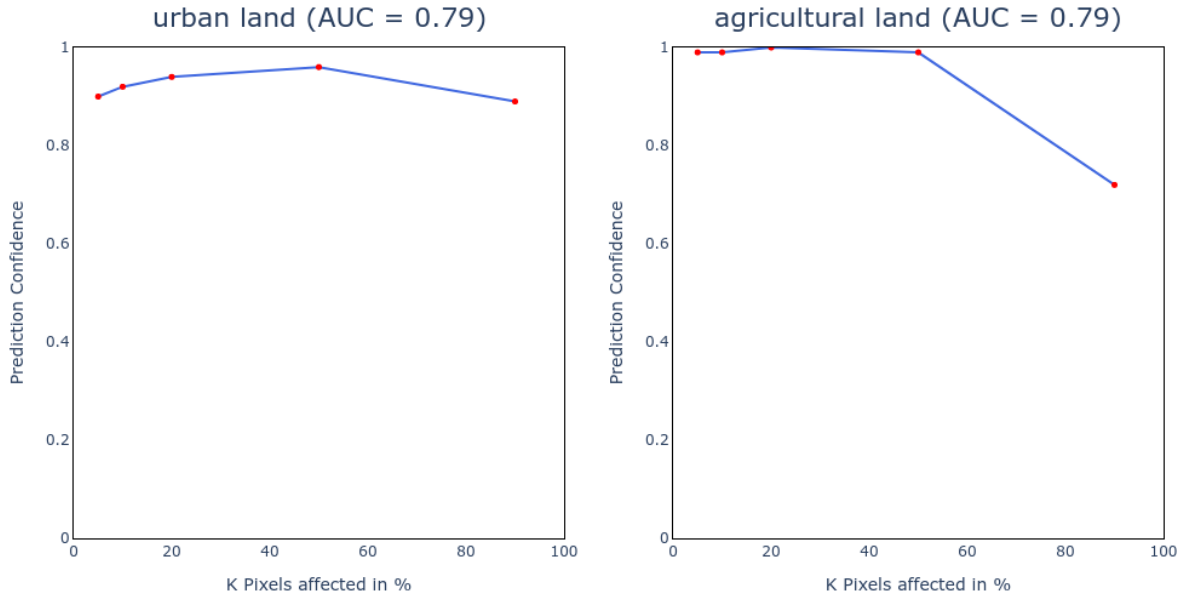


Figure 4.4: GradCAM: Prediction curve for the LeRF strategy applied to the same DeepGlobe dataset sample, illustrating the variance in model prediction between urban and agricultural land classes, for images masked with different k values.

ities, almost 0.99, up to $k = 0.9$. The main difference here is that the explanation provided by DeepLIFT compared to the GradCAM explanation (visualised in Figure 4.1) is not a coherent patch. DeepLIFT attributes most relevance to the edges in the image. As this leads to disjoint

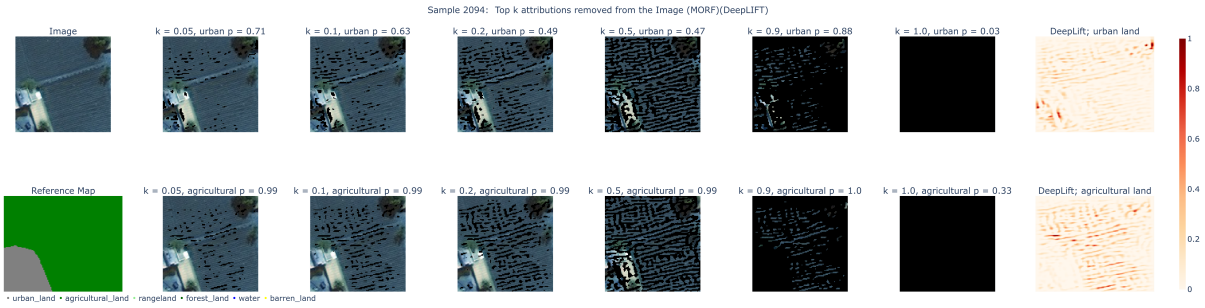


Figure 4.5: DeepLIFT: Visualisation of the MoRF strategy using a sample from the DeepGlobe dataset containing urban and agricultural land, with urban areas occupying a smaller portion. Illustrates prediction curve for varying k values in the set $[0.05, 0.1, 0.2, 0.5, 0.9, 1]$. The p values for Urban Land are: $p = 0.71$, $p = 0.63$, $p = 0.49$, $p = 0.47$, $p = 0.88$, $p = 0.03$ and for p for Agricultural Land: $p = 0.99$, $p = 0.99$, $p = 0.99$, $p = 0.99$, $p = 1$, $p = 0.33$. The first column shows the original image and its reference map. The top row presents the urban land class, and the bottom row the agricultural land class, each image is masked. The last column shows the attribution map. The subtitle of each plot shows the k value and the prediction logit for the corresponding class.

pixels being removed from the image, there are still enough relevant features contained in the image to predict with high confidence, e.g. the urban land prediction for $k = 0.5$ remains at a certainty of $p = 0.47$. Additionally, for $k = 0.9$ it even increases to $p = 0.88$, meaning that through the baseline perturbation, a discriminative artefact was introduced in the image. This also occurs for the agricultural class.

At $k = 1$, where all features are removed, the prediction remains at 0.33. This means that even from a completely black image, the model predicts the presence of agricultural land with a confidence level of 33%. This aligns with the findings of [88, 87], who showed that an MNIST classifier produces a prediction with 91% confidence on a Gaussian noise input and emphasizes the importance of a solid perturbation strategy and a carefully selected baseline.

Perturbation Strategy: The necessity of sophisticated perturbation strategies for explanatory methods has already been discussed in 3.7. However, the same arguments apply to explanation metrics. Usually, both Faithfulness and Robustness metrics, use perturbations to evaluate explanation methods. As shown in Figure 4.5, the classifier predicts the class agriculture in a black image with a confidence of 0.33. This can prevent metrics that use perturbations from being reliable. This issue is particularly prevalent with Faithfulness metrics. Robustness metrics, which also employ perturbations, usually insert noise to generate small perturbations. This is one of the reasons why Hooker et al. introduced ROAD [90]. This metric requires a retraining step to adapt to the distribution shift introduced by the perturbations, which is highly inefficient. For smaller perturbations, as typically used in Robustness metrics, there is an assumption that models trained on RS data will be more resilient to these perturbations. For CV datasets, perturbation of a unique descriptive feature can significantly alter the model's prediction. In contrast, for datasets that are biased towards classes that contain only descriptive textural features like in RS, such perturbations have less impact because the texture should be consistent and noise, such as clouds obscuring the view, is less problematic.

Efficiency: Lastly, when choosing appropriate evaluation metrics, the Efficiency of these metrics must be considered. The evaluation time of a metric often correlates with the efficiency

of the utilized explanation method and the inference time of the model. Additionally, the input size can increase the time needed for evaluation. Metrics that perturb the input often use a sliding window over the input image. While the size of this window is adjustable, higher granularity requires smaller windows and strides, which can be problematic due to the large size of RS images in terms of both spectral resolution and overall image size.

5 Guiding the Training with Explanations

Explainable Artificial Intelligence methods provide various insights into the reasoning of a model. However, most research in this area stops there: Decisions are explained and problems may be discovered, but the insights gained are rarely applied to achieve more trustworthy, fairer, or simply better-performing models [27].

Weber et al. [27] denotes the usage of explanations to improve models xAI-based augmentation and states that through it several properties of AI models can be improved improve: performance, convergence, robustness, efficiency, reasoning, and equality. They also identify five categories of xAI-based augmentations: 1) *Data Augmentation* modifies the input distribution using input layer explanations, which affects all components of the training loop. 2) *Feature Augmentation* uses intermediate explanations at specific layers to mask or transform corresponding intermediate features, affecting higher feature representations and subsequent training components. 3) *Loss augmentation* uses the power measures of the loss function to guide training, indirectly affecting all components of the backward pass. 4) *Gradient augmentation* has two sub-types: feature gradient augmentation, which masks or transforms feature gradients at a given layer, affecting parameter updates and feature gradients of lower layers, and parameter gradient augmentation, which alters parameter gradients at a given layer, affecting only parameter updates of that layer, requiring parameter-wise explanations. 5) *Model Augmentation* after training uses intermediate layer explanations to estimate the importance of neurons, filters or parameters, which can guide pruning or quantisation of the model [27].

In this thesis, the terms *xAI-guided training* and *xAI-based augmentation* are used interchangeably. The following discussion elaborates on two specific strategies used within xAI-guided training:

Loss augmentation modifies the behaviour of the loss function using explanations as feedback. Based on the unaugmented loss function $\mathcal{L}_{PRED}(\boldsymbol{f}(\boldsymbol{x}), \mathbf{y})$ and the explanation method Φ, the augmented loss function is

$$\mathcal{L}_{\text{AUG}}(\boldsymbol{f}(\boldsymbol{x}), \mathbf{y}) = \mathcal{L}_{\text{XAI}}(\boldsymbol{f}(\boldsymbol{x}), \mathbf{y}, \Phi(\boldsymbol{x})) \cdot (\boldsymbol{f}(\boldsymbol{x}), \mathbf{y}), \tag{5.1}$$

where $\mathcal{L}_{\text{XAI}}$ is added as a regularisation term and $\cdot$ can be any scaling operator.

Data augmentation uses explanations to change the structure of the data via the general function $\Theta(X, R^{l,t})$, which takes the original data X and the attributes $R^{l,t}$ as inputs to produce augmented data [27]:

$$(X')^{1,t} = \Theta(X, R^{l,t}). \tag{5.2}$$

For each of these categories, a method is selected and used in this thesis to evaluate the

effectiveness of different explanation methods in combination with the respective xAI-guided training method (see Objective B 1). For loss augmentation, the Right for the Right Reason method was selected, while CutMix with xAI LP was used as the data augmentation technique. Both methods are described in detail in the following sections.

5.1 Right for the Right Reasons

A popular approach for loss augmentation is the Right for the Right Reason (RRR) loss [30], a loss function that attempts to improve the reasoning of the model. Reasoning in this context means that the model's predictions are based on 'relevant' features rather than spurious correlations. Ross et al. argue that classical loss functions do not guarantee trustworthiness and that improved reasoning also increases the generalisation capabilities of the model [30].

To improve inference, they use an additional binary relevance mask m_i for each sample $x_i \in \boldsymbol{X}_{\text{tr}}$ and each input feature $\delta \in \{1, \ldots, D\}$, this mask defines whether it is relevant for the prediction of class c. In particular, if δ is relevant, $m_i[\delta] = 0$ and if it is irrelevant, $m_i[\delta] = 1$. To incorporate this knowledge of feature importance into the loss, they compare it with the results of the explanation. It is defined as

$$\mathcal{L}_{\text{REASON}}(\boldsymbol{e}_i, m_i) = \|m_i \odot \boldsymbol{e}_i\|_2^2, \tag{5.3}$$

where $\|\cdot\|_2^2$ is the ℓ_2 norm, $\odot$ is the Hadamard product, and $\boldsymbol{e}_i$ is the explanation for the sample $\boldsymbol{x}_i$. Now, the loss function can be extended by adding the additional loss regularisation term to align the prediction with the relevance mask:

$$\mathcal{L}_{\text{RRR}}(\boldsymbol{f}(\boldsymbol{x}_i), y_i) = \mathcal{L}_{\text{PRED}}(\boldsymbol{f}(\boldsymbol{x}_i), y_i) + \lambda \mathcal{L}_{\text{REASON}}(\boldsymbol{e}_i, m_i), \tag{5.4}$$

with λ as regularisation parameter.

One strategy to generate the relevance map is to use a reference map and assign positive relevance to all features that lie in the class area. However, this approach has two major disadvantages. First, the reference map must be known. For RS images, this may involve a costly manual labelling process. Second, class dependencies must be known or are not considered in this approach. This may not be critical for RS images, but for other data domains, AI models are valued for their ability to find patterns in data that humans might miss.

There is also a technical problem. Since the RRR loss only penalises incorrect explanations, it could lead the model to learn that it should only produce explanations that are close to 0 in terms of the size of the value assigned. Therefore, an extension of the RRR loss is introduced: Right for Right Reasons with Mean Squared Error (RRR MSE), denoted as $\mathcal{L}_{\text{RRR MSE}}$. The relevance map is reversed, i.e. if δ is relevant for the prediction, $m_i'[\delta] = 1$ holds. Then the MSE loss of the explanation and the relevance map is calculated to reward positive attributions and penalise negative ones. Mathematically, $\mathcal{L}_{\text{RRR MSE}}$ is defined as follows:

$$\mathcal{L}_{\text{RRR MSE}} = \frac{1}{n} \sum_{i=1}^{n} (\text{Clip}(\boldsymbol{e}_i) - m_i')^2, \tag{5.5}$$

where Clip is the clipping function to avoid penalising explanations that are too high:

$$\text{Clip}(x) = \begin{cases} x & \text{if } x \leq 1 \\ 1 & \text{else} \end{cases}. \tag{5.6}$$

5.2 CutMix with Label Propagation for Multi-Label Classification

One popular approach for classical data augmentation in image classification tasks is CutMix [66]. CutMix involves mixing parts of two existing images to generate an augmented image. The primary motivation behind CutMix is to improve model robustness and performance. This is achieved by filling informative parts from different images into uninformative cutout areas and updating the label accordingly. However, the direct application of CutMix in MLC can result in the erasure or addition of class labels in the augmented image, which can lead to the introduction of multi-label noise, as discussed by [65]. They propose a Label Propagation (LP) function to propagate pixel-wise label information into the label label of the augmented image.

The following notation is adapted from Burgert et al [65], except the notation for Φ as readout function, which denotes an explanation method in this thesis. For the LP readout function ϕ is used.

Let $\mathcal{D} := \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_N, \mathbf{y}_N)\}$ be a multi-label training set, where $N \in \mathbb{N}$ is the number of training images. Each pair of elements in the set $\mathcal{D}$ is constituted by an image, represented by a vector of real numbers of length C, and a vector of real numbers of length H and W, where C, H, and W are defined as follows: The number of image bands with a height of H and a width of W, and its multi-label vector $\mathbf{y}_i \in \{0,1\}^L$, where $L \in \mathbb{N}$ defines the number of classes. Each element of the vector $\mathbf{y}_i$ indicates the presence (1) or absence (0) of class c.

In the context of CutMix, a box $R = (r_a, r_b, r_c, r_d)$ is defined by its corners (r_a, r_b) and (r_c, r_d), and its area is $A_R = (r_c - r_a)(r_d - r_b)$. The binary mask $B_R = \left(b_{j_1 j_2}\right)_{j_1 j_2} \in \mathbb{R}^{H \times W}$ for R is defined as follows:

$$b_{j_1 j_2} = \begin{cases} 1, \text{if } r_a \leq j_1 \leq r_c \text{ and } r_b \leq j_2 \leq r_d \\ 0, \text{otherwise} \end{cases} \tag{5.7}$$

To generate an augmented image, $\tilde{\mathbf{x}}$, with label $\tilde{\mathbf{y}}$, we combine two images, $\mathbf{x}_1$ and $\mathbf{x}_2$, and their labels, $\mathbf{y}_1$ and $\mathbf{y}_2$. This results in the generation of two boxes, R_1 and R_2, and their masks, B_1 and B_2. The inverse mask $(1 - B_1)$ is combined with x_1 as $(1 - B_1) \odot x_1$, and R_2 is extracted from x_2 as $B_2 \odot x_2$. The image shift operator $T_{xR_2R_1}$ aligns the boxes.

$$\begin{aligned} B_1 \odot \mathbf{x}_1 &= \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{x}_{R_1} \end{pmatrix} \\ B_2 \odot \mathbf{x}_2 &= \begin{pmatrix} \mathbf{x}_{R_2} & 0 \\ 0 & 0 \end{pmatrix} \\ T^x_{R_2R_1}(B_2 \odot \mathbf{x}_2) &= \begin{pmatrix} 0 & 0 \\ 0 & \mathbf{x}_{R_2} \end{pmatrix} \end{aligned} \tag{5.8}$$

The augmented pair $(\tilde{\mathbf{x}}, \tilde{\mathbf{y}})$ is:

$$\begin{aligned} \tilde{\mathbf{x}} &= (1 - B_1) \odot \mathbf{x}_1 + T^x_{R_2 R_1} (B_2 \odot \mathbf{x}_2) \\ \tilde{\mathbf{y}} &= (1 - A_R)\, \mathbf{y}_1 + A_R \mathbf{y}_2 \end{aligned} \tag{5.9}$$

Similar to RRR, reference maps are required to propagate label information at the pixel level. For each pair of coordinates $(\mathbf{x}_i, \mathbf{y}_i)$ in the dataset $\mathcal{D}$, let $\mathbf{m}_i \in [0, L]^{H \times W}$ be a pixel-level reference map. The reference map shift operator $T^m_{R_2 R_1}$ and read-out function ϕ extract class positional information:

$$\begin{aligned} \tilde{\mathbf{x}} &= (1 - B_1) \odot \mathbf{x}_1 + T^x_{R_2 R_1} (B_2 \odot \mathbf{x}_2) \\ \tilde{\mathbf{m}} &= (1 - B_1) \odot \mathbf{m}_1 + T^m_{R_2 R_1} (B_2 \odot \mathbf{m}_2) \\ \tilde{\mathbf{y}} &= \phi(\tilde{\mathbf{m}}) \end{aligned} \tag{5.10}$$

In the absence of reference maps, the authors propose the use of explanation masks, making it a data augmentation strategy for xAI [65]. Let $\boldsymbol{e}_i \in \{0, 1\}^{L \times H \times W}$ be the masks. The class explanation mask shift operator $T^e_{R_2 R_1}$ and read-out function ψ generate the updated $\tilde{\mathbf{y}}$.

$$\begin{aligned} \tilde{\boldsymbol{e}} &= (1 - B_1) \odot \boldsymbol{e}_1 + T^e_{R_2 R_1} (B_2 \odot \boldsymbol{e}_2) \\ \tilde{\mathbf{y}} &= \psi(\tilde{\boldsymbol{e}}) \end{aligned} \tag{5.11}$$

A class c is considered present in $\tilde{\mathbf{y}}$ if the number of activating pixels exceeds a threshold t_{map}. If a class is not included in the multi-label of the image, the explanation mask is set to zero. The aforementioned masks can be converted to binary masks using a threshold t_{cam}.

The authors demonstrate that for ResNET [119] trained for 120 epochs with a box size range of $0.3 - 0.7$, CutMix with LP using reliable reference maps achieves the most optimal overall performance, with an 84.65% map macro, representing a 5.65% increase over the baseline without augmentation [65].

6 Datasets and Experimental Setup

6.1 Datasets

For the evaluation, three datasets were used: one single-label CV dataset: Caltech101, and two multi-label RS datasets: DeepGlobe and BigEarthNet-S2 (BEN). The key characteristics of the datasets are summarised in Table 6.1.

Table 6.1: Summary of key attributes of various scene classification datasets, including the number of images ($||D|$), unique labels (L), average labels per image, number of channels (C), image dimensions and spatial resolution, and the type of pixel-level reference maps, taken and adapted from [65].

Datasets	$\|D\|$	L	**Avg.** L **per Image**	C	**Image Size (Spatial Resolution)**	**Pixel-Level Reference Maps**
Caltech101	9,144	101	1	3	224×224 (-)	Manually Annotated
DeepGlobe	18,185	6	1.71	3	120×120 (0.5 m)	Manually Annotated
BEN	250,249	19	2.95	10	120×120 (10 m), 60×60 (20 m)	Thematic Product
BEN Lithuania	51,624	19	3.94	10	120×120 (10 m), 60×60 (20 m)	Thematic Product

The **Caltech101** [120, 121] dataset is a collection of digital images designed to facilitate computer vision research, particularly in the areas of object detection and image classification. It contains 9,144 images with 101 distinct categories of objects. The categories are diverse, including a variety of animals, vehicles, household objects, and other common items, each represented by around 50 but up to 800 images. Figure 6.1 shows four samples from the dataset.

The **DeepGlobe Land Cover Classification Challenge (DeepGlobe-LCCC)** dataset [122] includes 1,949 RGB tiles of size 2448×2448 pixels with a spatial resolution of 0.5 m, collected over Thailand, Indonesia, and India. Each tile is associated with a manually annotated ground reference map. The classes in this dataset are urban, agriculture, rangeland, forest, water, barren, and unknown. Figure 6.2 depicts four samples from the dataset.

In this thesis, the **DeepGlobe (DeepGlobe)** dataset is used, which Burgert et al. [65] derived from the original DeepGlobe-LCCC dataset to make it suitable for MLC. All tiles are divided into a grid of 120×120-pixel patches. Multi-labels are derived from the label information of the associated 120×120-pixel reference maps, excluding patches containing the class "unknown."

Figure 6.1: Four Examples from the Caltech101 dataset.

Only 20% of the patches with a single present class and all patches with more than one present class are included, resulting in an average of 1.71 present classes per patch. DeepGlobe includes 30,443 patches, split into a training set (60%), a validation set (20%), and a test set (20%).

Figure 6.2: Four Examples from the original DeepGlobe dataset

The **BigEarthNet-S2 (BEN)** dataset [123] is a multi-label dataset based on Sentinel-2 multi-spectral images acquired over ten countries in Europe. The class annotations were extracted from reference maps originating from the publicly available thematic product CORINE Land

Cover Map inventory of 2018. Each of the 590,326 patches is annotated with a subset of 19 LULC (Land Use and Land Cover) classes, including different types of forests, water bodies, and complex urban or agricultural classes. The dataset contains an average of 2.95 present classes per patch. Figure 6.3 visualised four samples from the dataset. The BEN data set contains noisy pixel-level reference maps, which are derived from a thematic product. However, due to the strong noise, these maps are not used for the evaluations.

For the experiments, only images from the country Lithuania were used, because this country encompasses a broad representation of all classes. This selection was strategically made to reduce the computational load required for running explanation metrics. The BEN-Lithuania dataset comprises a total of 51,624 samples, divided into training, testing, and validation subsets, making up about 46.8%, 28.2%, and 25% of the data respectively. Each patch in this dataset averages 3.8 classes, a higher count than the overall dataset, showcasing the diverse terrain and land cover types prevalent in this Baltic region. For simplicity, BEN lithuania is denoted as BEN

The label distribution for BEN-Lithuania, as visualised in Figure 6.4, reflects the geography of Lithuania, with a strong presence of agricultural and forested areas due to the country's extensive arable land and forests. Industrial areas are also well represented. However, some classes such as "Beaches, dunes and sands" are minimally represented, reflecting Lithuania's short coastline.

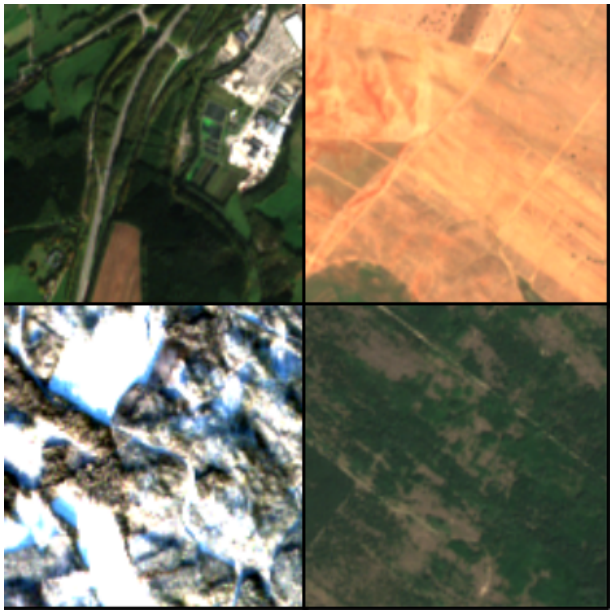


Figure 6.3: Four Examples from the BEN dataset.

6.2 Experimental Setup

This section describes the experimental setup, methodologies, and evaluation techniques used to investigate the capability of xAI methods on RS data and to assess their impact on improving classifier training.

The content is divided into two main parts. The first subsection defines the setup of the eval-

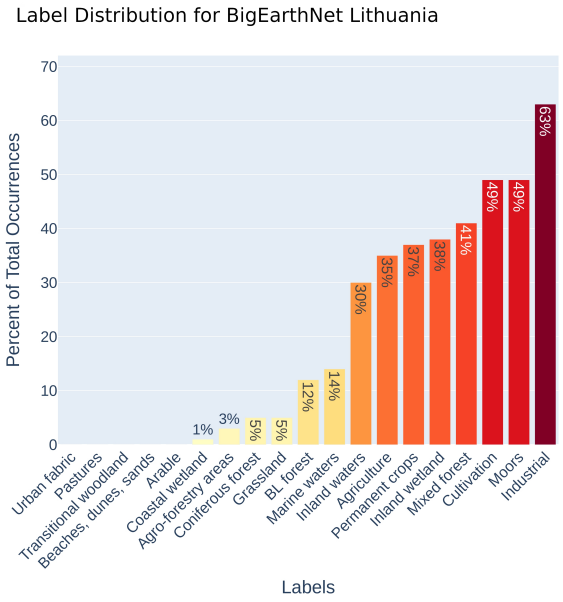


Figure 6.4: Label Distribution for BEN Lithuania, the label names are shortened.

uation for xAI methods and metrics on RS data, corresponding to Objective A 1. The second subsection defines the setup for the xAI-guided training, aligning with Objective B 1. To ensure reproducibility, each experiment was conducted three times using distinct random seeds $\rho \in \{42, 43, 44\}$. These iterations are henceforth referred to as *runs*. All code is published on GitHub in the main repository and the adapted explanation metrics repository.

Explanation Methods and Metrics

This section outlines the methodology used to compare the effectiveness of the different explanation methods across the explanation metrics. Each method was applied to models trained on the datasets (see Section 6.1). A detailed description of the methods can be found in Chapter 3, and a detailed description of the metrics in Chapter 4. The hyperparameters used for the metrics can be found in the Appendix 3. The aim is to assess whether the methods developed for RS data perform differently from those applied to classical CV tasks, and how they vary considering different RS data complexities.

In each run, a VGG16 classifier [79], pretrained on ImageNet [124] with weights from TorchVision [125], was fine-tuned on each dataset for 20 epochs. The hyperparameters used were: `Learning Rate`: 0.025, `Momentum`: 0.9, `Weight Decay`: 0.0005, `Optimizer`: 'Stochastic Gradient Descent'. The training outcomes are recorded as a baseline in Table 7.1. For SLC predictors, accuracy was the main metric; for MLC predictors, macro mean Average Precision (mAP) was the main metric, as it is well suited to MLC. To explain the predictions of a NN, it is necessary to evaluate whether a class is present in the image. For SLC this can be determined by using the maximum value from the prediction logits. However, for MLC tasks, where multiple classes may be present, a classification threshold is required. For the DeepGlobe datasets a classification threshold of 0.5 was used, while for BEN the optimal classification threshold

was approximated.

The implementation framework utilized was PyTorch [126], specifically leveraging the Lightning extension [127]. Explanation methods were adapted from Captum [128]; these methods, originally designed for SLC, required slight modifications:

Given a dataset of images and corresponding multi-labels, each image and label pair is denoted as $(\mathbf{x}_i, \mathbf{y}_i)$. Here, $\mathbf{x}_i \in \mathbb{R}^{C\times H\times W}$ represents an image with C channels (such as RGB bands), height H, and width W. The multi-label vector $\mathbf{y}_i \in \{0,1\}^L$ indicates the presence (1) or absence (0) of each of the L classes. For each class c, the entry $\mathbf{y}_i^c$ specifies whether class c is present in image $\mathbf{x}_i$. Let $\boldsymbol{e}_i^c \in \mathbb{R}^{C\times H\times W}$ represent the explanation for sample $\mathbf{x}_i$ concerning class c. Conversely, let $\boldsymbol{e}_i$ denote the explanation for the multi-label prediction $\boldsymbol{f}(\mathbf{x}_i)$, where $\boldsymbol{e}_i \in \mathbb{R}^{L\times C\times H\times W}$. For the experiments, only the classes predicted by the classifier were evaluated. If class c was not predicted, then $\boldsymbol{e}_i^c = 0$. Accordingly, a multi-label explanation method is a function Φ that produces a multi-label explanation:

$$\Phi : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{L\times C\times H\times W}, \tag{6.1}$$

which produces a vector of explanations $\boldsymbol{e}_i \in \mathbb{R}^{L\times C\times H\times W}$ for all classes $c \in L$. To generate $\boldsymbol{e}_i$, for each predicted c an explanation $\boldsymbol{e}_i^c \in \mathbb{R}^{C\times H\times W}$ was calculated and finally concatenated. For simplicity, the explanations were aggregated across channels by summing over the channels C:

$$\boldsymbol{e}_i^l = \sum_{k\in C, c\in L} \boldsymbol{e}_i^{k,c}, \tag{6.2}$$

where $\boldsymbol{e}_i^{k,c}$ denotes the explanation for class c and k-th channel.

The hyperparameters for the explanation methods were set as follows: For DeepLIFT, the baseline value was set to 0. In the case of GradCAM and Guided GradCAM, the selected layer was the last convolutional layer, and bilinear interpolation was used. For IG, the number of steps was set to 50 with a baseline value of 0, and the Gauss-Legendre method was utilized for the integral approximation. The LIME method employed a similarity measure using a Euclidean exponential kernel with a kernel width of 500. The interpretable model used was linear regression from [129], and 15 superpixels were generated using Simple Linear Iterative Clustering (SLIC) from [130]. For LRP, according to [2], the first third of the network utilized the LRP-γ rule, the second third utilized the LRP-ϵ rule, and the final third applied the LRP-ϵ 0 rule. Lastly, the occlusion method employed a sliding window of size $(50, 50)$ over the channels with strides of $(10, 10)$, and the baseline was set to 0.

Qualitative evaluation of explanations poses significant challenges. Ancona et al. [94] argue that user perceptions favour simpler, shape-focused explanations, which may not accurately reflect the internal processes of the network. Furthermore, most metric categories cannot be evaluated qualitatively; only Localisation and Complexity metrics are amenable to such evaluation. Categories such as Robustness or Randomisation, which require model modifications or input perturbations, must be evaluated quantitatively. Given the questionable value of simpler explanations for RS image (see Section 4.7), the focus in the qualitative evaluation lies on the Localisation of the OoI. Further details are discussed in the Chapter 4. For the visualisations, the VGG model trained with the random seed $\rho = 42$ was used, which showed approximately the same performance as the baselines.

As most metric categories cannot be evaluated qualitatively, the quantitative evaluation becomes even more important. However, the explanation metrics showed significant outliers and varying ranges between different metrics, making a direct comparison between different metrics difficult. To address these issues, several pre-processing steps were implemented.

Specifically, a logarithmic transformation, $\log_2(\Psi(\boldsymbol{e}_i))$, was applied to two metrics: RIS and ROS. The following steps were applied to all metrics. To further mitigate the influence of outliers, Robust Scaling was applied, by adjusting the scaling by using the median and the interquartile range (IQR), calculated as the difference between the 75th and 25th percentiles:

$$V''_\Psi = \frac{V'_\Psi - \text{median}(V'_\Psi)}{\text{IQR}}, \quad \text{with IQR} = Q_3 - Q_1, \tag{6.3}$$

with $V'_\Psi = \text{AVG}_{i \in \boldsymbol{X}_{\text{te}}} \Psi(\boldsymbol{e}_i)$ Furthermore, Min-Max Scaling was employed to normalize the values into a $[0, 1]$ range, to ensure comparability across all metrics:

$$V'''_\Psi = \frac{V''_\Psi - \min(V''_\Psi)}{\max(V''_\Psi) - \min(V''_\Psi)}. \tag{6.4}$$

Finally, for metrics where a lower value indicates a better outcome, the scores were adjusted to align all metrics such that higher values consistently represent better outcomes:

$$V_\Psi = 1 - V'''_\Psi, \tag{6.5}$$

where V_Ψ is the final value for metric Ψ.

Explanation-Guided Training

This section outlines the experimental design to assess whether two xAI-guided training methods, described in Chapter 5, can enhance the training process of NNs. As stated in Objective B 1, the primary goal for RRR is to improve reasoning, whereas the goal for CutMix with xAI LP is to improve performance and generalisation capabilities.

Right for the Right Reasons

For the RRR evaluation, VGG16 [79] models were trained using the specifications and hyperparameters listed above. The baseline as described in Table 7.1 was used. Experiments were conducted on the Caltech101 and DeepGlobe datasets to compare the SLC and MLC results. The BEN datasets was excluded as it does not provide a ground truth reference map. For RRR, a parameter λ can be used to scale the weighting of the RRR loss and the regular loss. Several values of λ have been used. Both, the original RRR and RRR MSE, RRR and RRR_{MSE} were utilised.

Evaluating improvements in model reasoning involves assessing its reliance on SC in the training data. For the Caltech101 dataset, which contains known SC, one can qualitatively assess whether the models rely on SC after training modifications. Additionally, improvements in Localisation metrics are measured. For a dataset containing SC, an improvement in these

metrics is expected; if the model uses SC, explanations will be localised in the region of the SC rather than on the OoI. However, these analyses were restricted to the Caltech101 dataset, as the multi-label datasets do not contain any known SC. As the model is encouraged not to use spurious features for prediction, the performance on the test metrics is expected to decrease. However, the method cannot be used if the performance of the model is substantially reduced. Therefore, two factors are evaluated: the test metrics and the reasoning.

CutMix with xAI Label Propagation for Multi-Label Classification

To evaluate CutMix with xAI LP guided training, VGG16 models were trained using the specifications and hyperparameters listed above. In addition, to compare the results with the original experiments by Burgert et al. [65], a ResNET34 [119] was trained using similar hyperparameters to the VGG models. To evaluate the model improvements, a baseline without augmentation was trained. As an upper bound, CutMix training was performed using LP with the original reference maps, as this is the optimal LP strategy. To evaluate the method the performance increase in the test metrics compared to this upper bound and the baseline without augmentations was evaluated.

The xAI integraton of CutMix utilizes three relevant hyperparameters: t_{map} and t_{cam} [65], and the box size [66]. The t_{map} threshold defines how many activating pixels are necessary for a label to be propagated, and the t_{cam} threshold defines the minimal relevance that pixel k needs to have to be considered in the binary explanation map. To choose thresholds t_{map} and t_{cam}, one can utilize the given reference maps. For a MLC task, let $x_i \in X$ be an input image, $y_i \in Y$ its corresponding multi-label vector, $\boldsymbol{e}_i$ its explanation mask, and s_i its reference map. Let $\tilde{x}$ be the CutMix augmentation of that image. Let $\phi : \{0,1\}^{L\times H\times W} \rightarrow \{0,1\}^L$ be the readout function to derive the new label $\tilde{y}$ from $\tilde{x}$. The readout function can be used with the augmented explanation masks $e\tilde{y}_i = \phi(\tilde{e}_i)$ or with the reference map to derive the true new label $s\tilde{y}_i = \phi(s_i)$. To approximate the optimal thresholds t_{map} and t_{cam} for a set of explanation masks E, one can maximize the accuracy between $e\tilde{y}_i$ and $s\tilde{y}_i$. The results for the DeepGlobe datasets, visualized in different matrices, for each explanation method, are shown in figure 1 in the appendix. The choice of the t_{cam} is best at 0, while t_{map} has a less significant impact. For the experiments, $t_{cam} = 0$ and $t_{map} = 10$ were selected, with t_{map} set similar to the original paper. For the last relevant parameter, the box size, experiments were conducted as suggested by the authors [65], testing different box sizes of $[0.1 - 0.5]$ and $[0.3 - 0.7]$.

Correlation Analysis

To examine the relationship between explanation metrics and performance in xAI-guided training, the correlation r between the results of each explanation metric Ψ for each explanation method Φ and the training success measured by a metric M on the test set $\boldsymbol{X}_{\text{te}}$ was evaluated. The correlations where calculated were the Pearson correlation with their correlation coefficient denoted as ρ. Formally:

$$r = \rho(V_\Psi), M_{\text{mAP}}(\boldsymbol{f}, \boldsymbol{X}_{\text{te}}), \tag{6.6}$$

with V_Ψ denoting the preprocessed value for metric Ψ, as described in Equations 6.3 to 6.5.

To consider not only correlations for individual metrics, the mean of the correlations was cal-

culated for the different metric categories. Correlation coefficients have a skewed distribution, which means that they are not symmetrically distributed around their mean, especially if they are close to -1 or 1. This skewness means that the simple arithmetic mean of these coefficients does not represent the 'central' value in their distribution. The simple mean can therefore give a distorted picture of the average correlation.

The Fisher z-transformation converts correlation coefficients into a value that is approximately normally distributed. This transformation thus equalises the skewness of the original values. Normally distributed values are easier to handle, mainly because the mean of a normally distributed variable is a good estimator for the central trend. The transformation is formally expressed as:

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right), \tag{6.7}$$

where r denotes the Pearson correlation coefficient.

Subsequently, the mean of the Z-values was calculated to establish a central metric of the transformed coefficients:

$$\overline{Z} = \frac{\sum_{i=1}^{n} Z_i}{n}, \tag{6.8}$$

here n represents the count of correlation coefficients, and Z_i indicates the Fisher-Z transformed values. To convert the mean of the Z-values back to a correlation coefficient, the inverse Fisher-Z transformation was applied:

$$\overline{r} = \frac{e^{2\overline{Z}} - 1}{e^{2\overline{Z}} + 1}. \tag{6.9}$$

This final step recalculates the average correlation coefficient in its original scale. Making it possible to directly compare the linear relationship strengths across different metric categories. However, it is important to note that the average of the correlations in a category is only an estimate of a trend across the category as a whole. Individual metrics within that category that have a strong positive correlation may be cancelled out by other metrics that have a strong negative correlation. Therefore, the average can only be used as a reliable measure if it is assumed that the metrics within a category provide approximately the same insight into the explanation method.

7 Experimental Results and Discussion

In this chapter, the experimental results and discussions are presented in three sections: first, a qualitative assessment of explanation methods; second, a quantitative analysis using explanation metrics; and finally, an evaluation of explanation-guided training contextualised by the quantitative analysis. To differentiate between the different explanation methods, a distinction is made between Backpropagation-based (BP-based) methods, which include LRP, DeepLIFT, IG and Guided GradCAM, and non-BP-based methods, which are further subdivided into CAM-based methods, such as GradCAM, and perturbation-based methods, such as LIME and Occlusion. The results are based on the baseline models, with the corresponding test metrics visualised in Table 7.1.

Table 7.1: Performance of the baseline models averaged over 3 runs. The metric used for Caltech101 was accuracy and mAP for the DeepGlobe and BEN datasets.

Model	**Dataset**		
	Caltech101	**DeepGlobe**	**BEN**
VGG16	97.39	83.81	60.47
ResNET34	-	83.84	61.68

7.1 Qualitative Assessment of Explanation Methods

This section provides a qualitative assessment of the explanation methods applied to the Caltech101, DeepGlobe and BEN datasets. By examining the visual quality of the generated explanations, the effectiveness of their corresponding explanation methods is analysed in the context of the SLC and MLC tasks. The focus is on understanding how well each method highlights relevant features given the structure of the task and nature of the data. As the Caltech101 dataset was designed for a SLC task, the methods are expected to work as intended. The focus is on understanding how well each method highlights relevant features, given the structure and nature of the data. As discussed in section 6.2, the qualitative analysis focuses on two properties. First, how well the method localises the Object of Interest (OoI) and, second, the visual appeal of the explanation, e.g. if it contains noise and is "understandable" from a human perspective.

Figure 7.1 shows a visual representation of an explanation for the class *motorbike* from the **Caltech101** dataset. The prediction of the VGG model was correct. The Figure includes the image, the reference map and the explanations for each explanation method as heatmaps $\boldsymbol{e}_i$. Each value in the heatmap $\delta_{j,k} \in \boldsymbol{e}_1$, ranges from 0 to 1, with $\delta_{j,k} = 1$ indicating the highest attributed relevance.

The visualisation shows that the BP-based methods, namely DeepLIFT, Guided GradCAM, IG and LRP, provide fine-grained explanations, with more focus on edges of the input image.

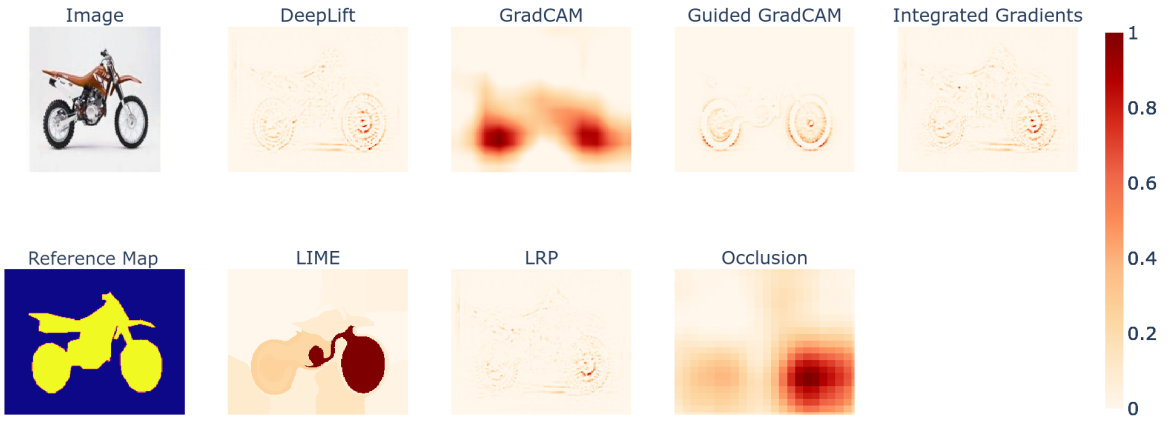


Figure 7.1: Caltech101: Single-label explanations for the correct prediction of the class *motorbike*. The plots in order are: image, explanations: DeepLIFT, GradCAM, Guided GradCAM, IG (1st row), reference map, explanations: LIME, LRP, Occlusion (2nd row). Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

In contrast, the explanations of LIME, GradCAM and Occlusion occupy a broader area. It can be seen that there is a focus on the front wheel of the motorcycle for all explanation methods. This pattern is evident in many of the other samples containing this class, making it the discriminative feature for this class.

The DeepLIFT method appears to identify all edges, including shadows on the ground and edges at the edges of the image that do not contribute to the explanation. The resulting image appears noisy, with the main focus on the front wheel. The GradCAM algorithm highlights both wheels and most other parts of the bike. The coherence of the explanation makes it seem less complex and easier to understand. The Guided GradCAM method also highlights the edges, but is more focused on the OoI, resulting in a clearer and less noisy explanation compared to DeepLIFT. The IG explanation is similar to DeepLIFT, with the same noise and problems. The LIME method places a clear emphasis on the front wheel, with a lesser focus on the rear. The explanation is easy to understand as only a few distinct image regions have been identified as significant. The LRP explanation looks analogous to the DeepLIFT and IG. For the Occlusion saliency map the sliding windows are visible, resulting in a more noisy image, however, also with a clear focus on the front wheel. From a subjective standpoint, GradCAM and Guided GradCAM perform best, depending on whether the user requires shape-focused or texture-focused explanations.

Figure 7.2 shows another sample originating from the Caltech101 dataset. The OoI in the image is a *leopard*, and the prediction is accurate. However, the explanation methods demonstrate that the black border around the image is the most predictive feature, meaning that the prediction was based on a Spurious Correlation. In general, all explanation methods are more noisy, and none highlight the object of interest. When analysing the images for the leopard class, it can be observed that the majority of images contain this black border. Consequently, the model utilised the border as a shortcut. Nevertheless, the leopard also is attributed some relevance. This aligns with the findings of [23], that the feature information of the OoI is also

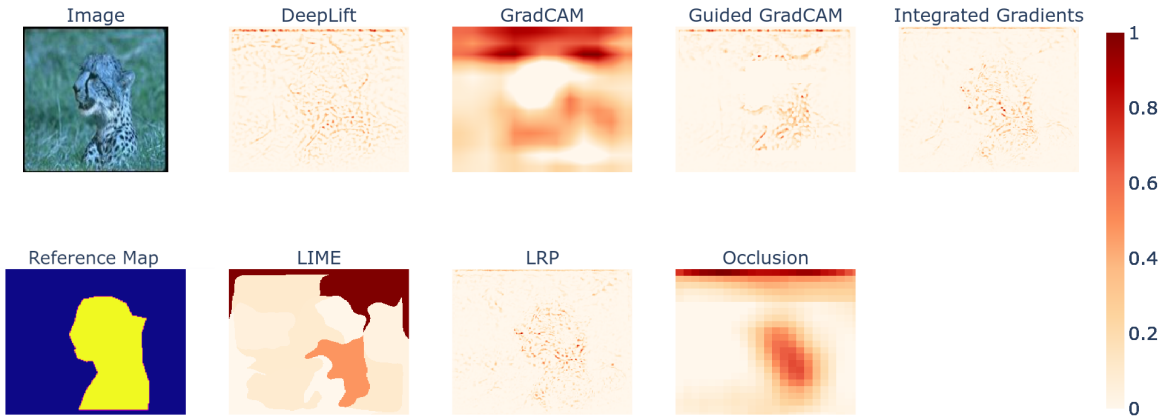


Figure 7.2: Caltech101: Single-label explanations for the correct prediction of the class *leopard* with a Spurious Correlation. The plots in order are: image, explanations: DeepLIFT, GradCAM, Guided GradCAM, IG (1st row), reference map, explanations: LIME, LRP, Occlusion (2nd row). Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

encoded in the internal network representations.

To evaluate the debugging qualities of the explanation methods, it is analysed if the methods successfully identify the SC. For LIME, this identification is somewhat dependent on the superpixel setup. The BP-based methods produce edge maps that are not easily distinguishable from their normal edge maps, making it more challenging for them to accurately identify the SC. GradCAM and Occlusion are particularly effective at highlighting the SC.

In the following, a qualitative analysis of selected samples from the **DeepGlobe** dataset is conducted. The analysis explores both well-executed and poor explanations across samples with correct and incorrect predictions.

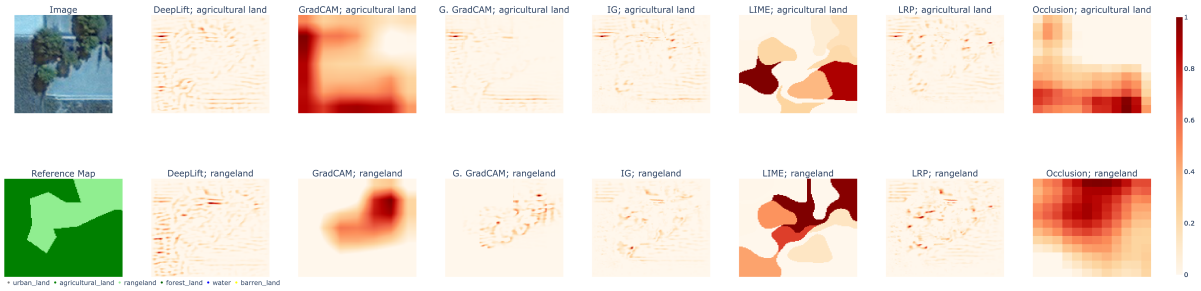


Figure 7.3: DeepGlobe: Multi-label explanations for the correct prediction of the classes agriculture land and range land. The first column shows the input image and its reference map. Then each row shows the class-wise explanations for the methods: DeepLIFT, GradCAM, Guided GradCAM, IG, LIME, LRP and Occlusion. Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

Figure 7.3 shows visual explanations for sample 2619 from the DeepGlobe dataset, high-

lighting the positive attribution across different explanation methods for each prediction. The structure of the plot is similar to the explanations visualised earlier, the only difference being that one set of explanations is provided per predicted class. The example contains the classes agricultural land and range land. From a human perspective, these two classes often have similar characteristics. Both ground types are often cultivated in rectangular shapes and have little prominent features.

The first method discussed is DeepLIFT, which does not focus well on the locality of the classes. It attributes pixels outside the reference map of the target objects and similarly highlights pixels for both agricultural and range land, with only the attributed relevance differing. The resulting image looks more like an edge map than an explanation. The second method, GradCAM, provides a more precise localisation of the OoI. For the agriculture class, it highlights the left and bottom parts of the image, where the agriculture area is located. Furthermore, it provides a more coherent visual explanation, although it assigns relevance to a greater number of pixels. Guided GradCAM produces an edge-like attribution map but correctly highlights only the relevant locations. The IG and LRP methods produce similar-looking explanations, highlighting the same edges as DeepLIFT, but distributing the relevance differently. For both methods, the most relevant pixels for range land are correctly localized, whereas, for agricultural land, the entire lower portion is incorrectly deemed irrelevant. LIME accurately locates the most relevant input features but misses many also containing the class. The Occlusion method also maintains a consistent localisation, with a similar relevance assignment to GradCAM, but appears slightly more noisy due to its low resolution.

An observation is that all BP-based methods tend to highlight the same pixels consistently, differing mainly in the assigned relevance values. However, except for Guided GradCAM, they all assign relevance to "wrong" pixels, pixels that do not contain the class of interest. For Guided GradCAM that is only due to its "guidance" via GradCAM.

GradCAM, LIME, and Occlusion each yield distinct results; among them, GradCAM is the most precise and comprehensible, LIME's attributions show some inconsistency with the ground truth, and Occlusion appears as a noisier variant of GradCAM, confirming GradCAM as the superior method for this sample.

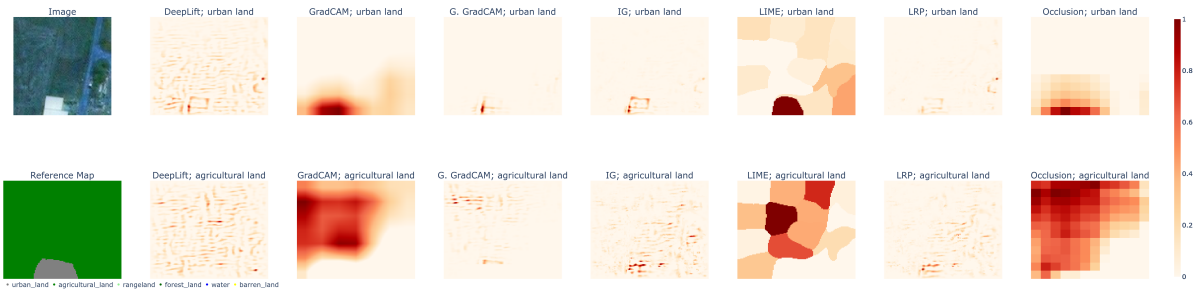


Figure 7.4: DeepGlobe: Multi-label explanations for the correct prediction of the classes agriculture land and urban land. The first column shows the input image and its reference map. Then each row shows the class-wise explanations for the methods: DeepLIFT, GradCAM, Guided GradCAM, IG, LIME, LRP and Occlusion. Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

Figure 7.4 shows the explanations for sample 2619 from the DeepGlobe dataset. The sample

contains agriculture and urban land and was also predicted correctly. The depiction is similar to the one in Figure 7.3. The class urban land only occupies a small area, compared to agriculture, which occupies most of the image. Compared to classes like agriculture, urban land has a more discriminative shape, as it usually contains rectangular houses.

In terms of explanation methods, DeepLIFT emphasises edges and lacks precision in class localisation, although the most relevant pixels are accurately located. However, it does not provide a good human understanding of the model's decision process, as the explanation contains strong noise. GradCAM performs well overall, but incorrectly assigns some relevance to rectangular shapes on the right side of the image. It locates urban areas and most agricultural areas accurately. Guided GradCAM similarly highlights both classes accurately, again with more focus on the edges. IG excels at identifying the shape of urban structures by focusing solely on this class, but it misses agricultural areas and provides noisy explanations by incorrectly attributing relevance to shapes in urban land. LIME effectively identifies the most relevant pixels, particularly for the urban land, but fails to distribute relevance to the large agricultural class map. LRP identifies urban land accurately but does not perform well with agricultural land, giving results similar to IG but with less focus on urban detail. Occlusion, like GradCAM, again provides a noisier but similar performance.

Overall, GradCAM and LIME emerge as the most effective method for both classes in this sample.

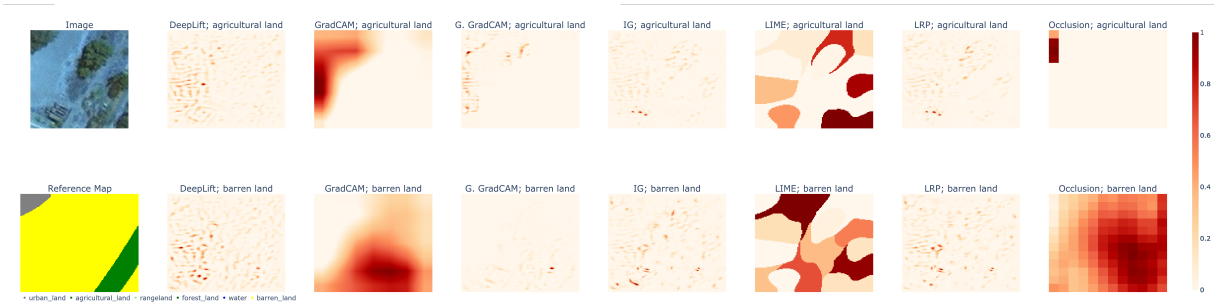


Figure 7.5: DeepGlobe: Multi-label explanations for the partly correct prediction of the classes agriculture land and barren land, urban land was not predicted correctly. The first column shows the input image and its reference map. Then each row shows the class-wise explanations for the methods: DeepLIFT, GradCAM, Guided GradCAM, IG, LIME, LRP Occlusion. Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

Figure 7.5 shows the explanations for the incorrectly predicted sample 5794. The layout of the Figure is similar to the previous figures. Investigating wrong predictions is a critical use case for explanation methods. For this sample, the classes barren land and agricultural land were predicted correctly, but the urban land class was predicted incorrectly. It is evident that the explanations for the agricultural class were inaccurate across all explanation methods, with a heavy focus on the left side of the image, whereas the actual land coverage for agriculture is in the bottom right. LIME is the only method that localizes this class more accurately. However, it does so in contrast to other methods, such as GradCAM and Occlusion, which attribute most relevance to the top left. This discrepancy raises concerns about why the methods interpret the model's decision so differently and which method explains the model correctly. The other class,

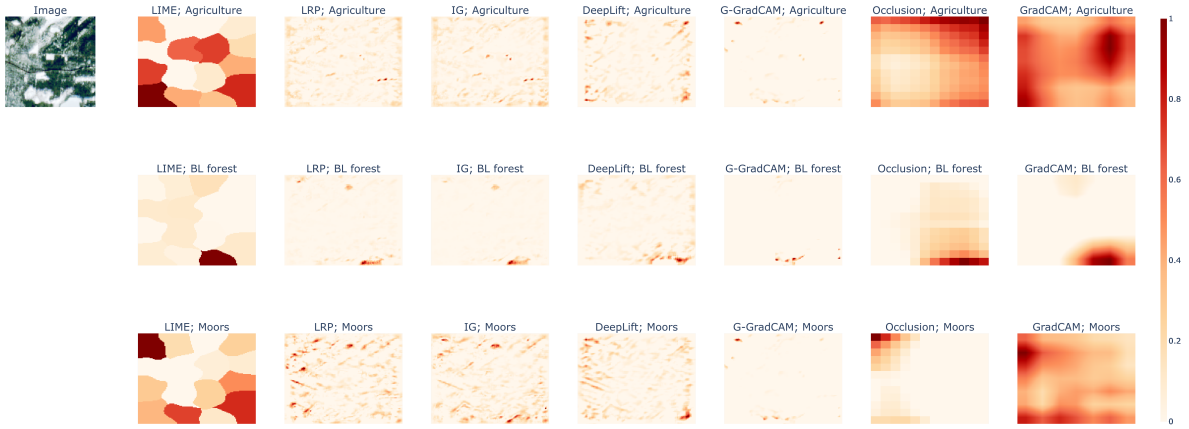


Figure 7.6: BEN: Multi-label explanations for the correct prediction of the classes agriculture, broad-leaved forest and moors. The first column shows the input image. Then each row shows the class-wise explanations for the methods: DeepLIFT, GradCAM, Guided GradCAM, IG, LIME, LRP and Occlusion. Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

barren land, is attributed more accurately, with explanation properties similar to the previous figures.

The patterns observed in the previous analysis of the DeepGlobe dataset persist in the BEN dataset. However, the absence of accurate ground truth complicates the visual evaluation of the explanations. Figure 7.6 visualizes explanations for correct predictions in the classes agriculture, broad-leaved forest, and moors. For the agriculture class, LIME attributes high relevance to three regions in the image, each consisting of two patches. These patches are located in the middle, bottom left, and bottom right of the image. BP-based methods (LRP, IG, DeepLIFT, Guided GradCAM) consistently highlight edges, similar to their behaviour in the DeepGlobe dataset. The Occlusion method assigns high importance to the top and top-right corner, while GradCAM attributes relevance to one patch in the top right and another on the left side. In the broad-leaved forest class, all methods attribute relevance to the bottom right, a patch occluded by a cloud. The visual evaluation suggests that the broad-leaved forest occupies almost the entire right half of the image. The consistent highlighting of the clouded patch by all methods indicates the model may have learned incorrect features. For the Moors class, all methods attribute relevance to the top left of the image. BP-based methods generate strong noise, while other methods agree on the moor's position. However, the absence of a ground truth map makes verifying the moor's actual location challenging.

Figure 7.8 illustrates the explanations for a BEN sample containing industrial land, complex cultivation patterns, mixed forest, and inland wetlands.

For the industrial land class, LIME and Occlusion attribute importance to the bottom-left corner, while GradCAM attributes importance to the entire left side and a patch in the bottom-right corner. Assessing the actual class location is challenging. Notably, the explanations from BP-based methods fail here, unlike in the DeepGlobe dataset. Urban and industrial lands share several features, such as high built-up area density, numerous buildings, and impervious surfaces like asphalt and concrete, with scarce vegetation. Despite these similarities, BP-based

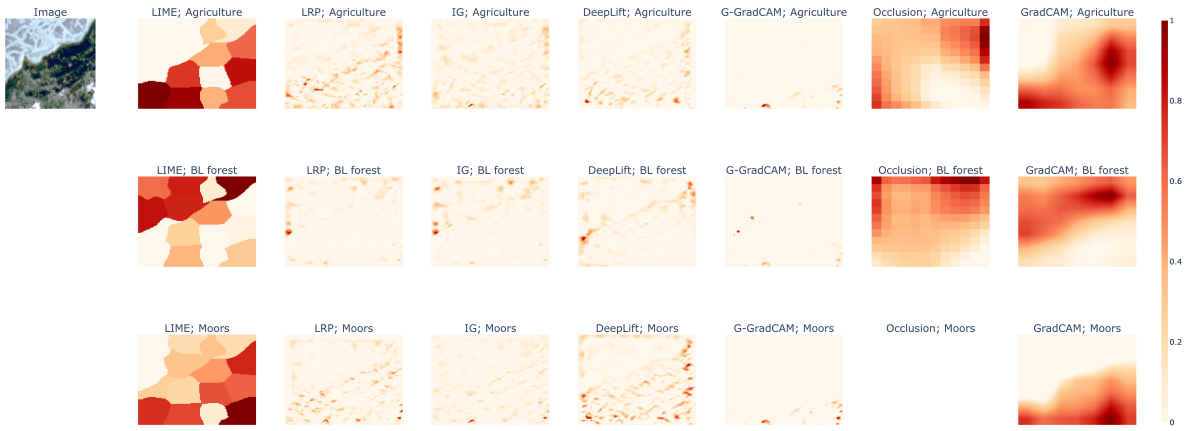


Figure 7.7: BEN: Multi-label explanations for the correct prediction of the classes agriculture, broad-leaved forest and moors. The first column shows the input image. Then each row shows the class-wise explanations for the methods: DeepLIFT, GradCAM, Guided GradCAM, IG, LIME, LRP and Occlusion. Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

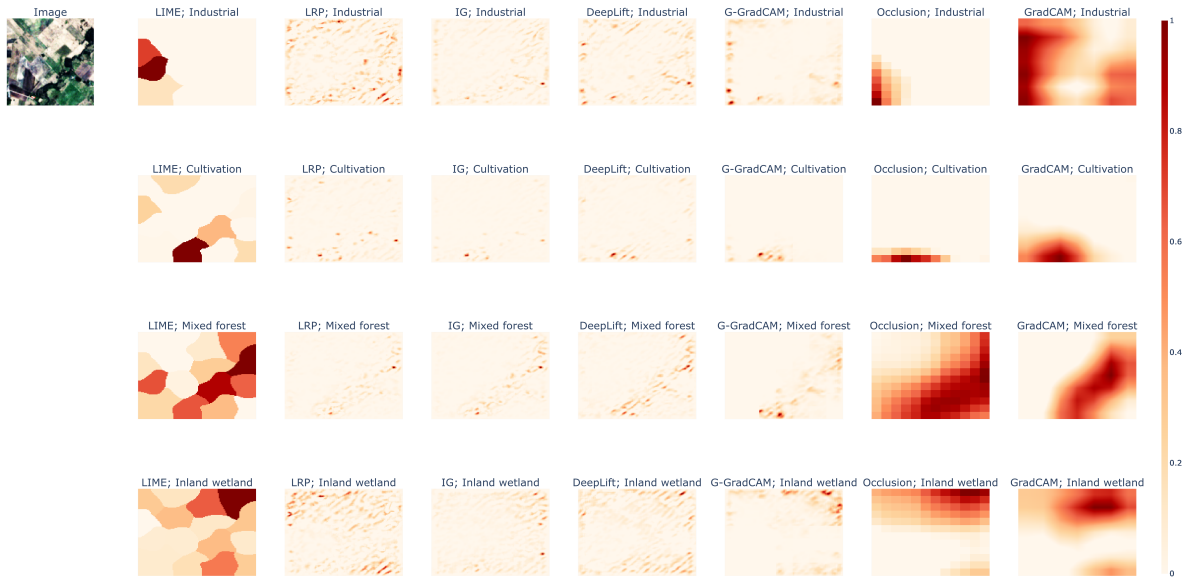


Figure 7.8: BEN: Multi-label explanations for the correct prediction of the classes industrial land, complex cultivation patterns, mixed forest and inland wetlands. The first column shows the input image. Then each row shows the class-wise explanations for the methods: DeepLIFT, GradCAM, Guided GradCAM, IG, LIME, LRP and Occlusion. Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

methods appear to visualize random edges for this class, possibly due to the overload of class information in MLC. For complex cultivation patterns and mixed forest classes, all methods seemingly attribute the locations correctly. The model identifies complex cultivation patterns in the bottom left and a mixed forest extending from the bottom left to the top right. The inland wetland class is also attributed correctly, although the explanations exhibit some noise.

The qualitative evaluation reveals significant differences in the effectiveness of explanation methods between the CV SLC dataset and the RS MLC datasets. For the Caltech101 dataset, explanations are generally well-focused on the OoI, with the methods performing effectively in highlighting relevant features. In contrast, explanations for the RS datasets appear more chaotic. BP-based models tend to perform better with unique, complex features, such as the urban land class in the DeepGlobe dataset, likely due to their ability to capture distinct edges and shapes and underperform with repetitive textures typical of RS data, such as the agriculture class. The performance of the evaluated model significantly impacts the quality of the explanations. The model performs best on the Caltech101 dataset and worst on the BEN dataset. This performance discrepancy is likely to bias the evaluation results, making it difficult to isolate the performance of the explanation methods from the overall accuracy of the model. Consequently, the evaluation remains unverifiable within the current scope. For the DeepGlobe dataset, there is a strong visual preference for non-BP-based methods, specifically GradCAM, LIME and Occlusion. These methods provide broader, more area-occupying explanations that better highlight the textures and are more consistent with the characteristics of the dataset. From a qualitative, subjective point of view, GradCAM is observed to be the best-performing method.

7.2 Quantitative Analysis with Explanation Metrics

This Section provides a quantitative analysis of the examined explanation methods across multiple datasets, evaluating their performance using a range of selected metrics. The datasets examined are Caltech101, DeepGlobe, and BEN. The explanation metrics are categorized into five categories: Complexity, Faithfulness, Localisation, Randomisation, and Robustness, allowing for a detailed comparison of each method's strengths and weaknesses.

Firstly, the overall results per metric for the **Caltech101** dataset are visualized in Figure 7.9. Here, the rows show the explanation methods: Occlusion, GradCAM, LRP, IG, DeepLIFT, LIME and Guided GradCAM. The description on the top of the columns shows the category of the metrics, while the bottom description shows the name of each metric. A preprocessing step was applied to the metrics, to make them comparable, as described in Section 6.2.

In the **Complexity** category, all BP-based methods (LRP, IG, DeepLIFT, Guided GradCAM) achieve a score of 0.99 in the ECO metric, significantly outperforming other methods which score between 0.59 and 0.66. In the Sparseness metric, Guided GradCAM leads with a score of 0.84. The other BP-based methods follow closely with scores around 0.64, whereas non-BP-based methods score approximately 0.37. This discrepancy is because methods such as Occlusion, LIME and GradCAM cover a larger area of explanation, as opposed to the finer details provided by BP-based methods. Complexity metrics value this as a more 'complex' explanation, however, this might not be the case for RS images, see Section 4.7.

For the **Faithfulness** category, the IROF metric sees Occlusion leading with a score of 0.69,

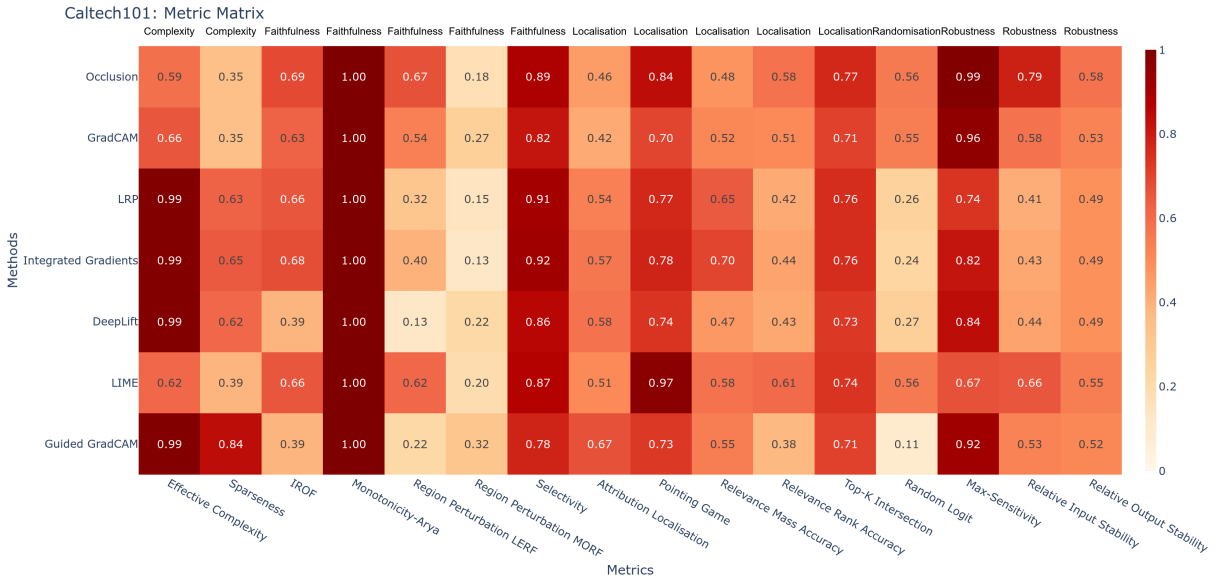


Figure 7.9: Caltech101: Matrix visualisation of the metrics for various explanation methods, including Occlusion, GradCAM, LRP, IG, DeepLIFT, and LIME, Guided GradCAM. The rows of the matrix represent different explanation methods, while the columns are categorized by metrics types at the top and specific metric names at the bottom

closely followed by IG, LIME, LRP and GradCAM. Both DeepLIFT and Guided GradCAM perform the worst, with an IROF of 0.39. All methods achieve the optimal score in the monotonicity metric (1), confirming that prediction confidence increases monotonically as more relevant features are included. In the SEL metric, IG scores highest with 0.92, while GradCAM and Guided GradCAM score lower with 0.82 and 0.78 respectively. When evaluating the RP metrics, particularly for the LeRF removal strategy, a notable outperformance of non-BP-based methods is observed. Occlusion emerges as the most effective method, achieving a score of 0.67, in contrast to DeepLIFT, which only achieves a score of 0.13. This disparity suggests that the removal strategy has a significant impact on performance outcomes. Conversely, when the MoRF strategy is used, all methods show low performance, suggesting that this strategy does not favour any particular type of explanation method within the dataset, with Guided GradCAM yielding the best results with a score of 0.32.

For the **Localisation** metric, AL the Guided GradCAM method performs best. Similar to the Complexity category, here, the BP-based methods perform better. Probably because the AL metric emphasises the ratio of relevance within the bounding box to total relevance, $\frac{R_{\text{in}}}{R_{\text{tot}}}$. For instance, the sliding window approach of Occlusion inherently attributes relevance outside the object, negatively affecting its score. In contrast, LIME leads the PG metric with 97% of the highest relevance contained within the OoI, followed by Occlusion with 0.84. The other methods range between 0.7 and 0.8. DeepLIFT consistently underperforms in the RMA and RRA metrics, while other methods show slightly better performance in these metrics. For the TKI metric, Occlusion scores the highest with 0.77, while GradCAM scores slightly lower with 0.71.

According to the results of the **Randomisation** metrics, and consistent with the results of

[83], BP-based methods underperform compared to LIME, GradCAM and Occlusion. In the **Robustness** category, Occlusion shows higher results in all metrics, particularly in RIS with a score of 0.79, significantly outperforming the other methods which average 0.50. For ROS the results are more uniform, ranging from 0.49 to 0.58. MS is dominated by Occlusion, GradCAM and Guided GradCAM with scores of 0.99, 0.96 and 0.92 respectively, with LIME scoring the lowest at 0.67.

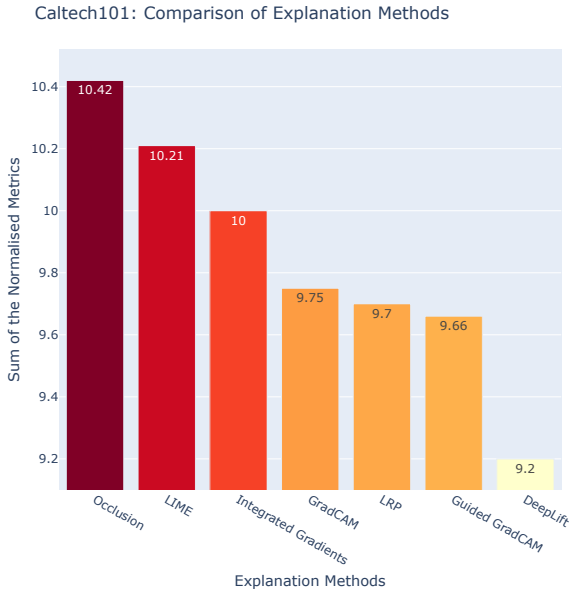


Figure 7.10: Caltech101: Comparison of explanation methods. The plot displays the sum of the normalised metrics for various explanation methods, including Occlusion, IG, LIME, LRP, Guided GradCAM, GradCAM, and DeepLIFT. The y-axis represents the sum of the normalised metrics, ranging from 9.1 to 10.4. The best possible value is 16.

The overall performance of explanation methods for the Caltech101 dataset is illustrated in Figure 7.10. The Figure shows the sum of the normalised explanation metrics for each method, as each metric was scaled from 0 to 1, the total performance of the methods can be compared. As there were 14 metrics in total the best possible value is 14. The Occlusion achieved the best overall result, scoring a sum of 10.42. This is due to its good performance in Robustness metrics and overall solid results. It is closely followed by LIME and IG, with scores of 10.21 and 10, respectively. TheDeepLIFT method performed worst, with a score of 9.2. However, the range of the results is not that large, with the best method only performing around 1.2 points better than the worst, which is about 10% from the total value.

Figure 7.11 shows the results for the quantitative evaluation for the **DeepGlobe** dataset. The structure of the Figure is similar to the previous matrix figure. The metrics are again analysed according to their respective categories. In the **Complexity** category, similar to the results for the Caltech101 dataset (7.2), all BP-based methods score 0.99 in ECO. In addition, Guided GradCAM achieves the highest score in the Sparseness metric at 0.78, with other BP-based methods also performing commendably.

In the **Faithfulness** category, Occlusion leads IROF with a score of 0.61, followed by GradCAM and LIME, unlike previous results where IG and LIME followed. All methods consis-

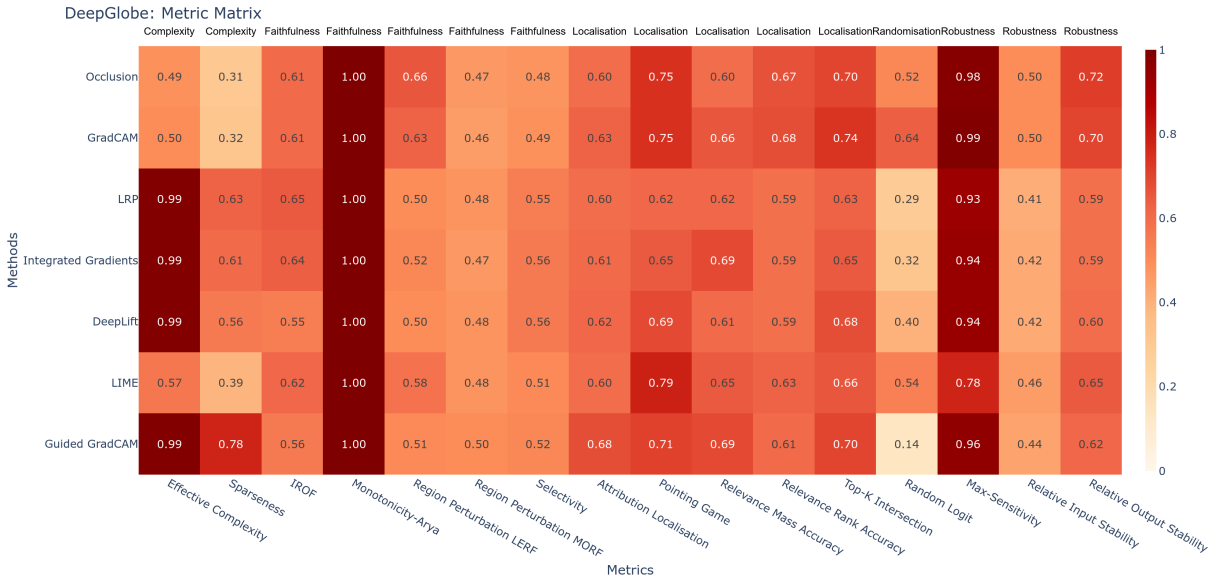


Figure 7.11: DeepGlobe: Matrix visualisation of the metrics for various explanation methods, including Occlusion, GradCAM, LRP, IG, DeepLIFT, and LIME, Guided GradCAM. The rows of the matrix represent different explanation methods, while the columns are categorized by metrics types at the top and specific metric names at the bottom.

tently achieve a perfect score of 1.00 in the Monotonicity metric. The RP metric, is considered using two removal strategies: MoRF and LeRF. The average score for LeRF is 55.7, higher than 47.7 for MoRF. The LeRF strategy, which is expected to perform better on RS images, does so because MoRF is more effective with unique, discriminative features, which are less prevalent in RS images, which are primarily distinguished by texture. This reasoning is further elaborated in the Section 4.7. The non-BP-based methods tend to outperform the LeRF strategy. For the SEL metric, IG and DeepLIFT rank highest, albeit closely, and Occlusion follows with a Selectivity score of 0.48.

In the **Localisation** category, LIME excels in the PG metric with a score of 0.79. The TKI metric shows competitive performance across all methods, with scores ranging from 0.65 to 0.74. AL sees Guided GradCAM in the lead with 0.68, with the other methods around 0.6. Interestingly, the BP-based methods now perform similarly to others, unlike the Caltech101 dataset. This could be due to the larger classes of interest, where broader explanations from non-BP-based methods yield a higher hit rate. For the PG metric, non-BP-based methods also perform better, with LIME leading, followed by GradCAM and Occlusion. The trend observed in the RMA metric, where BP-based methods were superior, does not hold, except for IG and Guided GradCAM, which maintain a high score of 0.69, with GradCAM at 0.66. In RRA, non-BP-based methods continue to outperform, and in the TKI metric, GradCAM scores the highest. LRP underperforms in all Localisation metrics.

In the **Randomisation** category, GradCAM and LIME show the highest performance in the RL test, with scores of 0.52 and 0.64 respectively. This pattern mirrors the results observed in the Caltech101 dataset, where BP-based methods significantly underperform compared to their counterparts. In the **Robustness** category, GradCAM leads with the highest Max-Sensitivity,

recording a score of 0.99, with Occlusion closely following. All methods except LIME show robust performance. In particular, GradCAM and Occlusion outperform the other methods on the additional Robustness metrics: RIS and ROS. They exceed the average performance of other methods by 0.07 for RIS and 0.10 for ROS.

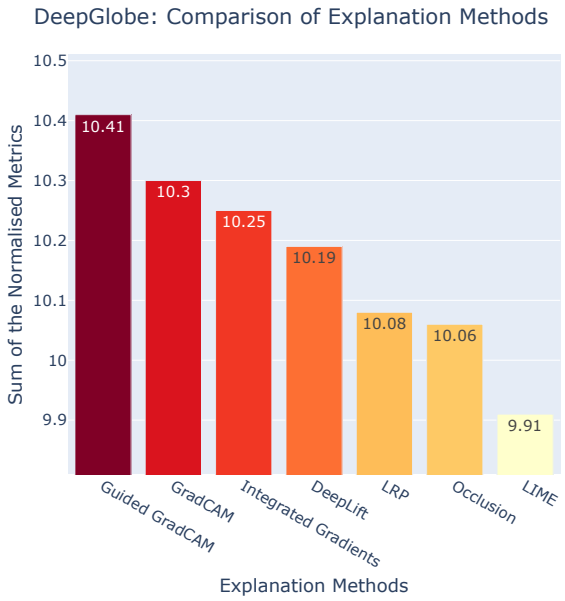


Figure 7.12: DeepGlobe: Comparison of explanation methods. The plot displays the sum of the normalised metrics for various explanation methods, including Occlusion, IG, LIME, LRP, Guided GradCAM, GradCAM, and DeepLIFT. The y-axis represents the sum of the normalised metrics, ranging from 9.9 to 10.5. The best possible value is 16.

The summed performance of the explanation methods for the DeepGlobe dataset is shown in Figure 7.12. The bar chart shows Guided GradCAM as the top performer with a total score of 10.41 out of a maximum of 16, closely followed by GradCAM with a score of 10.3. IG is the third best method, followed by DeepLIFT, LRP and Occlusion in descending order of effectiveness. LIME shows the weakest performance by these metrics. This visualisation highlights GradCAM and Guided GradCAM as the most effective methods overall in the DeepGlobe dataset. Interestingly, the Occlusion method performs significantly worse on the DeepGlobe dataset than on the Caltech101 dataset, where it was previously identified as the best performing method. This divergence in performance may be due to an inherent limitation of the method for RS images, which is discussed in more detail in Section 3.7.

Figure 7.13 illustrates the metric results for the BEN dataset, maintaining a structure similar to the previous matrices. Notably, the Monotonicity metric has been excluded from this analysis as its value is expected to be uniformly 1 across all methods and it has the longest computational time, approximately 65 seconds per class per sample per method. Additionally, the **Localisation** metrics were omitted due to the noisy reference maps associated with the BEN dataset, a condition discussed in more detail in Section 6.1.

The **Complexity** metrics show results consistent with previous analyses. For the ECO metric, all BP-based methods perform well, while other methods do not. This pattern is repeated in the Sparseness metric, with Guided GradCAM again performing best, closely followed by other

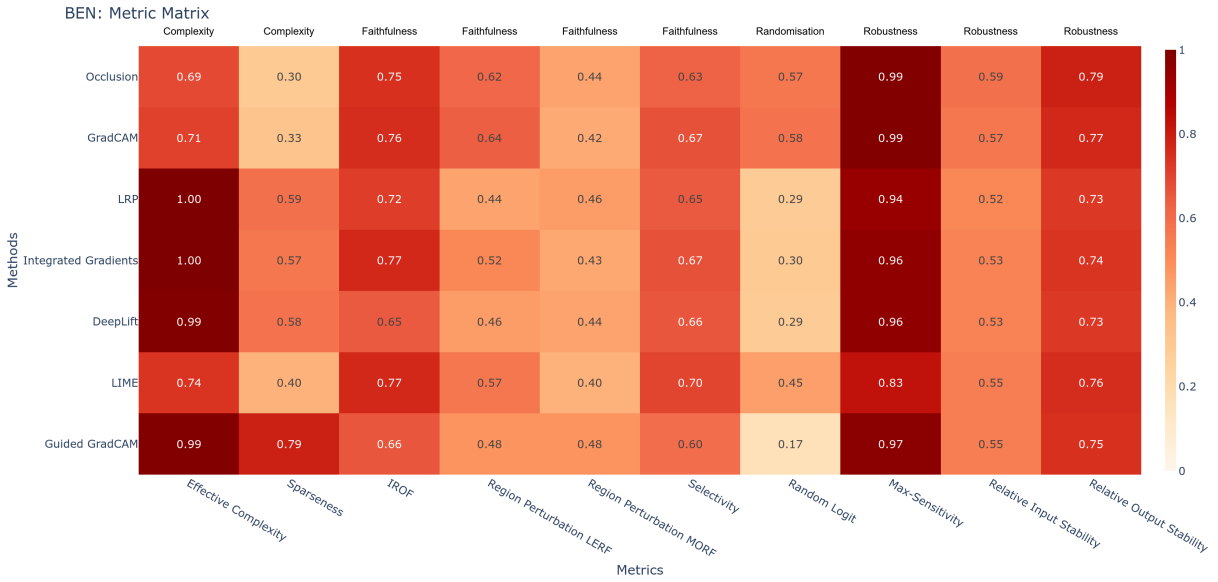


Figure 7.13: BEN: Matrix visualisation of the metrics for various explanation methods, including Occlusion, GradCAM, LRP, IG, DeepLIFT, LIME and Guided GradCAM. The rows of the matrix represent different explanation methods, while the columns are categorized by metrics types at the top and specific metric names at the bottom

BP-based methods.

In the **Faithfulness** category, the trend mirrors the one observed in the Caltech101 and DeepGlobe datasets, where DeepLIFT and Guided GradCAM underperform. Conversely, LIME and IG excel in the IROF metric with a score of 0.77. For the RP metric under the LeRF strategy, non-BP-based methods outperform others, consistent with the findings from the DeepGlobe dataset. The MoRF strategy shows a lower discrepancy between methods, with Guided GradCAM leading with 0.48. For Selectivity, LIME has the highest score of 0.70, with other methods performing similarly.

The **Randomisation** category continues the trend seen in the Caltech101 and DeepGlobe datasets, with non-BP-based methods performing better. Occlusion and GradCAM achieve the highest scores in the RL test with 0.57 and 0.58 respectively, while Guided GradCAM records the lowest score with 0.17. In the **Robustness** metrics, the Max-Sensitivity metric sees Occlusion and GradCAM leading with values of 0.99. All other methods are closely grouped, except LIME, which has the lowest score of 0.83. Occlusion and GradCAM also lead in all other Robustness metrics, although the other methods are closely matched in performance. The results are quite similar to the ones observed in the DeepGlobe dataset.

As shown in Figure 7.14, the performance of the explanation methods for the BEN dataset is close. IG leads slightly with a score of 6.49, closely followed by Guided GradCAM at 6.44 and GradCAM at 6.44. The methods LRP, Occlusion and LIME show similar effectiveness, with LIME scoring the lowest at 6.17. Despite these differences, the scores are very close, indicating that several methods show comparable performance across different metrics in the BEN dataset.

Figures 7.15 and 7.15 visualise the mean sum of the normalised explanation metrics on the

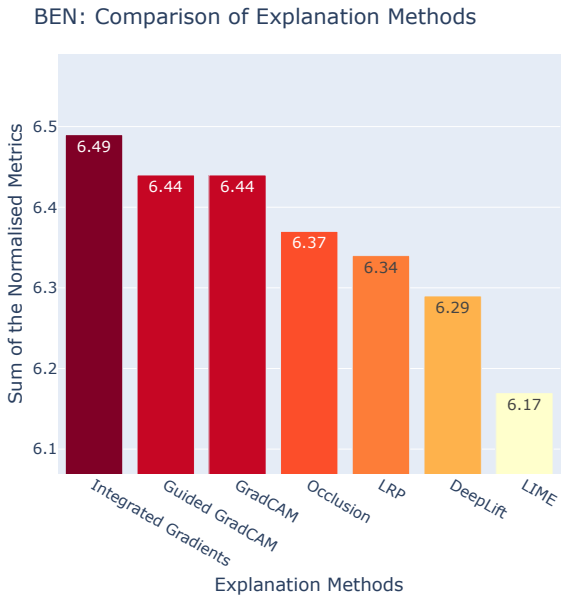


Figure 7.14: BEN: Comparison of explanation methods. The plot displays the sum of the normalised metrics for various explanation methods, including Occlusion, IG, LIME, LRP, Guided GradCAM, GradCAM, and DeepLIFT. The y-axis represents the sum of the normalised metrics, ranging from 6.25 to 6.55. The best possible value is 10.

y-axis, the corresponding explanation method used on the x-axis and the dataset used as color. The figures are used to compare the methods for natural and RS images, with a focus on the differences in data characteristics and task type.

In the **Complexity** category, as highlighted in Figure 7.15, BP-based methods (LRP, IG, DeepLIFT, Guided GradCAM) consistently outperform other methods across all datasets, indicating a general trend irrespective of task or dataset type. These methods provide simpler explanations, characterized by a lower number of pixels attributed as relevant. This uniform performance suggests that the inherent structure of the data does not significantly impact the metrics, despite the expectation that RS images might result in more complex explanations due to larger areas of interest. The results show a consistent trend across both image types. However, the suitability of these metrics for RS image data is questionable, as less complex explanations are not always better. In RS, detailed texture information can be crucial, and overly simplistic explanations may omit significant details. This is discussed in detail in Section 4.7.

In the **Faithfulness** category, the Monotonicity metric showed consistent results across methods, suggesting that this metric is more method-dependent than data-dependent. The SEL metric varied significantly between datasets: 86% for Caltech101, 65% for BEN and 52% for DeepGlobe, representing decreases of 21% and 34% respectively. This variation is partly due to lower prediction accuracies, but even if only correct predictions are considered, the average SEL for DeepGlobe increases by only 6%. Thus, for DeepGlobe, the remaining 28% discrepancy could be attributed to the characteristics of the dataset.

The discrepancy between the datasets becomes more pronounced when the SEL metric is contrasted with the IROF metric. Both metrics theoretically measure the decrease in prediction accuracy when the features considered most relevant are removed. However, SEL involves

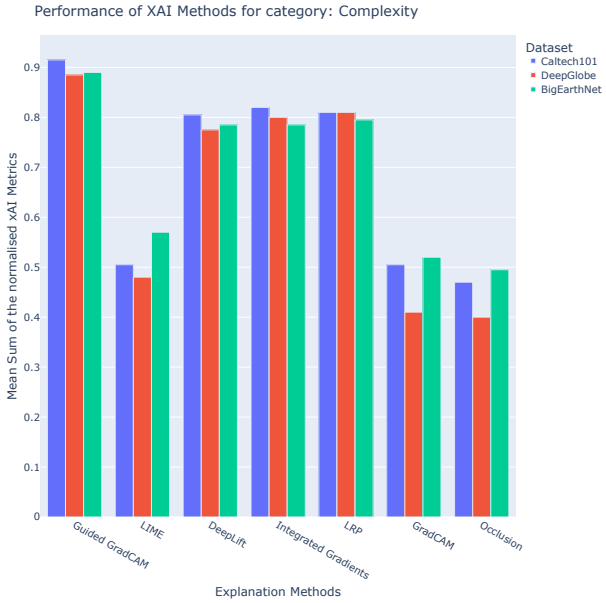


Figure 7.15: Performance of the explanation methods over all datasets in the metric category: Complexity. The x-axis enumerates different explanation methods. The y-axis measures the sum of normalized xAI metrics, scaled to 0 - 1. The color of the bars shows the dataset used.

the removal of individual pixels, whereas IROF involves the removal of entire segments. The average values for the IROF metric are as follows 0.58 for Caltech101, 0.60 for DeepGlobe and 0.72 for BEN. It is obvious that the RS MLC datasets, particularly DeepGlobe and BEN, perform better and show a similar difference, albeit more stable. This suggests that the IROF metric may be more appropriate for evaluating RS data.

Furthermore, for the RP metric, two removal strategies were evaluated: MoRF and LeRF. The purpose of this evaluation was to see if the choice of removal strategy affected the metric results, as described in Section 4.7. For RS images, the LeRF strategy results were consistently higher than for MoRF by approximately 10%, with the most notable differences observed in the non-BP-based metrics. However, this trend was also observed for the Caltech101 dataset, where the discrepancy was even higher. In particular, the non-BP-based methods averaged 0.61 for LeRF and only 0.21 for the MoRF strategy. It remains to be seen whether the MoRF strategy is viable at all. Future experiments could compare both strategies with the same evaluation as in Section 4.7. However, the results there suggest that LeRF may be the preferred removal strategy for RS data analysis.

The **Localisation** metrics show notable differences between the DeepGlobe and Caltech101 datasets. For the PG and TKI metrics, Caltech101 outperforms DeepGlobe by approximately 8% and 6% respectively. One factor contributing to this discrepancy is that Caltech101, as a SLC dataset, typically has a clearer distinction between the object of interest and the background. In contrast, MLC datasets such as DeepGlobe, which contain multiple objects, present a greater challenge in identifying the k most relevant pixels. This difficulty may be exacerbated by the nature of the PG metric, which tends to be trivial for images with large, dominant objects, making the metric less relevant for RS images, which typically contain a single class, and similarly for the Caltech101 dataset, where the presence of a single object against a background domi-

nates [110]. Furthermore, it is interesting to note that BP-based methods generally perform less effectively on these metrics, except Guided GradCAM for Caltech101. For example, the average scores for LRP, DeepLIFT and IG in the Caltech101 dataset are 0.76, whereas in DeepGlobe they drop to 0.65, reflecting a decrease of about 11% for the TKI metrics. Similarly, in the RMA metric, where BP-based methods typically excel for DeepGlobe, they show worse performance.

In the **Randomisation** category, the results are consistent across different datasets. As visualised in Figure 7.16 the BP-based (e.g. Guided GradCAM, IG, LRP) methods performed in this category for all datasets. As expected, this indicates that the nature of the data and the task does not substantially affect the outcomes in this context. In the **Robustness** category for the

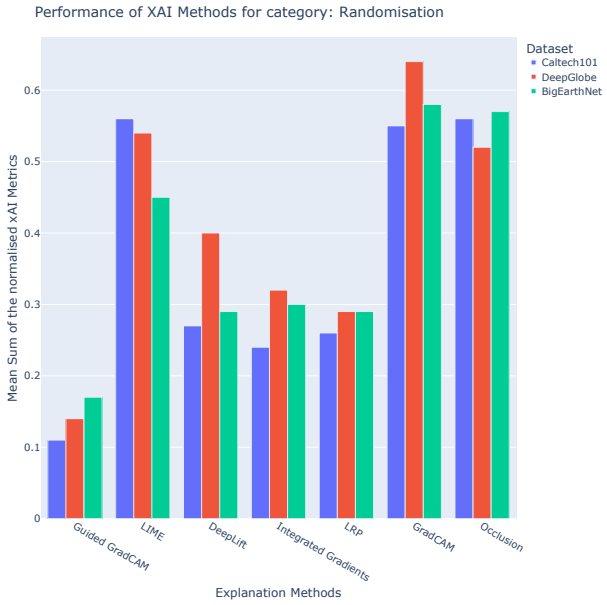


Figure 7.16: Performance of the explanation methods over all datasets in the metric category: Randomisation. The x-axis enumerates different explanation methods. The y-axis measures the sum of normalized xAI metrics, scaled to 0 - 1. The color of the bars shows the dataset used.

Caltech101 dataset, Occlusion and GradCAM significantly outperform other methods. However, for both RS datasets the performance spread is smaller, although the ranking remains the same: Occlusion and GradCAM perform best across all categories.

In summary, **Complexity** performance is higher for BP-based methods and lower for non-BP-based methods. However, lower performance in complexity for RS data might be beneficial for understanding, highlighting the need for new metrics. The same method-dependency is visible for **Randomisation** metrics. It has to be noted, that as there are two Complexity metrics and only a single Randomisation metric evaluated, the results of the summation might be biased towards BP-based methods. This is due to their theoretically founded strength in the former and weakness in the ladder category. **Faithfulness** is strongly related to data characteristics, with perturbation strategy playing an important role, as discussed in Section 3.7 and aligned with findings from [91]. **Robustness** shows less drastic changes, with Occlusion and GradCAM performing best. Significant changes are observed in **Localisation**, where metrics that rely on the top-k attributions (TKI, PG) perform worse on RS image data, but metrics that consider the size of the ground truth maps (AL, RRA, RMA) perform better and work well on RS image

data.

7.3 Evaluation of Explanation-Guided Training

This section evaluates the effectiveness of explanation-guided training methods in improving model performance. The focus is on two techniques: RRR loss and CutMix with xAI LP, analysing their impact on training results for different datasets and model architectures. Furthermore, the explanation metrics results were tested for correlations with xAI-guided training using the same explanation methods to determine if the explanation metrics could be used as predictors of xAI-guided training success.

Right for Right Reasons Loss

This subsection examines the application of the RRR loss, an explanation-guided training method that aims to improve model robustness by ensuring that the model's explanations align with human-annotated relevance maps. Furthermore, single and multi-label classification tasks are compared using on the Caltech101 and DeepGlobe datasets. As reference maps are a prerequisite for RRR, BEN was not considered. The analysis focuses on the test accuracy and mAP metrics, respectively, under various RRR parametrisations ($\lambda = 1$ and $\lambda = 10$) and explanation methods.

As shown in Table 7.2, LIME achieved the highest accuracy scores across all parametrisations (RRR MSE $\lambda = 1$: 98.08, $\lambda = 10$: 98.18; RRR $\lambda = 1$: 97.90, $\lambda = 10$: 98.06), outperforming the baseline and all other methods for the **Caltech101** dataset. IG and Guided GradCAM also showed strong performance. GradCAM's performance was notably variable, with a significant accuracy reduction under from RRR MSE with $\lambda = 10$: 96.85 to 21.54 forRRR with $\lambda = 10$.

Table 7.2: Test accuracy of various interpretability methods combined with RRR loss and the RRR MSE variant on the Caltech101 dataset across different parametrisations ($\lambda = 1$ and $\lambda = 10$).

Method	**RRR MSE**		**RRR**	
	$\lambda = 1$	$\lambda = 10$	$\lambda = 1$	$\lambda = 10$
Baseline	0.9739	0.9739	0.9739	0.9739
DeepLIFT	97.18	97.13	97.24	97.01
GradCAM	89.50	96.85	97.19	21.54
Guided GradCAM	96.62	96.80	97.21	97.26
Integrated Gradients	97.35	97.62	97.62	97.72
LIME	**98.08**	**98.18**	**97.90**	**98.06**
Occlusion	97.58	97.39	97.01	96.81

The Figures 7.17 and 7.18 show the explanations for the same example as in Figure 7.2. However, the difference with the previous visualisation is that the models used to generate the predictions and explanations were trained using RRR-guided approaches. In Figure 7.17, improvements can be seen in all the explanation methods, with improved localisation to the *leopard* class, rather than the spurious feature of the black background. In particular, the GradCAM

explanation is clearer and more focused on the object. The prediction remains unchanged. However, the explanation generated by the model trained using RRR with DeepLIFT in Figure

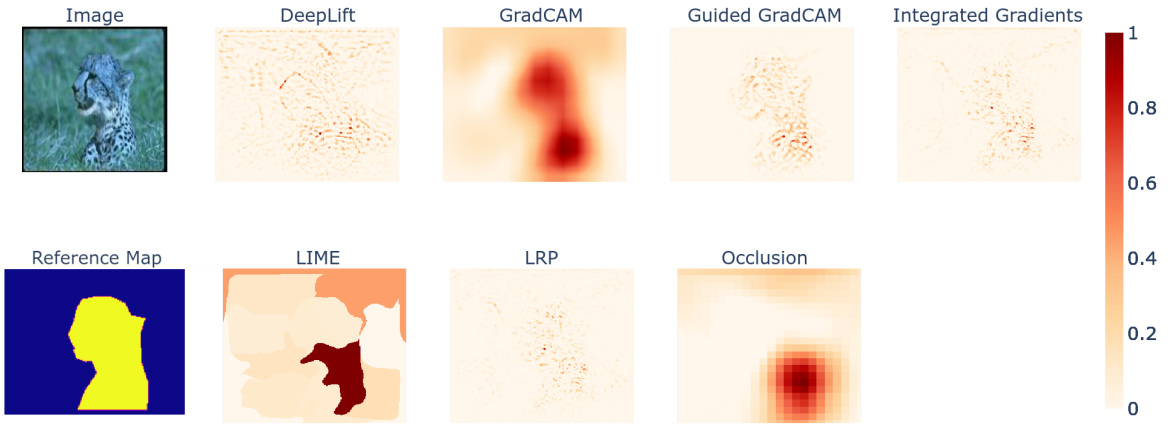


Figure 7.17: Caltech101: Single-label explanations for the correct prediction of the class *leopard* with a Spurious Correlation. The model was trained with RRR GradCAM using RRR MSE and λ = 10. The plots in order are: image, explanations: DeepLIFT, GradCAM, Guided GradCAM, IG (1st row), reference map, explanations: LIME, LRP, Occlusion (2nd row). Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

7.18 does not improve significantly. All explanation methods show that the model still uses the spurious feature for the prediction, even the explanation generated by DeepLIFT. This suggests that some methods, such as GradCAM, are more effective than others, like DeepLIFT, in reducing reliance on spuriousness through RRR-guided training approaches.

To evaluate the increase of reasoning quantitatively, for the RRR-trained model (GradCAM **MSE** λ **1**) the Localisation metrics were calculated again. The results are visualized in Figure 7.19, which has the same structure as the previous matrices. Given that the Caltech101 dataset includes SC, e.g. for the class leopard, it is expected that performance on localisation metrics will improve as reliance on SC is reduced. It is visible that there is an improvement for all Localisation metrics compared to Figure 7.9. The average increase over all Localisation metrics is about 7%, proving that RRR-guided training can improve the reasoning. However, it would be interesting to repeat the full quantitative evaluation.

The test mAP of the analysed explanation methods on the **DeepGlobe** dataset was evaluated across the same parametrisations ($\lambda = 1$ and $\lambda = 10$) using RRR loss and its RRR MSE extension metrics, as shown in Table 7.3. Occlusion achieved the highest scores for both RRR MSE $\lambda = 1$ (84.35) and RRR $\lambda = 10$ (84.38). DeepLIFT performed best for RRR MSE $\lambda = 10$ with a score of 83.98. LIME excelled in RRR $\lambda = 1$ with the highest score of 84.11. GradCAM exhibited poor performance, particularly under RRR $\lambda = 10$, where its score dropped significantly to 44.1. Overall, Occlusion and LIME demonstrated the most robust performance, particularly Occlusion, which topped two of the four metrics.

As described before, the absence of known SC in the DeepGlobe dataset precludes quantitative measurement of improved reasoning. Future research should focus on testing with a RS

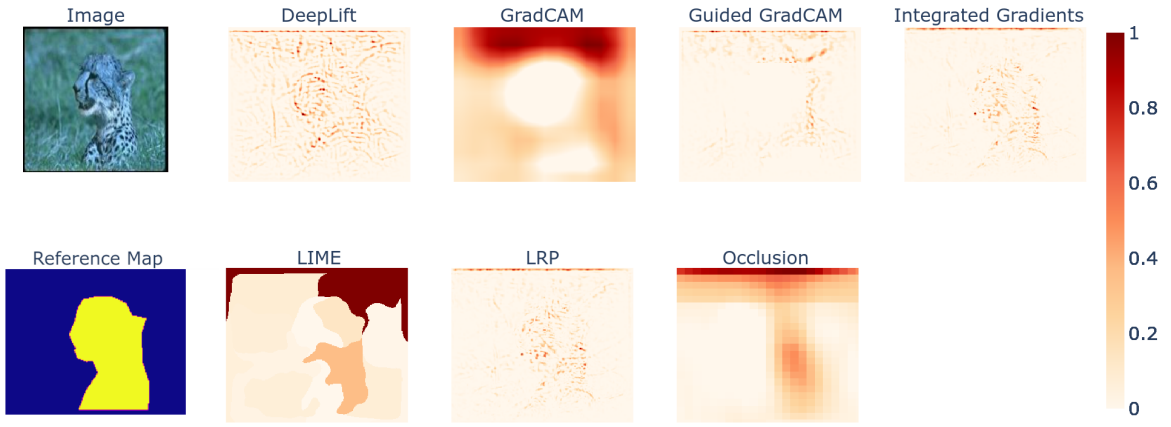


Figure 7.18: Caltech101: Single-label explanations for the correct prediction of the class *leopard* with a Spurious Correlation. The model was trained with RRR DeepLIFT using RRR MSE and λ = 10. The plots in order are: image, explanations: DeepLIFT, GradCAM, Guided GradCAM, IG (1st row), reference map, explanations: LIME, LRP, Occlusion (2nd row). Only positive contribution is visualized on a scale from 0 to 1, where higher means more contribution towards the class prediction.

dataset that includes known SC or integrating such SC into an existing RS dataset.

CutMix with Label Propagation for Multi-Label Classification

This subsection explores the CutMix with xAI LP method, another xAI-guided training technique designed to enhance model performance by leveraging mixed-sample data augmentation and propagating labels to improve generalisation. Its effectiveness is examined specifically for multi-label classification tasks, between VGG and ResNET models for DeepGlobe and BEN. Caltech101 was not considered, as it is a SLC dataset.

The Table 7.4 presents the test mAP results of various explanation methods applied to two different models, ResNET and VGG, trained on the DeepGlobe dataset. The explanation methods were evaluated using two different CutMix box sizes, $0.1 - 0.5$ and $0.3 - 0.7$. For each model and method, the table lists the test mAP achieved under the two CutMix box size conditions. The highest mAP values for each model and box size are highlighted in bold to emphasize the best-performing configurations. DeepLIFT and LIME achieved the highest performances for ResNET. Conversely, GradCAM showed the highest performance for the larger box size in VGG, suggesting that certain explanation methods may be more suitable for specific model architectures and parameter settings. The upper bound was the usage of CutMix with xAI LP using the reference maps instead of explanations, as described in Section 6.2.

The Table 7.5 presents the xAI-CutMix results for the BEN dataset. The structure of the Table is similar to the one above. GradCAM achieved the highest performances for ResNET with the smaller box size, while LIME showed the highest performance for the larger box size in ResNET. Conversely, IG showed the highest performance for both box sizes in VGG. For BEN no upper bound was provided, as reference maps are necessary to calculate it, and the reference

Table 7.3: Test mAP of various explanation methods combined with Right for the Right Reason (RRR) loss and the RRR MSE variant on the DeepGlobe dataset across different parametrisations ($\lambda = 1$ and $\lambda = 10$).

Method	**RRR MSE**		**RRR**	
	$\lambda = 1$	$\lambda = 10$	$\lambda = 1$	$\lambda = 10$
Baseline	83.81	83.81	83.81	83.81
DeepLIFT	83.91	**83.98**	83.85	83.88
GradCAM	84.33	70.29	51.30	44.10
Guided GradCAM	83.73	83.83	83.59	83.88
Integrated Gradients	83.87	83.89	83.86	83.81
LIME	83.78	83.21	**84.11**	83.49
Occlusion	**84.35**	83.85	83.71	**84.38**

Table 7.4: DeepGlobe: Training results for CutMix with xAI LP guided training, Test mAP per method for two different box sizes for VGG and ResNET. The upper bound was the usage of CutMix with xAI LP using the reference maps instead of explanations.

Model	Explanation Method	CutMix Boxsize	
		0.1-0.5	0.3-0.7
ResNET	Baseline	83.84	83.84
	Upper bound	86.28	86.28
	DeepLIFT	85.84	**86.08**
	GradCAM	85.91	85.92
	Guided GradCAM	85.89	86.05
	IG	85.38	84.97
	LIME	**85.99**	86.07
	LRP	85.94	85.98
	Occlusion	85.81	85.87
VGG	Baseline	83.81	83.81
	Upper bound	85.28	85.12
	DeepLIFT	84.29	84.46
	GradCAM	84.13	**84.58**
	Guided GradCAM	84.35	84.05
	IG	84.57	84.01
	LIME	**84.58**	84.01
	LRP	84.35	84.33
	Occlusion	83.99	84.18

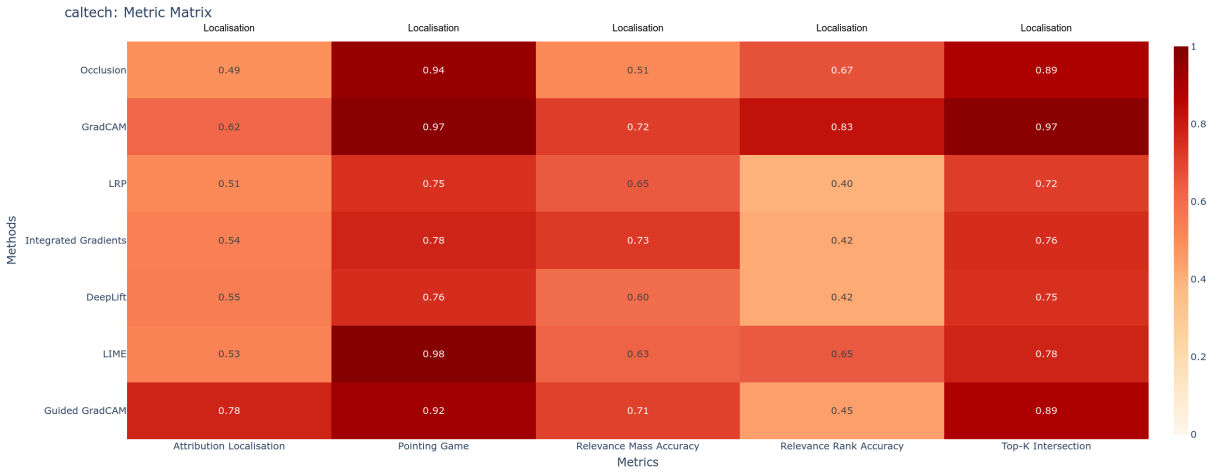


Figure 7.19: Caltech101: Localisation metrics after RRR training using GradCAM. The rows of the matrix represent different explanation methods, while the columns are categorized by metrics types at the top and specific metric names at the bottom

maps from BEN were not used.

Comparing the two Tables, we observe that for the DeepGlobe dataset, the highest performing methods for ResNET were LIME and DeepLIFT. LIME achieved a mAP of 85.99 with the smaller box size, representing an approximate 2% increase from the baseline of 83.84. DeepLIFT achieved a mAP of 86.08 with the larger box size, also an approximate 2% increase from the baseline. For VGG, LIME achieved a mAP of 84.58 for the smaller box size, which is approximately a 0.75% increase from the baseline of 83.81. Similarly, GradCAM achieved a mAP of 84.58 for the larger box size, also an approximate 0.75% increase from the baseline.

In contrast, for the BEN dataset, GradCAM achieved the highest performance for ResNET with the smaller box size, showing an improvement to 63.14 from the baseline of 61.68, approximately a 1.4% increase. LIME showed the highest performance for the larger box size in ResNET, with an improvement to 63.05 from the baseline, approximately a 1.2% increase. For VGG, IG showed the highest performance for both box sizes, with the mAP improving to 61.08 for the smaller box size and 61.12 for the larger box size, compared to the baseline of 60.47. This represents approximately a 1% increase for the smaller box size and a 1.1% increase for the larger box size. The results for VGG on the BEN dataset were all quite similar, indicating a consistent performance across different explanation methods.

The difference might arise because VGG, being a much larger model, can generalise more easily, but also has the potential to overfit due to its large number of weights. On the other hand, ResNET, being smaller, relies more on effective generalisation techniques to perform well. Therefore, ResNET might benefit more from techniques such as CutMix that help the model learn to generalise better. As a result, ResNET's performance improvements are about twice those of VGG, demonstrating the effectiveness of CutMix in improving ResNET's generalisation.

Table 7.5: BEN: Training results for CutMix with xAI LP guided training, Test mAP per method for two different box sizes for VGG and ResNET.

Model	Explanation Method	CutMix Boxsize	
		0.1-0.5	0.3-0.7
ResNET	Baseline	61.68	61.68
	DeepLIFT	62.89	62.84
	GradCAM	**63.14**	62.92
	Guided GradCAM	62.19	62.11
	IG	62.83	62.98
	LIME	63.08	**63.05**
	LRP	62.71	62.89
	Occlusion	63.12	62.77
VGG	Baseline	60.47	60.47
	DeepLIFT	60.12	60.48
	GradCAM	60.58	60.81
	Guided GradCAM	60.70	60.28
	IG	**61.08**	**61.12**
	LIME	60.77	61.10
	LRP	60.52	60.92
	Occlusion	60.88	60.35

Correlation Analysis between xAI Metrics and xAI-guided Training Success

xAI-guidance with RRR

In this subsection, the results of the xAI-guided training of a VGG model with the RRR loss are linked to the results of the metric analysis. The question examined here is if any of the metrics or metric categories correlate with increased training outcome, i.e. if for an explanation method, the success in any set of metrics can be used as a predictor to approximate better model training using this corresponding explanation method in combination with RRR guided training.

Figure 7.20 illustrates the correlation coefficients between selected xAI metrics and the effectiveness of RRR training on the **Caltech101** dataset. The calculations were performed using the optimal parameterisation for each explanation method within the xAI guided training framework, described in more detail in Section 6.2. Each subplot in Figure 7.20 represents a different metric category, as detailed in Section 4, and plots the Pearson correlation coefficients on the y-axis against the individual metrics on the x-axis. Full scatterplots of these correlations can be found in the Appendix 5.

The correlation analysis illustrates the relationships between training accuracy and various xAI metrics across different categories. In the first category, Complexity, visualised in subplot 1, a negative correlation is observed for both metrics. Sparseness and ECO have correlation coefficients of -0.31 and -0.39 respectively. This suggests that explanation methods that produce

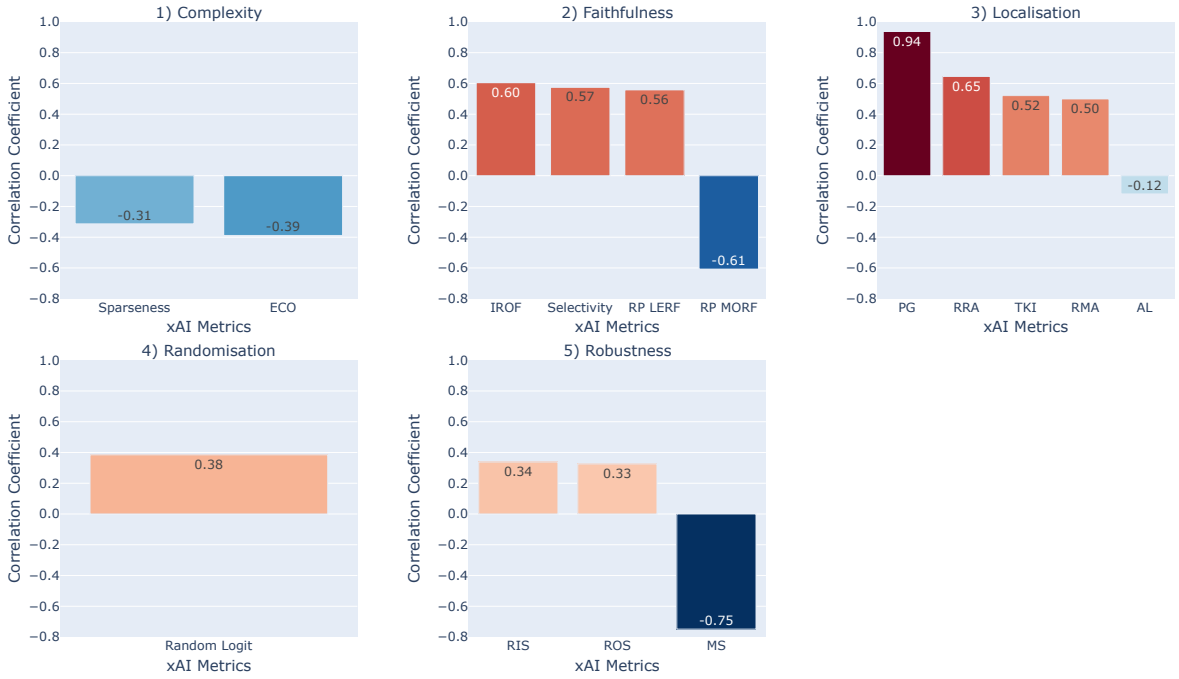


Figure 7.20: Caltech101: Correlation analysis between xAI metrics and test accuracy with RRR-training. The y-axis of each subplot represents xAI metric, while the x-axis represents the correlation coefficient. Each subplot visualises a category: 1) for Complexity, 2) for Faithfulness, 3) for Localisation, 4) for Randomisation and 5) for Robustness.

less complex metrics tend to undermine the effectiveness of RRR training for the Caltech101 dataset. Subplot 2 shows the Faithfulness metrics, which have varying correlations. Metrics such as IROF, Selectivity, and RP using the LeRF removal strategy show correlations ranging from 0.36 to 0.6. Conversely, RP MoRF shows a negative correlation with a coefficient of -0.37. Metrics in the Localisation category, in particular the PG metric, show the strongest positive correlation of all metrics at 0.94. Other Localisation metrics also show robust positive correlations, with RRA at 0.65 and RMA at 0.5. However, AL shows a slight negative correlation of -0.12. The only metric considered under Randomisation, the RL test, shows a moderate positive correlation of 0.38. For Robustness metrics, both relevance metrics, RIS and ROS, show moderate positive correlations. In contrast, the MS metric shows a significant negative correlation of -0.75, highlighting an inverse relationship with training success.

Figure 7.21 shows the correlation analysis for the **DeepGlobe** dataset, mirroring the structure of the previous graph. Detailed correlations and scatterplots related to this analysis are available in the Appendix 5. The correlation analysis for the DeepGlobe dataset shows similar trends as observed for the Caltech101 dataset, with even stronger correlations noted for certain metrics. Specifically, in subplot 1, both Complexity metrics show strong negative correlations: Sparseness shows a correlation of -0.94, while ECO correlates with -0.92. In subplot 2, representing the Faithfulness metrics, RP MoRF and SEL show negative correlations of -0.85 and -0.68 respectively. IROF and RP LeRF show moderate to strong positive correlations with values of 0.32 and 0.97 respectively. The Localisation metrics presented in subplot 3 show multiple

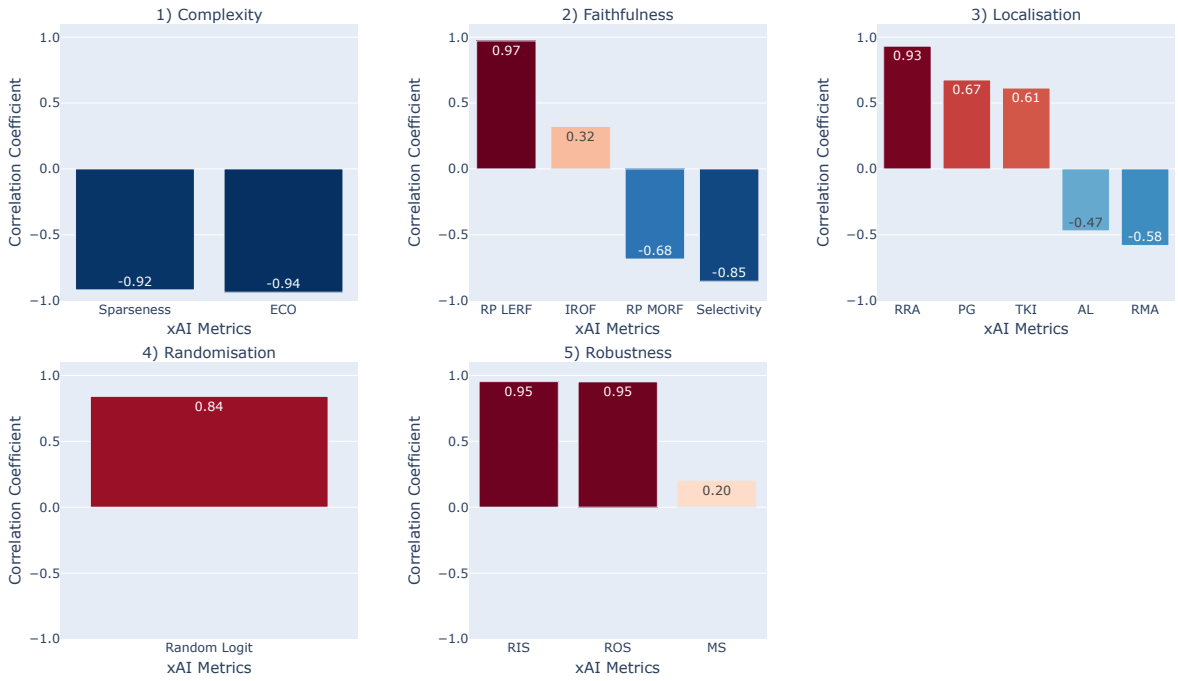


Figure 7.21: DeepGlobe: Correlation analysis between xAI metrics and test mAP with RRR-training. The y-axis of each subplot represents xAI metric, while the x-axis represents the correlation coefficient. Each subplot visualises a category: 1) for Complexity, 2) for Faithfulness, 3) for Localisation, 4) for Randomisation and 5) for Robustness. The coefficients range from -1 to 1, denoting a scale from perfect negative to perfect positive correlation, respectively

strong positive correlations. RRA is correlated with 0.93, PG with 0.67 and TKI with 0.61. However, there are also two strong negative correlations present: AL has a negative correlation of -0.47 and RMA is also negatively correlated at -0.58. The RL metric from the Randomisation category, shown in subplot 4, shows an even higher positive correlation for DeepGlobe at 0.95. In the Robustness metrics, both RIS and ROS continue to show high positive correlations, each at 0.95. In contrast, MS shows a moderate positive correlation of 0.2, reversing the previous negative trend observed in the Caltech101 dataset.

Figure 7.22 summarises the correlations for each metric categorised into different groups and presents the Fisher-Z means of the correlation coefficients, as described in Section 6.2. The x-axis categorises the metrics, while the y-axis represents the Fisher-Z means. Subplot 1, on the left, visualises the results for Caltech101 and subplot 2 visualises the results for DeepGlobe. The data within each subplot is ordered by the mean correlation values.

Localisation in Caltech101 has the highest positive mean correlation of 0.6, while in DeepGlobe it has a more moderate correlation of 0.38. This difference is logical since explanation methods that excel in Localisation metrics accurately localise the OoI. The RRR loss penalises explanations that assign high relevance to irrelevant pixels and rewards those that localise the OoI. This may contribute to a more stable training process. The Randomisation category shows a positive mean correlation of 0.38 for Caltech101, which is significantly lower than the strong positive mean of 0.84 observed for DeepGlobe. For the Faithfulness category, Caltech101 has

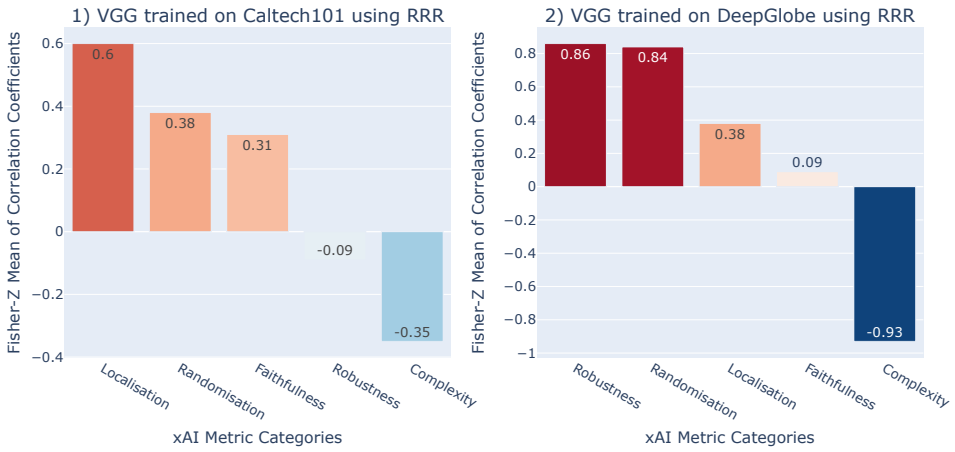


Figure 7.22: Correlations plotted between xAI metric categories and RRR-guided training success for VGG models. Each barplot shows the Fisher-Z Mean of the correlation coefficients (y-axis) for xAI metric categories (x-axis). 1): VGG trained on Caltech101; 2): VGG trained on DeepGlobe, both with RRR training.

a mean correlation of 0.32, indicating a moderate effect, while DeepGlobe has only a small positive mean of 0.09. In the Robustness category, Caltech101 has a slightly negative mean correlation of -0.09. In stark contrast, DeepGlobe has a positive mean correlation of 0.86 for the Robustness category. This notable disparity is due to the variation in the performance of the Selectivity metric, which shifts drastically from a correlation of 0.57 in Caltech101 to -0.87 in DeepGlobe. The performance of the selectivity metric drops significantly, with Caltech101 scoring 86%, which drops to 52% in DeepGlobe. Finally, the Complexity category shows a negative mean correlation of -0.35 for Caltech101, which contrasts sharply with a much stronger negative mean of -0.93 for DeepGlobe, indicating a pronounced negative trend for the latter.

xAI-guidance with CutMix with xAI LP

The following section examines the correlations between xAI-guided training metrics and the effectiveness of CutMix with xAI LP training enhancements for MLC tasks. As in the previous analysis, the aim is to determine whether any xAI metrics or their respective categories are associated with improved performance in xAI-guided training. These correlations are calculated as described in Section 6.2. In contrast to the previous section, which compared SLC and MLC tasks using RRR, this analysis deals exclusively with MLC tasks, reflecting the specificity of the CutMix with xAI LP and LP strategy applicable only to MLC tasks. In addition, this research compares two different model architectures: VGG and ResNET models.

Figure 7.23 illustrates the correlation coefficients for the DeepGlobe dataset, detailing the relationships between selected xAI metrics and the efficiency of CutMix with xAI LP training for VGG models. Correlations were calculated using the mean mAP as the test metric. The structure of this plot is consistent with previous Figures, and additional details along with scatter plots of these correlations are available in the Appendix 5.

The first metric category to be evaluated is Complexity, where both measures show moderate

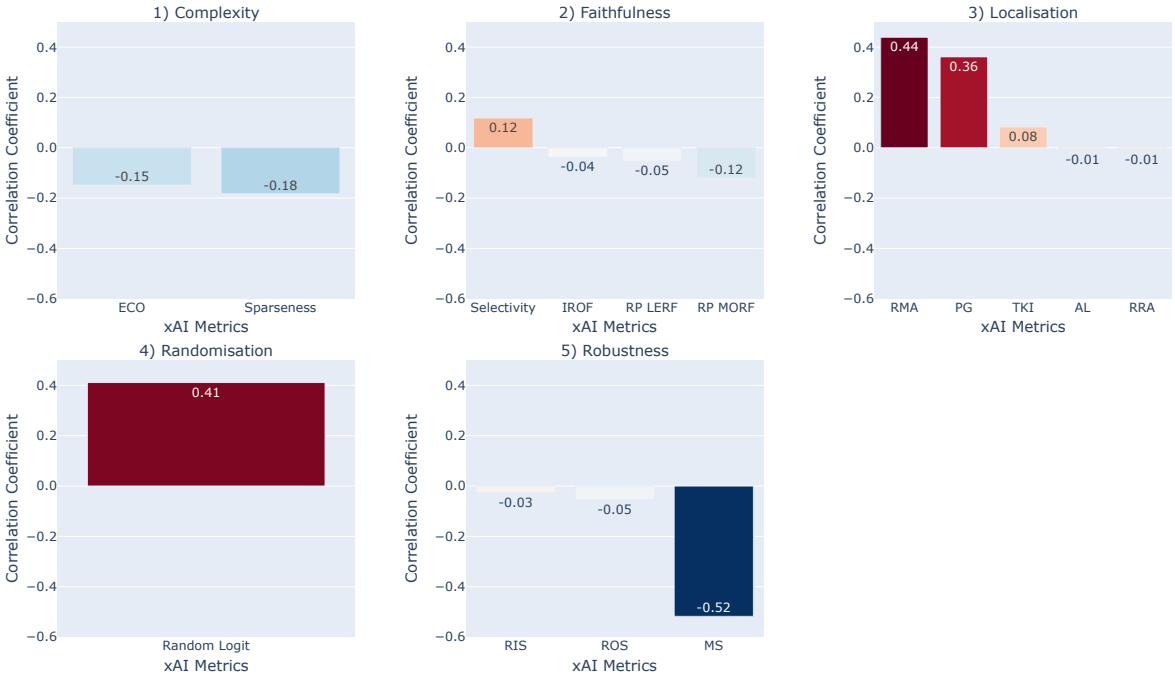


Figure 7.23: DeepGlobe: Correlation analysis between xAI metrics and training improvement using CutMix with xAI LP for VGG models. The y-axis of each subplot represents selected xAI metrics, while the x-axis shows the correlation coefficient. Each subplot visualises a category: 1) for Complexity, 2) for Faithfulness, 3) for Localisation, 4) for Randomisation and 5) for Robustness.

negative correlations with CutMix with xAI LP training success, extending previous trends observed with RRR-guided training. Specifically, Sparseness and ECO yield correlation values of -0.15 and -0.18 respectively. In the Faithfulness category, the correlations appear to be less substantial. Only SEL shows a slightly positive correlation of 0.12. Both IROF and RP with LeRF show coefficients around 0, indicating no correlation, while RP with MoRF shows a slight negative correlation of -0.12, suggesting minimal contributions from this category. For the Localisation metrics, RMA shows a strong positive correlation of 0.44 and PG shows a more moderate correlation of less than 0.4. TKI, AL and RRA register values around 0, indicating no significant correlation. Within the Robustness category, both relevance metrics show almost no correlation. However, the MS metric shows a significant negative correlation of -0.52.

Similar to prior figures, Figure 7.24 shows the correlation coefficients for the DeepGlobe dataset, but concerning the CutMix with xAI LP training success for the ResNET models. The correlations were analysed using the same methodological framework as before, with further details and scatter plots available in the Appendix 5. The structural layout of the plot remains consistent with the previous Figures. For the Complexity metrics in subplot 1, there is a slight negative correlation for ECO at -0.16, and virtually no correlation for Sparseness at -0.05. Both Complexity metrics show a smaller correlation than in the previous results. In the Faithfulness category, shown in subplot 2, there are significant changes. RP for the MoRF strategy now yields a strong positive correlation with a value of 0.48, while the LeRF strategy does not correlate with a result of 0.02. In contrast, to the previous results SEL, IROF show strong nega-

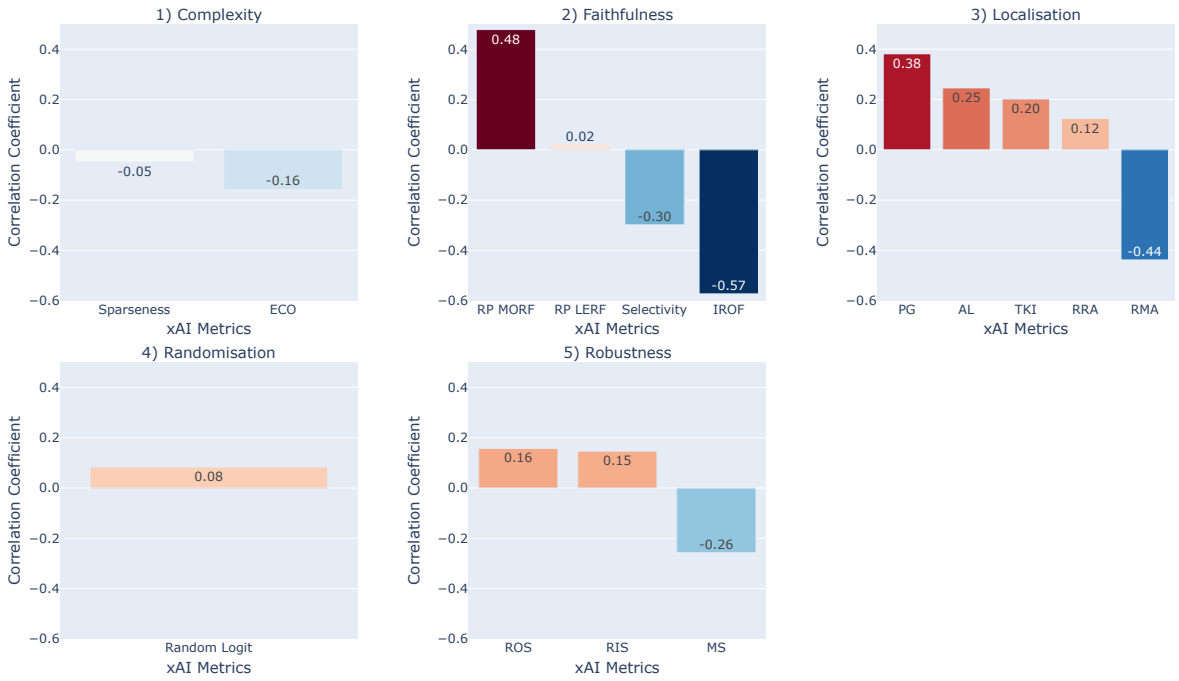


Figure 7.24: DeepGlobe: Correlation analysis between xAI metrics and training improvement using CutMix with xAI LP for ResNET models. The y-axis of each subplot represents selected xAI metrics, while the x-axis shows the correlation coefficient. Each subplot visualises a category: 1) for Complexity, 2) for Faithfulness, 3) for Localisation, 4) for Randomisation and 5) for Robustness.

tive correlations with -0.3 and -0.57, respectively. The Localisation metrics in subplot 3 also show drastic shifts. Previously, RMA had a positive correlation of 0.44 for the VGG model, but now it shows a strong negative correlation of -0.44. Conversely, the PG metric continues to show a strong positive correlation of 0.38. Other metrics such as TKI, AL and RRA show weak positive correlations of 0.25, 0.2 and 0.12 respectively. In subplot 4, the Randomisation metric RL now correlates 0.08, indicating a weaker correlation, similar to the changes seen in the Complexity metrics. Finally, subplot 5 for the Robustness category shows that the ResNET models show an increase in positive correlations for the relevancy metrics: RIS to 0.16 and ROS to 0.15. The negative correlation for the MS metric is now more moderate at -0.26, indicating increased positive correlations and decreased negative correlations compared to the VGG models.

Figure 7.25 summarises the correlations per metric, categorised into groups, and shows the Fisher Z means of the correlation coefficients, as described in Section 6.2. This plot mirrors the structure seen in Figure 7.22, where Caltech101 and DeepGlobe were compared, but the focus now shifts to examining differences between two architectures, with ResNET models in subplot 1 and VGG models in subplot 2. The x-axis categorises the metrics, while the y-axis represents the Fisher Z means. Within each subplot, the values are organised by the mean correlation values. For the ResNET results (subplot 1), the correlations appear to be rather moderate, mainly due to the effect of strong opposing correlations of different metrics neutralising each other. Within the Complexity and Randomisation categories, only small correlations are observed. In the Faithfulness category, a large positive correlation of 0.48 for RP with MoRF is offset by a

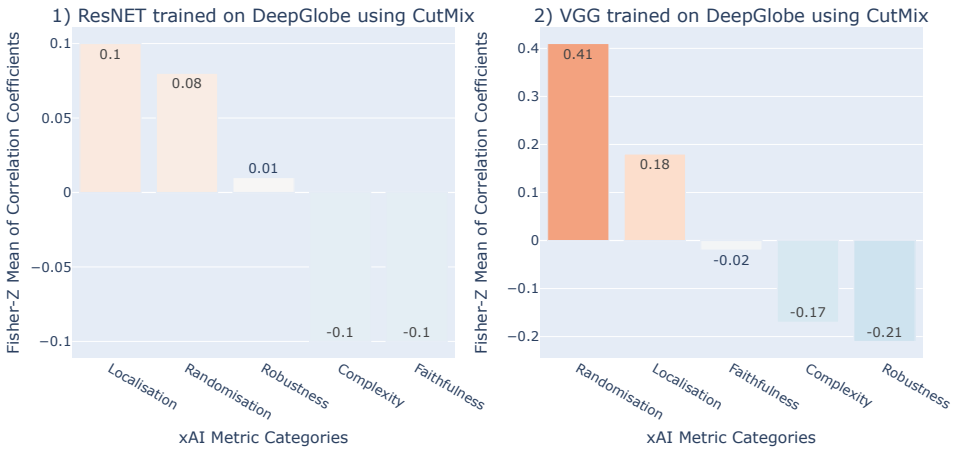


Figure 7.25: DeepGlobe: Correlations plotted between xAI metric categories and CutMix with xAI LP-guided training success for different models. Each barplot shows the Fisher-Z Mean of the correlation coefficients (y-axis) for xAI metric categories (x-axis). 1): VGG; 2): ResNET.

strong negative correlation for IROF at -0.57 and SEL at -0.3. Similarly, in the Localisation category, the positive correlation of the PG metric at 0.38 is negated by the negative correlation of the RMA metric at -0.44. For Robustness, while both relevancy metrics show correlations around 0.15, they are counteracted by the MS metric at -0.26. These findings suggest that summarising results at the category level may be misleading and that a deeper analysis of individual metrics is essential due to their high variability. Furthermore, this complexity makes it difficult to draw definitive conclusions about the relationship between xAI-guided training outcomes and the metrics.

For the VGG model, a stronger trend is observed per category. Randomisation has the highest correlation with 0.41, followed by Localisation with 0.18. It should be noted that the Randomisation category only contains a single metric, which limits the generalisability of these results. Faithfulness shows almost no correlation, as no single metric within this category is particularly influential. Complexity metrics again show a moderate negative correlation, and Robustness metrics, influenced by the strong negative correlation of the MS metric, also show a moderate negative trend.

Three overarching trends emerge from the analysis of the metrics for the DeepGlobe dataset forCutMix with xAI LP. The Localisation category consistently shows a moderate positive correlation, probably due to the nature of both xAI-guidance methods. Since CutMix with xAI LP uses a LP strategy that propagates labels to newly augmented images, it naturally benefits from explanations that accurately localise the object or class of interest. The second trend, observed in the Randomisation category, shows a moderate positive correlation, while the third trend, observed in the Complexity metrics, shows a moderate negative correlation.

Similar to previous figures, Figure 7.26 shows the correlation coefficients for the **BEN** dataset using VGG models. The correlations, calculated using the mean mAP as the test metric, follow the same structural layout as the previous analyses. Details and scatterplots are available in the appendix 5. Note that the Localisation category is excluded from this analysis due to the lack of quantitative metrics for the BEN dataset, as explained in Section 6.1.

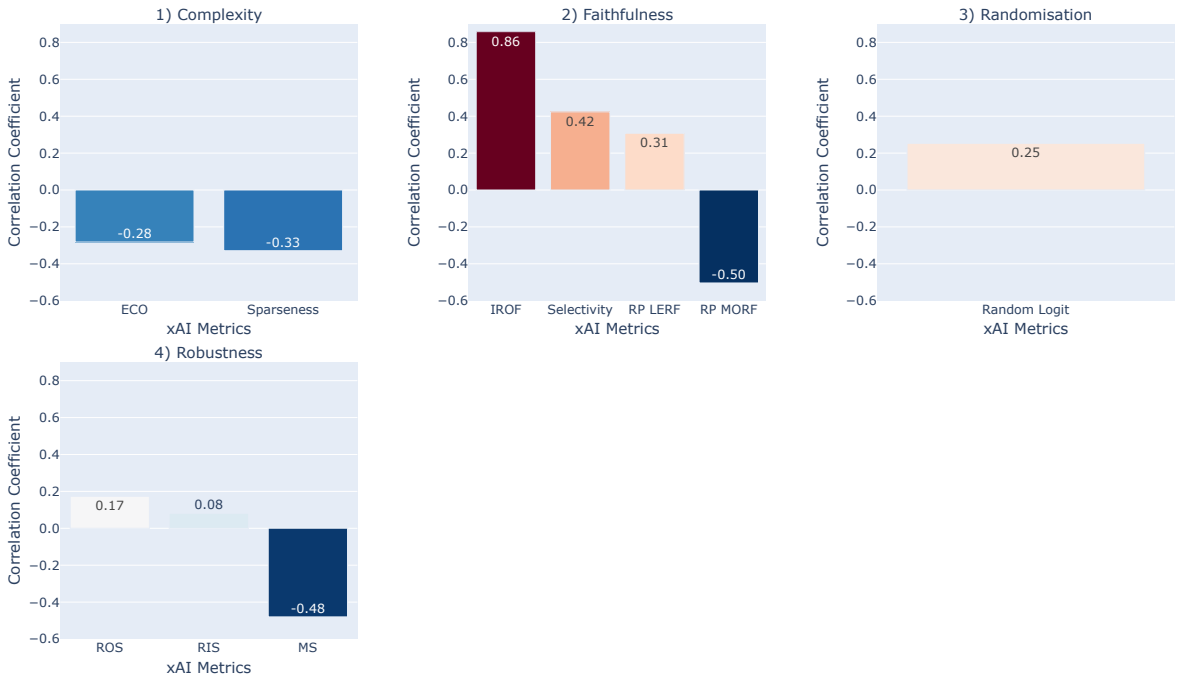


Figure 7.26: BEN: Correlation analysis between xAI metrics and training improvement using CutMix with xAI LP for VGG models. The y-axis of each subplot represents selected xAI metric, while the x-axis shows the correlation coefficient. Each subplot visualises a category: 1) for Complexity, 2) for Faithfulness, 3) for Localisation, 4) for Randomisation and 5) for Robustness.

Similar to the previous plot, both Complexity metrics in subplot 1 show a moderate negative correlation: Sparseness has a negative correlation of -0.33, while ECO is negatively correlated with -0.28. In the Faithfulness category, IROF has a strong positive correlation of 0.86, SEL has a correlation of 0.42, RP with LeRF has a positive correlation of 0.31, and RP MoRF has a negative correlation of -0.50. The Randomisation category in plot 4 shows a positive correlation for the RL metric at 0.25. In the Robustness category, shown in plot 5, the correlations for the ResNET models are similar to the VGG models shown previously: ROS at 0.17, RIS at 0.08 and a strong negative correlation for the MS metric at -0.48.

Figure 7.27 shows the correlation coefficients for the BEN dataset for ResNET models. The correlations, calculated using the mean mAP as the test metric, follow the same structural layout as the previous plots. Although the trends are similar, the Complexity category shows that ECO is negatively correlated with -0.61 and Sparseness with -0.89. In the Faithfulness category, some metrics show particularly strong correlations: SEL with 0.76, IROF with 0.71 and RP LeRF at 0.59. However, RP MoRF has a negative correlation of -0.78. The Randomisation metric RL has a strong correlation of 0.8. In the Robustness category, the relevance metrics ROS and RIS show positive correlations of 0.27 and 0.19 respectively, while the metric MS shows a moderate negative correlation of -0.19.

Figure 7.28 summarises the correlations per metric, categorised into groups, and shows the Fisher Z means of the correlation coefficients, as described in Section 6.2. The plot is structured similar to Figure 7.25. The figure shows a consistent trend across both model architectures for

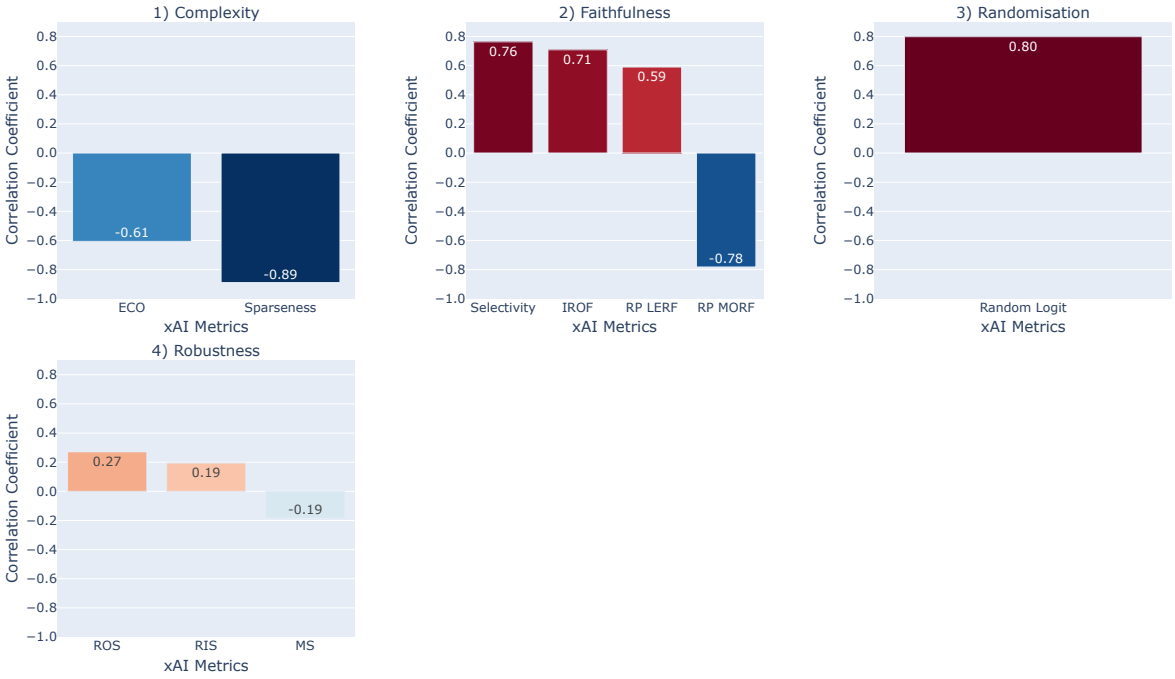


Figure 7.27: BEN: Correlation analysis between xAI metrics and training improvement using CutMix with xAI LP for ResNET models. The y-axis of each subplot represents selected xAI metric, while the x-axis shows the correlation coefficient. Each subplot visualises a category: 1) for Complexity, 2) for Faithfulness, 3) for Localisation, 4) for Randomisation and 5) for Robustness.

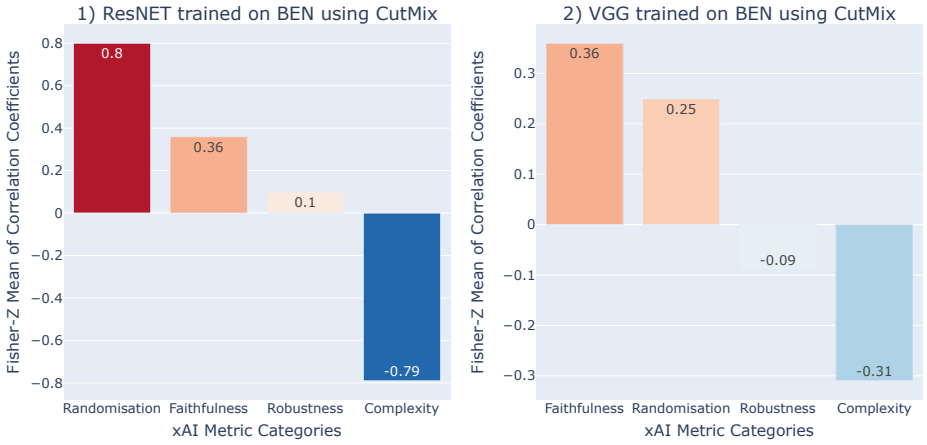


Figure 7.28: BEN: Correlations plotted between xAI metric categories and CutMix with xAI LP-guided training success for different models. Each barplot shows the Fisher-Z Mean of the correlation coefficients (y-axis) for xAI metric categories (x-axis). 1): VGG ; 2): ResNET.

the BEN dataset. In the Randomisation category, a consistent positive correlation is observed, indicating a strong and reliable influence of the RL metric on training success. Similarly, the Complexity category consistently shows a negative correlation, suggesting that less complex

explanations tend to be associated with better training results. In contrast, the Faithfulness category shows variability depending on specific metrics. RP MoRF consistently shows a negative correlation across both datasets and models, highlighting its limited effectiveness in improving training outcomes. Conversely, other metrics within the Faithfulness category, such as IROF, SEL and RP LeRF, show positive correlations.

Finally, Figure 7.29 provides an overview of the mean correlations across all datasets, models, and xAI guidance methods. This bar chart shows the Fisher Z-mean of the correlation coefficients for each xAI metric category, illustrating the general trends in how these metrics influence model performance improvement. The categories on the x-axis are ordered by their correlation strength, from left to right: Randomisation, Localisation, Faithfulness, Robustness, and Complexity. The Figure summarises the overall influence of each metric category, allowing a direct comparison of their impact on training effectiveness. However, it must be noted that categories can sometimes be misleading, as observed in the Faithfulness category where some metrics exhibit strong positive correlations while others show negative ones, effectively neutralizing each other.

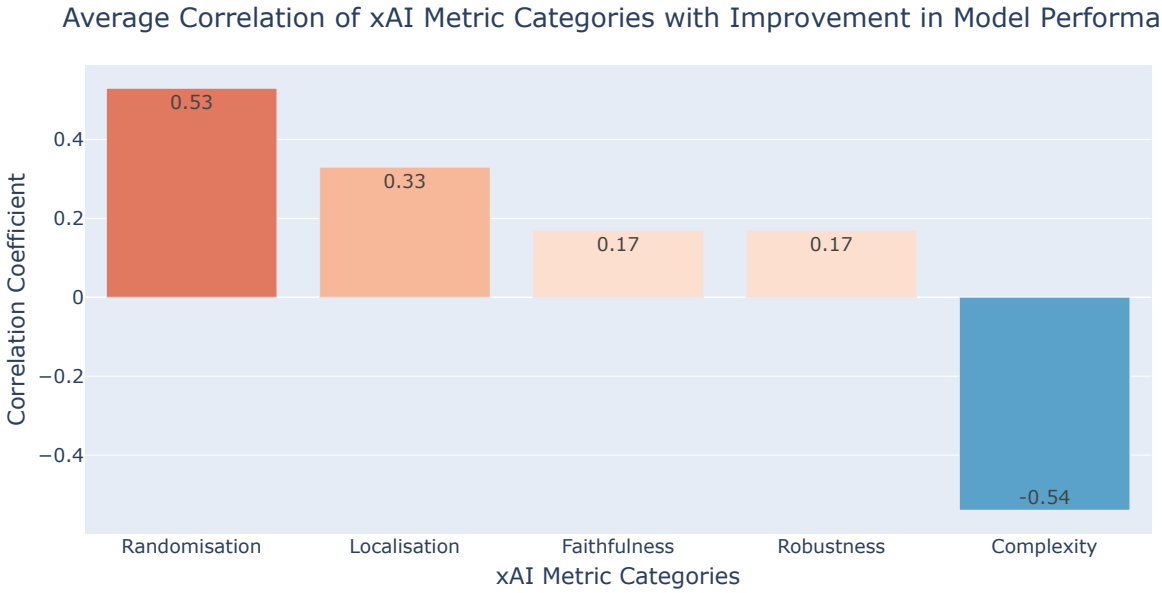


Figure 7.29: Mean correlation over all datasets, models and both xAI-guidance methods. The y-axis shows the Fisher-Z Mean of the correlation coefficients for each xAI metric category on the x-axis. The categories are ordered by their value; From left to right: Randomisation, Localisation, Faithfulness, Robustness, Complexity.

Notably, the Randomisation category shows a significant positive correlation at 0.53, suggesting a robust relationship with model performance improvement. In contrast, the Complexity category has the highest negative mean correlation at -0.54, indicating a strong negative impact on xAI-guided training. This pattern is consistent across all combinations of xAI guidance methods, datasets and models. As hypothesised in the theoretical analysis (see Section 3.7) and supported by the quantitative results, BP-based methods, such as LRP and IG, tend to underperform in the Randomisation category but overperform in the Complexity category. Gradient or perturbation-based methods generally show good performance in the Randomi-

sation category, while yielding more complex explanations, leading to worse performance in the Complexity category. As both categories show a strong correlation, this supports the hypothesis that xAI-guided training is more effective with non-BP-based methods. However, the singular inclusion of the RL test in the Randomisation may affect the reliability of these findings. Nevertheless, as shown by Adebayo et al. [81], BP-based methods generally underperform on Randomisation metrics, such as the MPRT, which increases the generalisability of the observed trends.

The Faithfulness category shows a dependence on certain metrics; RP MoRF often shows a negative correlation, in particular in 4 out of 6 cases, while metrics such as RP LeRF and IROF typically show positive correlations. In particular, in the two experiments (CutMix with xAI LP on DeepGlobe for VGG and ResNET) where RP MoRF is positive, RP LeRF and IROF are negative. For the Robustness metrics, a slight trend can be observed. The relevancy metrics RRA and RMA generally show moderate positive correlations, while the metric MS shows a moderate negative correlation. The Localisation category, assessed only with the Caltech101 and DeepGlobe datasets, shows a moderate to strong positive correlation with improved xAI-guided training, indicating its significant influence despite the limited dataset comparison. Localisation metrics that assess the alignment of explanations with ground truth are theoretically aligned with the goals of both xAI guidance methods employed. Specifically, RRR loss aims to improve the alignment of explanations with ground truth, suggesting that methods that excel in Localisation metrics would promote more effective and efficient training processes. Similarly, the CutMix with xAI LP employs a LP strategy that improves the accuracy of label propagation in augmented images by ensuring that explanations accurately localise the class of interest. Accurate localisation is critical in preventing the propagation of incorrect label information, thereby increasing the effectiveness of training.

Overall, despite the differences in dataset types, the observed correlation patterns are surprisingly consistent, highlighting the potential generalisability of certain xAI metric relationships across both SLC and MLC tasks. Of particular note is the consistent importance of Localisation metrics in predicting training success and the strong positive correlation of Randomisation metrics, contrasted with the negative effect of Complexity metrics. These findings suggest that RRR loss may be more compatible with non-BP-based metrics such as LIME, GradCAM and Occlusion.

8 Conclusions and Future Directions

This chapter summarises the key findings and insights from the study. In addition, the limitations of the current research of Explainable Artificial Intelligence (xAI) in Remote Sensing (RS) are discussed to highlight areas where the study may fall short or where further investigation is required. Finally, future research directions are outlined to guide further studies to build on the work presented here, address the identified gaps, and explore new opportunities to enhance the understanding and application of xAI in RS.

8.1 Summary of Key Findings

First, a theoretical analysis was conducted to determine the suitability of seven explanation methods (Deep Learning Important FeaTures (DeepLIFT), Gradient-weighted Class Activation Mapping (GradCAM), Guided Gradient-weighted Class Activation Mapping (Guided GradCAM), Integrated Gradients (IG), Local Interpretable Model-agnostic Explanations (LIME), Layer Relevance Propagation (LRP), and Occlusion Sensitivity (Occlusion)) and 34 explanation metrics for Multi-label Classification (MLC) RS image data (see objective A 1). The metrics are derived from six categories, namely: Axiomatic, Complexity, Faithfulness, Localisation, Randomisation, and Robustness. Furthermore, all methods and selected metrics were empirically examined with experiments on the Caltech101 dataset which is a Single-label Classification (SLC) Computer Vision (CV) dataset and two RS MLC datasets: DeepGlobe (DeepGlobe) and BigEarthNet-S2 (BEN). The results were thoroughly discussed and compared in terms of their respective image characteristics and task differences.

Secondly, for these explanation methods the outcome of xAI-guided training was examined using two methods: Right for the Right Reason (RRR) loss and CutMix with xAI Label Propagation for Multi-Label Classification (CutMix with xAI LP) (see objective B 1). The goal of this second set of experiments was to evaluate if there were improvements in model reasoning and performance. The results were contextualised within the previous quantitative evaluation, to determine whether success in one of the metrics or categories correlates with improved outcomes in xAI-guided training.

The first objective was to theoretically examine the efficacy of explanation methods on MLC RS image data. The theoretical analysis identified problems with backpropagation-based methods, namely LRP, IG, DeepLIFT and Guided GradCAM. These methods often highlight edges regardless of the underlying model's actual decision process [81]. This edge bias implies that they might struggle with repetitive texture-based RS images.

Furthermore, the analysis examined issues with perturbation-based methods, beginning with the Occlusion method in the context of MLC. A key issue arises when the area of a multi-label class is substantially large; the occluding patch may not cover a sufficient portion of the class area. If a significant number of input pixels representing that class remain visible, the predic-

tion logits are likely to remain relatively unchanged. This scenario could significantly reduce the effectiveness of the Occlusion method in accurately identifying the influence of specific class regions on the model's decision process. Another issue related to perturbations is defining a valid baseline, which is complicated by the presence of multiple bands in RS. For example, an all-zero baseline in thermal imagery is not meaningful and may lead to the misrepresentation of the importance of features. This might create Out-of-Distribution (OoD) images by introducing new artefacts, potentially even leaking class information to the classifier [102]. However, due to the lack of discriminative shapes in RS images, this issue is less problematic. Nonetheless, a carefully selected baseline is necessary, particularly for methods like LIME, DeepLIFT, Occlusion, and IG.

The theoretical analysis of the explanation metrics identified Localisation as a viable metric category for RS. Localisation metrics assess the accuracy with which the model's explanations align with specific areas of interest within these images. By measuring whether the explanations correctly attribute relevance to the pixels that "should" be relevant (based on ground truth data), these metrics ensure that the model's interpretative outputs are precise and trustworthy. However, this is also a disadvantage, as it requires reliable ground truth maps, which can be expensive to obtain in the RS setting. The metrics from the Faithfulness category are also considered important for RS MLC, as they assess the alignment of the explanation with the actual decision process of the model. However, it is argued that these metrics, if applied to RS images, are highly dependent on their parameterisation, in particular the removal strategy and the perturbation baseline. Robustness metrics slightly perturb the input features and measure differences in the prediction output of the model. Classifiers trained on MLC RS image data theoretically perform better with these metrics, because there are fewer unique complex features to be blurred than in natural images. Therefore, strong blurring is recommended to effectively evaluate robustness. Randomisation and Axiomatic metrics are largely method-dependent and therefore invariant to changes in the input data. However, they do reveal properties or limitations of methods that are of general importance. Complexity metrics prove to be not particularly useful for RS image data. A less 'complex' explanation implies that relevance is assigned to fewer features. However, often the texture is the discriminative characteristic in RS images, leading to naturally complex explanations.

To evaluate the methods and metrics, a VGG16 model was trained on all datasets. Explanations were then generated for each method and evaluated both qualitatively and quantitatively using the test sets of each dataset.

In the qualitative analysis, GradCAM and Guided GradCAM performed best for the Caltech101 dataset. For the DeepGlobe dataset, a visual preference for non-BP-based methods was found, specifically GradCAM, LIME and Occlusion. The BP-based methods seemed to struggle with RS image data, as they did not emphasise the importance of texture. The evaluation of the BEN dataset presented a significant challenge due to the absence of a known ground truth. Nevertheless, similar trends to those observed for DeepGlobe were noted. As the complexity of the task increased and the predictive performance of the model declined, the reliability of the explanations provided also decreased. This was particularly evident in the BEN dataset, where the baseline model had a mAP of ≈ 0.61, making it challenging to distinguish between the poorer perception of the explanations and the poor performance of the model. In the Caltech101 dataset, Spurious Correlations were identified by all methods, although backpropagation-based methods performed slightly worse at localising them. It was

not possible to conduct these evaluations for the MLC datasets because it is currently unknown whether any SCs exist. The quantitative analysis revealed that Occlusion performed best in most metrics for Caltech101, with a sum of normalised metrics of 10.42, followed by LIME and IG. For the DeepGlobe dataset, Guided GradCAM with 10.4 and GradCAM with 10.3 were the top performers, with best values in six metrics each. For the BEN dataset, IG performed best, closely followed by Guided GradCAM and GradCAM. However, no single method showed a clear superiority. A possible reason for this could be poor model performance leading to suboptimal explanations, as discussed above.

In terms of xAI-guided training, the RRR loss augmentation method was first evaluated. This method, as described by Ross et al. [30], incorporates explanations directly into the loss function to improve model reasoning. In addition, a proposed extension that combines the RRR method with the Right for Right Reasons with Mean Squared Error (RRR MSE) loss was also evaluated. This extension aims to strengthen positive relevance in the class ground truth. To evaluate the performance of RRR-guidance, models trained with RRR loss were compared to a baseline. The test metrics used for these comparisons were accuracy for SLC and mAP for MLC. To evaluate the influence of explanation methods on xAI-guided training, the training was conducted for every method. In addition, different parameterisations were examined for the classical RRR and the RRR MSE variants. Here λ is used to scale the influence of the xAI loss compared to the normal loss (see section 5.1). For the Caltech101 dataset, the models trained using the LIME method performed best for both loss variants and both parametrisations of $\lambda = 1$ and $\lambda = 10$. The best parametrisation was RRR MSE $\lambda = 10$, with ≈ 98.2, which surpassed the baseline by a margin of 0.8%. Although this increase is modest, it is significant as the method is primarily designed to enhance reasoning rather than performance. Additionally, in some RRR-trained models (e.g., GradCAM), the explanation methods indicated a reduced reliance on SC as a discriminative feature. However, other models (e.g. DeepLIFT) continued to use it. For the model retrained with GradCAM, this effect was quantified in terms of an improvement in Localisation metrics, showing an increase in reasoning capabilities. In the RRR-training on the DeepGlobe dataset, the optimal outcome was achieved using the RRR loss with $\lambda = 10$ and the Occlusion method. This configuration achieved a mAP of ≈ 84.1, marking an improvement of approximately 0.6% compared to the baseline. Here, the classical RRR outperformed its RRR MSE variants. As there are no known SC in the DeepGlobe dataset, the improvement of reasoning could not be quantified.

Overall, these results indicate that RRR can be an effective xAI-augmentation strategy, as it not only provides a small performance increase but also enhances reasoning capabilities, as demonstrated with SLC. However, one limitation of this method is its dependence on access to a ground truth map.

The second xAI-guided training technique under evaluation is CutMix with xAI LP, a data-augmentation strategy that enhances the classical CutMix method specifically for MLC applications. This approach leverages explanations to accurately propagate labels in CutMix-generated augmentations, aiming to improve the training performance of the models. The DeepGlobe and BEN datasets were utilized, with the introduction of a ResNET34 model to complement the VGG16 and make the results comparable to those of the original study by Burgert et al. [65]. Again multiple runs for each explanation method were conducted. The optimal parameters for t_{cam} and t_{map} were approximated, while multiple box sizes ($0.1 - 0.5$ and $0.3 - 0.7$) were tested. The ResNET training on the DeepGlobe dataset showed a consistent

performance improvement of about 2% over a baseline without augmentation, aligning with the results from [65]. Specifically, CutMix with xAI LP using the DeepLIFT method yielded the most favourable results, achieving a mAP of $\approx$ 86.1 for a box size of $0.3 - 0.7$. In contrast, the effect on the VGG16 models was less pronounced. For these models, the average increase in mAP was around 0.7%, with both LIME using a box size of $0.1 - 0.5$ and GradCAM using $0.3 - 0.7$ both achieving a mAP of $\approx$ 84.6. This discrepancy can be attributed to the potential of the larger VGG16 model to overfit the data, reducing its need for the additional generalisation provided by CutMix with xAI LP.

The results on the BEN dataset were quite similar to those on the DeepGlobe dataset, although the improvements were less pronounced. Using CutMix with xAI LP with GradCAM and a box size of $0.1 - 0.5$ gave the best results, achieving a mAP of $\approx$ 63.1, an increase of about 1.5% over the baseline, which had a mAP of $\approx$ 61.7. For the ResNET models, the overall improvement was approximately 1.1%. For the VGG16 models, performance gains were even more moderate, with an average increase of around 0.2%. Within this group, IG achieved the best results, with a mAP of $\approx$ 61.1. These results suggest that the effectiveness of CutMix with xAI LP can vary significantly between different model architectures.

To determine whether performance on an explanation metric can reliably predict success in specific techniques, a correlation analysis was conducted to assess whether improvements in model performance were associated with success in xAI metrics or categories. A positive correlation was observed between the metrics in the Randomisation category and Localisation metrics, while a negative correlation was observed with performance in Complexity metrics. These correlations held for both RRR-guided training and training utilizing CutMix with xAI LP over all datasets and models. The quantitative analysis of the explanation metrics revealed that backpropagation-based methods often underperform in Randomisation metrics but excel in Complexity metrics. Given that Randomisation metrics are positively correlated and Complexity metrics are negatively correlated, it was found that employing non-backpropagation methods for RRR and CutMix with xAI LP is preferable. In addition, the positive correlation with Localisation metrics suggests that these metrics can serve as an indicator of improved performance in xAI-guided training. This is because accurate highlighting of the OoI by the model and its explanations improves alignment for RRR, and aids in the correct propagation of labels for CutMix with xAI LP.

To summarise the results for MLC RS image data, GradCAM is recognised as the most effective explanation method. Explainable Artificial Intelligence guided training has been shown to improve model performance, especially when combined with non-backpropagation methods. In addition, the increase in performance can be measured using the performance of explanation methods in Localisation metrics as an indicator.

8.2 Limitations

However, while these results are promising, there is still room for improvement. The limitations of the current findings and the potential implications for wider applications are discussed in the following section. This discussion aims to address the nuances and challenges encountered in the study and provide a deeper understanding of where the methods may fall short.

The biggest challenge in the evaluation of explanation methods is the so-called "Challenge

of Unverifiability" [72], stating that an explanation cannot be verified due to missing ground truth. Additionally, many different variables can influence the quality of the explanations, with perhaps the most important factor being the performance of the model. Especially for complex datasets, this poses a huge problem. In the experiments, the accuracy for Caltech101 0.9739, for DeepGlobe $\approx$ 14% less with a mAP of 0.8381 and for BEN the mAP only was 0.6047. Thus, the results from the BEN dataset are considerably less reliable and require careful consideration in the evaluation. It is difficult to distinguish poor model performance as a separate factor influencing the results of the explanations. A possible solution to this problem would be to replicate the experiments using a model with better performance.

Another limitation was the small variety of methods used to evaluate explanation performance. While the study focused on seven well-known and established methods, a deeper evaluation would benefit from including a greater number of methods, particularly newer ones like ScoreCAM [15] or more advanced occlusion approaches such as RISE [16]. This is crucial as four of the seven methods were backpropagation-based, which appeared to underperform. Future studies should consider incorporating more non-backpropagation-based methods to provide a broader perspective.

A specific limitation for the RRR loss evaluations was the difficulty in measuring improved reasoning. For the Caltech101 dataset, a SC could be identified and the improved reasoning could be visualised and even quantified using localisation metrics; however, this was not possible for the DeepGlobe dataset.

The reliability of the correlation analysis for xAI metrics and xAI-guided training results is also limited. Although the results were averaged over three runs with different random seeds and over the full test set, the correlations often lacked statistical significance, as indicated by the p-value. It is important to note that these results may indicate a trend, but do not provide conclusive evidence or an optimal solution. Despite the fact, that the results can be supported theoretically, their validity in practice is not guaranteed.

Furthermore, the scope of this study was limited to older CNN architectures, such as VGG16 and ResNET. Investigating state-of-the-art architectures, such as vision transformers is not only a logical next step but could also prove to be particularly beneficial and insightful.

8.3 Future Research Opportunities

But it is not only because of these limitations that current research in xAI for RS offers significant opportunities for future work. Most common explanatory methods and metrics have been developed primarily for CV, leading to a noticeable research gap in their application to RS. However, explaining and understanding the decisions of complex ML models is critical in all domains. This is particularly true for tasks such as debugging - e.g. by identifying SCs - and for improving the performance, robustness and overall reliability of these models. Consequently, more research must be directed towards adapting and innovating xAI techniques that specifically address the unique challenges of RS domains.

One direction for future research is to develop further xAI methods specifically designed for the complex textural features characteristic of and the multi-label nature of RS images, as current explanation methods often fall short in accurately interpreting these.

Additionally, more experiments focusing on perturbation strategies and removal order for

RS images are essential, as these parameters are critical in shaping the effectiveness of explanation methods and metrics. Perturbation approaches often result in the creation of OoD samples that can significantly affect the prediction outcome and consequently the derived explanations.

A tailored perturbation strategy for MLC could involve perturbing samples with coherent patches rather than individual features. In cases where ground truth is available, these patches can be from the same image but should represent a different class from the one being explained. This approach would ensure that the patches do not affect the prediction result. Alternatively, when the ground truth is unknown, extraction from images that are measurably similar but have different labels could be a viable solution. This approach could ensure that perturbations remain relevant and reflect realistic variation within the dataset, thereby increasing the validity of the explanations generated.

The efficacy of these perturbation strategies can be evaluated using metrics specifically designed or adapted for MLC RS images. As discussed in the theoretical analysis of metrics (section 4), many metrics struggle with images where texture serves as the discriminative feature. For example, the faithfulness metric SEL iteratively removes an increasing number of input pixels from the input image, ordered by relevance, and measures the corresponding change in the prediction result. However, due to the repetitive textures and potentially large areas of interest in the images, the prediction accuracy may not decrease as quickly as it does in natural images.

Moreover, this metric does not take into account an appropriate baseline, as demonstrated in section 4.7, which can lead to artefacts that affect the prediction. For example, for a VGG16 model trained on the DeepGlobe dataset, the prediction logit for the class agricultural land remains at 0.33 even with a completely black baseline image. To quantify this behaviour, various studies with simulated noise could be conducted to determine whether models uniformly predict certain classes in the absence of visual input, or whether such predictions are specific to certain baseline colours (e.g., black, brown, white).

To address this issue, the perturbation strategy discussed above could be used. In addition, to adapt the current Faithfulness metrics to better fit MLC RS images, the slope of the prediction curve could be analysed concerning the class coverage indicated in the ground truth map instead of the AUC. In cases where no ground truth is available, considering the slope of the prediction function within a sliding window might provide a more accurate assessment of the relevance and faithfulness of the explanation.

Furthermore, current methods for measuring the complexity of explanations are significantly biased towards the shapes and number of features attributed as relevant, making popular complexity metrics less effective for RS images. An innovative approach would be to develop a new metric that combines complexity with knowledge of the ground truth. This metric could relate the number of highly attributed pixels to the class coverage specified in the ground truth map, providing a more contextual understanding of explanation complexity.

For classes where ground truth is not available, an extension of metrics such as Effective Complexity (ECO) could be considered. This could involve clustering neighbouring input features based on their relevance so that if many features in the same neighbourhood are considered relevant, they are counted as a single complex feature. This approach could reduce the overestimation of complexity by recognising coherent patches of relevance rather than treating each pixel independently.

However, the evaluation of explanation methods, particularly their localisation ability, often requires ground truth, which is a significant limitation. Therefore, it would be beneficial to explore the development of metrics that can assess these aspects without relying on the availability of ground truth. Investigating such metrics could broaden the applicability of explanation evaluation methods, especially in scenarios where ground truth data is scarce or non-existent.

The same challenge of missing ground truth applies to the usage of RRR loss. Although it has been shown to improve performance and reasoning, as seen in the case of SLC, its effectiveness is fundamentally dependent on the availability of segmentation maps. A promising strategy to mitigate this dependency, particularly relevant for MLC, is as follows RS images, where there is typically no discernible background, is to exploit the property that when attribution maps for all classes are aggregated, the result should be non-zero. Since explanations are generally normalised between 0 and 1, one can hypothesise that this aggregate should ideally sum to 1.

Implementing this strategy would involve modifying existing RRR loss formulations to effectively incorporate this summation property. By incorporating this principle into the loss function, explanations could be "spread" over a larger area, ensuring that each class's contribution is reflected in the model's output. This approach could potentially make the use of RRR loss viable even in scenarios where no ground truth is available.

To quantitatively evaluate the improved a models reasoning in RS image data, a method could be developed to insert SCs into datasets at different "correlation levels". For example, adding a watermark could represent a simple spurious correlation, while consistently pairing unrelated classes in similar images introduces a more complex correlation. By incorporating these intentional biases into a dataset such as the DeepGlobe dataset, it would be possible to measure the effectiveness of explanation methods in detecting these biases and to assess improvements in reasoning facilitated by xAI-guided training.

In summary, this thesis has rigorously investigated the application and effectiveness of various explanation methods and metrics in the context of RS and MLC image data. Through a combination of theoretical analysis and empirical evaluation, it has provided valuable insights into the capabilities and limitations of current xAI research for RS. Given the significant impact of AI on society and its potential to address critical global challenges such as climate change, improving the reliability of models in RS is imperative. However, verifying the reliability of these models remains a complex task. To ensure that AI systems in RS meet the stringent requirements for accuracy and robustness, continued advances in xAI for RS are essential. This research will not only improve model performance but also increase confidence in AI applications, which are crucial for informed decision-making in sensitive and high-impact areas.

Acronyms

AI Artificial Intelligence. 1–3, 23, 45, 46, 95

AL Attribution Localisation. 32, 33, 65, 67, 72, 79, 80, 82, 83, 116

AOC Area Over the Curve. 28

AS Average Sensitivity. 30, 31

AUC Area Under the Curve. 27, 28, 38, 39, 94

BEN BigEarthNet-S2. vii, ix, x, 3, 49–52, 54, 57, 62–64, 68–71, 73, 75, 77, 78, 84–86, 89–93, 118, 123, 124

BP-based Backpropagation-based. 11–14, 16, 19, 57, 59, 60, 62, 64–72, 87, 88, 90

Caltech101 Caltech101. vii, ix, x, 3, 49, 50, 54, 55, 57–59, 64–75, 77–81, 83, 88–91, 93, 118, 119

CAM Class Activation Mapping. 5, 6, 11, 16, 19, 57, 113

CNN Convolutional Neural Network. 7, 17, 19, 93

CO Complexity. 35

CRP Concept Relevance Propagation. 114

CutMix with xAI LP CutMix with xAI Label Propagation for Multi-Label Classification. vii, x, 3, 4, 7, 46, 54, 55, 73, 75, 76, 78, 81–86, 88, 89, 91, 92, 122–124

CV Computer Vision. 2, 3, 11, 20, 26, 30, 39, 42, 49, 52, 64, 89, 93

DeepGlobe DeepGlobe. vii, ix, x, 3, 38, 42, 49, 50, 52, 54, 55, 57, 59–62, 64, 66–77, 79–84, 88–95, 114, 115, 118, 120–122

DeepLIFT Deep Learning Important FeaTures. x, 2, 3, 6, 12, 14–16, 19, 20, 29, 40, 41, 53, 57–70, 72, 74–78, 89–92

DL Deep Learning. 1

DTD Deep Taylor Decomposition. 13

ECO Effective Complexity. 35, 37, 64, 66, 68, 78, 79, 82, 85, 94

EO Earth Observation. 1, 2, 5, 6

FC Faithfulness Correlation. 25–27, 38

FE Faithfulness Estimate. 25, 26, 28, 38

GBP Guided Backpropagation. 6, 15, 17, 19

GradCAM Gradient-weighted Class Activation Mapping. x, 3, 5–7, 16, 17, 19, 38, 41, 53, 57–70, 72–78, 88–92

Guided GradCAM Guided Gradient-weighted Class Activation Mapping. 3, 6, 17, 53, 57–70, 72–76, 78, 89–91

HiResCAM High-Resolution Class Activation Mapping. 7

IG Integrated Gradients. 3, 6, 12, 13, 20, 29, 53, 57–70, 72–78, 87, 89–92

IROF Iterative Removal Of Features. 26, 28, 38, 64–66, 69–71, 79, 82, 84, 85, 87, 88, 114

LeRF Least Relevant First. ix, 21, 27, 38–41, 65, 67, 69, 71, 79, 82, 85, 87, 88

LIME Local Interpretable Model-agnostic Explanations. 3, 6, 7, 17, 18, 20, 30, 53, 57–70, 73–78, 88–92

LLE Local Lipschitz Estimate. 30, 31

LP Label Propagation. 47, 55, 81, 84, 88

LRP Layer Relevance Propagation. 3, 6, 12–15, 19, 29, 53, 57–70, 72, 74–76, 78, 87, 89, 114

mAP mean Average Precision. vii, x, 52, 57, 73–78, 80, 81, 84, 85, 90–93

ML Machine Learning. 3, 5, 7, 18, 93

MLC Multi-label Classification. 2–4, 6, 7, 9, 19, 25, 27, 30, 35, 37–39, 47, 49, 52, 54, 55, 57, 64, 71, 81, 88–92, 94, 95, 114, 118

MoRF Most Relevant First. ix, 6, 21, 27, 28, 38–42, 65, 67, 69, 71, 79, 82, 83, 85, 87, 88

MPRT Model Parameter Randomization Test. 36, 88

MS Maximum Sensitivity. 30, 31, 66, 79, 80, 82–85, 88, 116

NN Neural Networks. 5, 8, 11–15, 17, 30, 52, 54, 113, 114

Occlusion Occlusion Sensitivity. 3, 6, 7, 11, 13, 19, 20, 57–60, 62, 64–70, 72, 74–76, 88–91

OoD Out-of-Distribution. 20, 21, 90, 94

OoI Object of Interest. 37, 53, 55, 57, 58, 60, 64, 65, 80, 92

PF Pixel-Flipping. 25, 27, 38

PG Pointing-Game. 32, 33, 65, 67, 71, 72, 79, 80, 82–84

RIS Relative Input Stability. 30–32, 54, 66, 68, 79, 80, 83, 85, 116

RISE Randomised Input Sampling to provide Explanations. 20, 93, 113

RL Random Logit. 36, 67, 69, 79, 80, 83, 85, 86, 88, 116

RMA Relevance Mass Accuracy. 32–34, 65, 67, 72, 79, 80, 82–84, 88

ROAD RemOve And Debias. 21, 26, 28, 42

ROS Relative Output Stability. 30, 32, 54, 66, 68, 79, 80, 83, 85, 116

RP Region Perturbation. 25, 27, 28, 38, 65, 67, 69, 71, 79, 82, 83, 85, 87, 88

RRA Relevance Rank Accuracy. 32–34, 65, 67, 72, 79, 80, 82, 83, 88

RRR Right for the Right Reason. vii, x, 3, 4, 7, 46, 48, 54, 73–82, 88, 89, 91–93, 95, 119, 120

RRR MSE Right for Right Reasons with Mean Squared Error. 46, 54, 73–76, 91

RRS Relative Representation Stability. 30, 32

RS Remote Sensing. 1–9, 12, 18–21, 25–28, 30, 33, 35, 37–39, 42, 43, 46, 49, 51–53, 64, 67, 68, 70–72, 74, 75, 89, 90, 92–95

SAR Synthetic Aperture Radar. 5

SC Spurious Correlation. ix, x, 2, 54, 55, 58, 59, 74, 75, 90, 91, 93, 95

ScoreCAM Score-weighted Class Activation Mapping. 6, 93, 113

SEL Selectivity. 25, 27, 28, 38, 65, 67, 70, 79, 82, 84, 85, 87, 94

SENS-N Sensitivity-n. 25, 28, 29, 37, 38

SHAP SHapley Additive exPlanations. 6, 20, 30, 113

SLC Single-label Classification. 2, 3, 9, 52–54, 57, 64, 71, 75, 81, 88, 89, 91, 95

SLIC Simple Linear Iterative Clustering. 53, 114

SMCAM Self-Matching CAM. 5

SP Sparseness. 35

SSIM Structural Similarity Index Measure. 36, 116

TKI Top-K Intersection. 32, 33, 65, 67, 71, 72, 80, 82, 83

xAI Explainable Artificial Intelligence. ix, x, 1–5, 7, 8, 11, 19, 23, 24, 38, 45, 46, 48, 51, 52, 54, 55, 71–73, 75, 78–89, 91–93, 95, 114, 119

Bibliography

[1] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, and C. Seifert, "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–42, 2023.

[2] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K.-R. Müller, "Layer-Wise Relevance Propagation: An Overview," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, ser. Lecture Notes in Computer Science, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, Eds. Cham: Springer International Publishing, 2019, pp. 193–209. [Online]. Available: https://doi.org/10.1007/978-3-030-28954-6_10

[3] X. Chen, C. Liang, D. Huang, E. Real, K. Wang, H. Pham, X. Dong, T. Luong, C.-J. Hsieh, Y. Lu *et al.*, "Symbolic discovery of optimization algorithms," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[4] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[5] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 408–14 419.

[6] K. Karalas, G. Tsagkatakis, M. Zervakis, and P. Tsakalides, "Deep learning for multi-label land cover classification," in *Image and signal processing for remote sensing XXI*, vol. 9643. SPIE, 2015, pp. 244–257.

[7] L. Ma, Y. Liu, X. Zhang, Y. Ye, G. Yin, and B. A. Johnson, "Deep learning in remote sensing applications: A meta-analysis and review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 152, pp. 166–177, Jun. 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271619301108

[8] M. E. Paoletti, J. M. Haut, J. Plaza, and A. Plaza, "Deep learning classifiers for hyperspectral imaging: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 158, pp. 279–317, Dec. 2019, publisher: Elsevier B.V.

[9] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[10] W. Saeed and C. Omlin, "Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities," *Knowledge-Based Systems*, vol. 263, p. 110273, 2023.

[11] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI magazine*, vol. 38, no. 3, pp. 50–57, 2017.

[12] Council of European Union, "Regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act)," 2023,

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[13] R. Achtibat, M. Dreyer, I. Eisenbraun, S. Bosse, T. Wiegand, W. Samek, and S. Lapuschkin, "From attribution maps to human-understandable explanations through concept relevance propagation," *Nature Machine Intelligence*, vol. 5, no. 9, pp. 1006–1019, 2023.

[14] K. Schulz, L. Sixt, F. Tombari, and T. Landgraf, "Restricting the flow: Information bottlenecks for attribution," 2019.

[15] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, "Score-cam: Score-weighted visual explanations for convolutional neural networks," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2020, pp. 111–119.

[16] V. Petsiuk, A. Das, and K. Saenko, "Rise: Randomized input sampling for explanation of black-box models," 2018.

[17] M. T. Ribeiro, S. Singh, and C. Guestrin, "Model-Agnostic Interpretability of Machine Learning," Jun. 2016, arXiv: 1606.05386. [Online]. Available: https://arxiv.org/abs/1606.05386v1

[18] M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8689 LNCS, no. PART 1, pp. 818–833, Nov. 2013, arXiv: 1311.2901 Publisher: Springer Verlag ISBN: 9783319105895. [Online]. Available: https://arxiv.org/abs/1311.2901v3

[19] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[20] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *stat*, vol. 1050, p. 2, 2017.

[21] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *stat*, vol. 1050, p. 27, 2020.

[22] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K.-R. Müller, "Unmasking Clever Hans Predictors and Assessing What Machines Really Learn," *Nature Communications*, vol. 10, no. 1, p. 1096, Mar. 2019, arXiv:1902.10178 [cs, stat]. [Online]. Available: http://arxiv.org/abs/1902.10178

[23] P. Kirichenko, P. Izmailov, and A. Gordon Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," *ICLR 2023*, 2023.

[24] A. J. DeGrave, J. D. Janizek, and S.-I. Lee, "Ai for radiographic covid-19 detection selects shortcuts over signal," *Nature Machine Intelligence*, vol. 3, no. 7, pp. 610–619, 2021.

[25] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the isic image datasets: Usage, benchmarks and recommendations," *Medical image analysis*, vol. 75, p. 102305, 2022.

[26] A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, "xxAI - Beyond Explainable Artificial Intelligence," in *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, ser. Lecture Notes in Computer Science, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham: Springer International Publishing, 2022, pp. 3–10. [Online].

Available: https://doi.org/10.1007/978-3-031-04083-2_1

[27] L. Weber, S. Lapuschkin, A. Binder, and W. Samek, "Beyond explaining: Opportunities and challenges of XAI-based model improvement," *Information Fusion*, vol. 92, pp. 154–176, Apr. 2023, arXiv: 2203.08008 Publisher: Elsevier.

[28] D. Tuia, R. Roscher, J. D. Wegner, N. Jacobs, X. Zhu, and G. Camps-Valls, "Toward a collective agenda on ai for earth science data analysis," *IEEE Geoscience and Remote Sensing Magazine*, vol. 9, no. 2, pp. 88–104, 2021.

[29] C. M. Gevaert, "Explainable ai for earth observation: A review including societal and regulatory perspectives," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, p. 102869, 2022.

[30] A. S. Ross, M. C. Hughes, and F. Doshi-Velez, "Right for the right reasons: Training differentiable models by constraining their explanations," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization, 2017.

[31] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (xai)," *IEEE access*, vol. 6, pp. 52 138–52 160, 2018.

[32] A. Höhl, I. Obadic, M. Á. F. Torres, H. Najjar, D. Oliveira, Z. Akata, A. Dengel, and X. X. Zhu, "Opening the black-box: A systematic review on explainable ai in remote sensing," *arXiv preprint arXiv:2402.13791*, 2024.

[33] C. Leluschko and C. Tholen, "Goals and stakeholder involvement in xai for remote sensing: A structured literature review," in *International Conference on Innovative Techniques and Applications of Artificial Intelligence*. Springer, 2023, pp. 519–525.

[34] Z. Feng, M. Zhu, L. Stanković, and H. Ji, "Self-matching cam: A novel accurate visual explanation of cnns for sar image interpretation," *Remote Sensing*, vol. 13, no. 9, p. 1772, 2021.

[35] X. Huang, Y. Sun, S. Feng, Y. Ye, and X. Li, "Better visual interpretation for remote sensing scene classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[36] G. De Lucia, M. Lapegna, and D. Romano, "Towards explainable ai for hyperspectral image classification in edge computing environments," *Computers and Electrical Engineering*, vol. 103, p. 108381, 2022.

[37] I. Kakogeorgiou and K. Karantzalos, "Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 103, Dec. 2021, arXiv: 2104.01375 Publisher: Elsevier B.V.

[38] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps," *2nd International Conference on Learning Representations, ICLR 2014 - Workshop Track Proceedings*, Dec. 2013, arXiv: 1312.6034 Publisher: International Conference on Learning Representations, ICLR. [Online]. Available: https://arxiv.org/abs/1312.6034v2

[39] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning Important Features Through Propagating Activation Differences," *34th International Conference on Machine Learning, ICML 2017*, vol. 7, pp. 4844–4866, Apr. 2017, arXiv: 1704.02685 Publisher: International Machine Learning Society (IMLS) ISBN: 9781510855144. [Online]. Available: https:

//arxiv.org/abs/1704.02685v2

[40] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic Attribution for Deep Networks," in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 3319–3328, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v70/sundararajan17a.html

[41] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," in *ICLR (workshop track)*, 2015.

[42] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why Should I Trust You?": Explaining the Predictions of Any Classifier," *NAACL-HLT 2016 - 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Demonstrations Session*, pp. 97–101, Feb. 2016, arXiv: 1602.04938 Publisher: Association for Computational Linguistics (ACL) ISBN: 9781450342322. [Online]. Available: https://arxiv.org/abs/1602.04938v3

[43] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," Jun. 2017.

[44] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar, "On the (In)fidelity and Sensitivity of Explanations," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/a7471fdc77b3435276507cc8f2dc2569-Abstract.html

[45] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 11, pp. 2660–2673, 2016.

[46] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K. R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proceedings of the IEEE*, vol. 109, no. 3, pp. 247–278, Mar. 2021, arXiv: 2003.07631 Publisher: Institute of Electrical and Electronics Engineers Inc.

[47] J.-P. Kucklick and O. Müller, "Tackling the accuracy-interpretability trade-off: Interpretable deep learning models for satellite image-based real estate appraisal," *ACM Transactions on Management Information Systems*, vol. 14, no. 1, pp. 1–24, 2023.

[48] A. Abdollahi and B. Pradhan, "Urban Vegetation Mapping from Aerial Imagery Using Explainable AI (XAI)," *Sensors*, vol. 21, no. 14, p. 4738, Jan. 2021, number: 14 Publisher: Multidisciplinary Digital Publishing Institute. [Online]. Available: https://www.mdpi.com/1424-8220/21/14/4738

[49] S. M. Lundberg and S. I. Lee, "A Unified Approach to Interpreting Model Predictions," *Advances in Neural Information Processing Systems*, vol. 2017-December, pp. 4766–4775, May 2017, arXiv: 1705.07874 Publisher: Neural information processing systems foundation. [Online]. Available: https://arxiv.org/abs/1705.07874v2

[50] A. Wolanin, G. Mateo-García, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi, and L. Guanter, "Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt," *Environmental Research Letters*, vol. 15, no. 2, p. 024019, Feb. 2020, publisher: IOP Publishing. [Online]. Available: https://dx.doi.org/10.1088/1748-9326/ab68ac

[51] H. Yessou, G. Sumbul, and B. Demir, "A Comparative Study of Deep Learning Loss Functions for Multi-Label Remote Sensing Image Classification," in *IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium*, Sep. 2020, pp. 1349–1352, arXiv:2009.13935 [cs]. [Online]. Available: http://arxiv.org/abs/2009.13935

[52] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation," *PLOS ONE*, vol. 10, no. 7, p. e0130140, Jul. 2015, publisher: Public Library of Science. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0130140

[53] Q. Su, X. Zhang, P. Xiao, Z. Li, and W. Wang, "Which cam is better for extracting geographic objects? a perspective from principles and experiments," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 15, pp. 5623–5635, 2022.

[54] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks," in *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2018, pp. 839–847.

[55] D. Omeiza, S. Speakman, C. Cintas, and K. Weldermariam, "Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models," *arXiv preprint arXiv:1908.01224*, 2019.

[56] A. Mohan and J. Peeples, "Quantitative analysis of primary attribution explainable artificial intelligence methods for remote sensing image classification," in *2023 IEEE International Geoscience and Remote Sensing Symposium IGARSS*. IEEE, 2023, pp. 950–953.

[57] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 976–11 986.

[58] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2020.

[59] J. Yang, C. Li, X. Dai, and J. Gao, "Focal modulation networks," *Advances in Neural Information Processing Systems*, vol. 35, pp. 4203–4217, 2022.

[60] R. L. Draelos and L. Carin, "Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks," *arXiv preprint arXiv:2011.08891*, 2020.

[61] A. Hedström, L. Weber, D. Krakowczyk, D. Bareeva, F. Motzkus, W. Samek, S. Lapuschkin, and M. M. M. Höhne, "Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond," *Journal of Machine Learning Research*, vol. 24, no. 34, pp. 1–11, 2023.

[62] X. Cheng, A. Doosthosseini, and J. Kunkel, "Improve the Deep Learning Models in Forestry Based on Explanations and Expertise," *Frontiers in Plant Science*, vol. 13, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpls.2022.902105

[63] T. Beker, Q. Song, and X. X. Zhu, "An analysis of the gap between hybrid and real data for volcanic deformation detection," in *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2023, pp. 825–828.

[64] W. Xiong, Z. Xiong, Y. Cui, L. Huang, and R. Yang, "An interpretable fusion siamese network for multi-modality remote sensing ship image retrieval," *IEEE Transactions on Circuits and Sys-*

tems for Video Technology, 2022.

[65] T. Burgert, T. Siebert, K. N. Clasen, and B. Demir, "A label propagation strategy for cutmix in multi-label remote sensing image classification," *arXiv preprint arXiv:2405.13451*, 2024.

[66] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6023–6032.

[67] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal, "Explaining explanations: An overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE, 2018, pp. 80–89.

[68] S. Bruckert, B. Finzel, and U. Schmid, "The next generation of medical decision support: A roadmap toward transparent expert companions," *Frontiers in artificial intelligence*, vol. 3, p. 507973, 2020.

[69] A. Páez, "The pragmatic turn in explainable artificial intelligence (xai)," *Minds and Machines*, vol. 29, no. 3, pp. 441–459, 2019.

[70] G. Schwalbe and B. Finzel, "A Comprehensive Taxonomy for Explainable Artificial Intelligence: A Systematic Survey of Surveys on Methods and Concepts," *Data Mining and Knowledge Discovery*, Jan. 2023, arXiv:2105.07190 [cs]. [Online]. Available: http://arxiv.org/abs/2105.07190

[71] N. Burkart and M. F. Huber, "A survey on the explainability of supervised machine learning," *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.

[72] A. Hedström, P. Bommer, K. K. Wickstrøm, W. Samek, S. Lapuschkin, and M. M.-C. Höhne, "The meta-evaluation problem in explainable ai: identifying reliable estimators with metaquantus," *Transactions on Machine Learning Research (TMLR)*, 2023.

[73] X. Bai, X. Wang, X. Liu, Q. Liu, J. Song, N. Sebe, and B. Kim, "Explainable deep learning for efficient and robust pattern recognition: A survey of recent developments," *Pattern Recognition*, vol. 120, p. 108102, Dec. 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320321002892

[74] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[75] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, "Explainable AI Methods - A Brief Overview," in *xxAI - Beyond Explainable AI*, A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek, Eds. Cham: Springer International Publishing, 2022, vol. 13200, pp. 13–38, series Title: Lecture Notes in Computer Science. [Online]. Available: https://link.springer.com/10.1007/978-3-031-04083-2_2

[76] 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.

[77] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[78] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern*

recognition, 2016, pp. 2921–2929.

[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations (ICLR 2015)*. Computational and Biological Learning Society, 2015, pp. 1–14.

[80] J. Xing and R. Sieber, "The challenges of integrating explainable artificial intelligence into geoai," *Transactions in GIS*, vol. 27, no. 3, pp. 626–645, 2023.

[81] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity Checks for Saliency Maps," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html

[82] W. Nie, Y. Zhang, and A. Patel, "A theoretical explanation for perplexing behaviors of backpropagation-based visualizations," in *International conference on machine learning*. PMLR, 2018, pp. 3809–3818.

[83] L. Sixt, M. Granz, and T. Landgraf, "When Explanations Lie: Why Many Modified BP Attributions Fail," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 9046–9057, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v119/sixt20a.html

[84] Q. Yang, X. Zhu, J.-K. Fwu, Y. Ye, G. You, and Y. Zhu, "Mfpp: Morphological fragmental perturbation pyramid for black-box model explanations," in *2020 25th International conference on pattern recognition (ICPR)*. IEEE, 2021, pp. 1376–1383.

[85] S.-H. Kang, H.-G. Jung, and S.-W. Lee, "Interpreting undesirable pixels for image classification on black-box models," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2019, pp. 4250–4254.

[86] C. Agarwal, N. Johnson, M. Pawelczyk, S. Krishna, E. Saxena, M. Zitnik, and H. Lakkaraju, "Rethinking stability for attribution-based explanations," in *ICLR 2022 Workshop on PAIR {\textasciicircum} 2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*, 2022.

[87] L. Qiu, Y. Yang, C. C. Cao, Y. Zheng, H. Ngai, J. Hsiao, and L. Chen, "Generating perturbation-based explanations with robustness to out-of-distribution data," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 3594–3605.

[88] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 427–436.

[89] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.

[90] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," *Advances in neural information processing systems*, vol. 32, 2019.

[91] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A consistent and efficient evaluation strategy for attribution methods," in *International Conference on Machine Learning*. PMLR, 2022, pp. 18 770–18 795.

[92] L. Brocki and N. C. Chung, "Evaluation of interpretability methods and perturbation artifacts

in deep neural networks," *arXiv preprint arXiv:2203.02928*, 2022.

[93] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[94] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, "Towards better understanding of gradient-based attribution methods for Deep Neural Networks," *6th International Conference on Learning Representations (ICLR)*, no. 1711.06104, pp. 0–0, 2018, number: 1711.06104 Place: Ithaca, NY, USA Publisher: Arxiv - Computer Science. [Online]. Available: https://arxiv.org/abs/1711.06104

[95] P. Wang and N. Vasconcelos, "Scout: Self-aware discriminant counterfactual explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8981–8990.

[96] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of biomedical informatics*, vol. 113, p. 103655, 2021.

[97] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 44, pp. 22 071–22 080, 2019.

[98] M. L. Leavitt and A. Morcos, "Towards falsifiable interpretability research," *arXiv preprint arXiv:2010.12016*, 2020.

[99] P. Q. Le, M. Nauta, V. B. Nguyen, S. Pathak, J. Schlötterer, and C. Seifert, "Benchmarking explainable ai: a survey on available toolkits and open challenges," in *International Joint Conference on Artificial Intelligence*, 2023.

[100] A. Ghorbani, A. Abid, and J. Zou, "Interpretation of neural networks is fragile," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 3681–3688.

[101] U. Bhatt, A. Weller, and J. M. F. Moura, "Evaluating and aggregating feature-based model explanations," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, ser. IJCAI'20, Yokohama, Yokohama, Japan, Jan. 2021, pp. 3016–3022.

[102] Y. Rong, T. Leemann, V. Borisov, G. Kasneci, and E. Kasneci, "A Consistent and Efficient Evaluation Strategy for Attribution Methods," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, Jun. 2022, pp. 18 770–18 795, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v162/rong22a.html

[103] D. Alvarez Melis and T. Jaakkola, "Towards Robust Interpretability with Self-Explaining Neural Networks," in *Advances in Neural Information Processing Systems*, vol. 31. Curran Associates, Inc., 2018. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2018/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html

[104] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilovic, S. Mourad, P. Pedemonte, R. Raghavendra, J. Richards, P. Sattigeri, K. Shanmugam, M. Singh, K. R. Varshney, D. Wei, and Y. Zhang, "One Explanation Does Not Fit All: A Toolkit And Taxonomy Of AI Explainability Techniques," Oct. 2021. [Online]. Available: https://research.ibm.com/publications/one-explanation-does-not-fit-all-a-toolkit-and-taxonomy-of-ai-explainability-techniques

[105] G. Montavon, W. Samek, and K.-R. Müller, "Methods for interpreting and understanding deep neural networks," *Digital Signal Processing*, vol. 73, pp. 1–15, Feb. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1051200417302385

[106] L. Rieger and L. K. Hansen, "IROF: a low resource evaluation metric for explanation methods: Workshop AI for Affordable Healthcare at ICLR 2020," *Proceedings of the Workshop AI for Affordable Healthcare at ICLR 2020*, 2020.

[107] S. Dasgupta, N. Frost, and M. Moshkovitz, "Framework for Evaluating Faithfulness of Local Explanations," in *Proceedings of the 39th International Conference on Machine Learning*. PMLR, Jun. 2022, pp. 4794–4815, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v162/dasgupta22a.html

[108] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *stat*, vol. 1050, p. 20, 2015.

[109] A. Arias-Duart, F. Parés, D. Garcia-Gasulla, and V. Giménez-Ábalos, "Focus! Rating XAI Methods and Finding Biases," in *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, Jul. 2022, pp. 1–8, iSSN: 1558-4739. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9882821

[110] J. Zhang, S. A. Bargal, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-Down Neural Attention by Excitation Backprop," *International Journal of Computer Vision*, vol. 126, no. 10, pp. 1084–1102, Oct. 2018. [Online]. Available: https://doi.org/10.1007/s11263-017-1059-x

[111] M. Kohlbrenner, A. Bauer, S. Nakajima, A. Binder, W. Samek, and S. Lapuschkin, "Towards Best Practice in Explaining Neural Network Decisions with LRP," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2020, pp. 1–7, iSSN: 2161-4407. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9206975

[112] J. Theiner, E. Müller-Budack, and R. Ewerth, "Interpretable Semantic Photo Geolocation." IEEE Computer Society, Jan. 2022, pp. 1474–1484. [Online]. Available: https://www.computer.org/csdl/proceedings-article/wacv/2022/091500b474/1B12Fl0FVAc

[113] L. Arras, A. Osman, and W. Samek, "CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations," *Information Fusion*, vol. 81, pp. 14–40, May 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1566253521002335

[114] P. Chalasani, J. Chen, A. R. Chowdhury, X. Wu, and S. Jha, "Concise Explanations of Neural Networks using Adversarial Training," in *Proceedings of the 37th International Conference on Machine Learning*. PMLR, Nov. 2020, pp. 1383–1391, iSSN: 2640-3498. [Online]. Available: https://proceedings.mlr.press/v119/chalasani20a.html

[115] A.-p. Nguyen and M. R. Martínez, "On quantitative aspects of model interpretability," Jul. 2020, arXiv:2007.07584 [cs, stat]. [Online]. Available: http://arxiv.org/abs/2007.07584

[116] A. Binder, L. Weber, S. Lapuschkin, G. Montavon, K.-R. Müller, and W. Samek, "Shortcomings of top-down randomization-based sanity checks for evaluations of deep neural network explanations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 143–16 152.

[117] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim, "The (un) reliability of saliency methods," *Explainable AI: Interpreting, explaining and visualizing deep learning*, pp. 267–280, 2019.

[118] R. Tomsett, D. Harborne, S. Chakraborty, P. Gurram, and A. Preece, "Sanity checks for saliency metrics," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 6021–6029.

[119] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[120] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," in *2004 conference on computer vision and pattern recognition workshop*. IEEE, 2004, pp. 178–178.

[121] F.-F. Li, M. Andreeto, M. Ranzato, and P. Perona, "Caltech 101," Apr 2022.

[122] I. Demir, K. Koperski, D. Lindenbaum, G. Pang, J. Huang, S. Basu, F. Hughes, D. Tuia, and R. Raskar, "Deepglobe 2018: A challenge to parse the earth through satellite images," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[123] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl, "Bigearthnet: A large-scale benchmark archive for remote sensing image understanding," in *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2019, pp. 5901–5904.

[124] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[125] T. maintainers and contributors, "TorchVision: PyTorch's Computer Vision library," Nov. 2016. [Online]. Available: https://github.com/pytorch/vision

[126] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[127] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," Mar. 2019. [Online]. Available: https://github.com/Lightning-AI/lightning

[128] N. Kokhlikyan, V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, and O. Reblitz-Richardson, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[129] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[130] S. van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Gouillart, T. Yu, and the scikit-image contributors, "scikit-image: image processing in Python," *PeerJ*, vol. 2, p. e453, 6 2014. [Online]. Available: https://doi.org/10.7717/peerj.453

Appendix

1 Further Explanation Methods

Below is a brief overview of some prominent explanation methods that were not included in the detailed theoretical analysis and experiments.

A simple but effective approach is to visualise the saliency map [38], $\frac{\partial y}{\partial \boldsymbol{x}}$, i.e. the partial derivative of the network output y with respect to the input features $\boldsymbol{x}$. Since the derivative value represents the degree to which the output is affected by small changes in the input features, the saliency map quantifies the sensitivity of the output value to the input features. The Deconvolution method [18] visualises the activations of a NN by reversing the convolution operations, thereby highlighting which parts of the input image are most influential.

Gradient x Input is a backpropagation-based method that multiplies each input feature $\boldsymbol{x}$ by the gradient of the output with respect to the input, formulated as $\boldsymbol{x} \cdot \nabla_{f(\boldsymbol{x})}$ [39].

SmoothGrad reduces noise in attribution maps, which are often noisy due to sharp gradient fluctuations, by averaging the gradients of multiple noisy versions of the input image. This is defined as

$$\Phi_{\text{sm}}(\boldsymbol{x}) = \frac{1}{n}\sum_{1}^{n} \mathcal{M}_n\left(\boldsymbol{x} + \mathcal{G}(0, \sigma^2)\right),$$

where n is the number of instances, $\boldsymbol{x}$ is the input image, and $\mathcal{G}$ is Gaussian noise with standard deviation σ [43].

Randomised Input Sampling to provide Explanations (RISE) [16] is a perturbation-based method that estimates the value of important pixels in an image by reducing the brightness of pixels to zero in random combinations. This effect is mimicked by multiplying an input image I element by element with a randomly generated binary mask $\mathcal{M}$. The confidence score is computed using the masked images by feeding them to a DNN. A heatmap is produced by a linear combination of the masks, with the confidence score derived from the target class for the masked input.

Score-weighted Class Activation Mapping (ScoreCAM) [15] combines perturbation and gradient-based methods. It occludes the CAM maps for a convolutional layer $\mathcal{N}_L$ with a baseline $\overline{\boldsymbol{x}}$ and measures the change in outcome. Let $\boldsymbol{f}$ be a model that takes $\boldsymbol{x}$ as input image and produces logits $\boldsymbol{f}' \cdot \mathcal{A}^i_{\mathcal{N}_L}$, where $\mathcal{A}^i_{\mathcal{N}_L}$ denotes the i-th channel of the convolutional layer $\mathcal{N}_L$. The contribution of $\mathcal{A}^i_{\mathcal{N}_L}$ to $\boldsymbol{f}'$ with $\overline{\boldsymbol{x}}$ as the baseline for class c is

$$\beta^c_i = \mathrm{C}\left(\mathcal{A}^i_{\mathcal{N}_L}\right) = \boldsymbol{f}^c\left(\boldsymbol{x} \cdot \mathbb{H}^i_l\right) - \boldsymbol{f}^c\left(\overline{\boldsymbol{x}}\right),$$

where $\mathbb{H}^i_{\mathcal{N}_L} = s\left(U_p\left(\mathcal{A}^i_{\mathcal{N}_L}\right)\right)$, $U_p(\cdot)$ upsamples $\mathcal{A}^i_{\mathcal{N}_L}$, and s normalises each element to $[0, 1]$.

Formally, ScoreCAM is represented as:

$$\Phi^c_{\text{ScoreCAM}} = \text{ReLU}\left(\sum_i \beta^c_i \mathcal{A}^i_{\mathcal{N}_L}\right).$$

SHapley Additive exPlanations [49], derived from cooperative game theory, fairly distribute the "payout" (model prediction) among the "players" (input features). For a given prediction, the Shapley

value of each feature represents its average contribution across all possible combinations of features. This approach ensures that contributions are fairly allocated and provides insight into how each feature influences the output of the model. The method is theoretically grounded, providing axiomatic guarantees of fairness, efficiency and additivity, making it a powerful tool for interpreting complex models and understanding the importance of features in machine learning predictions.

One of the most recent approaches is Concept Relevance Propagation [13], which is an extension of LRP using conditional relevance propagation. CRP provides detailed, concept-specific explanations by propagating relevance through NN layers conditioned on specific concepts. This approach makes it possible to identify the contribution and location of learned concepts (such as 'fur' or 'eye') in the model's decision-making process, thus answering both the 'where' and 'what' questions about the model's reasoning.

2 Experimental Setup

The xAI integraton of CutMix utilizes three relevant hyperparameters: t_{map} and t_{cam} [65], and the box size [66]. The t_{map} threshold defines how many activating pixels are necessary for a label to be propagated, and the t_{cam} threshold defines the minimal relevance that pixel k needs to have to be considered in the binary explanation map.

To choose thresholds t_{map} and t_{cam}, one can utilize the given reference maps. For a MLC task, let $x_i \in X$ be an input image, $y_i \in Y$ its corresponding multi-label vector, $\boldsymbol{e}_i$ its explanation mask, and s_i its reference or reference map. Let $\tilde{x}$ be the CutMix augmentation of that image. Let $\phi : \{0,1\}^{L\times H\times W} \to \{0,1\}^L$ be the readout function to derive the new label $\tilde{y}$ from $\tilde{x}$.

The readout function can be used with the augmented explanation masks $e\tilde{y}_i = \phi(\tilde{e}_i)$ or with the reference map to derive the true new label $s\tilde{y}_i = \phi(s_i)$.

To approximate the optimal thresholds t_{map} and t_{cam} for a set of explanation masks E, one can maximize the accuracy between $e\tilde{y}_i$ and $s\tilde{y}_i$. The results for the DeepGlobe datasets, visualized in different matrices, for each explanation method, are shown in figure 1.

3 Hyperparameters for Explanation Metrics

The explanation metrics utilised in this study are implemented using the Quantus framework [61]. The default parameters from the framework are used, with specific hyperparameters detailed below for each metric to ensure reproducibility.

Faithfulness

- **Iterative Removal Of Features**:
  - `Segmentation Method = SLIC`
  - `Perturbation Baseline = black`
- **Monotonicity**:
  - `features_in_step`
  - `Perturbation Baseline = black`

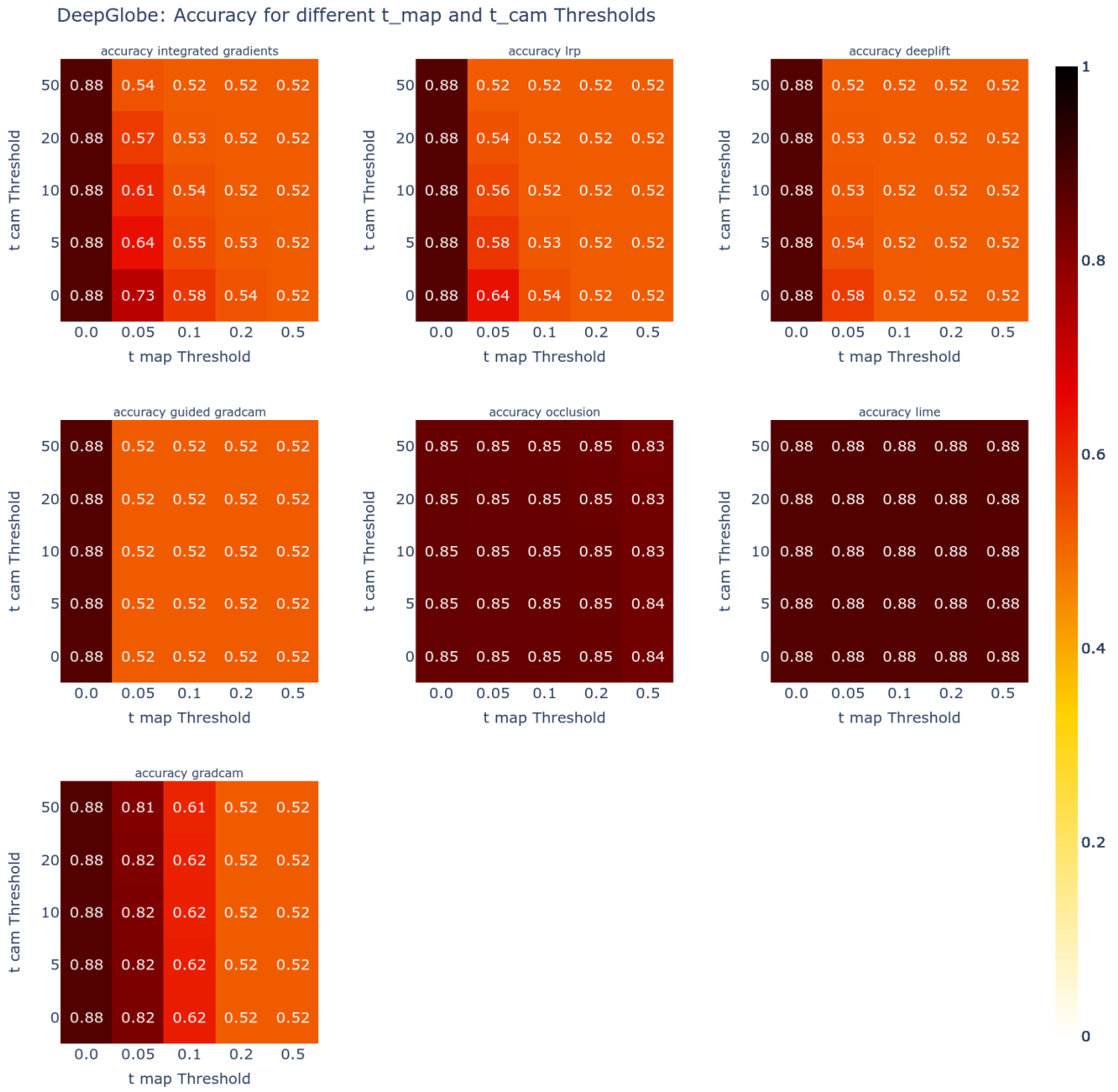


Figure 1: DeepGlobe: Accuracy of the "true" new labels using Label Propagation from reference maps and the labels approximated explanation maps of different explanation methods for different thresholds: t_{cam} and t_{map}

- **Selectivity**:
  - `Patch size = int(height / 4)`
  - `Perturbation Baseline = black`
- **Region Perturbation MORF**:

    - `Patch size = int(height / 4)`
    - `Number of Regions = 10`
    - `Perturbation Baseline = black`
- **Region Perturbation LERF**:
    - `Patch size = int(height / 4)`
    - `Number of Regions = 10`
    - `Perturbation Baseline = black`

Robustness

- **Maximum Sensitivity**:
    - `Number of Samples = 5`
    - `Perturbation std = 0.2`
    - `Perturbation mean = 0.0`
    - `Perturbation type = Uniform Noise`
    - `Similarity Function = Difference`
- **Relative Input Stability**:
    - `Number of Samples = 5`
    - $\epsilon_{\text{min}} = 1e - 6$
- **Relative Output Stability**:
    - `Number of Samples = 5`
    - $\epsilon_{\text{min}} = 1e - 6$

Randomisation

- **Random Logit**:
    - `Similarity Function = SSIM`

Localisation

- **Pointing Game (PG)**: No additional hyperparameters used.
- **Top-k Intersection (TKI)**:
    - `k = 100`
- **Relevance Mass Accuracy (RMA)**: No additional hyperparameters used.
- **Relevance Rank Accuracy (RRA)**: No additional hyperparameters used.
- **Attribution Localisation**: No additional hyperparameters used.

Complexity

- **Sparseness (SP)**: No additional hyperparameters used.
- **Effective Complexity (ECO)**:
  - $\epsilon = 0.3$

These metrics, as implemented in the Quantus framework, are reproducible using the specified parameters and the default settings provided by the framework. All metric-specific hyperparameters have been listed to ensure clarity and reproducibility.

4 Quantitative Results

Efficiency of Metrics

When evaluating methods and metrics, also the efficiency plays an important role. Figure 2 illustrates the average seconds per sample per explanation method. As shown in the figure, the rows represent the explanation methods, while the descriptions at the top of the columns indicate the category of the metrics and the descriptions at the bottom show the names of the metrics. The majority of calculations occur in constant time, indicating efficiency and predictability in computational performance.

However, certain metrics, such as Monotonicity and Selectivity, deviate from this norm, each requiring approximately 60 seconds to compute. Furthermore, the performance of metrics in the robustness and randomisation category is heavily influenced by the performance of the explanation method, as these tests assess the stability and consistency of explanations under varying conditions and inputs. Among the various explanation methods employed, Occlusion stands out as notably slower in these categories. Occlusion is a method that systematically obscures parts of the input data to determine their impact on the output. As a result, it is computationally intensive, requiring multiple iterations of the model for each occluded version of the input, which significantly increases the overall computation time.

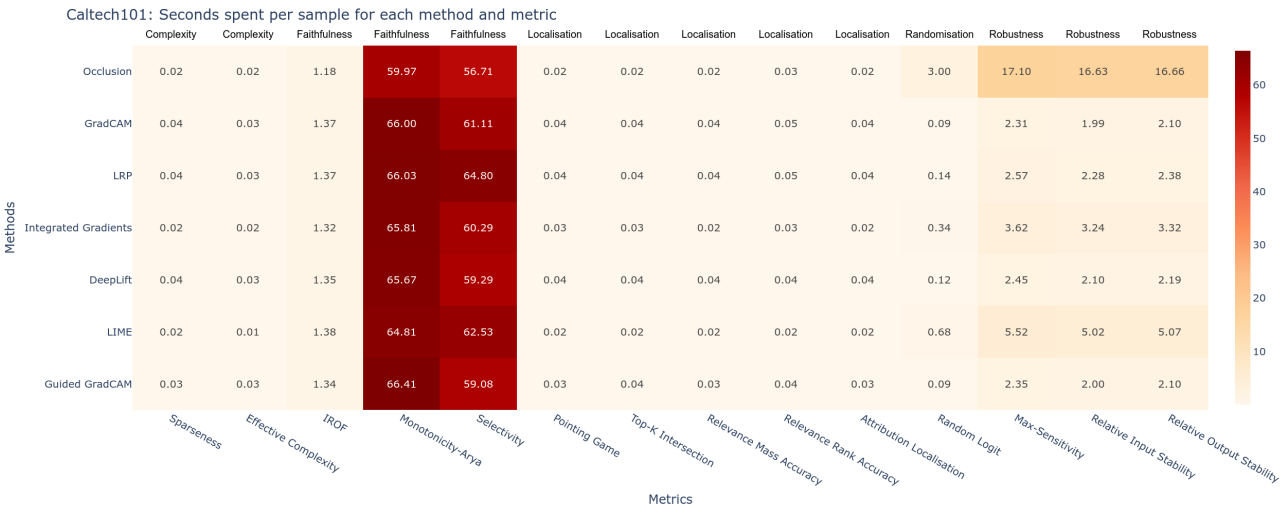


Figure 2: Caltech101: Seconds spent per sample for each method and metric

One possible reason for the decrease per explanation visible in Figure 3 is that the image size for DeepGlobe is smaller than for Caltech101 (see Table 6.1). However, it has to be taken into account that the table visualizes the time per class-wise explanation. This means that for the DeepGlobe data set where we have 1.71 labels per sample, the time has to be multiplied by this amount.

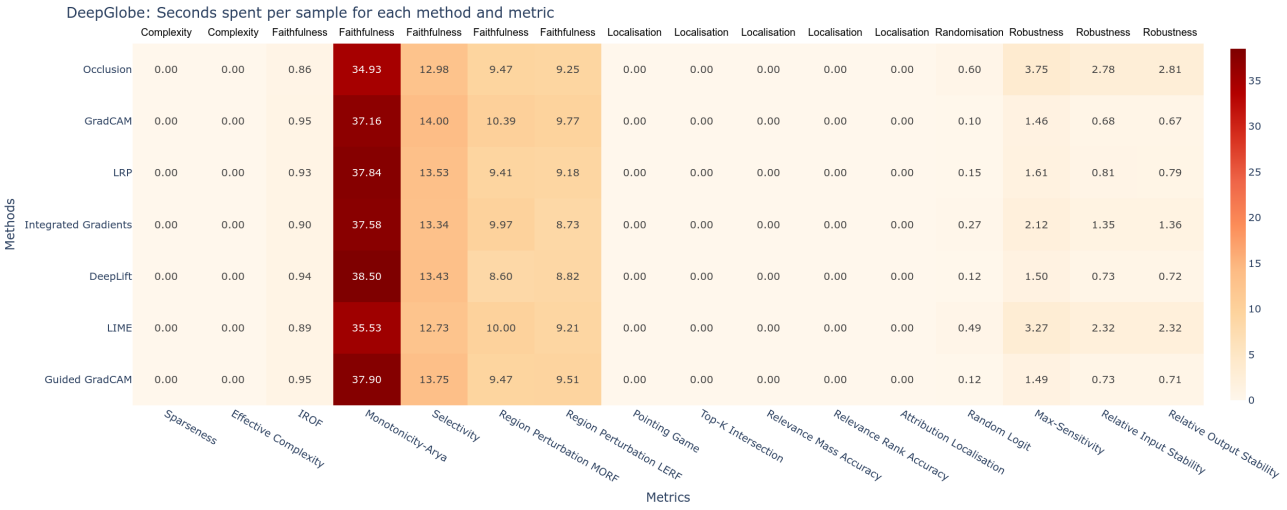


Figure 3: DeepGlobe: Seconds spent per sample for each method, metric and class

Similar for the BEN dataset, visualized in Figure 4, where there are even 3.94 average labels per sample. Thus, for MLC data the efficiency of the explanation method is even more important. Notably, Occlusion is much slower for BEN dataset. This is due to its sliding window approach.

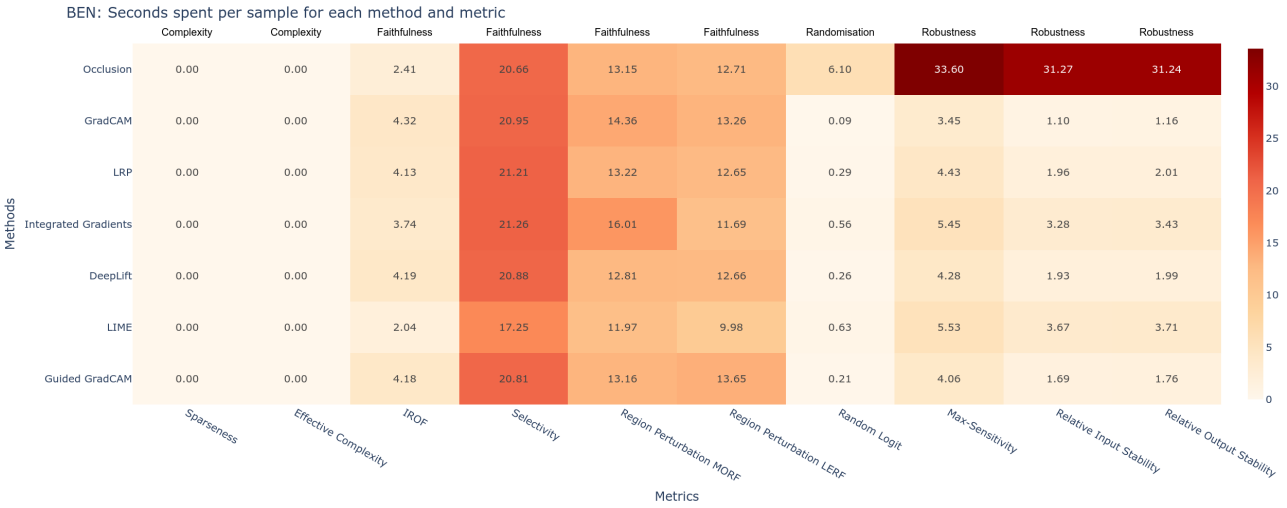


Figure 4: BEN: Seconds spent per sample for each method and metric

5 Detailed Correlational Analysis

Figures 5,6, 7, 8, 9 and 10, present the detailed correlations of individual xAI metrics with the training success for each xAI-guidance method, dataset and model. Each subplot represents a different xAI metric, with the y-axis showing the metric performance and the x-axis showing the training metric. The colour of the markers indicates the explanation method, and the shape of the markers indicates the parameters for the xAI-guidance method. The training metric displayed is the maximum performance from the hyperparameters. The title of each subplot includes the metric name, the calculated Pearson correlation, and the corresponding p-value.

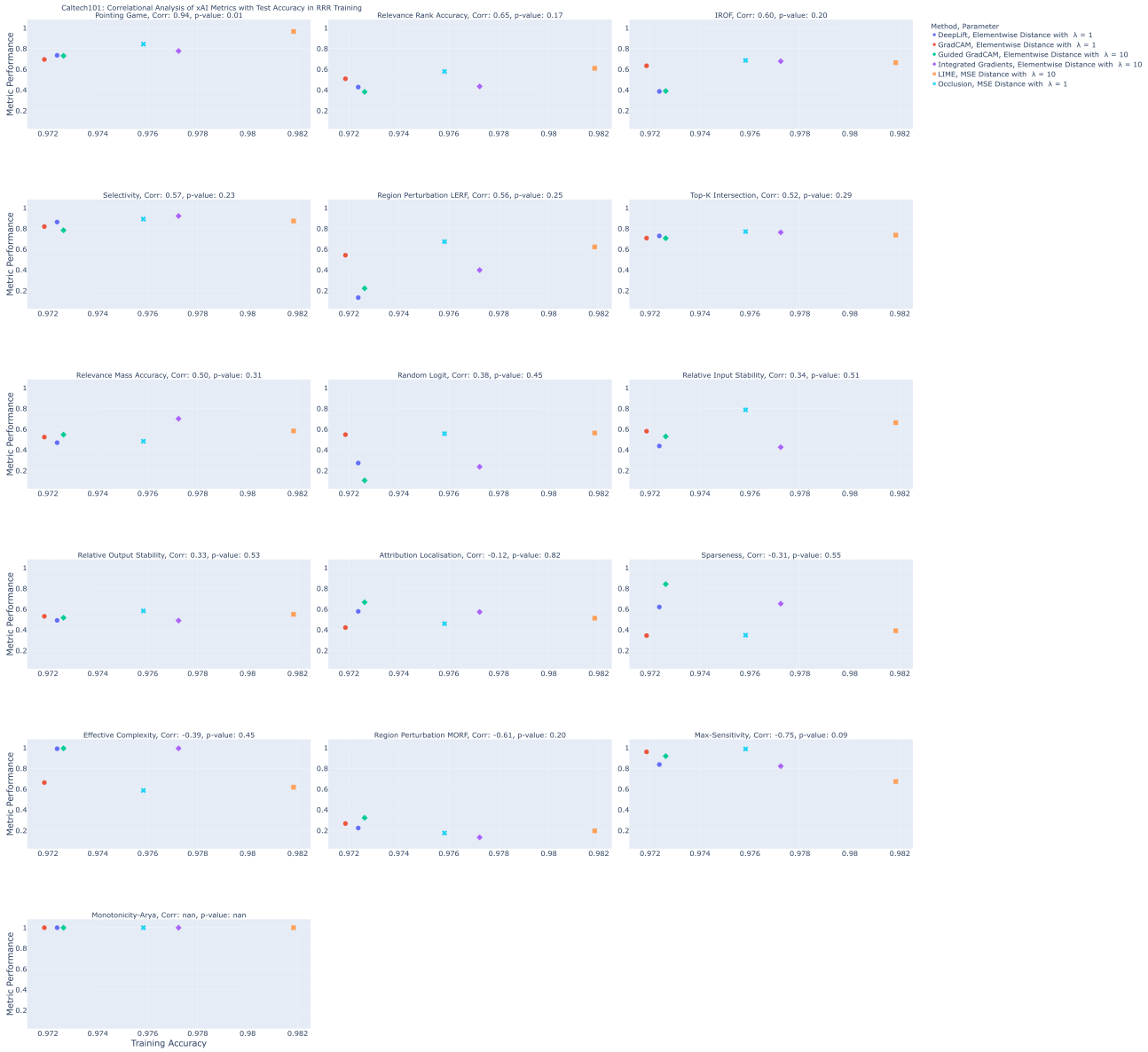


Figure 5: Caltech101 VGG: Correlation analysis between xAI metrics and xAI guided training using RRR. (Test Accuracy)

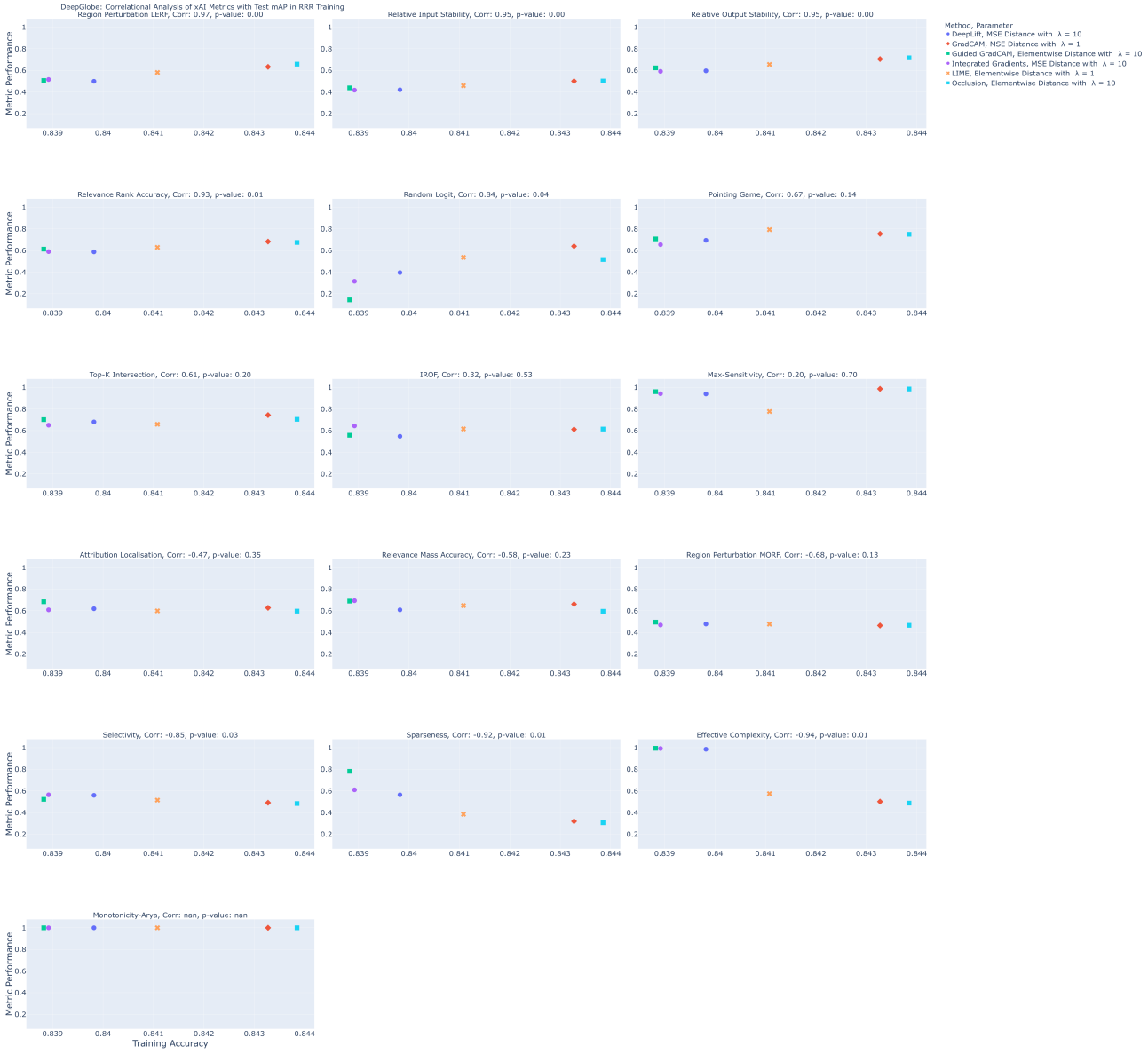


Figure 6: DeepGlobe, VGG: Correlation analysis between xAI metrics and xAI guided training using RRR. (Test mAP)

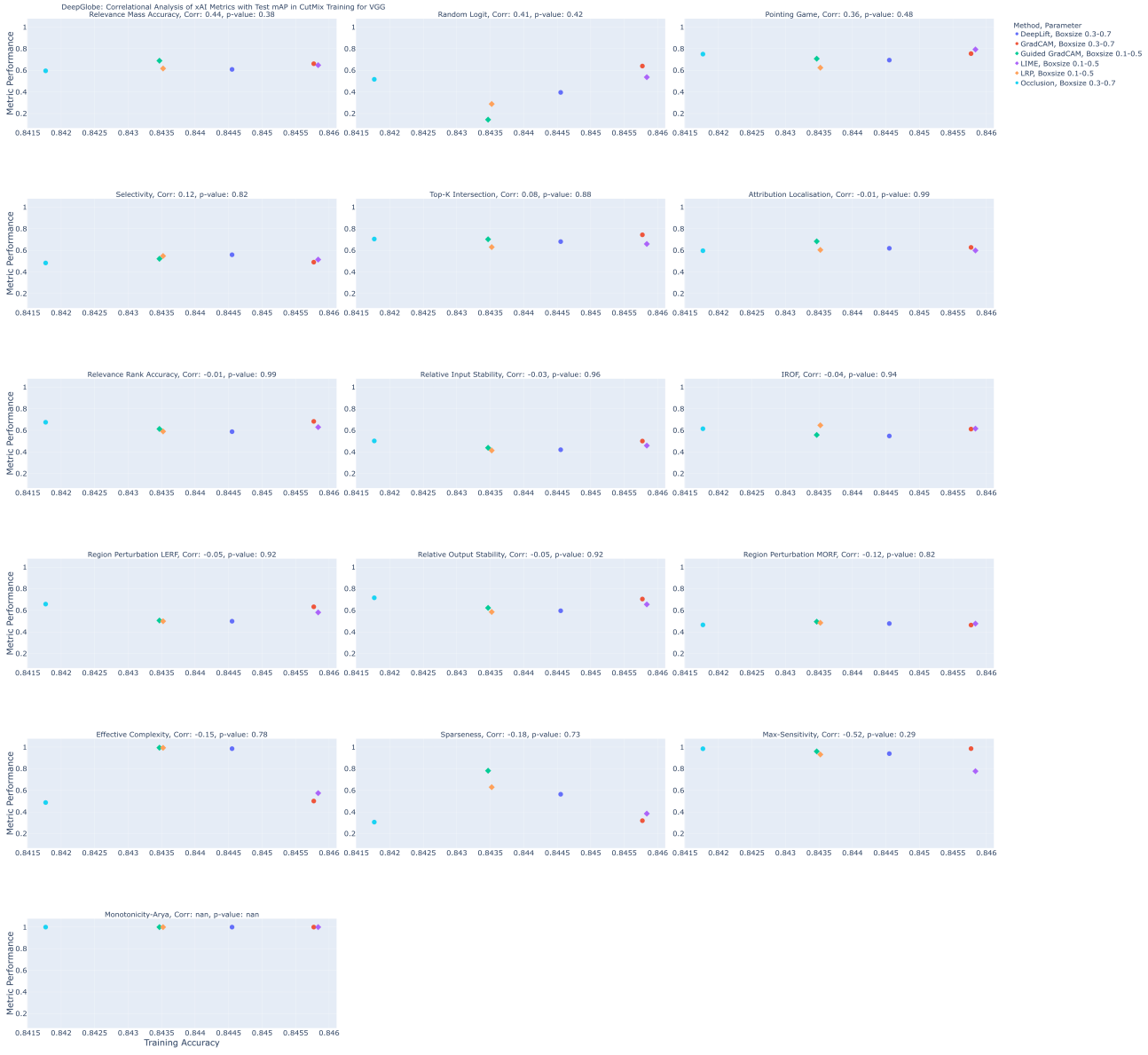


Figure 7: DeepGlobe, VGG: Correlation analysis between xAI metrics and xAI guided training using CutMix. (Test mAP)

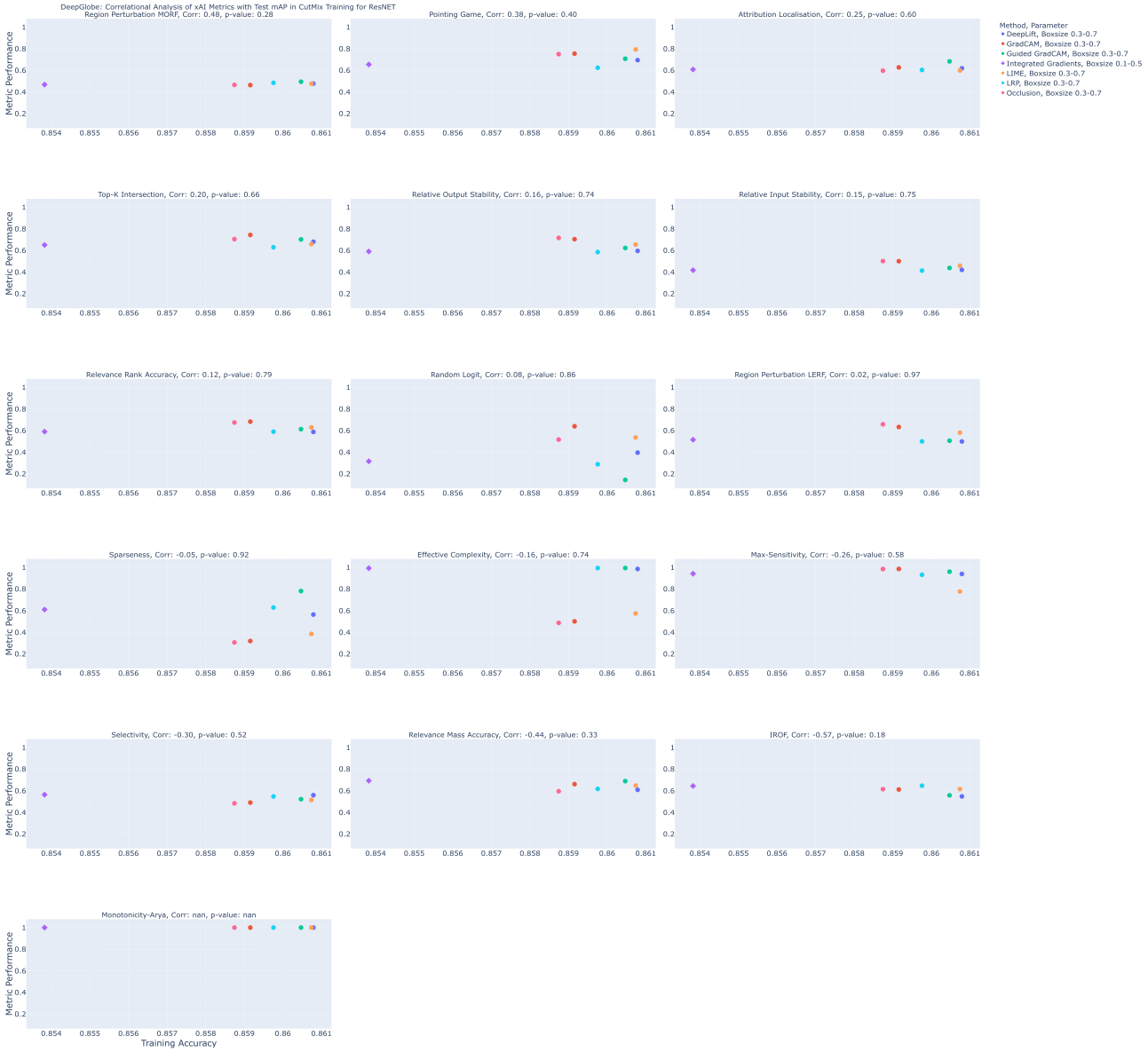


Figure 8: DeepGlobe, ResNET: Correlation analysis between xAI metrics and xAI guided training using CutMix with xAI LP. (Test mAP)

Figure 9: BEN ResNET: Correlation analysis between xAI metrics and xAI guided training using CutMix with xAI LP. (Test mAP)

Figure 10: BEN VGG: Correlation analysis between xAI metrics and xAI guided training using CutMix with xAI LP. (Test mAP)