Performance of stacking machine learning model for reducing coral reef bleaching

Ngoc Nhat Pham

1 Introduction

Coral reefs, often described as the "rainforests of the sea," are among the most biodiverse and productive ecosystems on Earth. Although they occupy less than 0.1% of the ocean floor, they are associated with roughly 25% of all marine life, providing critical habitat, nursery grounds, and refugia for fishes and invertebrates (8). In addition to their inherent natural value, reefs support around one billion people in coastal areas by providing food and jobs (6). They help protect coastlines by absorbing wave energy, reducing the impact of storms and preventing erosion (17).

Figure 1: Colorful coral reef ecosystem. (Credit: Chris Newbert, Minden Pictures.)

Despite their importance, coral reefs face severe threats, with coral bleaching being a primary

concern. Bleaching occurs when corals, stressed by environmental changes such as rising sea temperatures, ocean acidification, or pollution, expel their symbiotic algae (zooxanthellae), leading to whitening and potential mortality due to the loss of these algae's pigments or the algae themselves. The Great Barrier Reef, which is the world's largest coral reef ecosystem, is an example of this phenomenon. The reef has experienced global mass bleaching events recorded in 1998, 2002, 2006, 2016, 2017, 2020, 2022, and now currently unfolding again in 2024 (Fig.2) (19). The loss of coral reefs would disrupt marine ecosystems, coastal protection, and economic stability, underscoring the urgency of addressing this crisis

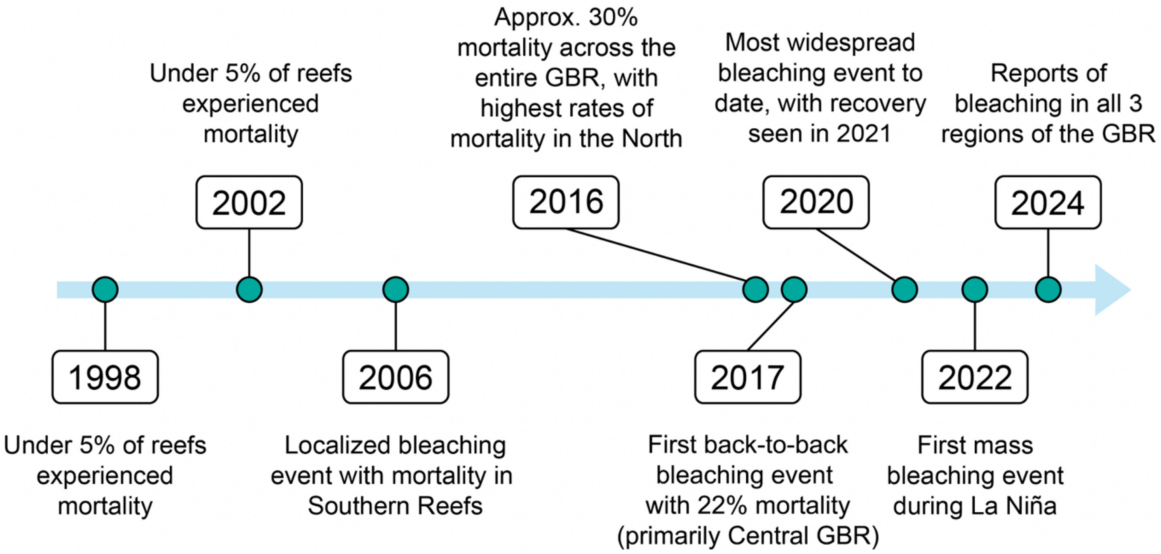

Figure 2: Coral bleaching history on the Great Barrier Reef from 1998 to 2024 (Credit: Dr. Emma Rehn)

One way to combat coral reef bleaching is to make data driven decisions. Advanced models that predict coral reef outcomes often use thermal stress metrics, local environmental factors, and large observational datasets. Ensemble and machine learning approaches have proven effective, as they can integrate diverse predictors and capture complex patterns and interactions that are challenging to model explicitly. However, decision-makers require more than just accurate predictions, they need clear explanations of what a model predicts, why it makes those predictions, and, most importantly, how practical management actions could reduce risks. The demand for improved reef-scale predictions and usable tools highlights the need to balance model performance with clear, interpretable results (5)

Therefore, this study evaluates a stacking machine learning framework for predicting coral bleaching severity (classified as Mild, Moderate and Severe) using a comprehensive dataset from the Biological and Chemical Oceanography Data Management Office (BCO-DMO), covering the period from 1980 to 2020 (21). The dataset involves environmental features such as Distance to Shore, Turbidity, Cyclone Frequency, Depth, Sea Surface Temperature (SST), etc. I compared the stacked model against strong tree-based baselines (Random Forest and XGBoost), using a consistent training steps with stratified splits, standardlization, class balancing via SMOTE (Synthetic Minority Oversampling Technique), and hyperparameter search. In the experiment, Stacking model delivers the best overall performance among the evaluated models and improves balanced performance across

classes

A key contribution of this work is to pair performance with transparent model reasoning via SHAP (SHapley Additive exPlanations). Local (waterfall) and global (Beeswarm and mean SHAP) are provided explanations to identify how features push predictions toward or away from each class. These insights are then applied in a practical "what-if" intervention analysis where only modifiable features (e.g., percent coral cover via restoration or small turbidity changes where ecologically acceptable) are nudged within realistic bounds to quantify expected reductions in Moderate/Severe probabilities under management scenarios. This approach quantifies how such actions could reduce the probability of moderate to severe bleaching, while keeping unchangeable factors, such as reef depth, distance from shore, or long term climate trends, constant.

The findings emphasize three key points:

- The effectiveness of combining models (Random Forests and XGBoost) under a simple logistic meta-learner to leverage their strengths
- Thhe critical influence of coral cover and background sea surface temperature on distinguishing mild from severe bleaching outcomes
- The value of SHAP-guided "what-if" scenarios to estimate how management actions could lower predicted risks.

While my analysis isn't causal and relies on the model's accuracy and data quality, it provides a clear framework for prioritizing management actions and communicating expected changes in bleaching risk at the reef or site level.

2 Data and Study Area

2.1 Study scope and observational units

This study uses a global site-level dataset of coral reef conditions and bleaching observations. The raw table contains **41,361 rows** and **62 columns** covering locations across major oceans with observation years from 1980-2000. Each record represents a site-date observation with geolocation, ocean basin, environmental variables, and bleaching information

The key fields are:

- **Location**: Latitude_Degrees, Longitude_Degrees, Ocean_Name (Atlantic, Indian, Pacific, Red Sea, Arabian Gulf), Distance_to_Shore.
- **Exposure and Habitat**: Exposure (Sheltered, Sometimes, Exposed), Depth_m, Percent_Cover.
- **Climate and Oceanography**: Temperature_Kelvin, ClimSST, Windspeed, SSTA, TSA, Cyclone_Frequency.
- **Bleaching**: Percent_Bleaching (continuous, used to derive the categorical target).
- **Date Fields**: Date_Day, Date_Month, Date_Year (original Date column removed during data cleaning).

Where units are not explicit in the metadata, I retain variables as provided and avoid re-scaling to assumed physical units.

2.2 Target definition (Mild/Moderate/Severe)

The continuous bleaching percentage is converted into a 3-class outcome:

- **Mild**: $0 - 10\%$
- **Moderate**: $11 - 50\%$
- **Severe**: $> 50\%$

2.3 Data cleaning and preprocessing

Column pruning: I dropped redundant identifiers to streamline the dataset and optimize memory usage (e.g., Reef_ID, State_Island_Province_Name, City_Town_Name, Site_Name, Site_Comments, Sample_Comments, Bleaching_Comments, Site_ID, Sample_ID, Data_Source, Realm_Name, Ecoregion_Name, Country_Name, and the original string Date). Bleaching_Level and Substrate_Name are also dropped as they contained excessive missing data.

Missing values: The string sentinel "nd" was treated as missing (NaN) across all fields. I then used simple column-wise strategies consistent with robust practice:

- Mean imputation for SSTA, TSA, Cyclone_Frequency
- Median imputation for Distance_to_Shore, Turbidity, Depth_m, Percent_Cover, Percent_Bleaching, ClimSST, Temperature_Kelvin, Windspeed.

Type Casting: The following variables are cast to float64: Distance_to_Shore, Turbidity, Cyclone_Frequency, Depth_m, Percent_Cover, Percent_Bleaching, ClimSST, Temperature_Kelvin, Windspeed, SSTA, TSA; Exposure and Percent_Bleaching_Category were treated as categorical.

2.4 Encoding and feature types

Ordinal Encoding and One-Hot Encoding are data preprocessing techniques used to convert categorical variables into numerical formats suitable for machine learning models. While both serve the purpose of transforming categorical data, they are fundamentally different in their approaches and applications. In this case, both Ordinal Encoding and One-Hot Encoding are required before moving to next machine learning steps. One-Hot encoding involves creating a new binary feature for each category in the categorical variable. This technique is suitable for categorical variables with a small number of unique categories (20). On the other hand, Ordinal encoding assigns a unique integer value to each category in a categorical variable based on the order or rank of the categories, with lower integer values assigned to categories that are considered "lower" or "earlier" in the order (20). Therefore, the following methods are applied for these category features:

- **Ordinal encoding for:**
  - `Exposure`: Sheltered < Sometimes < Exposed.
  - `Percent_Bleaching_Category`: Mild < Moderate < Severe.
- **One-hot encoding:** Ocean_Name was expanded into basin indicators Ocean_Name_Atlantic, `Ocean_Name_Indian`, Ocean_Name_Pacific, Ocean_Name_Red Sea, Ocean_Name_Arabian Gulf)

2.5 Final analytic table

After cleaning and typing, the modelling table contained **41,181** rows and **24** columns. Train/test split and Standardisation (z-scoring) were applied after splitting, while Standardlisation used a StandardScaler fitted on the training set only.

2.6 Feature selection

Random forest (RF) is a machine-learning method that may be a good candidate for integrating omics data as it generally works well with high-dimensional problems and can identify strong predictors of a specified outcome without making assumptions about an underlying model (4). However, a common problem of high-dimensional data sets is the presence of correlated predictors, which impact RF's ability to identify the strongest predictors by decreasing the estimated importance scores of correlated variables (11). In this data set, applying all 24 features to predict coral bleaching class may indicate a sign of overfitting. A suggested solution is the Random-Forest-Recursive Feature Elimination (RF-RFE) algorithm (11). RFE was initially proposed to enable support vector machines to perform feature selection by iteratively training a model, ranking features, and then removing the lowest ranking features (12). This method has been similarly applied to RF and found to be beneficial in the presence of correlated features

Algorithm. Let $\mathcal{F}$ denote the current feature set and k the target number of features. Recursive feature elimination (RFE) proceeds as:

1. Fit a Random Forest (RF) classifier on the current $\mathcal{F}$ using the training data.
2. Compute model-based importances $I(f)$ for each $f \in \mathcal{F}$ (here, impurity-based RF importances).
3. Remove the least important feature(s) (step size = 1).
4. Repeat steps (1)–(3) until $|\mathcal{F}| = k$.

I set $k = 10$ to balance parsimony and performance. The RF used within RFE is initialised with a fixed random seed to stabilise the ranking.

Data-leakage control. Feature selection is performed strictly within the training data to avoid optimistic bias. In cross-validated experiments, RFE is nested inside the model-selection loop so that the test fold remains untouched by both the estimator and the selector. If a single train/test split is used, the selector is fit only on the training split and then applied to the test split.

Selected features. RFE identifies the following top 10 predictors:

- Distance_to_Shore
- Turbidity
- Cyclone_Frequency
- Depth_m
- Percent_Cover

- ClimSST
- Windspeed
- Temperature_Kelvin
- SSTA
- TSA

2.7 Exploratory data analysis (EDA)

2.7.1 Continuous variables

To visualize skew and inform robust modeling decisions, I generated histograms using the log1p transformation for ten continuous predictor variables (Figure 3).

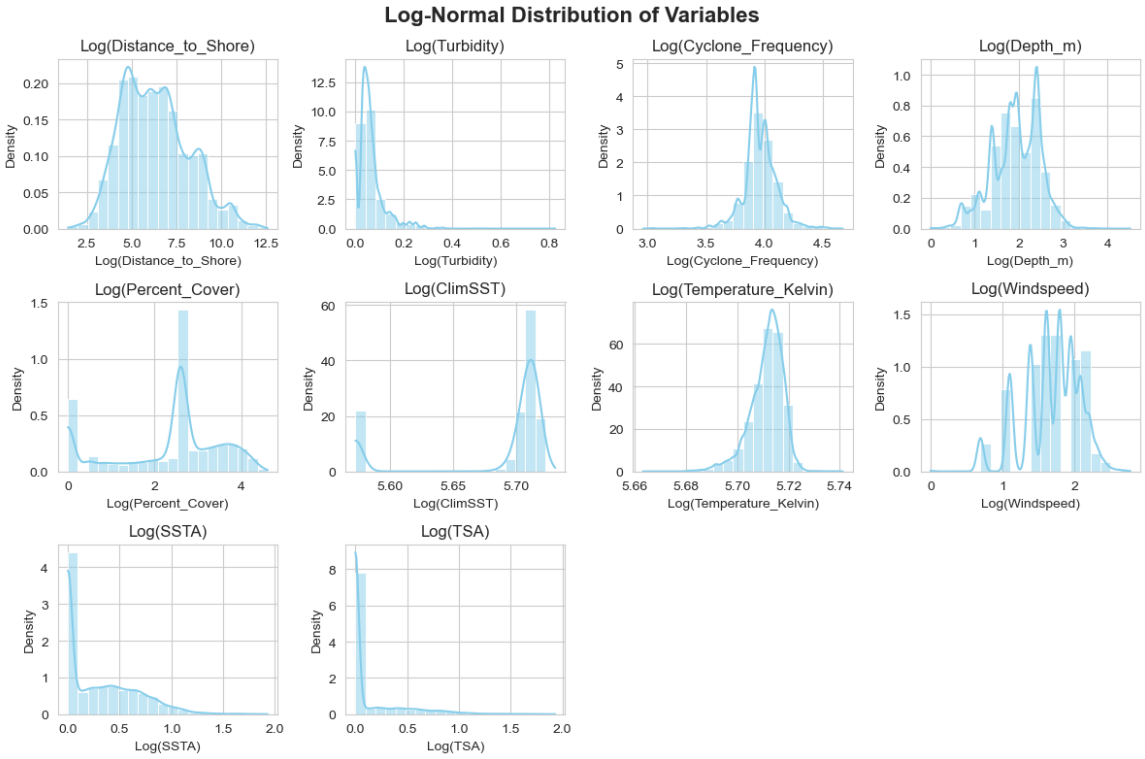

Figure 3: Log-Normal Distribution of Variables

Several patterns are across variables:

- **Distance_to_Shore** shows a right skew on the original scale that becomes near-unimodal on the log scale, which indicates many near-shore observations with a thinner tail of offshore sites
- **Turbidity** is highly right-skewed; most observations cluster close to zero with a long tail of higher values.

- **Cyclone_Frequency** is relatively concentrated and approximately symmetric after log transform
- **Depth_m** remains multimodal even on the log scale
- **Percent_Cover** exhibits a mixed/bimodal pattern: substantial mass at low cover and a secondary mode at moderate cover.
- **ClimSST** and **Temperature_Kelvin**, are tightly distributed and close to Gaussian on the log scale, reflecting small spatial gradients in background temperature compared to other factors.
- **Windspeed**is unimodal with mild right skew.
- **SSTA** and **TSA** concentrate near zero with a positive tail, consistent with many sites experiencing small anomalies punctuated by fewer high-anomaly events.

2.7.2 Categorical variables

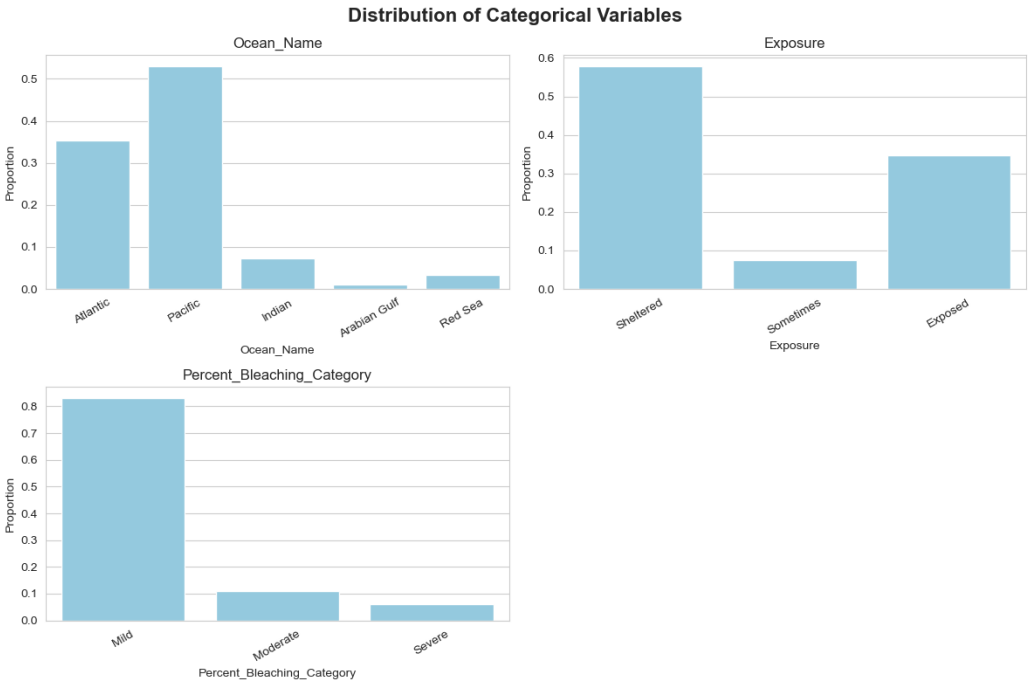

Figure 4: Distribution of Categorical Variables

Category distributions plot (Figure 4) reveal the sample composition:

- **Ocean_Name** is dominated by the Pacific Ocean dominates with the highest count, followed by the Atlantic, Indian, Arabian Gulf, and Red Sea, suggesting the Pacific is the most studied or affected region.
- **Exposure:** Most sites are classified as Sheltered, with fewer Sometimes or Exposed sites, indicating a bias toward protected reef environments.

- The target **Percent_Bleaching_Category**: The majority of observations fall into the "Slow" category, with fewer "Mild," "Moderate," and minimal "Severe" cases, suggesting that severe bleaching is less frequent but still significant.

2.7.3 Bleaching by ocean (Box Plot)

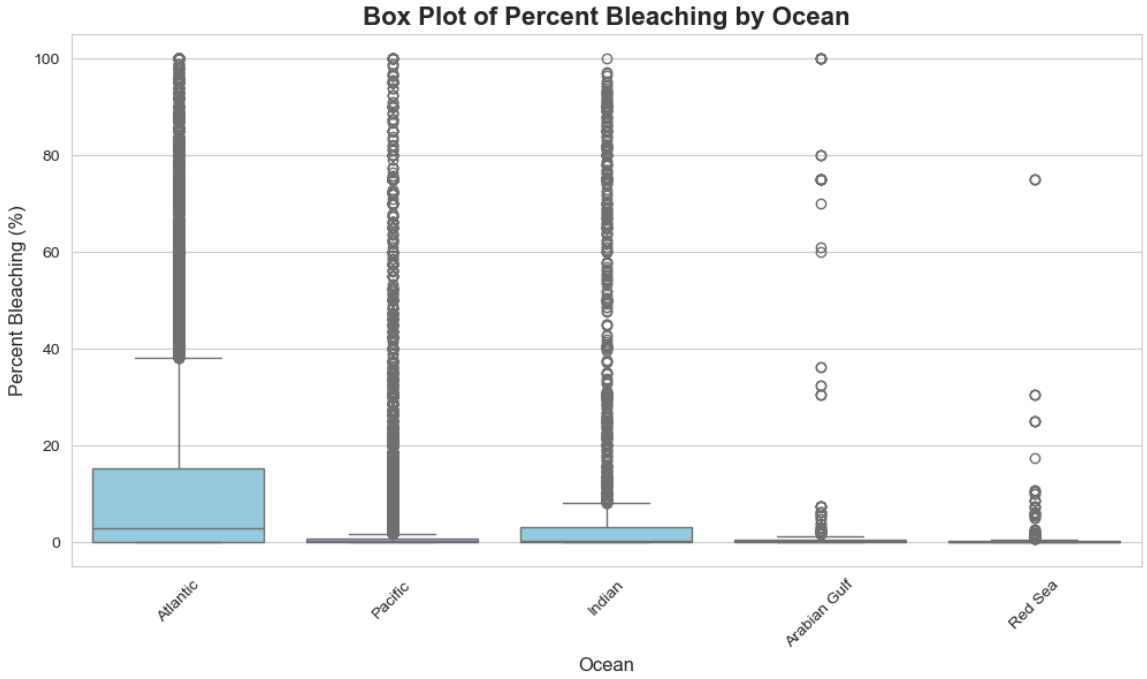

Figure 5: Box Plot of Percent Bleaching by Ocean

The median Percent_Bleaching (Figure 5) being close to 0 % for all oceans indicates that, typically, coral reefs in these regions experience little to no bleaching. This could mean that the data observations focused on periods or locations with low bleaching. However, the box plots show large variability and outliers in the Atlantic, Pacific and Indian Oceans. The Arabian Gulf and Red Sea exhibit lower median bleaching percentages but still show extreme outliers despite smaller sample sizes.

2.7.4 Bleaching by exposure (Violin Plots)

Distributions of Percent_Bleaching across Sheltered, Sometimes, and Exposed classes are zero-inflated with pronounced upper tails (Figure 6). Median differences between classes are modest, although Exposed conditions show slightly more frequent high-bleaching events in the tail. This supports treating Exposure as ordinal while recognising its relatively limited global effect (it was deprioritised by RFE).

2.7.5 Temporal patterns (global and by ocean)

Two complementary plots highlight time structure:

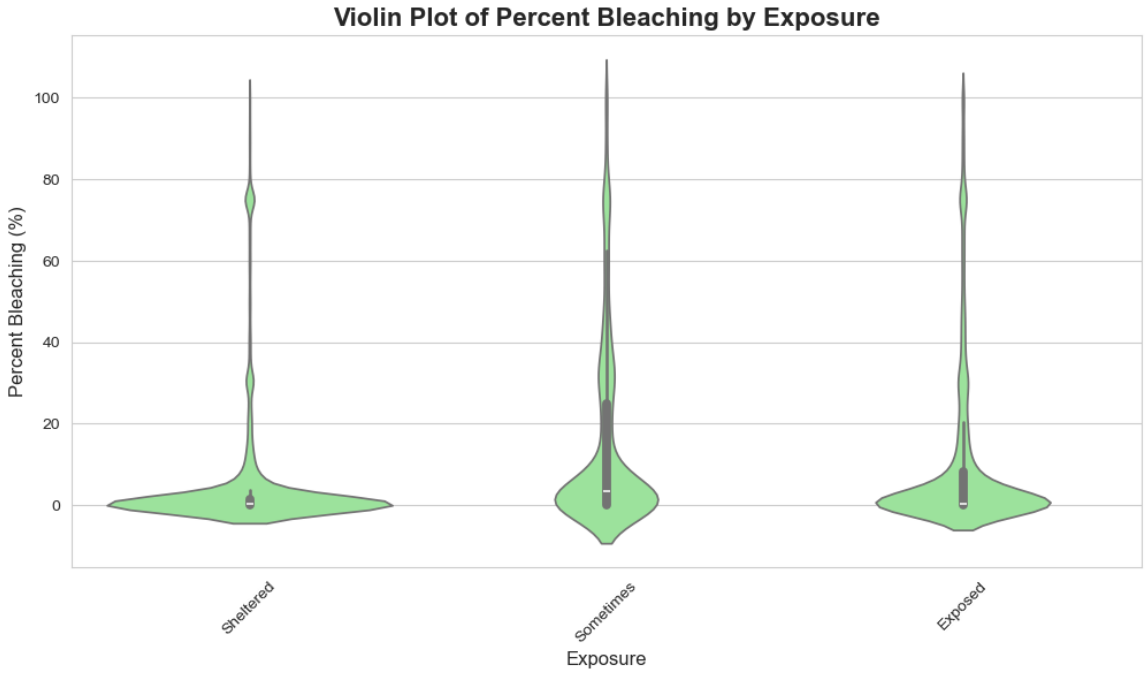

Figure 6: Violin Plot of Percent Bleaching by Exposure

- **Global mean over time (Figure 7)** The figure plot illustrates notable bleaching percentages from 1980 to 2020, with pronounced peaks in the mid-1990s and early 2000s, corresponding to major global bleaching events, such as the one in 1998. Following the early 2000s peak, the data shows a general decline, with a noticeable dip during the 2010s. This temporal trend highlights the recurring nature of environmental stress events and emphasizes the increasing impact of factors like rising sea surface temperatures over time.
- **By ocean over time (Figure 8)** Although from Figure 4 above, the Pacific Ocean experienced the most occurrence of bleaching events, from this graph, Indian and Arabian Gulf actually exhibit the highest bleaching levels, with peaks in Indian around 75% and close to 80% (the highest of all time) in Arabian Gulf around the mid-1990s and early 2000s, aligning with global mass bleaching events. The Pacific shows moderate peaks (around 50%) in the same periods, while the Atlantic and Red Sea have lower and more sporadic peaks (below 50%). Post-2000, bleaching levels generally decline across all oceans but show some spikes, particularly in the Indian Ocean around 2015–2020.

2.7.6 Correlations (heatmap)

There is a strong positive correlation between Temperature_Kelvin and Thermal Stress Anomaly (TSA) (0.85), indicating that higher baseline temperatures are closely associated with increased thermal stress. Similarly, Sea Surface Temperature Anomaly (SSTA) shows a strong correlation with TSA (0.55), reinforcing their combined role as key indicators of thermal anomalies that contribute to coral bleaching. Additionally, SSTA has a moderate correlation with Temperature_Kelvin

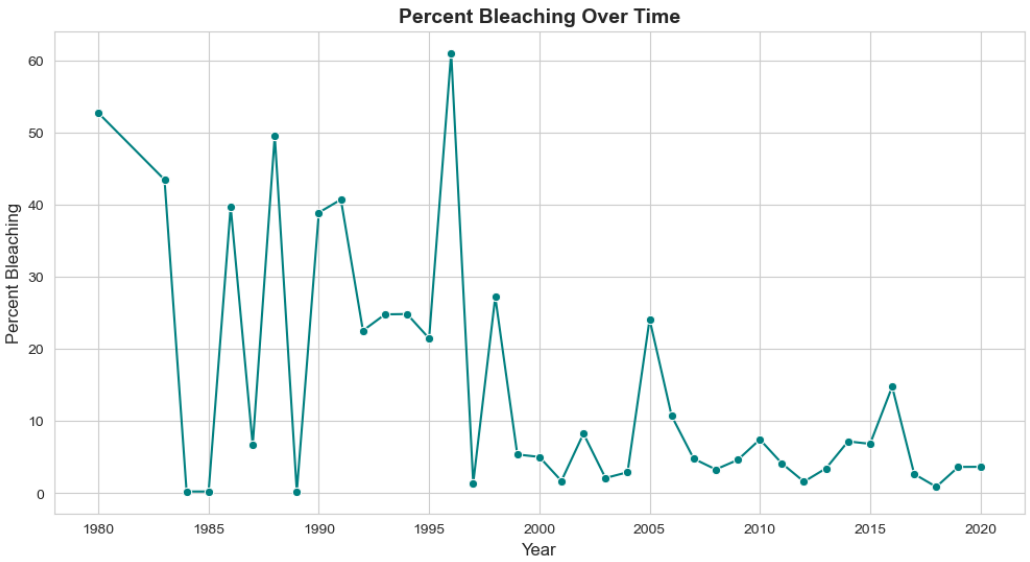

Figure 7: Percent Bleaching Over Time

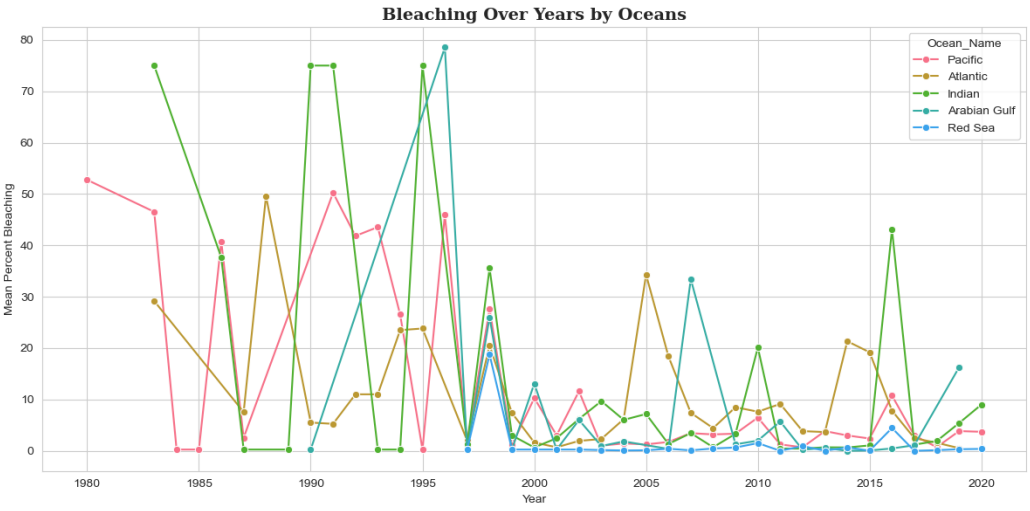

Figure 8: Percent Bleaching Over Time by Oceans

(0.42), suggesting a meaningful relationship between general sea temperature levels and short-term temperature deviations. In contrast, the remaining variables exhibit relatively weak or negligible correlations with each other, indicating limited linear associations.

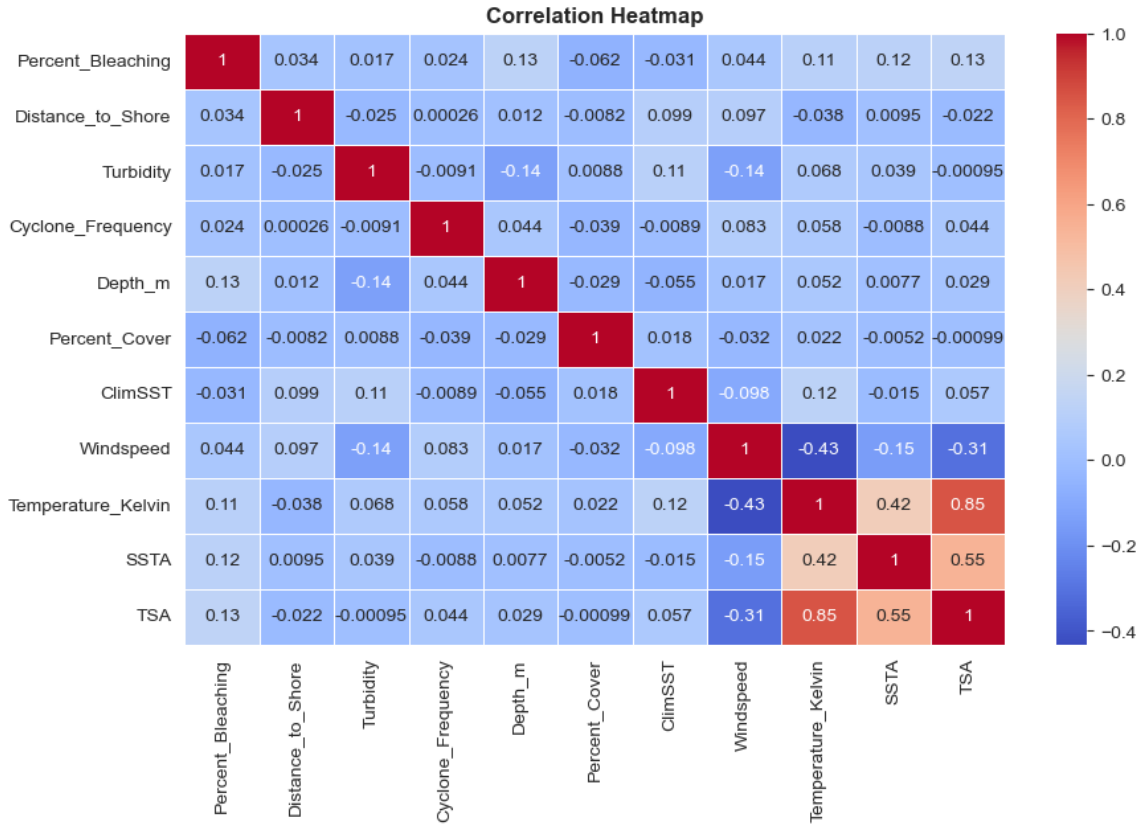

Figure 9: Correlation Heatmap

3 Methods

3.1 Problem Formulation and Class Imbalance

Let $x_i \in R^p$ denote the vector of predictors for site–date i and $y_i \in \{0, 1, 2\}$ the bleaching class, where 0 = Mild, 1 = Moderate, and 2 = Severe. After the preprocessing and RFE described in Section 3, $p = 10$. I have a probabilistic classifier

$$f : R^p \to \Delta^2, \qquad f(x) = \big(P(Y = 0 \mid x),\, P(Y = 1 \mid x),\, P(Y = 2 \mid x)\big),$$

where Δ^2 is the 3-simplex.

The label distribution is highly imbalanced (Mild $\gg$ Moderate $>$ Severe). To mitigate the tendency of learners to overfit the majority class, the data are split into train/test with stratification on y, and the training set is rebalanced using SMOTE (Synthetic Minority Oversampling Technique).

SMOTE (Synthetic Minority Over-sampling Technique) algorithm is an extended algorithm for imbalanced data (Nitesh and Kevin). In essence, SMOTE algorithm obtains new samples by random linear interpolation between a few samples and their neighboring samples. The data imbalance ratio is increased by generating a certain number of artificial minority samples, so that the classification effect of the imbalanced data set is improved (14). The specific process of SMOTE (22) is as follows:

- Step 1: For each minority sample x_i $(i = 1, 2, \ldots, n)$, calculating its distance to other samples in minority sample according to certain rules to obtain its k nearest neighbors.
- Step 2: According to the over-sampling magnification, the random m nearest neighbors, as a subset of k nearest neighbors set, of each sample x_i are selected and denoted as x_{ij} $(j = 1, 2, \ldots, m)$, then an artificially constructed minority sample p_{ij} is calculated by:

$$p_{ij} = x_i + \text{rand}(0, 1) \times (x_{ij} - x_i) \tag{1}$$

  where $\text{rand}(0, 1)$ is a random number uniformly distributed within the range of $[0, 1]$. The operation of the formula (1)is stopped until the fused data reaches a certain imbalance ratio

Moreover, the test set remains unchanged. This decision ensures that the evaluation remains stable and reliable, allowing for a fair comparison of model performance across different classess without the influence of varying test data.

3.2 Baseline Models: Random Forest and XGBoost

Random Forest (RF) Random Forest is a popular machine learning algorithm used for several types of classification tasks (7), (10). A Random Forest is an ensemble of tree-structured classifiers. Every tree of the forest gives a unit vote, assigning each input to the most probable class label. It is a fast method, robust to noise and it is a successful ensemble which can identify non-linear patterns in the data. It can easily handle both numerical and categorical data. One of the major advantages of Random Forest is that it does not suffer from over fitting, even if more trees are appended to the forest. To increase the model accuracy, hyperparameters are also tuned by cross-validation including the number of trees, maximum depth, and minimum split/leaf sizes.

Extreme Gradient Boosting (XGBoost) XGBoost is another powerful algorithm used in many real-world scenarios (1), (2). It is a stage-wise additive model in which each new tree fits the gradient of the multiclass log-loss with regularisation to control complexity. Boosting emphasises difficult cases and often excels when signal is sparse or subtle. Same with Random Forest, hyperparameters in XGBoost are also tuned including the number of trees, depth, learning rate, and subsampling/column sampling rates. To predict bleaching classes (Mild, Moderate, Severe), multiclass log-loss (mlogloss) is used as the objective function in this setting. Mlogloss measures the performance of a classification model whose output is a probability value between 0 and 1, comparing predicted probability distributions to the true labels across multiple classes. It penalizes confident but incorrect predictions more heavily, making it suitable for tasks where distinguishing between multiple classes is critical.

These two baselines possess complementary inductive biases (bagging vs. boosting). Their diversity is the motivation for stacking.

3.3 Stacking Architecture (RF + XGB → Logistic Regression)

One of effective approaches in machine learning classification problems is stacking. The main idea of stacking is using predictions of machine learning models from the previous level as input variables for models on the next level (16). In this case study, the stack is a two-level ensemble:

On level 0: The tuned RF and tuned XGB are trained on the same training data. For each training fold, I generate out-of-fold (OOF) predicted probabilities from each base learner. Hence, concatenating these OOF probabilities produces a meta-feature matrix $Z \in R^{n\times 6}$ (two models $\times$ three class probabilities).

On level 1: A multinomial logistic regression learns weights that combine the base probabilities into final class probabilities. The logistic meta-learner is simple, well-calibrated, and less prone to overfitting than another tree method at the meta level. Using OOF predictions prevents information leakage from the base models into the meta-learner.

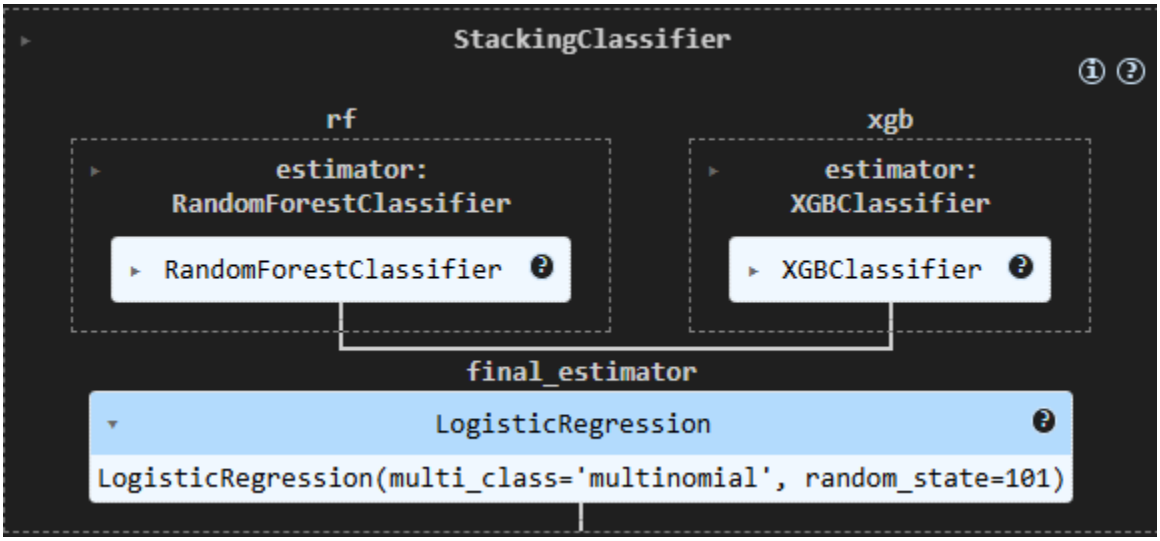

Figure 10: Stacking Classifier Architecture with RandomForest and XGBoost Estimators

3.4 Training Protocol (Split, Standardisation, Grid Search)

Data are split into 70% training and 30% testing with stratification. As many machine learning algorithms perform better when numerical input variables are scaled to a standard range, Standardization technique is also used in this study. Standardization scales each input variable separately by subtracting the mean and dividing by the standard deviation to shift the distribution to have a mean of zero and a standard deviation of one (18). StandardScaler is fitted on the training data only and applied to both train and test predictors. While trees are scale insensitive, scaling benefits the logistic meta-learner and SMOTE's distance computations.

Class imbalance is handled by applying SMOTE to the scaled training set prior to model fitting. For RF and XGB, hyperparameters are tuned with 5-fold cross-validated grid search on the

rebalanced training data. The best base learners are then assembled in the stack; internal cross-validation within the StackingClassifier generates the OOF meta-features for the logistic regression. All random states are fixed for reproducibility.

3.5 Evaluation Metrics

In classification models, key performance metrics include Accuracy, Precision, Recall, and the F1 score. Accuracy measures the overall correctness of the model by calculating the proportion of correct predictions (both positive and negative) out of all predictions, making it a good general indicator when classes are balanced. Precision evaluates how reliable the model is when it predicts a positive outcome, answering the question: "Of all the instances labeled as positive, how many were actually correct?". On the other hand, Recall assesses how well the model identifies all actual positive cases, asking: "Of all the true positive instances, how many did the model correctly detect?". The F1 score combines precision and recall into a single value using their harmonic mean, providing a balanced measure of the model's performance when both false positives and false negatives need to be considered

In order to assess the performance of 2 baseline (RandomForest and XGBoost) and stacking models, the key metrics were calculated and can be expressed as below:

Let $C \in N^{3\times 3}$ be the confusion matrix on the test set with entries C_{ij} counting predictions $\hat{y} = j$ for true class i. For class c,

$$\mathrm{TP}_c = C_{cc}, \quad \mathrm{FP}_c = \sum_{i\neq c} C_{ic}, \quad \mathrm{FN}_c = \sum_{j\neq c} C_{cj}, \quad \mathrm{TN}_c = \sum_{i\neq c}\sum_{j\neq c} C_{ij}.$$

(TP, FP, FN, TN means True Positive, False Postive, False Negative and True Negative, respectively)

Therefore, in each class, the Precision, Recall, F1 scores are:

$$\mathrm{Precision}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FP}_c}, \qquad \mathrm{Recall}_c = \frac{\mathrm{TP}_c}{\mathrm{TP}_c + \mathrm{FN}_c}, \qquad \mathrm{F1}_c = \frac{2\,\mathrm{Precision}_c\,\mathrm{Recall}_c}{\mathrm{Precision}_c + \mathrm{Recall}_c}.$$

I also calculated macro average scores, which give each class equal weight:

$$\text{Macro-Precision} = \frac{1}{3}\sum_{c=0}^{2}\mathrm{Precision}_c, \quad \text{Macro-Recall} = \frac{1}{3}\sum_{c=0}^{2}\mathrm{Recall}_c, \quad \text{Macro-F1} = \frac{1}{3}\sum_{c=0}^{2}\mathrm{F1}_c.$$

The overall Accuracy is calculated as follows:

$$\mathrm{Accuracy} = \frac{\sum_{c=0}^{2}\mathrm{TP}_c}{\sum_{i=0}^{2}\sum_{j=0}^{2} C_{ij}}.$$

3.6 Explainability with SHAP

To connect predictions to mechanisms, we use SHAP (Shapley Additive Explanations). For a model f and input x, SHAP represents the deviation from a baseline as an additive decomposition

$$f(x) - E[f(X)] = \sum_{j=1}^{p} \phi_j(x),$$

where $\phi_j(x)$ is the contribution of feature j computed from Shapley values. (13)

SHAP values are a way to explain the output of a machine learning model by breaking down the contribution of each feature to the prediction, fairly distributing the "credit" for the prediction across all features based on their individual impact, considering all possible combinations of features. (9)

For RF and XGB we compute SHAP values with TreeExplainer, enabling local explanations (waterfall plots for single observations and classes) and global summaries (beeswarm and mean SHAP by class).

For the stacking model, which lacks a native tree explainer, we employ

- Kernel SHAP, a model-agnostic estimator using a local weighted linear surrogate on a background sample to obtain class-wise SHAP values
- A surrogate Random Forest trained to mimic the stack's probabilities, enabling fast TreeExplainer-based waterfalls while validating global trends against Kernel SHAP

In this case, SHAP helps connect the model's predictions of coral severity class to the underlying mechanisms by identifying which features drive the outcome. Together, these tools reveal which variables push a site toward Mild, Moderate, or Severe in both local cases and in aggregate, offering insights into the ecological or environmental mechanisms behind coral bleaching severity.

3.7 Counterfactual "What-If" Intervention Analysis

The "what-if" intervention analysis, conducted as a feasible counterfactual analysis, explores how small, realistic changes to modifiable features might alter predicted outcomes, thereby providing actionable insights (3). In this context, it involves adjusting features like Percent_Cover (which can be influenced through restoration or reseeding efforts) and, where ecologically viable, Turbidity (manageable via water-quality interventions), while keeping unchangeable factors like Depth, Distance to shore, Cyclone climatology, SSTA, etc) constant. Guided by the SHAP results, I treat Percent_Cover and where ecologically acceptable, Turbidity as levers. Factors that cannot be locally changed (e.g., depth, distance to shore, cyclone climatology) are held fixed.

For each test observation I form a counterfactual x' by adding small deltas to the modifiable coordinates and clipping to bounds derived from the training distribution (or i.e. the inter-quantile ranges). The same Standardlisation method then is applied to x', and the stacking model produces new probabilities $f(x')$. Then I compare mean class probabilities before and after the intervention,

$$\bar{p}_c = \frac{1}{n}\sum_{i=1}^{n} f_c(x_i) \quad \text{vs.} \quad \bar{p}'_c = \frac{1}{n}\sum_{i=1}^{n} f_c(x'_i),$$

to quantify the expected reduction in Moderate and Severe under those scenarios.

Nevertheless, this approach is prescriptive rather than causal. It reflects what the trained model (with the highest accuracy) expects under feasible edits consistent with SHAP directions. Consequently, it serves best as a tool to prioritize management strategies and identify potential sites for field validation, rather than functioning as a definitive policy directive on its own.

4 Results

4.1 Model performance (baselines vs. stacking)

After training and predicting on test set, all models achieved strong predictive accuracy, with the Stacking model performing best overall. Random Forest (RF) reached a test accuracy of 86.05%

(training accuracy 97.87%), while XGBoost (XGB) achieved 81.63% on test set and 82.07% on train. The stacking model, which combines RF and XGB through a multinomial logistic meta-learner, attained 86.44% on test set and 98.54% on traiing set. Macro-averaged scores, which weight Mild, Moderate, and Severe equally, show a similar picture: RF obtained a macro precision of 64.47%, macro recall of 69.93%, and macro F1 of 66.63%; XGB followed with 57.48%, 67/69%, and 60.99%, respectively; The stack recorded 65.64%, 67.35%, and 66.31%.

In short, the Stacking model slightly improves test accuracy and macro precision over RF while essentially matching its macro F1; XGB remains competitive but lower on all three macro metrics. The larger gaps between training and testing set for RF and the stack reflect their higher capacity and the use of SMOTE, yet the generalisation remains stable.

	Model	Train Accuracy	Test Accuracy	Mean Precision	Mean Recall	Mean F1-Score
0	Random Forest	0.978734	0.860542	0.644682	0.693316	0.666325
1	XGBoost	0.820657	0.816350	0.574829	0.676870	0.609944
2	Stacking	0.985364	0.864427	0.655392	0.673465	0.663093

Figure 11: Summary of Machine Learning Models

The feature-importance Figure 11 reveals key insights: Percent_Cover emerges as the dominant predictor by a substantial margin, with Turbidity, Cyclone_Frequency, Windspeed, and TSA following in significance. Secondary factors contributing to the predictive signal include Distance_to_Shore, Depth_m, ClimSST, Temperature_Kelvin, and SSTA.. Although impurity-based importances are not causal, the ranking is consistent with the SHAP analyses presented below and with ecological expectations that healthier, higher-cover reefs are more resilient.

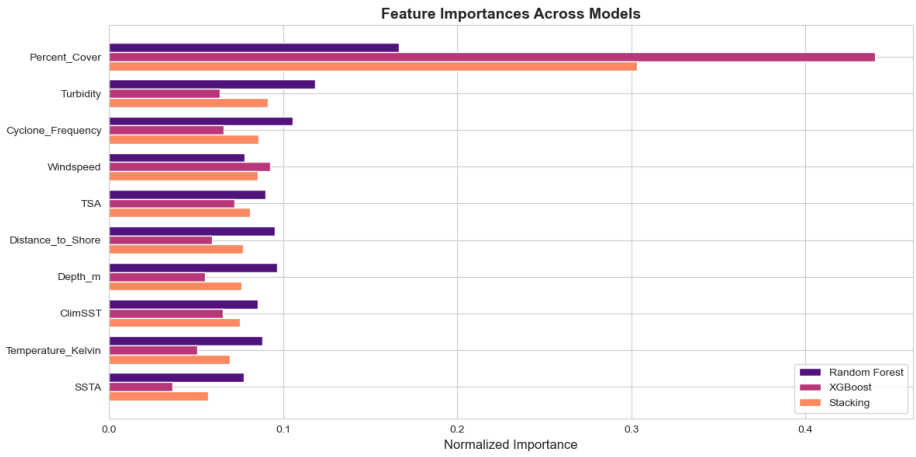

Figure 12: Feature Importance Across Models

4.2 Class probabilities and confusion matrices

In Figure 13, the Stacking model assigns most probability to Mild bleaching (79.55%), with smaller mass on Moderate (13.96%) and Severe (6.49%). This distribution mirrors the underlying class imbalance while still leaving room to detect rarer severe events.

```
Class probabilities using Stacking Model (on TEST set)
          Mean_Prob
Mild       0.795522
Moderate   0.139588
Severe     0.064890
```

Figure 13: Class probabilities using Stacking Model (on TEST set)

Moreover, the confusion matrix (Figure 14) clarifies where the model succeeds and where ambiguity remains. For Mild cases, Recall is 92.64%, indicating that the Ensemble method almost always recognises low-severity outcomes. However. performance on the minority classes is necessarily lower: Moderate is correctly identified 58.46% of the time, with most errors falling back to Mild (29.04%) and a smaller ones mislabelled as Severe (12.50%). Severe is recovered with 50.95% Recall; but its remaining errors split between Moderate (29.19%) and Mild (19.86%). These patterns are typical of imbalanced multiclass scenarios, where adjacent categories share overlapping covariate spaces, and where the cost of false positives is naturally weighed against the aim of achieving high overall accuracy. Notably, the stacking model demonstrates a substantially better balance compared to XGB and a slightly improved balance over RF, as assessed by macro precision and overall accuracy.

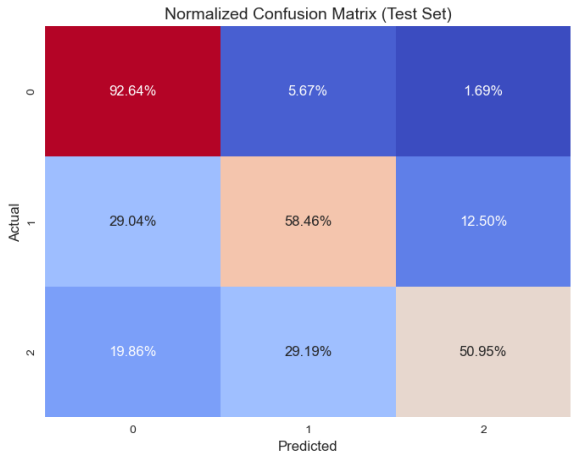

Figure 14: Normalized Confusion Matrix (Test Set)

4.3 Global and local SHAP explanations

Global explanations derived from Kernel SHAP applied to the stacking model, alongside TreeExplainer applied to RF, XGB, and a surrogate tree fitted to the stack's probabilities, consistently point to the same underlying mechanism.

- In the mean SHAP by class summary (Figure 15), Percent_Civer stands out as a strong positive influence on the Mild log odds while negatively affecting the Moderate and Severe log odds.Turbidity exhibits a smaller but similarly directional pattern within the dataset's empirical range, gently moving predictions toward Mild. Other features, ClimSST, Temperature_Kelvi, SSTA, TSA, Windspeed, Cyclone_Frequency, Distance_to_Shore and Depth_m play a vital role in modulating risk, although their impact is notably less

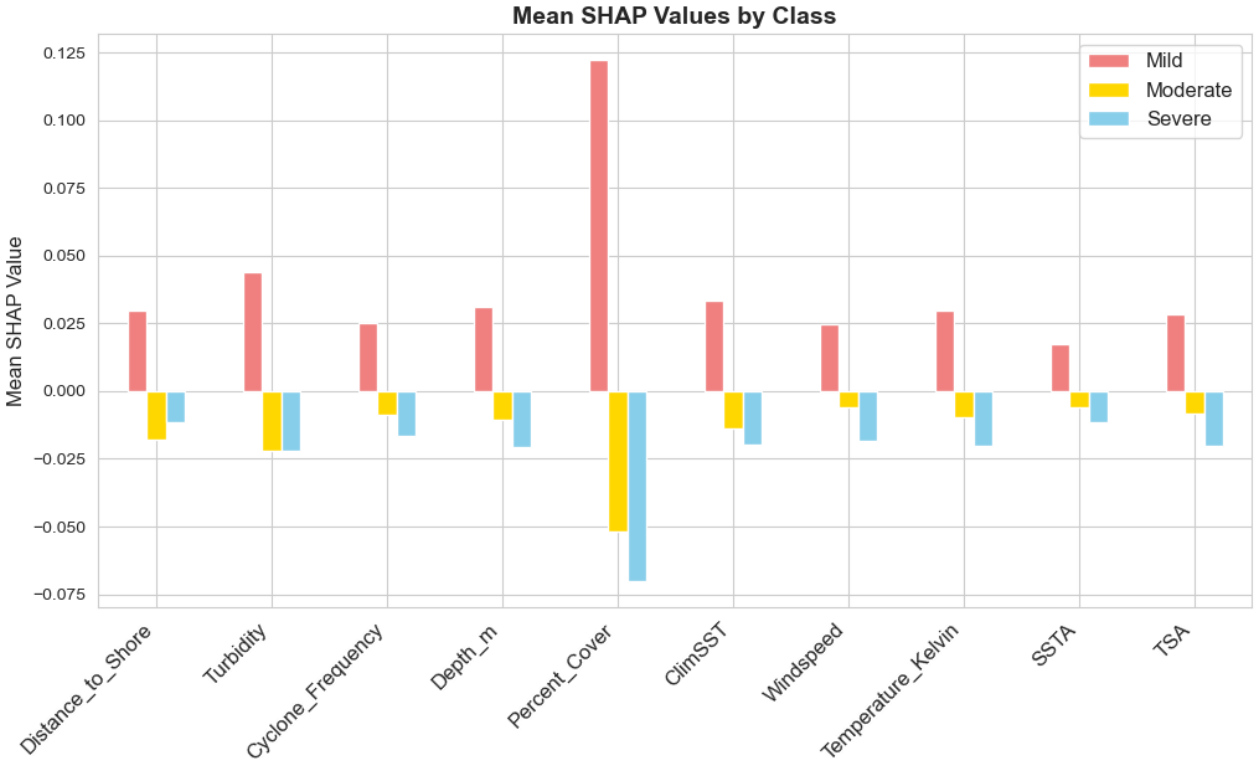

Figure 15: Mean SHAP Values by Class

- The Beeswarm plot (Figure 16) also highlights these patterns: Percent_Cover has a huge spread of SHAP values (up to +0.6), strongly pushing predictions towards Mild Bleaching. In ClimSST, high value strongly increases Mild bleaching classification while low ClimSST tends to reduce probability of Mild Bleaching. This would indicate that reefs under higher ClimSST regimes are not experiencing Severe bleaching, but instead Mild. On the other hand, Turbidity with higher values tend to increase mild positive SHAP values, showing cloudier water may reduce light stress and helping corals avoid severe bleaching. Other factors have either mixed or week influence

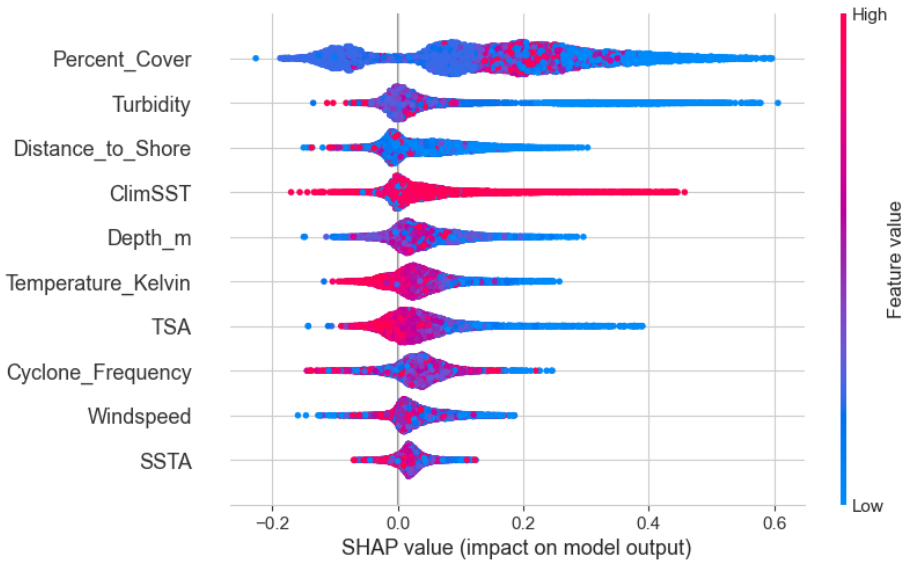

Figure 16: Beeswarm plot (Global Explaination)

Local waterfall plots reveal how these effects accumulate for individual sites.

- For a Mild site (Figure 17), the waterfall plot highlights Percent_Cover as the primary positive shift from the baseline expectation, with ClimSST and Windspeed contributing smaller additional increase. There are also negative adjustments from Temperature_Kelvin, Distance_to_Shore, and Turbidity, however they fail to outweigh the strong positive influence of cover. Hence, they result in a final probability ($f(x)$ =0.88) well above the baseline ($E[f(X)]$ = 0.325)

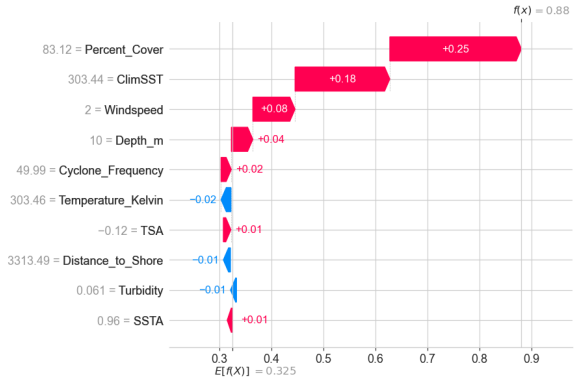

Figure 17: Waterfall plot for class Mild)

- In contrast, for a site with a very low probability of either being Moderate or Severe (Figure 18, 19), the same features work in the opposite direction: Lower coverage and other factors

suppress the Mild log odds, causing the final probability to drop below the baseline. These local insights reinforce the global finding that Percent_Cover serves as the critical driver in the model, with other variables offering secondary refinements.

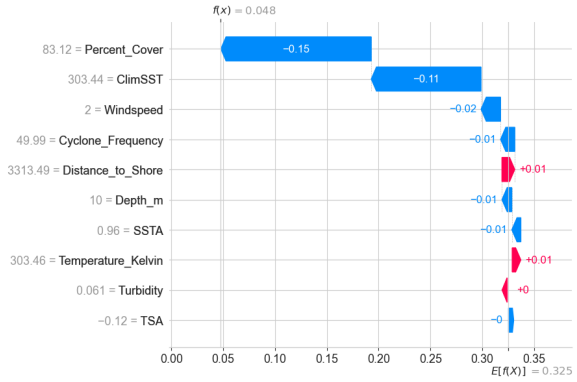

Figure 18: Waterfall plot for class Moderate)

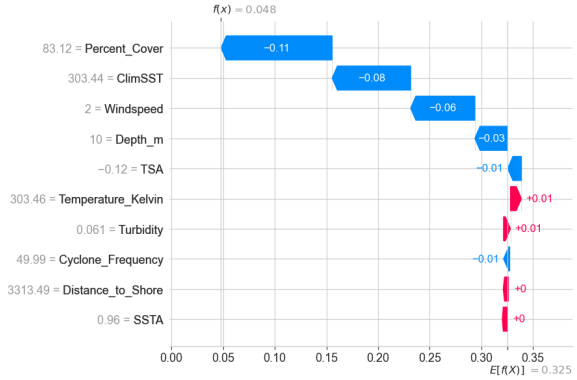

Figure 19: Waterfall plot for class Severe

4.4 Intervention outcomes (before/after class probabilities)

Informed by SHAP insights, I decided to make feasible adjustments. I increased Percent_Cover by an ecologically plausible amount in areas where cover is low, and making small, appropriate tweaks to Turbidity where suitable. All changes were constrained within data-derived limits, while immovable covariates (e.g., depth, distance to shore) remained unchanged. The intervention magnitudes were set at +10.0 percentage points for Percent_Cover (representing re-seeding or restoration efforts) and +0.01 for Turbidity (a small increase if ecologically acceptable)

The adjusted test set with the stacking model reveals a noticeable shift in predicted risks (Figure 20, 21): The mean probability of Mild bleaching rises from 79.55% to 89.89%, marking an absolute increase of 0.1033, which translates to a 12.99% relative improvement. The probability of Moderate bleaching drops from 13.96% to 9.25% (with an absolute decrease of 0.0470, or a -33.71% relative

change), while Severe bleaching falls from 6.49% to less than 1% (with an absolute reduction of 0.0563, or a -86.75% relative decrease).

Class probabilities BEFORE vs AFTER feasible interventions

	Baseline	After_Intervention	Abs_Change	Rel_Change%
Mild	0.7955	0.8989	0.1033	12.9904
Moderate	0.1396	0.0925	-0.0470	-33.7061
Severe	0.0649	0.0086	-0.0563	-86.7495

Figure 20: Class probabilities BEFORE vs AFTER feasible interventions

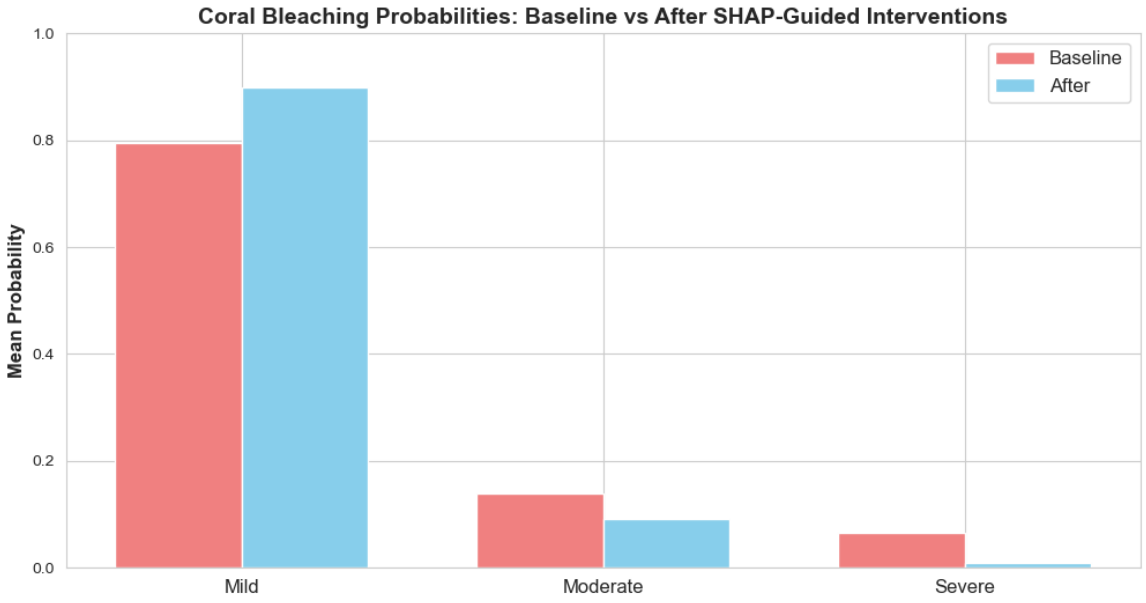

Figure 21: Before and After Interventions Bar Plot

These results should be interpreted as model prescriptions rather than causal effects. They quantify what the trained stacking models predict to happen under a very small changes that move features in directions indicated by SHAP. Therefore, the predictions serve as a starting point for prioritizing management actions and identifying sites for further investigation, providing a basis for decision making without assuming direct causality.

5 Discussion

5.1 Key Drivers of Bleaching Severity in the Model

With the three complementary approaches: stacking model feature importance, Kernel SHAP, and TreeExplainer used on the baseline models as well as a surrogate has pointed some key insights. Live

Percent_Cover emerges as the primary driver of predicted severity. On a global scale, Percent_Cover stands out with the largest mean SHAP magnitude, contributing positively to the Mild log odds and negatively to the Moderate and Severe log odds. At the local level, waterfall plots consistently highlight Percent_Cover as the feature exerting the strongest shift away from the baseline expectation. This suggests the model isn't just amplifying covariate correlations, but in fact cover delivers the most reliable explanatory signal across different sites and years.

A second, weaker factor is Turbidity. Within the empirical range of this dataset, a slight increase in turbidity tends to push predictions toward Mild. It is recognizable that Turbidity might ease photic stress during heat stress, with other factors. The overall effect in the data leans protective but remains modest, with clear boundaries: extreme turbidity is neither recommended nor included in the counterfactual scenarios. Thermal metrics (ClimSST, Temperature_Kelvin, SSTA, TSA) and exposure variables (Windspeed, Cyclone_Frequency, Distance_to_Shore, Depth_m) fine-tune the risk around the dominant cover signal. Their SHAP distributions are narrower and often show mixed directions, reflecting dependent interactions. For example, a specific temperature anomaly might carry different risks depending on depth or background conditions. Overall, the model operates with notable simplicity: live cover bears the brunt of the explanatory weight, while the remaining predictors offer marginal adjustments to the risk profile.

5.2 Management Levers: Modifiable versus Non-Modifiable Features

The model highlights live coral cover as the clearest lever for reducing bleaching risk. In practical terms, this points toward actions that protect and rebuild cover: reducing local stressors, maintaining good water quality, supporting restoration and reseeding, and safeguarding nursery and recruitment habitats. The SHAP guided counterfactuals back this up, showing that even modest increases in cover shift a large share of predicted outcomes away from Moderate and Severe bleaching toward Mild.

Water quality emerges as a second, but more qualified, lever. In the data, turbidity is only a rough stand-in, since it mixes light attenuation with other processes. The management lesson is not to increase turbidity itself, but to maintain light and water-quality conditions that reduce stress on corals without creating new risks. For that reason, interventions in this study kept turbidity within conservative limits and treated it as a supporting factor rather than a primary one.

Other predictors act more as background context than as things managers can directly influence at project timescales. Depth and distance to shore are fixed site features, while cyclone activity and regional thermal conditions operate at scales beyond local control. Short-lived anomalies are better handled through monitoring, early warning, and tactical responses than through long term prevention. In practice, these contextual factors are most useful for setting priorities, which helps to identify sites and seasons where natural conditions already favor resilience, and where targeted investments in coral cover and water-quality management are most likely to deliver meaningful reductions in bleaching risk.

5.3 Robustness and Sensitivity

Looking closely at two potential weak points: how well the stacked model's explanations hold up, and how sampling choices affect results in an imbalanced setting. For the first, I compared Kernel SHAP applied to the stacking classifier with TreeExplainer applied to a Random Forest surrogate trained to mimic the stack's class probabilities. The two were in strong agreement. At the global level, both highlighted the same drivers in the same order and direction; at the local level, the waterfall explanations told the same story. This suggests that Kernel SHAP confirms the surrogate's

explanations are faithful to the ensemble's decision surface, while the surrogate adds speed and produces clearer, class-specific plots.

For the second, I tested sensitivity to class imbalance and to the specification of the SHAP background distribution. Stratified splitting and applying SMOTE solely to the training partition improved balance, though the expected asymmetry persisted: Recall was highest for Mild cases and lower for the minority classes. Reporting macro-averaged metrics and examining the normalised confusion matrix helped temper optimism and reveal the trade-offs more clearly. Changing the size of the Kernel SHAP background sample shifted the absolute values but not the ranking of features. Likewise, moderate tweaks to SMOTE parameters led to small changes in minority recall but did not alter the main conclusions. The ensemble model reveals that live cover is the strongest driver, and that even modest increases reduce the chance of severe outcomes. Future refinements could include the use of class balanced or focal loss functions or probability calibration to stabilise thresholds for Moderate and Severe classes.

5.4 Limitations and Assumptions

There are several important caveats to keep in mind. This study is observational rather than experimental, which means it identifies associations rather than proving causal effects. The counterfactual experiments show what the trained model expects under small, feasible changes, but they should be used to guide hypotheses and highlight sites for field validation, not as stand-alone prescriptions. Both the labels and environmental covariates come from mixed data sources, and the timing between bleaching observations and environmental drivers is not perfectly aligned. The models also treat covariates mostly as snapshots rather than full time histories. Similarly, the feature selection results represent a snapshot in time: the top 10 predictors identified here reflect the current dataset and period, but the ranking could shift with new covariates, better sensors, or a different regional focus.

The use of SMOTE introduces synthetic samples that cannot perfectly capture the complexity of the real data. I limited this risk by scaling before resampling and by testing only on untouched data, but some inflation of variance is unavoidable. Another limitation is that the training data are dominated by observations from the Pacific and Atlantic, raising the familiar concern of domain shift when applying the model to under-represented regions or to future climate conditions that go beyond the historical record.

Even with these limitations, the framework is practically useful. It combines strong predictive performance with explanations that are consistent across methods and easy to translate into management action. The core message is stable and actionable: protecting and rebuilding live coral cover, while maintaining suitable water-quality conditions, offers the clearest path, according to the data and this model—toward lowering the risk of Moderate and Severe bleaching.

6 Conclusion

This study shows that a simple, disciplined stacking ensemble can predict coral bleaching severity accurately while remaining interpretable and operational. Using a global dataset and a compact, RFE-selected set of predictors, the stacked model (RF + XGB with a multinomial logistic meta-learner) achieved the best test performance among the models I evaluated (Accuracy 86.44%, Macro F1 66.3%). More importantly, SHAP analyses, which are computed directly on the stack and cross-checked with a tree-based surrogate, reveal a coherent mechanism behind those predictions: live coral cover is the dominant driver, with smaller contributions from turbidity, thermal descriptors, and exposure-related variables.

I used these explanations not only to describe model behavior but also to ask "What-If" questions under realistic constraints. Modest, feasible increases in live cover produced a large reallocation of probability mass away from Moderate and Severe outcomes and toward Mild (Mild +0.1033 absolute; Moderate 0.0470; Severe 0.0563). While not causal, these scenario results are directionally robust across explanation methods and provide a practical basis for prioritizing management

The work has limits. It is observational: labels and covariates come from heterogeneous sources, and imbalance, domain shift, and measurement error inevitably shape the learned decision surface. Nonetheless, the central message is consistent across models, plots, and sensitivity checks. For managers and practitioners, the research indicates that actions protecting and rebuilding live coral cover, alongside maintaining appropriate water-quality regimes, are the most reliable levers, which are essential for reducing the likelihood of higher-severity bleaching.

Methodologically, these underlined techniques offer a reproducible template that can be extended. Future work could incorporate richer spatiotemporal signals (e.g., DHW trajectories, hydrodynamics), cost-sensitive training and calibration for decision thresholds, and targeted validation in under-represented regions. Even in its current form, the framework is ready to support screening and prioritization, linking predictive performance with explanations that translate into concrete, testable actions on the reef.

References

[1] Bakasa, W. and Viriri, S. (2023). Vgg16 feature extractor with extreme gradient boost classifier for pancreas cancer prediction. *Journal of Imaging*, 9(7):138.

[2] Bansal, A. and Kaur, S. (2018). Extreme gradient boosting based tuning for classification in intrusion detection systems. In *International conference on advances in computing and data sciences*, pages 372–380. Springer.

[3] Box, G. E. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical association*, 70(349):70–79.

[4] Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

[5] Cornet, V. J., Cantin, N. E., Joyce, K. E., Leggat, W., Ainsworth, T. D., and Heron, S. F. (2025). Enhancing coral bleaching predictive tools through integrating sensitivity to heat exposure. *Biological Conservation*, 302:110958.

[6] Costanza, R., De Groot, R., Sutton, P., Van der Ploeg, S., Anderson, S. J., Kubiszewski, I., Farber, S., and Turner, R. K. (2014). Changes in the global value of ecosystem services. *Global environmental change*, 26:152–158.

[7] Cutler, D. R., Edwards Jr, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., and Lawler, J. J. (2007). Random forests for classification in ecology. *Ecology*, 88(11):2783–2792.

[8] El-Naggar, H. A. (2020). Human impacts on coral reef ecosystem. In *Natural resources management and biological sciences*. IntechOpen.

[9] Fabian, P. (2011). Scikit-learn: Machine learning in python. *Journal of machine learning research 12*, page 2825.

[10] Ghimire, B., Rogan, J., and Miller, J. (2010). Contextual land-cover classification: incorporating spatial dependence in land-cover classification models using random forests and the getis statistic. *Remote Sensing Letters*, 1(1):45–54.

[11] Gregorutti, B., Michel, B., and Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27(3):659–678.

[12] Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422.

[13] Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

[14] Mi, Y. (2013). Imbalanced classification based on active learning smote. *Research Journal of Applied Science Engineering and Technology*, 5:944–949.

[Nitesh and Kevin] Nitesh, V. and Kevin, W. Lokw (2002). smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, pages 321–357.

[16] Pavlyshenko, B. (2018). Using stacking approaches for machine learning models. In *2018 IEEE second international conference on data stream mining & processing (DSMP)*, pages 255–258. IEEE.

[17] Sheppard, C., Dixon, D. J., Gourlay, M., Sheppard, A., and Payet, R. (2005). Coral mortality increases wave energy reaching shores protected by reef flats: examples from the seychelles. *Estuarine, Coastal and Shelf Science*, 64(2-3):223–234.

[18] Thakker, Z. L. and Buch, S. H. (2024). Effect of feature scaling pre-processing techniques on machine learning algorithms to predict particulate matter concentration for gandhinagar, gujarat, india. *Int. J. Sci. Res. Sci. Technol*, 11(1):410–419.

[19] Thiault, L., Curnock, M. I., Gurney, G. G., Heron, S. F., Marshall, N. A., Bohensky, E., Nakamura, N., Pert, P. L., and Claudet, J. (2021). Convergence of stakeholders' environmental threat perceptions following mass coral bleaching of the great barrier reef. *Conservation Biology*, 35(2):598–609.

[20] Udilă, A. (2023). Encoding methods for categorical data.

[21] van Woesik, R. and Burkepile, D. (2022). Bleaching and environmental data for global coral reef sites from 1980-2020. https://www.bco-dmo.org/dataset/773466. Version 2, Version Date 2022-10-14, doi:10.26008/1912/bco-dmo.773466.2, accessed 2025-10-02.

[22] Wang, S., Dai, Y., Shen, J., and Xuan, J. (2021). Research on expansion and classification of imbalanced data based on smote algorithm. *Scientific reports*, 11(1):24039.