Transfer Learning Under the Surface: Explainable Coral Bleaching Classification Across Datasets

Samantha Estrada
Stanford University
estradas@stanford.edu

Abstract

Accurate coral bleaching classification is essential for effective reef monitoring and conservation. This project investigates the robustness and transferability of transfer learning models—ResNet18, MobileNetV2, and baseline CNN and logistic regression classifiers—across datasets that vary notably in image saturation and size. Models trained on more vibrant, high-quality images generally achieve higher accuracy but show reduced generalization when tested on datasets with less saturated, muted tones. Combining datasets with different saturation qualities often leads to diminished overall performance, highlighting the challenges posed by heterogeneous real-world data. Applying explainability methods such as Grad-CAM reveals that these models consistently focus on ecologically meaningful features, enhancing trust and interpretability. These insights are important for guiding the development of robust, transferable coral bleaching classifiers capable of performing across diverse environmental conditions.

1. Introduction

Coral reefs are vital to marine ecosystems and global biodiversity. But as ocean temperatures rise and environmental conditions shift, corals experience stress, resulting in the loss of their symbiotic algae and turn white—threatening the survival of the reef. Accurate and efficient automated classification of coral bleaching can enable quick conservation actions and large-scale monitoring, critical for maintaining these fragile ecosystems healthy.

Automated classification of coral bleaching using underwater imagery has the potential to largely enhance monitoring efforts, offering scalable and efficient alternatives to manual surveys. Transfer learning techniques, which adapt pre-trained deep learning models to specific tasks, are particularly promising due to limited labeled coral datasets and the complexity of underwater images.

However, datasets used for coral bleaching classification often vary widely in image quality, particularly in color saturation and vibrancy. These differences pose challenges to model robustness and generalization, especially when models trained on one dataset are applied to another with very distinct visual characteristics.

In this project, I investigated the robustness of several classification models—including ResNet18, MobileNetV2, a baseline CNN, and logistic regression—trained on small, medium, and large coral datasets with differing saturation levels. I assessed how these models generalize across datasets, highlighting the effects of combining datasets with heterogeneous image quality. To add interpretability, I applied Grad-CAM explainability methods to visualize model attention, confirming that classification decisions are grounded in relevant coral features. These findings offer valuable insights for developing reliable, transferable models that can support coral reef conservation efforts under varying real-world conditions.

2. Related Works

Automated classification of coral bleaching has gained increasing interest as the need for scalable and automated monitoring grows. Early approaches relied heavily on traditional machine learning methods using handcrafted features. Jamil et al. (2021) proposed a Bag of Features (BoF) based deep learning framework to detect bleached corals, demonstrating that carefully engineered feature extraction can yield good classification performance, though a shortcoming is that such methods often struggle when dataset variability is introduced [2]. In contrast, more recent work has shifted towards deep convolutional neural networks (CNNs), which automatically learn relevant features from data.

A variety of CNN architectures have been explored for coral reef classification. Karthik et al. (2024) investigated several machine learning algorithms for coral classification, emphasizing the need for robust model performance across different datasets and imaging conditions [3]. Similarly, Wang et al. (2024) introduced ML-Net, a multi-local per-

ception network designed to classify healthy and bleached corals by capturing both global and local image features, highlighting the importance of spatial attention to improve accuracy [6]. These approaches underscore the potential of deep learning but often lack thorough evaluation of model transferability across datasets with varying image qualities.

Transfer learning has emerged as a powerful tool to leverage pre-trained models on large-scale datasets like ImageNet, fine-tuning them for coral bleaching classification with limited labeled data. Recent studies have demonstrated the effectiveness of architectures such as ResNet and MobileNet in this domain, enabling improved accuracy and reduced training times [1][4]. However, the influence of dataset characteristics—such as coral saturation, color vibrancy, and dataset size—on model robustness remains largely underexplored. My work builds on these insights by systematically evaluating transfer learning models across small, medium, large , and combined visual quality datasets.

Explainability in coral classification models is also essential to consider when evaluating robustness. Techniques like Grad-CAM provide visual explanations of model predictions, allowing researchers to verify that models focus on biologically relevant features such as bleached coral regions [5]. While prior works have applied Grad-CAM to general image classification, its integration in coral bleaching studies remains limited. This approach incorporates Grad-CAM to validate and interpret model decisions.

In summary, this project differentiates itself by combining transfer learning across multiple coral datasets with heterogeneous image qualities and leveraging explainability methods to assess model reliability. This holistic approach addresses gaps in robustness and interpretability that are critical for practical use in marine conservation contexts.

3. Datasets

This project uses three publicly available coral image datasets on Kaggle for binary classification of bleached vs. healthy corals. These datasets differ in size, visual quality, and how easy the classes are to distinguish. Together, they allow for testing the generalizability and robustness of transfer learning models across varying image conditions.

The small dataset, titled 'Bleached and Unbleached Corals Classification' [2], contains 342 high-quality, balanced images. The photos tend to be vibrant, with a clear visual contrast between bleached (white/pale) and healthy (colorful) coral classes, making it relatively easy for models to learn from.

The medium dataset, titled Healthy and Bleached Corals Image Classification, includes 923 images with lower color saturation. Many of the images are close-ups, and class distinction is more subtle. This dataset includes naturally white or earth-toned corals, making the labeling task more difficult and closer to real-world conditions.

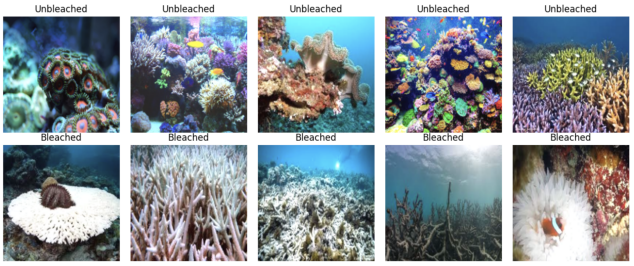

Figure 1. Example images from the small coral dataset.

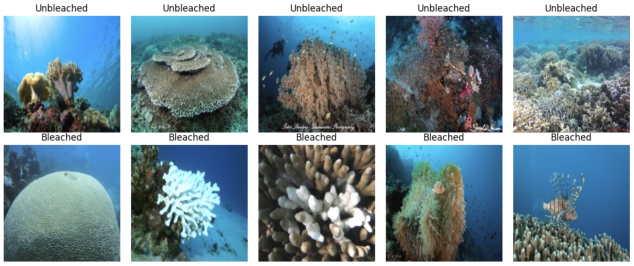

Figure 2. Example images from the medium coral dataset.

The large dataset, titled 'BHD Corals'[2], consists of 1,432 images and includes some overlap with the small dataset (approximately one-third). It is augmented with image rotations and flips and reflects underwater photography conditions with more noise and format variability. This dataset serves as a more comprehensive and realistic benchmark for evaluating model robustness.

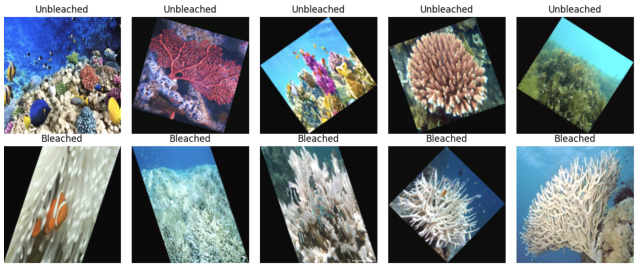

Figure 3. Example images from the large coral dataset.

A combined dataset (comb) was created by merging the small and medium datasets to test the effects of mixed-quality training data. Interestingly, combining datasets with different saturation levels and coral types created challenges for generalization, as the model sometimes struggled to reconcile differences between vibrant and duller coral images.

All datasets were preprocessed using TensorFlow pipelines. Images were resized to 224×224 pixels and batched for efficient loading. For training sets, light augmentations were applied, including random flips, subtle brightness shifts, and contrast and saturation adjustments to simulate murky underwater conditions. Initial experiments also included rotations, but I found that this was not reflec-

tive of how reef images are captured, and thus harmed learning more than provided a regularization effect. Light augmentation helped models become more resilient to lighting and visual noise while avoiding overfitting to overly bright or sharp features.

4. Methods

To classify coral bleaching and evaluate model robustness across varying datasets, I implemented and compared four different model architectures: logistic regression and a custom CNN as baseline models, and two pretrained convolutional neural networks (ResNet18 and MobileNetV2) for transfer learning.

4.1. Baseline Models

Our logistic regression baseline serves as a simple, interpretable starting point. It flattens input images and applies a single dense sigmoid layer to predict binary bleaching labels. While it lacks spatial feature extraction, it provides a useful performance benchmark.

We also built a basic 3-layer convolutional neural network (CNN) as a stronger non-transfer baseline. This model consists of three convolutional layers with ReLU activations and max pooling, followed by a fully connected layer and sigmoid output. This architecture allows the model to learn spatial patterns relevant to coral bleaching without relying on pretrained features.

To assess how dataset size affects generalization and overfitting, I trained both baseline models on the small and medium datasets and evaluated their performance not only on their respective test sets, but also on the larger and more diverse large dataset. This allowed us to explore whether models trained on limited or lower-quality data can still generalize to broader distributions. In particular, this cross-evaluation aimed to show how overfitting on small, saturated datasets manifests when tested on images with more variability in lighting, formatting, and coloration.

4.2. Transfer Learning Models

For more powerful and generalizable representations, I used transfer learning with two pretrained architectures: ResNet18 and MobileNetV2, both trained on ImageNet. I froze the base convolutional layers and only adjusted the last classification layers, using global average pooling, dropout, and dense layers. ResNet18 was chosen for its shallower depth and residual connections, which help maintain feature propagation and reduce vanishing gradients. MobileNetV2 offers a lightweight architecture that has proven to be effective in many classification tasks, making it an efficient and powerful alternative for feature extraction. Both transfer models were fine-tuned to classify coral images as bleached or unbleached, leveraging learned visual features from large-scale image data.

4.3. Evaluation and Explainability

I evaluated all models using accuracy, precision, recall, and F1 score across different training and testing dataset combinations. In addition to within-dataset performance, I tested cross-dataset generalization—particularly for models trained on smaller datasets but evaluated on the larger, more diverse test set.

For model interpretability, I applied Grad-CAM visualizations using the tf-keras-vis library. This technique highlights class-relevant regions of each image that influenced the model's decision, offering insights into which visual cues (e.g., coral texture, saturation, or shape) each model attended to. I visualized both correctly and incorrectly classified examples to understand failure modes and investigate how dataset quality impacted model focus.

This methodological framework enabled a comprehensive investigation of classification accuracy, robustness, and explainability across different data conditions and model complexities.

5. Experiments

To evaluate the effectiveness and robustness of the coral bleaching classifiers, I conducted a series of experiments using four model architectures across multiple datasets of varying size and quality. These experiments focus on assessing performance, understanding generalization across datasets, and exploring how dataset characteristics such as saturation and diversity influence outcomes.

5.1. Baseline Models: Logistic Regression and CNN

We began by establishing two baseline models: a logistic regression classifier and a shallow convolutional neural network (CNN). These were trained on the small and medium datasets and evaluated on both their respective test sets and on the large test set to test cross-dataset generalization.

5.1.1 Logistic Regression Baseline

On the small dataset, the logistic regression model performed reasonably well with an accuracy of 78.4% and precision of 85.7%, but its recall was lower (57.1%), indicating it struggled to detect some positive (unbleached) cases. On the medium dataset, while recall improved significantly to 90.9%, precision dropped to 51.7%, suggesting the model over-predicted unbleached coral when trained on this lower-quality dataset. Both the small and medium models saw a substantial drop in performance when evaluated on the large test set, with identical accuracy (53.6%) and F1 score (55.7%). This suggests that simple models overfit to the stylistic and color saturation biases of their training data and do not generalize well to more complex or diverse inputs.

Table 1. Baseline Model Performance (Logistic Regression and CNN)

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression (Small)	0.7843	0.8571	0.5714	0.6857
Logistic Regression (Medium)	0.5507	0.5172	0.9091	0.6593
Logistic Regression (Small → Large Test)	0.5364	0.5039	0.6214	0.5565
Logistic Regression (Medium → Large Test)	0.5364	0.5039	0.6214	0.5565
CNN (Small)	0.7843	0.7778	0.6667	0.7179
CNN (Medium)	0.5797	0.5500	0.6667	0.6027
CNN (Small → Large Test)	0.7409	0.8485	0.5437	0.6627
CNN (Medium → Large Test)	0.5727	0.5542	0.4466	0.4946

5.1.2 CNN Baseline

On the small dataset, the CNN baseline slightly improved recall (66.7%) while maintaining a similar accuracy (78.4%) and a higher F1 score (71.8%), suggesting better spatial pattern recognition than logistic regression On the medium dataset, performance dropped across all metrics (F1: 60.3%), again indicating the lower visual contrast of this dataset affected the model's ability to distinguish between classes. Interestingly, the CNN trained on the small dataset performed better on the large dataset (F1: 66.3%) than the one trained on the medium dataset (F1: 49.5%). This may reflect how visual clarity and contrast in the small dataset enabled more transferable features, even if the quantity of data was lower.

5.2. Transfer Learning Models: ResNet18 and MobileNetV2

Next, I evaluated transfer learning models, which were pretrained on ImageNet and fine-tuned for binary coral classification.

5.2.1 ResNet18

Small and large datasets both led to strong performance (F1: 93.0% and 93.3%, respectively), indicating that pretrained representations paired with clear training images can yield high-quality classifiers even with limited data. Medium dataset performance lagged behind (F1: 67.2%), consistent with earlier trends suggesting that subtle visual features and lower saturation in this dataset were harder for models to learn from. Combined dataset (small + medium) performance was lower (F1: 71.2%) than either dataset alone. This suggests that combining datasets with different visual styles may introduce noise or confusion, harming generalization instead of improving it.

5.2.2 MobileNetV2

Large dataset training yielded good results (F1: 77.7%), though lower than ResNet18, consistent with MobileNet's shallower architecture and reduced parameter count. Small dataset training produced comparable performance (F1:

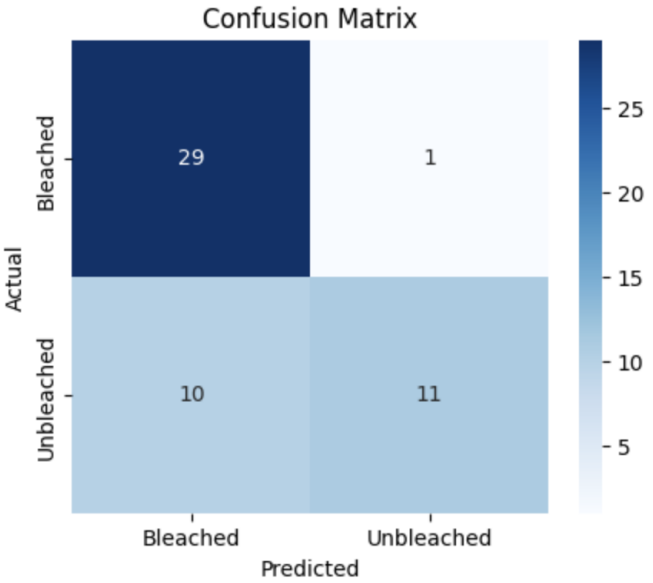

Figure 4. Logistic Regression Small

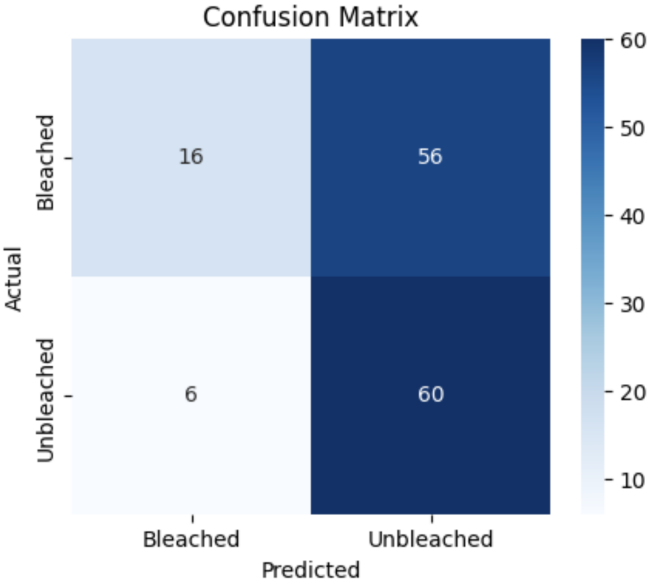

Figure 5. Logistic Regression Medium

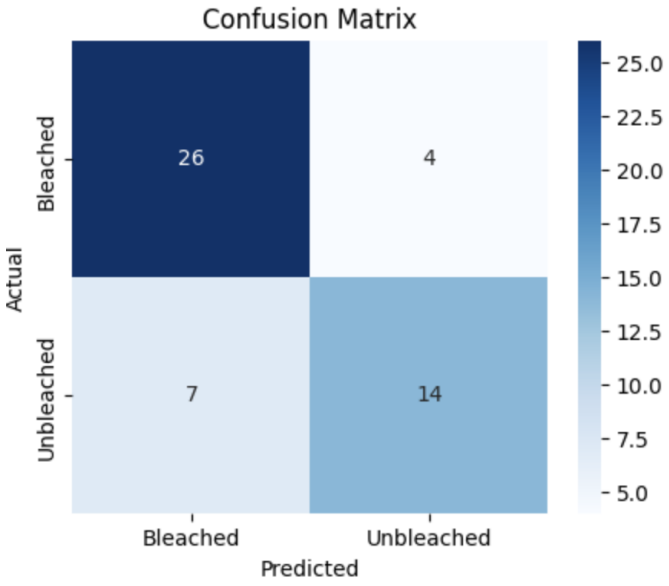

Figure 6. CNN Small

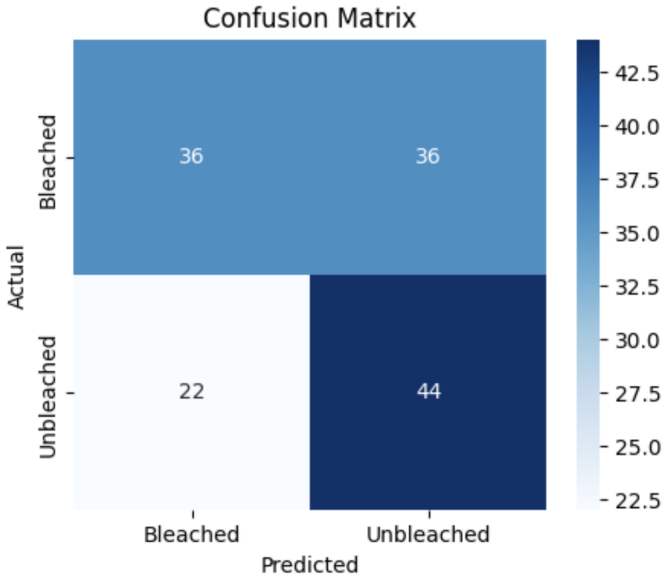

Figure 7. CNN Medium

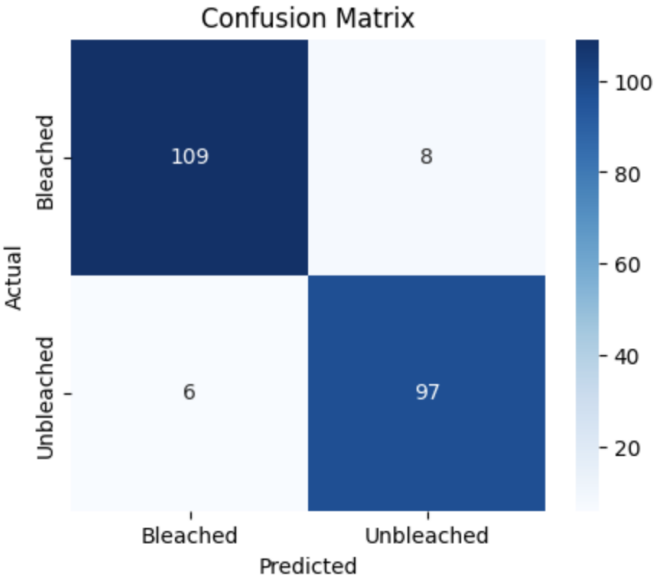

Figure 8. ResNet18 on large dataset

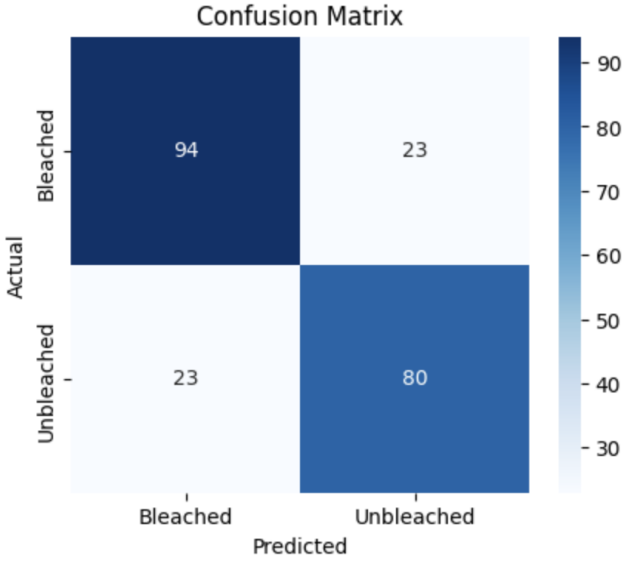

Figure 9. MobileNetV2 on large dataset

71.2%), again showing that small but high-contrast data can be highly effective with transfer learning. Medium dataset and combined dataset performance was lower (F1: 59.7% and 58.8%, respectively), reinforcing the theme that visual variability and lower saturation challenge model learning and generalization.

5.3. Using Grad-CAM For Model Explainability

To better understand model behavior and gain insight into the decision-making process behind coral bleaching classification, I applied Grad-CAM to visualize class activation maps for both ResNet18 and MobileNetV2. These visualizations were particularly useful in highlighting what parts of the image were driving classification outcomes, especially for images that were incorrectly labeled.

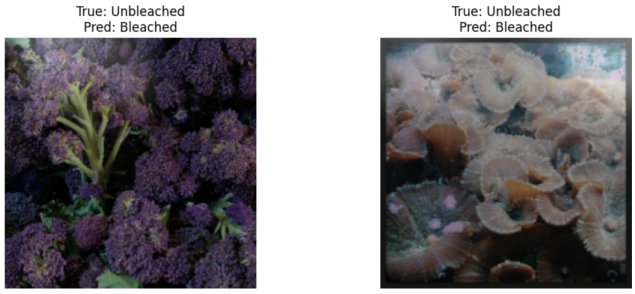

Figure 10. Low activation for closeups.

Table 2. Transfer Learning Performance Using ResNet18 and MobileNetV2

Model	Accuracy	Precision	Recall	F1 Score
ResNet18 (Small)	0.9412	0.9091	0.9524	0.9302
ResNet18 (Medium)	0.6884	0.6769	0.6667	0.6718
ResNet18 (Large)	0.9364	0.9238	0.9417	0.9327
ResNet18 (Combined)	0.7302	0.7000	0.7241	0.7119
MobileNetV2 (Small)	0.7302	0.7000	0.7241	0.7119
MobileNetV2 (Medium)	0.6377	0.6379	0.5606	0.5968
MobileNetV2 (Large)	0.7909	0.7767	0.7767	0.7767
MobileNetV2 (Combined)	0.6667	0.6818	0.5172	0.5882

Across both models, I observed a consistent trend: images where the coral or reef filled most of the frame—especially close-up shots with little to no visible blue water—tended to have very weak or unfocused activation. This lack of broader spatial context may be limiting the model's ability to infer bleaching status, particularly if key visual cues like color contrast with surrounding water or structural patterns are absent.

Figure 11. High activation along reef horizon lines

Figure 12. Bleached images focus on horizon line and have low activation for closeups (ResNet18)

Figure 13. ResNet18 Incorrect Classifications

In the ResNet18 model, many of the correctly classified images showed strong activation along the top and side thirds of the image. These areas often correspond to horizon lines or boundaries between reef structures and open water. This suggests the model may be relying on large-scale contrast features or environmental cues that appear in these regions. However, this also led to misclassification when such a boundary was present but not informative—for example, a diver in the frame, light reflections, or open water with no reef intersection. In several failure cases, the model focused on these misleading elements rather than on the coral features themselves.

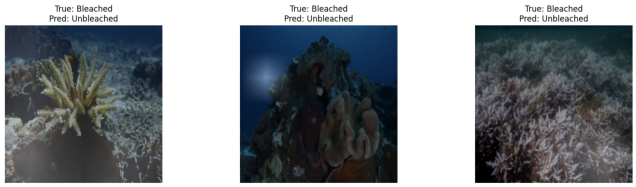

Figure 14. Localized focus, weaker activation with heterogenous corals

The MobileNetV2 model demonstrated a slightly different behavior. Incorrectly classified images often showed more diffuse or less sharply defined activation, sometimes focusing on multiple unimportant regions or on isolated texture features rather than the broader reef structure. In some cases, it concentrated on smaller contrast spots or isolated coral patches without picking up the larger context of bleaching. This may suggest that MobileNetV2, with its lighter architecture, is less able to capture global spatial relationships and instead relies more on local texture features.

These explainability findings reveal that both models are sensitive not only to coral features but also to background and contextual elements such as water coloration, horizon lines, or foreign objects like divers. This has important implications: for robust coral bleaching classifiers, input data should ideally capture consistent visual cues while minimizing distracting or irrelevant features. Future work could explore preprocessing or attention-guided architectures to reduce spurious focus and improve generalization across image styles and compositions.

6. Conclusion

This project explored the use of transfer learning and explainability techniques for coral bleaching classification across datasets of varying size and visual quality. I found that transfer learning models like ResNet18 and Mo-

bileNetV2 significantly outperformed traditional baselines such as logistic regression and shallow CNNs, especially when trained on larger and more diverse datasets. However, combining datasets with different visual characteristics—such as color saturation and image framing—did not always improve generalization, highlighting the challenges of dataset shift.

Grad-CAM visualizations revealed that model attention often focused on contextual features like horizon lines or background contrast, rather than solely on coral structures. This suggests that classifier performance can be sensitive to framing, scene composition, and non-coral elements within the image.

These findings emphasize the importance of dataset consistency and the need for robust explainability tools in environmental image classification. Future work could explore fine-tuning feature attention, improving data augmentation to simulate natural variation more effectively, or incorporating domain-specific knowledge into model training.

References

[1] 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*, pages 770–778, 2016. 2

[2] S. Jamil, M. Rahman, and A. Haider. Bag of features (bof) based deep learning framework for bleached corals detection. *Big Data and Cognitive Computing*, 5(4):53, 2021. 1, 2

[3] S. N. Karthik, M. Hariharasudhan, and M. A. Devi. An investigation on coral reef classification using machine learning algorithms. In *Innovative Computing and Communications*. Springer, 2025. 1

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 2

[5] 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*, pages 618–626, 2017. 2

[6] S. Wang, N.-L. Chen, Y.-D. Song, T.-T. Wang, J. Wen, T.-Q. Guo, H.-J. Zhang, L. Mo, H.-R. Ma, and L. Xiang. Ml-net: A multi-local perception network for healthy and bleached coral image classification. *Journal of Marine Science and Engineering*, 12(8):1266, 2024. 2