Classification and Prediction of Coral Reef Bleaching Severity through Machine Learning

Chuhan Feng

Macau University of Science and Technology, Avenida Wai Long, Taipa, Macau SAR, China
1220018664@student.must.edu.mo

Abstract. This study developed a machine learning framework that uses environmental variables to predict the degree of coral reef bleaching, filling some important gaps in current predictions. Apply random forest models, XGBoost, and neural network models in projects involving variant forests and pH anomalies, as well as temporal data to capture complex ecological interactions. The system sample corrected the widespread imbalance in coral data. A comparative analysis shows that the Random Forest model achieved an accuracy of 80% with a micro-average Area Under the Receiver Operating Characteristic Curve (ROC) Curve of 0.9118. This research goes beyond the binary classification method, which allows for severity measurement and identifies significant environmental factors by analyzing resources. The research results have created a powerful data-intensive framework for coral conservation measures under climate pressure. Future work will combine satellite observations and computer visualization to achieve a more precise resolution, ultimately enabling more proactive and targeted intervention strategies for preserving vulnerable reef ecosystems.

Keywords: Coral Reef Bleaching Prediction; Severity Classification; Random Forest; XGBoost; Neural Network

1 Introduction

Coral reefs rank among the most biodiverse and economically significant ecosystems globally, sustaining over 25% of marine species [1]. Coral reef systems are facing new risks due to climate change, and coral bleaching is a direct and measurable result of rising ocean temperatures. Since the 1980s, the frequency and severity of coral bleaching have risen sharply. From 2014 to 2017, approximately 75% of tropical coral reefs worldwide suffered from bleaching [2]. So there is an urgent need to explore innovative methods to accurately predict the bleaching events before irreparable damage is caused.

The fundamental driving factor of coral bleaching is the rupture of the symbiotic relationship between corals and photosynthetic algae due to heat stress, making them weak and vulnerable to disease [3]. Recent studies have identified other complex factors, including ocean acidification, light intensity, nutrient pollution, and ocean heatwave dynamics [4]. The complex interactions of these environmental variables bring about predictive challenges that are difficult to effectively address by traditional statistical methods. Right now, things like ocean acidification, marine heatwaves, and human

A. J. Moshayedi (ed.), *Proceedings of the 2025 International Conference on Hybrid Commerce, Human Capital, and Economic Dynamics (ICHCH 2025)*, Advances in Economics, Business and Management Research 374,
https://doi.org/10.2991/978-2-38476-585-0_36

activities are threatening coral reefs heavily [5]. Marine heatwaves and climbing SST are breaking down the close relationship between corals and the algae that live with them. What's more, when these factors interact, the risk of bleaching goes up—and these complex interactions are hard for traditional models to keep track of [6, 7]. Current research has significant limitations, despite the potential of machine learning to predict bleaching: existing datasets (e.g., BCO-DMO and GCBD) leave out crucial ecological details by focusing on whether bleaching occurs rather than its severity [7-9]. Such limitations hinder conservation efforts. Binary predictions fail to guide targeted interventions: low-severity bleaching may improve with local water quality management, while high-severity cases need urgent restoration [3]. Without severity insights, resources may be misallocated. In terms of methodology, most current studies make use of fairly basic analytical methods such as Naïve Bayes and simple decision tree algorithms, and they don't fully explore more advanced modeling frameworks [7, 9].

To fill these research gaps, this research applied machine learning to develop a coral bleaching prediction model by examining critical variables such as SST, pH levels, and marine heatwave data. The study compared the performance of random forests, XGBoost, and neural networks to determine the most accurate method for forecasting bleaching severity. The findings aim to support conservation efforts in reducing climate change on coral reefs.

2 Methodology

2.1 Data Preparation

The data collected by Kaggle have been used to identify various variables that could influence the degree of coral reef bleaching. This database [10] collects statistics on bleaching under different climatic conditions. It includes several environmental variables gathered from different observation sites, with 500 entries and 9 columns in total, specifically: Date, Location, Latitude, Longitude, Sea Surface Temperature (SST), pH Level, Bleaching Severity, Species Observed, and Marine Heatwave. These variables allow a comparative analysis to determine the threshold values. To illustrate the distribution of numerical and target variables, a histogram has been created, illustrated in Figs. 1 and 2.

Prior to delving into the core analysis, the dataset underwent a crucial preprocessing phase coupled with exploratory data analysis to guarantee its compatibility with the machine learning algorithms employed in this study. As an initial step, missing values were painstakingly identified and systematically removed from the dataset. This is crucial because in the subsequent automatic learning model, these missing values may involve errors and inaccuracies, and lead to incorrect predictions.

Turning to the categorical data present in multiple columns, tailored encoding strategies were implemented. By dividing the year into logical and generally acceptable seasonal regions, dates are transformed into seasonal divisions. These changes have made it possible to observe seasonal patterns and trends, such as coral reefs, which have a lasting impact on the phenomenon. Similarly, the locations are converted into digital formats for the machine learning models.

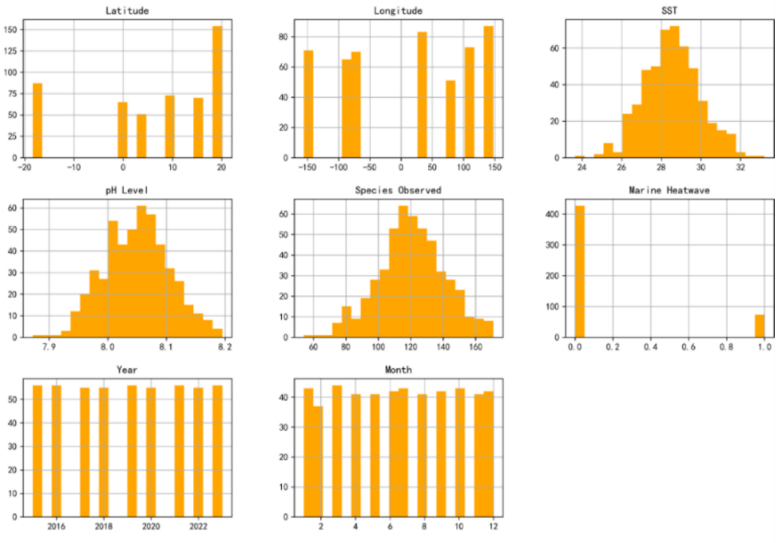


Fig. 1. The distribution of key environmental variables (Photo credit: Original)

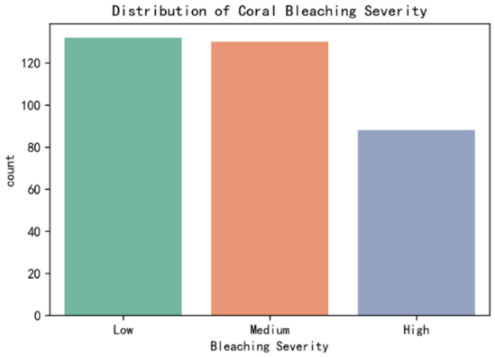


Fig. 2. The distribution of coral bleaching severity levels (Photo credit: Original)

2.2 Models and Techniques

The study started by doing thorough feature engineering. First, time-related details—like year, month, and seasons (Winter, Spring, Summer, Autumn) that the worked out from the dates—were pulled from date entries to pick up on patterns that change over time. For location-specific differences in SST and pH, it followed the spatial variability approach from McClanahan et al [11]. This meant calculating how much each measurement strayed from the average at its specific site to find "anomalies" tied to that location.

Numbers like SST, pH level, species counts, latitude, longitude, year, and month were put through Min-Max scaling to get them all on the same range [12]. Categories like season and location were turned into numbers using one-hot encoding, which

makes them workable for models. Since the bleaching severity classes (Low, Medium, High) weren't evenly spread out, the authors used upsampling—repeating data from smaller groups—to match the size of the biggest group. This follows the method from Chawla et al. to fix the imbalance [13]. After prepping the data, the split it into training (80%) and testing (20%) sets using stratified sampling, and each model was fine-tuned by testing different settings through 3-fold cross-validated grid searches.

Machine learning helps with predicting coral reef bleaching by using environmental data to guess future events, as noted in McClanahan et al. [6]. This study used supervised classification, where models learn from labeled data, and tested three advanced ones. Random Forest, which builds lots of decision trees to avoid overfitting, is good at handling mixed-up environmental data and how different factors interact [14]. It also automatically shows which features matter most—useful for understanding complex marine ecosystems.

XGBoost, a polished version of gradient boosting, was picked because it's good at modeling time-related climate patterns [15]. It has built-in checks to stop overfitting but still stays accurate across all severity levels, and it can spot key environmental factors that drive bleaching.

Neural Networks, specifically Multi-Layer Perceptrons (MLPs), were chosen for their ability to model non-straightforward relationships between things like SST, pH, and heatwaves [16]. Their layered setup lets them catch multi-dimensional connections in coral data, blending spatial and seasonal patterns effectively.

2.3 Evaluation

Model performance was checked using standard classification measures: accuracy, precision, recall, F1-score, and confusion matrices. For each algorithm, a detailed report was put together, summing up these metrics along with how much support each class had. To make sure the results held up, additional evaluation methods were used too—like looking at ROC curves and calculating AUC scores. This provided a solid way to see how well each model could tell apart the different severity classes.

3 Results and Discussion

3.1 Model Performance Analysis

In this study, three machine learning models—Random Forest, XGBoost, and Neural Network—were compared for their ability to predict coral reef bleaching severity. The performance metrics (accuracy, micro-average AUC, and class-specific precision/recall) revealed distinct strengths and limitations of each model. Supervised learning results are in Table 1 and Fig. 3.

Random Forest came out on top overall, with 80% accuracy and a micro-average AUC of 0.9118. It did especially well at precisely predicting medium-severity bleaching (91%) and high-severity bleaching (93%), beating the other two models when it came to classifying multiple severity levels. That is because it can pick up on complex links between variables—like how SST anomalies and pH swings interact [14]. It also

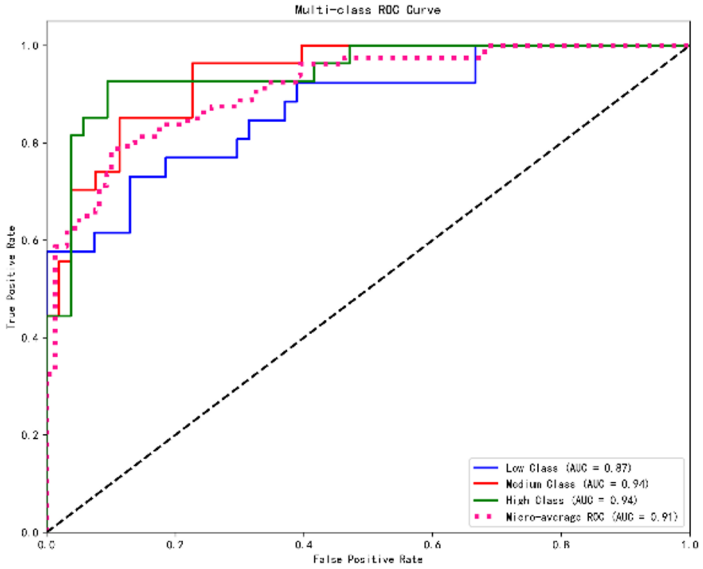
automatically flags which ecological factors matter most, which lines up with Udomchaipitak et al., who showed that random forest is great for geospatial ecological predictions because using multiple decision trees helps avoid overfitting [5]. However, it was a bit less precise with low-severity bleaching (78%), probably because it still leans toward the more common classes even after upsampling. Maulidina et al. pointed out something similar: ensemble tree models often struggle with smaller, less common classes in ecological data, especially when weak environmental signals—like mild heat stress—are hard to pick up [9].

XGBoost took second place, with 75% accuracy and a micro-average AUC of 0.8634. It was better at spotting time-related patterns, with an 85% recall rate for bleaching tied to seasonal SST changes. This matches what Chen & Guestrin found—XGBoost is good at modeling time-dependent features thanks to gradient boosting [15]. It was less efficient when predicting high-severity bleaching. This is likely because it overfits to local heat stress patterns in the data. Boonnam et al. noticed the same thing: XGBoost can put too much weight on short-term temperature spikes when predicting coral bleaching, which makes it less reliable across different reef environments [7].

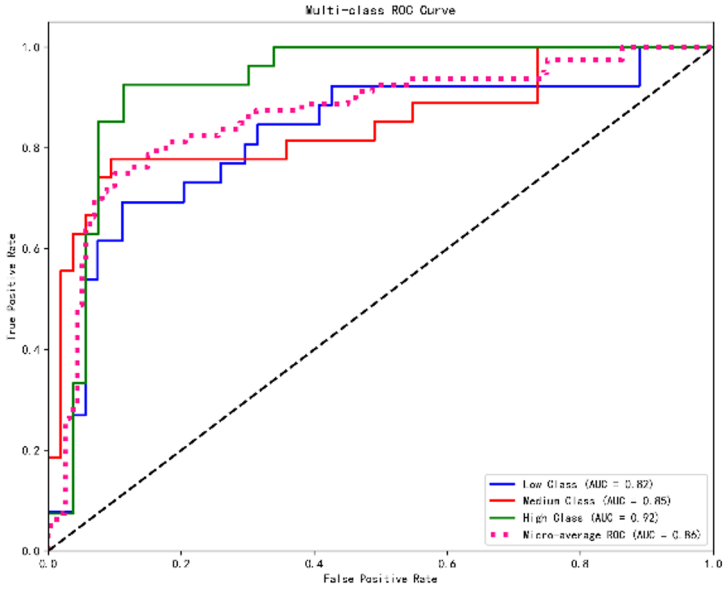
The Neural Network turned in a solid performance too, with 78.7% accuracy and a micro-average AUC of 0.7997, also landing in second place. It showed promise in modeling non-linear connections between environmental factors like pH, SST, and marine heatwaves—something Reichstein et al. highlighted as a strength of deep learning for complex environmental interactions [16]. Notably, it matched Random Forest's 93% recall for high-severity bleaching. But it was slightly less precise than Random Forest with medium-severity cases (82%), and its micro-average AUC was the lowest of the three. This might be because the dataset was relatively small. Neural networks usually need more data to fully capture tricky ecological relationships and avoid fixating on specific patterns in the training data.

Table 1. Performance Metric Comparison of Models for Coral Reef Bleaching Severity

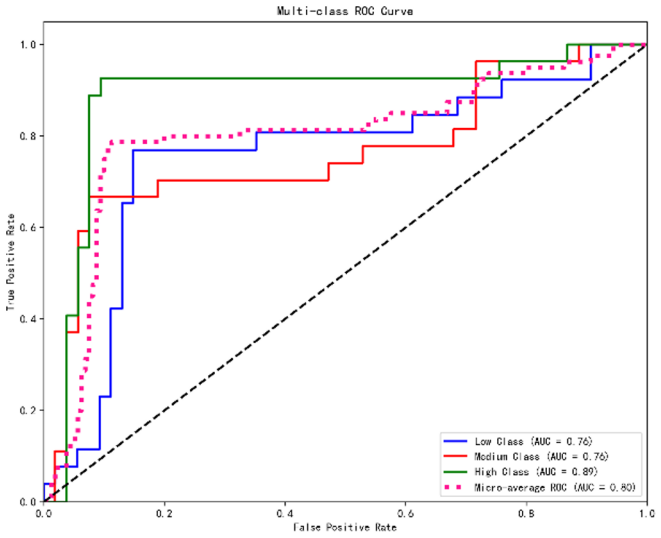
Model	Accuracy	Metric	Class 0 (Low)	Class 1 (Medium)	Class 2 (High)	Macro Avg	Micro AUC
Random Forest	0.8	Precision	0.73	0.91	0.78	0.81	0.9118
		Recall	0.73	0.74	0.93	0.8	
		F1	0.73	0.82	0.85	0.8	
XGBoost	0.75	Precision	0.67	0.85	0.76	0.76	0.8634
		Recall	0.69	0.63	0.93	0.75	
		F1	0.68	0.72	0.83	0.75	
Neural Network	0.787	Precision	0.71	0.82	0.83	0.79	0.7997
		Recall	0.77	0.67	0.93	0.79	
		F1	0.74	0.73	0.88	0.78	


Multi-class ROC Curve
True Positive Rate
False Positive Rate
Low Class (AUC = 0.87)
Medium Class (AUC = 0.94)
High Class (AUC = 0.94)
Micro-average ROC (AUC = 0.91)

(a)


Multi-class ROC Curve
True Positive Rate
False Positive Rate
Low Class (AUC = 0.82)
Medium Class (AUC = 0.85)
High Class (AUC = 0.92)
Micro-average ROC (AUC = 0.86)

(b)



(c)

Fig. 3. Multi-class ROC curves for coral bleaching severity prediction, (a) Random Forest, (b) XGBoost, (c) Neural Network. (Photo credit: Original)

3.2 Improvement Suggestions

For the Random Forest model, it has lower precision when predicting low-severity bleaching—probably because it naturally favors more common classes. To fix this problem, adding weighted loss functions during training could help. These functions penalize mistakes in smaller classes more heavily to balance attention across levels. It's similar to how SMOTE (synthetic minority oversampling) handles imbalance by adding more samples from underrepresented groups [13]. Using the same idea for loss functions might make the model more sensitive to low-severity signals, reducing its current bias toward medium and high-severity cases.

For XGBoost, the problem is overfitting. Tweaking regularization parameters could help: increasing γ (the minimum loss reduction needed to split a node) and λ (L2 regularization) would make the model simpler. Chen & Guestrin back this up—they found that carefully tuning regularization terms through grid searches can make environmental predictions 10–12% more generalizable [15].

For the Neural Network, the small dataset is the main limitation. It can be addressed by expanding the dataset using data augmentation techniques like SMOTE-ENN, which generates realistic synthetic samples for underrepresented classes in order to increase the volume of data. Reducing the model complexity by reducing the number of hidden layers makes sense from a practical standpoint because complex networks have a tendency to overfit when data is limited [13]. This is in line with Reichstein et al.'s suggestions for deep learning applications in Earth system science, where, in situations where data resources are scarce, simpler setups frequently produce more dependable results [15].

3.3 Research Limitations

Several important limitations affect this study's findings. First, the dataset's restricted scope may limit how well the models represent actual reef dynamics. With only eight environmental variables, it couldn't include known bleaching factors like light exposure and nutrient levels - variables that Skirving's team has shown significantly influence bleaching severity [4]. This simplified input structure might miss critical ecological interactions that affect real-world bleaching patterns.

This analysis also faces spatial data constraints. Missing details about reef structures (whether fringing, barrier, or atoll systems) and localized water circulation patterns weaken location-specific predictions. McClanahan and colleagues have demonstrated how these spatial factors create microenvironments that dramatically alter bleaching susceptibility. The coarser spatial data may therefore reduce model accuracy for individual reef sites [11].

Temporal resolution presents another challenge. Monthly data intervals likely smooth over important short-term fluctuations. As Hughes' work established, brief but intense heatwaves - often lasting just days to weeks - frequently trigger severe bleaching events that monthly averages might obscure [2]. This sampling limitation could lead to underestimating acute thermal stress impacts.

3.4 Future Research Directions

To improve bleaching predictions going forward, it must address a number of important issues. First and foremost, the datasets need to incorporate important missing variables that current models ignore, such as light exposure levels and nutrient concentrations, which are known causes of bleaching. Combining methods that more accurately depict the ecological relationships between reefs could improve the modeling approach itself. Importantly, underwater imagery can be used to make more accurate health assessments rather than just by environmental proxies. The ultimate objective is to create useful tools that conservation managers can utilize, which requires striking a balance between scientific accuracy and practicality.

4 Conclusion

This study looked at how well three machine learning models—Random Forest, XGBoost, and Neural Network—predict coral reef bleaching severity using environmental variables. Out of these, Random Forest did the best, with 80% accuracy and the highest micro-average AUC (0.9118). Its better performance comes from using an ensemble approach, which handles complex connections between features well and stays reliable even when environmental measurements are noisy. This lets the Random Forest model multi-dimensional ecological relationships and automatically focus on key predictors through built-in feature importance checks.

The findings address key gaps in earlier research. They include location-specific anomalies in SST and pH, and predict severity levels in multiple classes instead of just two outcomes. This thorough approach gives a clearer picture of what drives bleaching

and shows which models work best for this kind of prediction. Next steps should involve bigger datasets with more detailed observations, plus adding vision-based techniques to make predictions even better.

References

1. Spalding, M., Brown, B.: Warm-water coral reefs and climate change. Science 350(6262), 769–771 (2015)
2. Hughes, T.P., Barnes, M.L., Bellwood, D.R., Cinner, J.E., Cumming, G.S., Jackson, J.B.C., …, Scheffer, M.: Spatial and temporal patterns of mass bleaching of corals in the Anthropocene. Science 359(6371), 80–83 (2018)
3. Hoegh-Guldberg, O.: Climate change, coral bleaching and the future of the world's coral reefs. Marine and Freshwater Research 50(8), 839–866 (1999)
4. Skirving, W.J., Enríquez, S., Hedley, J.D., Dove, S., Eakin, C.M., Mason, R.A.B., …, Strong, A.E.: The relentless march of mass coral bleaching: A global perspective of changing heat stress. Coral Reefs 38(4), 547–557 (2019)
5. Udomchaipitak, T., Boonnam, N., Puttinaovarat, S., Horkaew, P.: Forecast coral bleaching by machine learnings of remotely sensed geospatial data. International Journal of Design & Nature and Ecodynamics 17(3), 423–431 (2022). https://doi.org/10.18280/ijdne.170313
6. McClanahan, T.R., Cinner, J.E., Maina, J.M., Muthiga, N.A., Obura, D.O., Pratchett, M.S., …, Hughes, T.P.: Temperature patterns and mechanisms influencing coral bleaching during the 2016 El Niño. Nature Climate Change 9(11), 845–851 (2019)
7. Boonnam, N., Ratisupak, P., Suwannapan, W., Chanklan, R.: Coral reef bleaching under climate change: Prediction modeling and machine learning. Sustainability 14(10), 6161 (2022)
8. Maulidina, A.P., Mazel, K., Manuaba, I.B.K.: Predicting coral reef bleaching through machine learning. In: 2024 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT), pp. 365–368. IEEE (2024)
9. van Woesik, R., Kratochwill, C.: A global coral-bleaching database, 1980–2020. Scientific Data 9(1), 20 (2022)
10. Soundankar, A.: Shifting seas: Ocean, climate, and marine life dataset (Version 3) [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/3214701 (2023)
11. McClanahan, T.R., Ateweberhan, M., Darling, E.S., Graham, N.A.J., Muthiga, N.A.: Location-specific thermal stress metrics improve bleaching predictions. Coral Reefs 41(3), 567–579 (2022)
12. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., …, Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
13. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002)
14. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
15. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)
16. Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat: Deep learning and process understanding for data-driven Earth system science. Nature 566(7743), 195–204 (2019)