Interpretability Analysis in the Classification of Nature-Related Images on Social Media

Paula Feliu Criado

June 30, 2025

Abstract

Urbanization and screen-based lifestyles are thought to erode the nature values that underlie environmental concern, yet social media's role in fostering these values remains largely overlooked. To explore this potential, we developed a multitask AI pipeline fine-tuned on manually coded X posts. Built on a backbone, the model classifies images across four tasks: Nature/Non-Nature, Biotic/Abiotic, Material/Immaterial, and Landscape Type. Of four interpretability methods, Grad-CAM and LIME showed the most human-aligned heatmaps, remaining robust to perturbations. Binary tasks focuses on faces or animals, while landscape predictions drew on broader context. EfficientNet-B0 matched ResNet-18's accuracy but emphasized textures, showing that architecture shapes explanation quality as much as metrics. The pipeline preselects relevant social-media content, reducing workload and analyzing DRVs at scale to design online experiences that promote real-world environmental stewardship.

Keywords: Digital Relational Values (DRV), Human-nature connection, Social-media imagery, Transfer learning, Multitask image classification, Explainable AI, GradCam, LIME.

1 Introduction and Context

The rapid interweaving of everyday life with digital infrastructures has created a human–digital nexus—a continuous feedback loop in which meanings, values, and behaviors flow between on-screen representations and off-screen realities. Understanding this nexus is not just a technical challenge for computer science; it is a pressing societal issue. We need data-intensive methods and analytical tools that reveal how online interactions reshape our relationships with the physical world.

One of the most concerning manifestations of this shift is its impact on our connection to nature. Urbanization and digitization have weakened direct interactions with the environment, leading to a disconnect and a decline in nature values (Miller, 2005). This "extinction of experience" describes a cycle where reduced exposure to nature across generations erodes familiarity, appreciation, and engagement (Soga and Gaston, 2016).

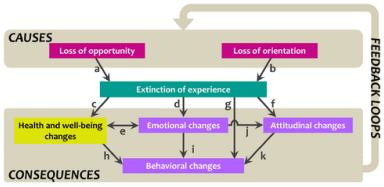


Fig. 1: "Extinction of Experience". (Soga and Gaston, 2016)

Soga and Gaston (2016) (Figure 1) illustrate how stressors such as climate change, urban living, and increased screen time drive a feedback loop that diminishes opportunities and motivation to engage with nature, further reducing environmental concern and stewardship.

Psychological studies confirm that reduced contact with nature not only diminishes environmental concern but also shifts individual values toward self-centered aspirations, consequently impacting generosity, community involvement, and overall social connectedness (Weinstein et al., 2009). Soga and Gaston (2016) compare rates of children's outdoor activities at different points in time and highlights a marked decline in participation and time spent outside. In a related vein, they demonstrate the downward trend in per capita visits to natural areas, underscoring the broader societal shift away from meaningful nature experiences.

However, recent work hypothesizes that virtual interactions, primarily on social media, not only can compensate for lost physical contact with nature but also strengthen environmental appreciation, care, and stewardship and may even motivate real-world conservation behaviors (Calcagni

• Contact E-mail: 1630423@uab.cat
• Supervised by: Johannes Langemeyer & Ramin Solyemani
• Academic Year 2024/25

et al., 2019; Langemeyer and Calcagni, 2022). Introducing the core concept of Digital Relational Values (DRV) defined as the nature-related values and attachments that arise within online communities, Langemeyer and Calcagni (2022) suggest a research agenda that explores how digital platforms can foster human–nature connections by offering virtual encounters with the natural world. Yet the field remains largely theory-driven; few studies leverage large-scale, data-driven evidence to examine how DRVs form, evolve, and spill over into offline action.

This bachelor's thesis addresses that gap. It is embedded within this research endeavor, more precisely within the ERC-project BIG-5 (https://big-5.eu/), and supports that framework by deepening the understanding of how DRVs form and evolve in digital contexts. In support of the BIG-5 research objectives, this thesis project designed and implemented a first-stage filtering module that automatically extracts relevant nature-related content from social media. It thereby aims at increasing the research efficiency and reducing the reliance on time and work intensive manual annotations.

The project uses artificial intelligence (AI) to analyze social media images (Väisänen et al., 2021). Applying AI classification techniques is ultimately supposed not only to identify but also to understand DRVs. As a first step toward this ultimate objective, this study conducts an exploratory analysis, based on a representative data sample, to automatize the assessment of nature elements, human–nature interactions, and emotional context. Because trust in these insights hinges on transparency, techniques as Grad-CAM and LIME will be applied to explain which visual features drive the model's decisions, aligning algorithmic outputs with human perception and strengthening the bridge between digital and real-world nature experiences.

2 Research Objectives

Following this overarching goal, this project is divided into two main objectives, each playing a crucial role in achieving the overall goal of understanding and explaining the decision-making process of deep learning models in the context of nature-related image classification.

This project focuses on the understanding of how DRVs emerge and can be explained through AI-driven analysis of nature-related social media content. To achieve this, the work is organized into two sequential phases: building a robust, transparent image classification system for extracting DRV relevant information; and applying interpretability techniques to align model decisions with human perceptions.

The specific objectives of the project are:

1. **Automated Extraction of Nature-Related DRVs**: Develop a deep learning–based image classification framework that can reliably identify and categorize nature-related content in large social-media datasets, thereby creating an evidence-based, reproducible pipeline for detecting Digital Relational Values in online images.

2. **Explainability and Human-Aligned Interpretation**: Evaluate and compare state-of-the-art interpretability techniques to reveal which visual features drive the classifier's decisions and fostering transparency and trust in DRV analytics.

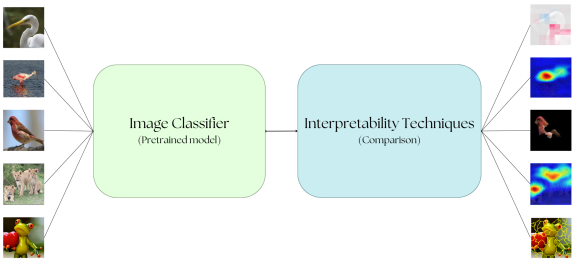


Fig. 2: Approach

2.1 Development of an Image Classification System

The development of an image classifier system will allow to categorize nature-related images into predefined classes, including:

1. **Nature vs. Non-Nature**: Identifying whether the content represents nature or not.

2. **Biotic vs. Abiotic**: Distinguishing between images containing living organisms (biotic) and non-living elements (abiotic).

3. **Material vs. Immaterial** : Classifying the content as either physically tangible (material) or abstract/intangible representations (immaterial).

4. **Landscape Type** : Categorizing images into predefined landscape types (artificial surfaces, agricultural areas, forests and semi-natural areas, wetlands, water bodies, other or none of these).

Each image will be assigned one and only one label per class, ensuring exclusivity within each category. This approach follows a multiclass classification structure, where the classes are mutually exclusive, but independent of one another.

To develop the image classifier, deep learning techniques will be employed, particularly Convolutional Neural Networks (CNNs), which are highly effective for image classification tasks. The model will be trained to predict the most likely label for each of the four categories based on the visual content of the images.

2.2 Interpretability Techniques

The second objective of the project focuses on the interpretability of the AI approach, addressing these research questions.

1. What regions of the image does the model consider most relevant for classification?

2. How do these regions vary across different interpretability techniques?

3. How does the model's classification criteria compare to human classification criteria?

To achieve this, the trained classifier will be used as a foundation, and various interpretability techniques will be applied to analyze the model's decision-making process. Techniques such as Grad-CAM and LIME will be used to highlight the regions of the images that are most influential in the classifier's predictions.

The goal is to evaluate how effectively these interpretability methods can explain the model's outputs and how well they align with human classification criteria. By comparing the results across these techniques, the study aims to identify which methods provide the most transparent insights into the model's behavior.

3 State of the Art

3.1 Image Classification and Transfer Learning

Image classification with deep learning has advanced through pretrained models and transfer learning. Pretrained networks, trained on large datasets like ImageNet, learn versatile features, from simple edges to complex object parts. Reusing these models for new tasks provides a strong initialization, cutting down on task-specific data needs and speeding up convergence (Bengio, 2012).

Transfer learning adapts a pretrained model to a new domain by using it as a fixed feature extractor and retraining or fine-tuning its higher layers. Low-level filters (e.g., edge detectors) remain mostly unchanged, while deeper layers adjust to the target data. This method excels when labeled data are scarce, delivering high performance in specialized applications (Pan and Yang, 2010).

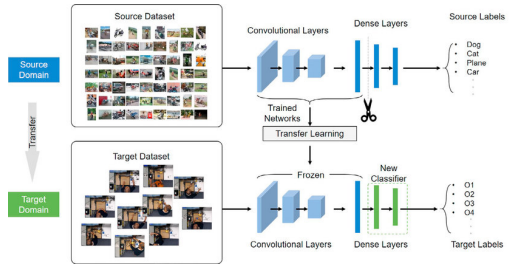


Fig. 3: Transfer Learning Schema (Kumar, 2024)

Several landmark models underpin modern image classification through transfer learning such as **AlexNet** (Krizhevsky et al., 2012), **VGGNet** (Simonyan and Zisserman, 2015), **ResNet** (He et al., 2016) and **GoogLeNet** (Szegedy et al., 2015).

By leveraging these pretrained backbones, researchers can quickly fine-tune models for specialized tasks, cutting down on both computational cost and training time, especially valuable when labeled data are scarce, as is the case in the BIG-5 projects, characterized by the specific annotation framework of DRV.

3.2 Interpretability Techniques in Classification Models

Despite high accuracy, CNNs often operate as "black boxes," prompting the need for interpretability methods to reveal decision-making processes. This is particularly important to facilitate the integration of human and machine interpretation, and specific interpretation objectives.

3.2.1 Class Activation Mapping (CAM) and Grad-CAM

CAM and Grad-CAM are widely used for visually interpreting CNN decisions by generating heatmaps highlighting the regions influential for predictions (Selvaraju et al., 2017). Grad-CAM uses gradients from the final convolutional layer to identify discriminative regions, effectively aligning the model's focus with human attention. Grad-CAM's effectiveness in revealing biases or dataset anomalies has made it essential for transparent AI evaluations.

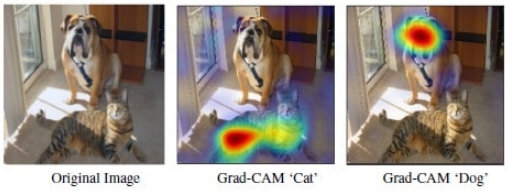


Fig. 4: Grad-CAM. Adapted from Selvaraju et al. (2017).

3.2.2 Local Interpretable Model-Agnostic Explanations (LIME)

LIME provides local, detailed explanations by segmenting images into superpixels and evaluating their influence on model predictions through perturbations (Ribeiro et al., 2016). Despite its granularity, LIME faces stability and variability challenges, though it remains valuable for individual prediction explanations.

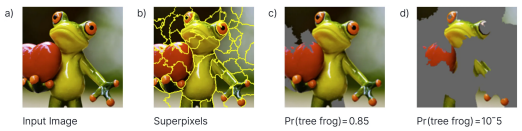


Fig. 5: LIME. Adapted from Ribeiro et al. (2016).

3.2.3 SHapley Additive exPlanations (SHAP)

SHAP uses game theory to quantify pixel-level feature contributions to predictions, ensuring consistency and robustness in interpretations (Lundberg and Lee, 2017). Its unified approach integrates local and global interpretability, making it effective for validating model behavior.

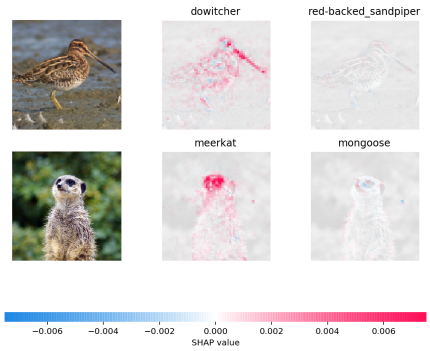


Fig. 6: SHAP. Adapted from Lundberg and Lee (2017).

3.2.4 Occlusion Sensitivity

Occlusion Sensitivity is a perturbation-based interpretability technique for CNN image classification, originally introduced by Zeiler and Fergus (2014). The approach systematically occludes (masks or replaces) different portions of an input image to measure changes in the model's prediction confidence. If covering a particular region causes a significant drop in the target class probability, that region is inferred to be important for the model's decision. They used this method to verify that their ImageNet CNN was focusing on the object itself rather than spurious background context.

Fig. 7: Occlusion Sensitivity. Adapted from Zeiler and Fergus (2014).

This method is model-agnostic and intuitive, though it requires many forward passes (one per occluded position) and its resolution is limited by the occlusion window size. For example, in their results occluding a dog's face drastically reduced the "Pomeranian" class score (while occluding background had little effect), indicating that the face region was crucial for the prediction.

3.2.5 XRAI

XRAI (Kapishnikov et al., 2019) is a region-based attribution method that produces more coherent visual explanations for CNN image classification. Unlike pixel-level saliency approaches, XRAI builds upon Integrated Gradients by aggregating attributions over meaningful image segments. The algorithm first over-segments the input image (e.g., into superpixels) and computes an attribution score (via integrated gradients) for each segment. It then iteratively adds the most salient segments, merging them into larger highlighted regions that strongly influence the prediction. This yields saliency maps that emphasize entire objects or important parts of the image, rather than scattered individual pixels.

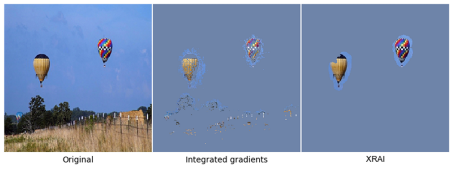


Fig. 8: XRAI. Adapted from Kapishnikov et al. (2019).

All figures presented in this section are adapted from the official papers of each method.

In sum, this work shows that transfer learning on ImageNet backbones is the most efficient route to a strong nature-image classifier, allowing the project to concentrate on the second question of *why* and letting the model decide as it does. The literature also makes clear that no single interpretability technique is definitive: Grad-CAM, LIME, SHAP, Occlusion Sensitivity, and XRAI each expose a different layer of the model's reasoning, from coarse class-discriminative regions to fine-grained, theoretically grounded attributions.

Building on these insights, the forthcoming steps will involve fine-tuning a pretrained CNN and analyzing the different methods to establish practical guidelines on how each technique can be trusted. In this way, the state of the art not only informs but actively structures the experimental workflow that follows.

4 Methodology

The project is structured into clearly defined phases over approximately five months, each designed to ensure a comprehensive approach from research through analysis. The timeline and tasks for each phase are summarized below and visually represented in A.1, Figure 23.

Literature Review *(Weeks 1-3, March)*: Conduct a targeted literature review on image classification and interpretability to build a solid theoretical base.

Data Acquisition and Processing *(Week 2 of March – Mid-April)*: Collect, label and clean a diverse nature-themed image dataset from Platform X to ensure high quality and representativeness.

Classifier Development *(End of March – April)*: Fine-tune pretrained CNNs, including hyperparameter tuning to create an accurate classifier.

Interpretability Techniques *(Mid-April – Mid-May)*: Apply different techniques to visualize and dissect the model's decisions.

Results Analysis *(May – June)*: Analyze and compare the outputs, draw conclusions, and prepare a final report and presentation summarizing findings and future directions.

This integrated approach ensures a logical progression through each stage, optimizing both methodological rigor and effective time allocation.

4.1 Data Collection and Dataset Creation

Creating a comprehensive and high-quality dataset tailored specifically to the project's objectives involved meticulous planning, execution, and rigorous standardization. In contrast to state-of-the-art projects that rely on publicly available datasets, this dataset is manually created and built entirely by BIG-5 team. This approach ensures that data precisely captures the diversity and authenticity of social media content related to nature interactions. The entire process was organized into three key stages: data collection, data labeling, and protocol implementation.

4.1.1 Data Collection

The dataset consists exclusively of visual content sourced from social media, specifically from the platform X. While more data were originally collected, only the subset suitable for training the CNN was used, approximately 1700 high-quality images. To ensure diversity and temporal balance, a stratified sampling strategy was applied across different time periods, and posts were selected randomly without the use of keyword-based searches.

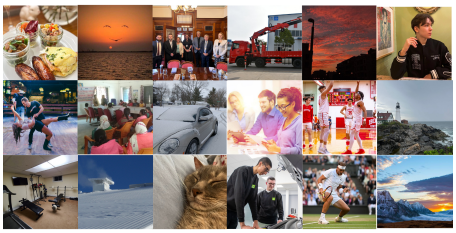


Fig. 9: Examples of images in the dataset

4.1.2 Data Labeling

Once the images were collected, the next critical step involved labeling them according to defined classification tasks. To achieve consistency and accuracy in this process, the labeling was performed using LabelStudio, a specialized tool designed for structured and reproducible labeling of image data. Each image was carefully examined and assigned labels across four specific, mutually exclusive categories:

1. **Nature vs. Non-Nature**
2. **Biotic vs. Abiotic**
3. **Material vs. Immaterial**
4. **Landscape Type**

To ensure high standards and reliability, all coders were required to agree on the assigned labels, as this step is critical for data consistency, reduced ambiguity, and accurate model training. Importantly, only images depicting nature were further classified into the additional three categories, while non-nature images were excluded from these classifications. The final dataset includes a unique identifier per image and the full set of applicable labels, ensuring traceability, consistency, and robustness for downstream modeling and evaluation.

4.1.3 Protocol Implementation

To ensure clear and consistent labeling, the BIG-5 group established a detailed coding protocol to standardize image classification.

Thorough, unambiguous guidelines minimize labeling errors and ensure consistency across coders. Without them, inconsistent labels degrade classifier accuracy, compromise interpretability analyses, and undermine the project's validity.

4.2 Image Classifier Development

This section describes the design, customization, and fine-tuning of the multitask image classification system. Because the classification tasks are nuanced and labeled data are limited, a flexible framework was developed that integrates several state-of-the-art pretrained models (Convolutional neural network, CNN). This section outlines the key design choices and justifies the selection of these models as robust feature extractors for the project.

4.2.1 Pretrained Models Selection

For this approach, a custom backbone module has been implemented that allows the selection among four widely recognized pre-trained architectures:

DenseNet121: Characterized by dense connectivity between layers, DenseNet121 helps in better gradient flow and feature reuse. This architecture is especially advantageous when training data are limited because its design facilitates learning more diversified features with fewer parameters. (Architecture in Figure 24a).

ResNet18: A lighter variant of the ResNet family, ResNet18 offers a balance between computational efficiency and representational power. Its reduced complexity makes it suitable for projects with scarce data, as it mitigates the risk of overfitting while still capturing essential visual patterns. (Architecture in Figure 24b).

EfficientNetB0: EfficientNetB0 is known for its compound scaling method, which optimizes depth, width, and resolution in a balanced manner. This results in a model that is both resource-efficient and effective in extracting salient features, making it a strong candidate when using limited datasets. (Architecture in Figure 24c).

ResNet50: Although deeper and more parameter-intensive, ResNet50's robust architecture, utilizing residual connections, has proven successful for large-scale image recognition tasks. Its high-dimensional feature representations can be beneficial in capturing subtle nuances; however, care must be taken to avoid overfitting with scarce data. (Architecture in Figure 24d).

All these models come pretrained on ImageNet dataset, ensuring that the learned features are general enough to transfer effectively to the target tasks. By comparing these options, the workflow facilitates empirical investigation of how the choice of backbone affects not only the classification performance but also downstream interpretability analyses.

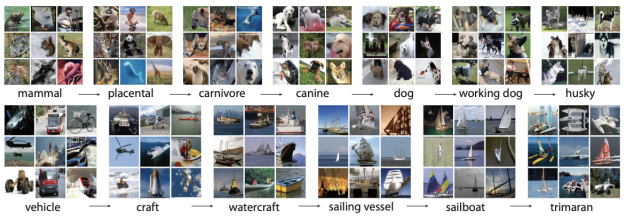


Fig. 10: Imagenet

4.2.2 Dataset Preparation

As it has seen previously, the dataset for training and evaluating the model was created from images collected from the social media X.

All labels were standardized and mapped into numerical format to be compatible with neural network training. A structured data loader was created to efficiently handle and preprocess images, applying standard transformations such as resizing to 224x224 pixels, normalization, and conversion to tensor format.

4.2.3 Multitask Classifier Architecture

The network separates feature extraction from task-specific decisions to handle four simultaneous classification tasks (nature vs. non-nature, materiality, biological content, landscape type):

1. **Feature Extraction:** Remove the final layer of a chosen backbone, letting it output a high-dimensional feature vector (512, 1024, 1280, or 2048 dims, respectively).

2. **Task-Specific Classifier Heads:** From this shared feature vector, the architecture branches into four fully connected heads: *Nature Visual*, *Materiality*, *Biological Content*, *Landscape Type*.

By decoupling these tasks, the model can learn task-specific decision boundaries from a shared feature space, boosting performance in data-scarce settings.

4.2.4 Model Training, Fine-Tuning, and Comparative Evaluation

To make the most of limited data, each multitask backbone end-to-end was finetuned using the Adam optimizer with an adaptive learning rate and a composite, task-balanced cross-entropy loss. Extensive data augmentation (random cropping, flips, color jitter) combats overfitting, while the weighted loss ensures no single task dominates training.

The data were split 90/10 into training and test sets, and task-specific accuracy, loss curves, and confusion matrices were tracked to compare architectures quantitatively.

4.2.5 Backbone Selection

To choose the most suitable model for applying interpretability methods, it has been evaluated not only average performance but also stability across tasks and efficiency (inference time and complexity) with the mean accuracy and weighted F1 scores for each model.

4.3 Interpretability Techniques

Only the techniques that produced meaningful, interpretable results are presented and exlained in this chapter. While several additional methods were implemented during the course of this study, their outputs did not meet the criteria for meaningful analysis and are therefore excluded from the main text. Detailed descriptions of these techniques and observed outcomes, can be found in A.3.

4.3.1 GradCAM

GradCAM (Gradient-weighted Class Activation Mapping) is an interpretability method designed to produce localization maps, highlighting which regions of an input image are most influential for a particular class prediction. It combines the spatial information from deep convolutional layer activations with gradient signals associated with the predicted class, producing a heatmap highlighting important areas.

The implementation involves capturing activations through a forward pass and gradients via a backward pass on a targeted convolutional layer. These gradients are used to weight activations, forming a heatmap that's resized and overlaid onto the original image, providing clear insights into the model's decision-making without altering the existing model structure or training process.

4.3.2 LIME

Local Interpretable Model-agnostic Explanations (LIME) provides human-interpretable, localized explanations for predictions made by black-box classifiers. It works by segmenting the input image into distinct regions called superpixels, then generates variations of the original image by randomly hiding combinations of these superpixels. These perturbed images are evaluated by the classifier, producing predictions that help fit a simplified linear model reflecting how each superpixel influences the final decision.

In the implementation, the classifier was wrapped in a LIME explainer; images were perturbed through standard preprocessing and the resulting explanations highlighted the most influential super-pixels, generating an interpretable visualization clearly illustrating the model's decision-making process.

4.3.3 Occlusion Sensitivity

Occlusion Sensitivity measures the importance of different image regions by systematically occluding (covering) parts of the image and observing how the classifier's predictions change. In practice, a fixed-size patch is slided across the image, temporarily replacing each region with a baseline value, and measure the resulting drop in the predicted class score.

The target model is wrapped, predictions are computed for every occluded image, and the resulting prediction differences are recorded. These importance scores are aggregated into a heatmap that highlights the areas most influential to the model's decision, clearly visualizing regions critical for classification.

4.3.4 XRAI

XRAI (eXplanation with Ranked Area Integrals) enhances pixel-level attributions by producing region-focused explanations of model predictions. It begins by calculating Integrated Gradients, which quantify pixel importance based on gradual interpolation from a baseline image to the input.

These pixel-wise attributions are then grouped into meaningful image segments using a region segmentation algorithm *(Felzenszwalb's method)*. Each segment's importance is computed by averaging its pixel attributions, and the segments are ranked accordingly.

Predictions are isolated through a wrapper around the model, Integrated Gradients are computed, regions are segmented and ranked by importance, and a coherent heatmap is generated to intuitively highlight influential image areas.

4.3.5 Comparison of Interpretability Across Models

For this test, the outputs produced by the explained techniques will be compared on three models: ResNet18, EfficientB0 and DenseNet121.

The evaluation is grounded in human criteria, searching for the salient regions that align with those a human would attend to when classifying the image. This analysis allows to determine not only which model yields the most human-aligned explanations, but also how interpretability performance degrades (or perhaps remains robust) when moving from the top-performing ResNet-18 to EfficientNet-B0 and then to DenseNet-121.

4.3.6 Robustness under Perturbations

In this section, the goal is to evaluate both the prediction robustness of the models and the stability of the interpretability techniques implemented. Controlled perturbations were applied to a set of test images, altering color balance, rotation, and adding filters and then run them through the different techniques. Comparing output images and correlating saliency maps between original and modified inputs assesses the resilience of the methods to common visual distortions. These results will help us understand the reliability of the models and highlight areas for improvement.

5 Results

This section presents the key findings of the image classification study, followed by an outline of the interpretability analyses applied to the trained model. Performance metrics are first reported to establish predictive accuracy and efficiency and then visual explanation methods are introduced to contextualize the decision-making process of the model.

5.1 Image Classifier

Below there are the quantitative results from fine-tuning four pretrained backbones, ResNet-18, ResNet-50, EfficientNet-B0, and DenseNet-121, on the multitask classification tasks.

Model	Test Loss	Nature (%)	Materiality (%)	Biological (%)	Landscape (%)
ResNet18	4.2291	78.92	76.05	78.86	66.68
ResNet50	4.2544	79.04	74.85	77.25	64.07
EfficientNetB0	3.9973	77.25	76.65	75.45	66.47
DenseNet121	3.8308	80.84	78.44	79.04	56.89

TABLE 1: Test Loss and Task Accuracies for Different Pretrained Backbones

Across the four evaluated backbones, test-loss values span a narrow band of 3.8308–4.2544, with DenseNet121 reporting the lowest loss and ResNet50 the highest. Nature-class accuracy ranges from 77.25 % (EfficientNetB0) to 80.84 % (DenseNet121), while Materiality accuracy varies between 74.85 % for ResNet50 and 78.44 % for DenseNet121. In the Biological class, accuracies lie between 75.45 % (EfficientNetB0) and 79.04 % (DenseNet121). Landscape accuracy shows the widest spread: ResNet18 and EfficientNetB0 achieve similar scores around 66.6 %, ResNet50 follows at 64.07 %, and DenseNet121 records 56.89 %.

5.1.1 Backbone Selection

The table below offers a concise overview of each backbone's aggregate performance, reporting mean accuracy alongside the corresponding weighted-F1 score obtained on the test set.

Model	Mean Acc. (%)	Mean F1
ResNet18	73.62	0.7360
ResNet50	73.30	0.7375
EfficientNetB0	73.46	0.735
DenseNet121	73.80	0.725

TABLE 2: Mean accuracy and weighted F1 by model

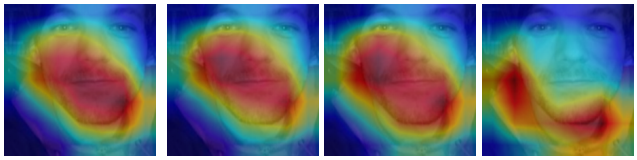
Mean-accuracy values sit in a tight band between 73.30 % and 73.80 %, with DenseNet121 at the upper end and ResNet50 at the lower. Weighted-F1 scores follow a similarly narrow spread, ranging from 0.725 for DenseNet121 to 0.7375 for ResNet50, while ResNet18 and EfficientNetB0 register intermediate values of 0.7360 and 0.735, respectively.

5.2 Interpretability Techniques

The next part of the report shows what the interpretability techniques reveal about the model's decisions. For each method, it will be presented the original and untouched input image. This is followed by four heat-map pictures, one for each task (Nature, Materiality, Biological and Landscape), that highlight the parts of the image the model focused on. Additionally, a few extra tests are included to provide a complete view of how robust the techniques remain across different models and perturbations.

5.2.1 GradCam

Fig. 11: Original image



(a) Nature (b) Biological (c) Materiality (d) Landscape

Fig. 12: GradCAM results

The four Grad-CAM maps reveal distinct attention patterns by task: for the three binary classifiers (Nature, Biological, Materiality), all high-intensity regions align on the man's face—skin texture, hair contours and other facial features. By contrast, the Landscape Type head shifts to background structures, using contextual spatial patterns.

5.2.2 LIME

Fig. 13: Original image

(a) Nature (b) Biological (c) Materiality (d) Landscape

Fig. 14: LIME results

The LIME visualizations for the four heads reveal that each prediction is driven by small, well-delimited clusters of super-pixels. For the Nature head the most influential regions lie on the surrounding grass, with a scattering of additional importance patches on the face and muzzle of the front goat. In the Biological and Materiality heads the highlighted zones trace the full silhouettes of both goats, covering their heads, torsos and legs while leaving the background almost untouched. Finally, the Landscape-Type head ignores the animals altogether and concentrates on broad background areas: the grass and the thin band of trees and sea at the horizon.

5.2.3 Occlusion Sensitivity

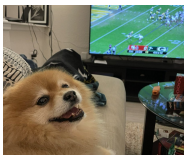
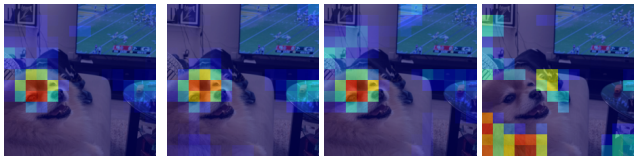


Fig. 15: Original image



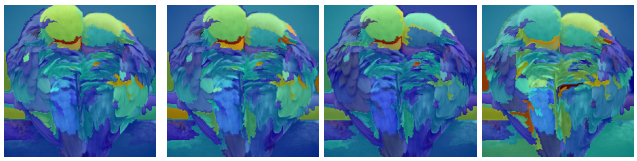
(a) Nature (b) Biological (c) Materiality (d) Landscape

Fig. 16: Occlusion Sensitivity results

The occlusion maps in Figure 16 reveal that masking the dog's face causes the largest confidence drop for the Nature and Biological heads. In Materiality, the most important patch spans the fur and muzzle, but also pays attention to a secondary hotspot on the tabletop. Finally, Landscape Type shifts attention to the background TV and surroundings.

5.2.4 XRAI

Fig. 17: Original image



(a) Nature (b) Biological (c) Materiality (d) Landscape

Fig. 18: XRAI results

Figure 18 presents the XRAI region-based saliency maps generated for the four tasks. In the Nature map, a single contiguous region highlights the birds silhouettes; the Biological map illuminates nearly the same area. The Materiality map highlights the bird's plumage together with the nearby branches and perch, while the Landscape map shows also activated regions in the image corners, with a few small fragments along the borders.

5.2.5 Comparison of Interpretability Across Models

To facilitate clearer analysis, the results are considered in two different groups: the binary tasks—Nature, Biological, and Materiality—and the multiclass task, Landscape Type.

Each group is examined separately before drawing broader comparisons.

Binary "Nature" Task

For binary tasks it will be shown only the Nature task since the others display very similar patterns.

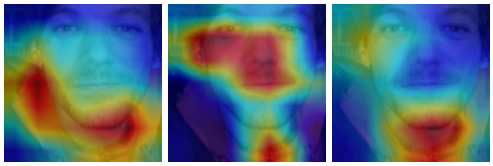
(a) ResNet18 (b) EfficientB0 (c) DenseNet121

Fig. 19: LIME results on Nature

These results in Figure 19, generated by LIME exhibit a high degree of consistency across all three architectures. Each model highlights essentially the same regions: the center of the man's face. This uniformity likely arises from two complementary factors. First, all models achieve similar performance on the Nature task (approximately 75–80 % accuracy), suggesting that they rely on a common set of discriminative features. Second, binary tasks usually hinge on a few key features that any model might detect equally well.

Multiclass "Landscape" Task

By contrast, the multiclass Landscape task interpreted via GradCAM, reveals marked differences in how each model allocates attention:



(a) ResNet18 (b) EfficientB0 (c) DenseNet121

Fig. 20: GradCam results on Landscape

Figure 21 shows that ResNet-18 (highest accuracy) produces heatmaps that most closely align with human expectations, concentrating on the background textures.

EfficientNet-B0 (second-best accuracy) diverges sharply, focusing its attention on the man's face and neck instead of on the background that a human would prioritize.

DenseNet-121 (lowest accuracy) unexpectedly mirrors ResNet-18's focus patterns emphasizing the same background regions despite its overall performance.

5.2.6 Robustness under Perturbations

(a) Rotated (b) Grey-scaled (c) Rotated

Fig. 21: Test images

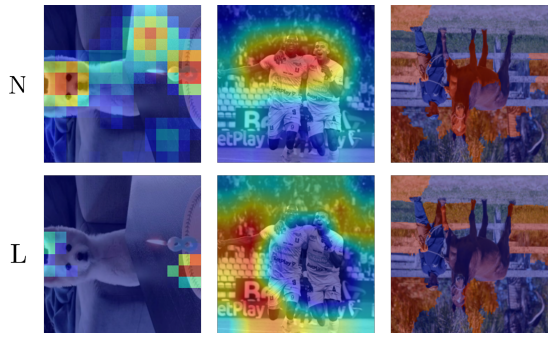


Fig. 22: Results for Nature and Landscape Tasks

After the perturbations, the binary classifier's saliency maps for the Nature task resemble those obtained from the unaltered images shown in Figure 22. Occlusion centers on the rotated dog's face with only faint background activation. Grayscale Grad-CAM concentrates on the football players' faces and torsos, and LIME continues to highlight the rotated horse's head and body contours.

In the multiclass task, perturbations cause only limited shifts. Occlusion on the rotated dog moves attention toward both the candle and the dog's face rather than the broader room. Grayscale Grad-CAM on the football image emphasizes the grass field and background, while LIME on the rotated horse directs almost all importance to the distant trees and sky. Across all tasks, core features such as animal faces, bodies, and field textures remain visible under rotation or grayscale transformation.

6 Discussion

This section interprets the empirical results, relating saliency behaviors and performance metrics to theoretical expectations about model robustness and generalization. Strengths and limitations are examined.

6.1 Image Classifier

Table 1 highlights a clear efficiency–accuracy spectrum across the four backbones. ResNet18, the lightest and fastest, achieves moderate performance on all tasks, making it suitable for quick prototyping or edge deployment but less reliable on fine-grained tasks, especially on biological content.

ResNet50 leverages a richer feature space but comes with a slightly higher loss. It boosts nature task but on other ones falls short of ResNet18.

EfficientNetB0 strikes the best balance: it matches or exceeds ResNet18 on binary tasks, and leads on the multiclass landscape task thanks to its compound scaling with fewer parameters.

DenseNet121 achieves the lowest loss and highest accuracies in three of four tasks due to its dense connectivity and feature reuse, although it underperforms in the landscape task, suggesting that even high-capacity models can require targeted augmentation for fine-grained categories.

This efficiency–accuracy trade-off across backbones aligns closely with established trends in transfer learning. Literature confirms that ResNet18, despite its lightweight structure, often outperforms deeper models on domain-specific tasks with limited data due to better generalization and lower overfitting (Kornblith et al., 2019).

Conversely, EfficientNet-B0—building on compound scaling—achieves strong efficiency and accuracy, as noted in landmark studies. Meanwhile, the best performance from DenseNet121 in your binary tasks echoes findings that dense connectivity helps feature reuse and robustness (Dastour and Hassan, 2023).

Overall, for purely binary distinctions (nature, materiality, biological) the ranking is *ResNet18* $>$ *DenseNet121* $>$ *ResNet50* $>$ *EfficientNetB0*, whereas for the multiclass landscape task it is *ResNet18* $>$ *EfficientNetB0* $>$ *ResNet50* $>$ *DenseNet121*.

6.1.1 Backbone Selection

As shown in Table 2, ResNet50 and EfficientNetB0 deliver nearly identical overall performance, but neither stands out on any specific task. In contrast, DenseNet121 achieves the highest mean accuracy (73.80 %) driven by its strong performance on the three binary classification tasks while ResNet18 combines very competitive mean accuracy (73.62 %) with the best capacity to discriminate fine-grained landscape images.

Therefore, ResNet50, EfficientNetB0, and even DenseNet121 were discarded for interpretability experiments, and ResNet18 was selected as the final backbone. Its balanced and more stable performance across both binary and multiclass classification makes it the best candidate for generating robust, generalizable interpretation maps.

6.2 Interpretability Techniques

6.2.1 GradCam

This technique reveals that overlapping activations in binary tasks indicate feature entanglement, with attention converging on similar facial patterns and even treating the presence of a human subject as a cue for the broader "nature" class—blending living beings with material scene elements and limiting fine-grained distinctions. For landscape classification, the diffuse heat map, despite the correct orientation, still fails to pinpoint key background structures such as trees or buildings. These insights highlight two improvement paths: sharper separation of facial features from other semantic cues in binary tasks and more precise localization of background context, both of which would elevate interpretability and predictive performance.

Consistent with Selvaraju et al. (2017) and recent user-study findings (Kornblith et al., 2019), these Grad-CAM maps reveal a coarse, human-aligned pattern, but still masked biases (e.g. faces conflated with "nature" cues). These limitations align with observations that Grad-CAM may lack precise fine-grained localization.

6.2.2 LIME

These patterns show that the network shifts its focus to suit the semantic scope of each task. When judging Nature, it treats the goat's face and adjacent grass as joint indicators of natural content. In the Biological and Materiality heads the attention collapses onto the goats' entire shapes, which is consistent with tasks that require recognizing living organisms (biological) or tangible elements (materiality). Although both heads highlight the animals, the exact pixels differ, implying that each classifier relies on slightly different visual cues rather than a single, shared template. By contrast, the Landscape head looks past the foreground subjects and relies on the large-scale layout of grass, trees and sea, precisely the contextual information needed to decide scene type.

Our saliency maps align with Ribeiro et al. (2016) original LIME study, which shows the explainer naturally gravitates to the strongest class-discriminative superpixels—typically faces or other central objects—mirroring our focus on goat heads and full silhouettes.

6.2.3 Occlusion Sensitivity

Occlusion sensitivity reveals class-specific dependencies that align with intuitive indicators. In the Nature and Biological heads, masking the dog's face triggers the largest confidence drop, confirming that both heads hinge on the central animal. For Landscape Type, the greatest decline appears when the background TV and surrounding room are occluded, indicating that this head leans on broader environmental context rather than the subject itself. In the Materiality head, suppressing the fur and muzzle lowers confidence most, yet a secondary hotspot along the tabletop edge also proves influential—showing that the model contrasts soft fur with the hard surface to judge material class. Together, these patterns validate occlusion as a perturbation-based check: binary heads concentrate on the animal's face, the landscape head on scene cues, and the materiality head on textural contrasts. They also suggest that using smaller or adaptive patches would yield crisper importance maps for finer structures.

These observations, dominant signal drops when occluding central features, match studies on max-sensitivity confirming that Occlusion is robust in coarse attribution, though slow and low-res resolution (Höhl et al., 2024).

6.2.4 XRAI

The shared highlight over the birds in both the Nature and Biological maps suggests that the model relies on the same "animate form" region for these two tasks. Materiality's focus on both feathers and wooden branches indicates that

texture details and surrounding context jointly drive material judgments. The Landscape map's emphasis on the outer zones points to a dependence on broader environmental patterns, yet the additional edge fragments hint at an over-segmented response. The near-identical Nature and Biological maps further indicate that separate tuning may be needed to tease apart closely related concepts.

The aggregation of segment-level attributions into coherent regions matches XRAI's design. The explained findings, shared regions between "nature" and "biological", reinforces the necessity of task-specific fine-tuning to disentangle semantically close classes.

6.2.5 Comparison of Interpretability Across Models

Across the Nature task, the Grad-CAMs produced by ResNet-18, EfficientNet-B0 and DenseNet-121 look strikingly alike. As Nature task is binary (if the content on the image represents nature or not), all three architectures can reach high confidence by focusing on to the same regions.

The Landscape task seems to be harder: it is multiclass and demands fine-grained discrimination among several scene types. Here the networks' architectural differences become prominent, so their heat-maps diverge. Each model settles on different context clues reflecting the specific features its layers have become sensitive to.

The differences in how each model allocates attention can be explained by their difference in architectures. In ResNet-18, the use of straightforward residual blocks creates an equilibrium between low-level features such as edges and textures and high-level abstractions like shapes and semantic regions. As a result, its Grad-CAM visualizations often mirror the way a human examines an image: outlining contours then shifting to semantically rich areas like faces or objects.

By contrast, EfficientNet-B0's architecture is built around a compound scaling strategy that simultaneously increases network width, depth, and input resolution, while leveraging mobile convolutions to keep the overall parameter count low. Although this makes the model highly efficient, it also encourages attention to very fine-grained patterns that may boost test performance but diverge from what a human would consider meaningful.

DenseNet-121's dense connections reuse features across multiple depths. These connections consistently reinforce the same visual cues such as background or secondary regions, resulting in heatmaps that appear particularly stable and coherent to a human.

Overall, interpretability quality does not always track classification performance. These observations suggest a decoupling between raw classification accuracy and the human-aligned quality of explanations: despite its lower accuracy, DenseNet-121 generates heatmaps that may appear more "intuitive" than those of EfficientNet-B0. Consequently, it is important to evaluate both metrics when selecting models for applications requiring human-aligned explanations. These results and conclusions are consistent with recent analyses highlighting this decoupling (Kansal et al., 2024).

6.2.6 Robustness under Perturbations

The robustness of these attention patterns after rotation and grayscale conversion suggests that the model's internal representations are largely insensitive to simple geometric or color changes. Grad-CAM and LIME retain spatial focus on cues that align with human criteria, reinforcing their usefulness as post-hoc explanation tools under moderate perturbations. Occlusion, however, shows a greater dependence on scene context: its more diffuse activation on the rotated dog in the multiclass setting implies that geometric distortions can disrupt its reliability.

Overall, the stability of Grad-CAM and LIME across tasks supports their suitability for interpretability and it aligns with studies demonstrating stability of backpropagation-based methods under image corruptions (Ihongbe et al., 2024), whereas the behavior of occlusion highlights the need for caution when applying it to images in which contextual information is critical.

7 Limitations

Despite the successful completion of this project, several constraints should be acknowledged. The main limitation of this project was the scarcity of data since the collected dataset was limited in size, and some images did not meet the quality standards necessary to fully achieve the study's objectives. Additionally, classes were unevenly represented, leading to an imbalance that may have affected model performance. Finally, a small number of coders worked on labeling, which could introduce minor inconsistencies, especially for cases where category boundaries are subtle, possibly affecting the model's performance.

8 Further Work

To build on this project, data augmentation can mitigate the limited dataset size and introduce variability even though it is assumed to collect more data as the project continues. Ensuring balanced class representation is also vital for developing robust, generalizable models. To validate the hypotheses presented in this report regarding human visual attention, it is necessary to review or conduct studies identifying where coders focus when labeling; the results would confirm (or refute) these proposed criteria.

Future efforts should adopt a targeted data collection strategy rather than random sampling, getting specialized images and using the current model as a prefilter before experts validate its predictions; this approach will reduce manual workload while expanding a high-quality, balanced dataset. Finally, applying the interpretability techniques to social media, by examining how and where nature appears, can connect model insights with real-world exposure patterns. Together, these steps will refine model performance and broaden the project's impact into studies of visual attention and nature exposure on social media.

9 Conclusions

This project set out to build and validate an AI-powered preprocessing framework that prepares social-media images

and, crucially, whether those techniques can be transparent enough for researchers to trust the resulting insights. To that end it has been (i) built a multitask image-classification pipeline that filters nature-related content and (ii) compared state-of-the-art interpretability methods to expose what the model highlights when classifying an image.

The study shows that a lightweight ResNet-18 pipeline is enough for analyzing DRVs on social-media imagery, achieving 75 % accuracy while running faster and with lower resource demand than deeper networks. Crucially, model choice also governs transparency: ResNet-18 paired with Grad-CAM or LIME delivers saliency maps that seemed to match human criteria, making its outputs easier to audit than those from EfficientNetB0 or DenseNet121. Across binary (Nature, Biological, Materiality) and multi-class (Landscape) tasks, the model focuses on semantically meaningful indicators as faces, animal features, and horizon context, showing that high interpretability does not have to come at the expense of speed. Therefore, it is recommended that lightweight, interpretable architectures be adopted as the default for large-scale monitoring on social platforms and suggest that future work prioritize explanation quality alongside accuracy.

The results demonstrate that DRV analysis is technically feasible at scale: a lightweight network can classify raw social-media with acceptable accuracy, reducing manual coding workload. Importantly, it was proved that model architecture is a primary driver of explanation quality as ResNet18's residual design produces human-redeable attention patterns.

These interpretable maps confirm that the chosen pipeline captures nature-related content and the maps reveal which visual features underpin online expressions of relational values, giving the first empirical window onto the ways social-media pictures shape, and are shaped by, human connection to nature.

Deploying the pipeline as a prefilter for ongoing BIG-5 data collection can both scale up the DRV corpus and provide a live test-bed for human-in-the-loop refinement of saliency maps.

Taken together, the work advances the BIG-5 project from concept to operational prototype: it shows that transparent AI can reconnect people to nature even when that nature is found on a screen and lays the groundwork for evidence-based stewardship campaigns in an increasingly digital world.

REFERENCES

Bengio, Y. (2012, July). Deep learning of representations for unsupervised and transfer learning. In I. Guyon, G. Dror, V. Lemaire, G. Taylor, & D. Silver (Eds.), *Proceedings of icml workshop on unsupervised and transfer learning* (pp. 17–36, Vol. 27). PMLR. https://doi.org/10.5555/3044805.3044807

Calcagni, F., Amorim Maia, A. T., Connolly, J. J. T., & Langemeyer, J. (2019). Digital co-construction of relational values: Understanding the role of social media for sustainability [Systematic review showing social media can reveal plural and relational dimensions of cultural ecosystem services]. *Sustainability Science*, *14*(5), 1309–1321. https://doi.org/10.1007/s11625-019-00672-1

Dastour, H., & Hassan, Q. K. (2023). A comparison of deep transfer learning methods for land use and land cover classification [Evaluates 39 pretrained deep learning models for LULC classification; ResNet50, EfficientNetV2B0, and ResNet152 were top performers]. *Sustainability*, *15*(10), 7854. https://doi.org/10.3390/su15107854

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Höhl, A., Obadic, I., Fernández-Torres, M.-Á., Najjar, H., Oliveira, D., Akata, Z., Dengel, A., & Zhu, X. X. (2024). Opening the black-box: A systematic review on explainable ai in remote sensing [Comprehensive survey of xAI methodologies, objectives, findings, and challenges in remote sensing]. *IEEE Geoscience and Remote Sensing Magazine*, 2–45. https://doi.org/10.1109/MGRS.2024.3467001

Ihongbe, E. I., Fouad, S., Mahmoud, T. F., Rajasekaran, A., & Bhatia, B. (2024). Evaluating explainable artificial intelligence (xai) techniques in chest radiology imaging through a human-centered lens [User-study comparing Grad-CAM and LIME on chest X-rays and CT for pneumonia/COVID-19; participants favored Grad-CAM]. *PLoS ONE*, *19*(10), e0308758. https://doi.org/10.1371/journal.pone.0308758

Kansal, K., Chandra, T. B., & Singh, A. (2024). Resnet-50 vs. efficientnet-b0: Multi-centric classification of various lung abnormalities using deep learning [Compares ResNet-50 and EfficientNet-B0 on multicenter chest X-ray datasets, showing slight superiority of EfficientNet-B0]. *Procedia Computer Science*, *235*, 70–80. https://doi.org/10.1016/j.procs.2024.04.007

Kapishnikov, A., Sipper, M., Ishay, M., & Keller, Y. (2019). Xrai: Better attributions through regions. *Proceedings of the IEEE International Conference on Computer Vision*, 2955–2964. https://doi.org/10.1109/ICCV.2019.00305

Kornblith, S., Shlens, J., & Le, Q. V. (2019). Do better imagenet models transfer better? [Investigates the transfer learning performance of models pretrained on ImageNet]. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2661–2671. https://doi.org/10.1109/CVPR.2019.00277

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1097–1105. https://doi.org/10.5555/2999134.2999257

Kumar, S. (2024). *Transfer learning and data augmentation in deep learning* [Accessed: 2025-05-18]. https://www.tredence.com/blog/transfer-learning-and-data-augmentation-in-deep-learning

Langemeyer, J., & Calcagni, F. (2022). Virtual spill-over effects: What social media has to do with rela-

tional values and global environmental stewardship [Conceptualises 'digital relational values' and virtual spill-over effects linking social media engagement to environmental stewardship]. *Ecosystem Services*, *53*, 101400. https://doi.org/10.1016/j.ecoser.2021.101400

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*, 4765–4774. https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html

Miller, J. R. (2005). Biodiversity conservation and the extinction of experience [Introduces the 'extinction of experience' concept and its implications for biodiversity conservation]. *Trends in Ecology & Evolution*, *20*(8), 430–434. https://doi.org/10.1016/j.tree.2005.05.013

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, June). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In J. DeNero, M. Finlayson, & S. Reddy (Eds.), *Proceedings of the 2016 conference of the north american chapter of the association for computational linguistics: Demonstrations* (pp. 97–101). Association for Computational Linguistics. https://doi.org/10.18653/v1/N16-3020

Selvaraju, R. R., Cogswell, M., Das, A., et al. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. *ICCV*.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*. https://arxiv.org/abs/1409.1556

Soga, M., & Gaston, K. J. (2016). Extinction of experience: The loss of human–nature interactions [Introduces and reviews the "extinction of experience" concept]. *Frontiers in Ecology and the Environment*, *14*(2), 94–101. https://doi.org/10.1002/fee.1225

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S. E., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Väisänen, T., Heikinheimo, V., Hiippala, T., & Toivonen, T. (2021). Exploring human–nature interactions in national parks with social media photographs and computer vision [Applied AI to Flickr photos to monitor public engagement with nature]. *Conservation Biology*, *35*(2), 424–436. https://doi.org/10.1111/cobi.13704

Weinstein, N., Przybylski, A. K., & Ryan, R. M. (2009). Can nature make us more caring? effects of immersion in nature on intrinsic aspirations and generosity [Experimental evidence that nature exposure shifts values toward prosocial behavior]. *Personality and Social Psychology Bulletin*, *35*(10), 1315–1329. https://doi.org/10.1177/0146167209341649

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European Conference on Computer Vision*, *8689*, 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

APPENDIX

A.1 Gantt Chart

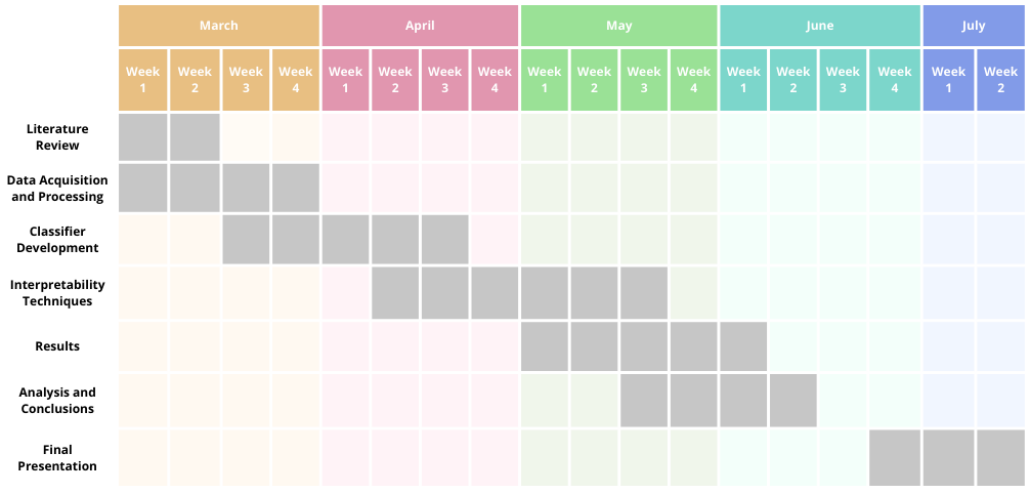
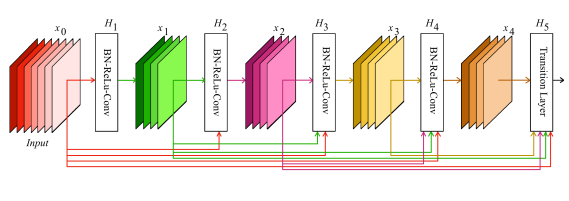
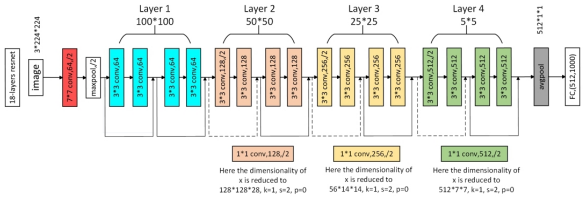


Fig. 23: Gantt Chart

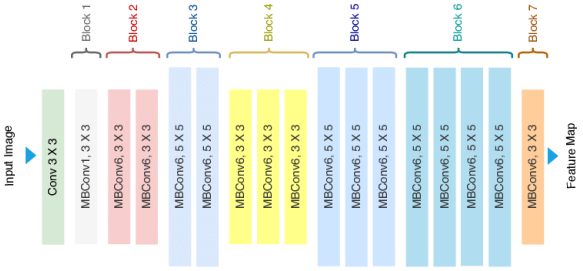
A.2 Pretrained models; Architectures



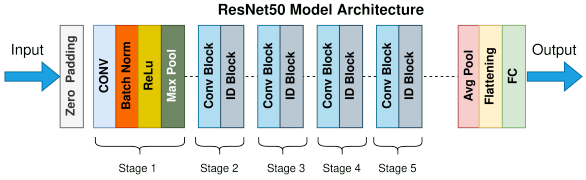
(a) DenseNet Architecture



(b) ResNet18 Architecture



(c) EfficientNetB0 Architecture



(d) ResNet50 Architecture

Fig. 24: Architectures for pretrained models

A.3 Additional Interpretability Techniques

A.3.1 Guided GradCAM

Guided GradCAM combines the strengths of GradCAM, which identifies broad influential regions (”where”), and guided backpropagation, which captures detailed features like edges and textures (”what”), to produce precise, visually rich explanations of model predictions. Implementation involves wrapping the model to isolate the task-specific output and registering two sets of hooks: one captures activations and gradients at a convolutional layer to compute the GradCAM mask; the other ensures only positive gradients flow through ReLU layers (guided backpropagation). These outputs are then combined—multiplying the coarse GradCAM localization map with detailed guided backpropagation gradients—to form a refined, detailed saliency map. This process yields clear, intuitive visual explanations without altering the underlying model structure.

Results on Guided GradCAM

Fig. 25: Original image

(a) Nature (b) Biological (c) Materiality (d) Landscape

Fig. 26: Guided GradCAM results

A.3.2 SHAP

SHapley Additive exPlanations (SHAP) provides theoretically grounded explanations by attributing a model’s prediction to individual input features, based on cooperative game theory. In the context of images, it estimates how much each pixel or region contributes to increasing or decreasing the predicted class score relative to a baseline. SHAP was implemented using a gradient-based explainer, wrapping the model to isolate the target task and selecting a representative set of background images. For each test image, SHAP computes contribution maps by averaging gradients over small perturbations, producing an attribution heatmap that reflects the influence of each region. This heatmap is then normalized, visualized using a colormap, and blended with the input image for an intuitive explanation of the model’s decision.

Results on SHAP

Fig. 27: Original image

(a) Nature (b) Biological (c) Materiality (d) Landscape

Fig. 28: SHAP results