ReefCast: Prediction of Coral Bleaching using Machine Learning in India

Rhythm Mehra, Rishit Anand and Shrijak Kumar

Abstract

Coral reefs in India are increasingly threatened by climate change, yet most existing bleaching assessments are retrospective and overly reliant on sea surface temperature (SST) or image-based methods. These approaches often ignore region-specific ecological and climatic factors. This study introduces a predictive machine learning framework tailored to Indian reef systems, integrating variables such as SST, pH, salinity, fCO2, turbidity, coral genera, Degree Heating Weeks (DHW), and climate indices like IOD and ENSO on a compiled dataset covering the coral reefs in the Andaman & Nicobar Islands, Lakshadweep, Gulf of Mannar, and Gulf of Kachchh. By incorporating localized, multi-variable data, our model improves the accuracy of bleaching forecasts and provides a scalable frame- work for proactive reef management in India and similarly underrepresented reef regions worldwide.

Keywords:

Coral Bleaching, Machine Learning, LSTM, XGBoost, LightGBM, Random Forest, Climate Change, India, Environmental AI

Introduction

Coral bleaching prediction is a critical challenge in marine conservation, particularly for India's reef ecosystems, where environmental stressors are complex and region-specific. Coral reefs in the Andaman & Nicobar Islands, Lakshadweep, Gulf of Mannar, and Gulf of Kachchh play a vital role in supporting marine biodiversity and sustaining local economies. However, these ecosystems are increasingly affected by rising sea surface temperatures (SST), ocean acidification, salinity fluctuations, and other climatic stressors.

The core problem addressed is the inadequacy of traditional SST-threshold models, which fail to capture the complex, nonlinear interactions between multiple climatic and oceanographic variables. While remote sensing and statistical methods have expanded monitoring capabilities, these approaches are often limited by data resolution and single-variable dependencies.

A key gap in the current literature is the lack of India-specific, machine learning (ML) based forecasting frameworks that incorporate a diverse set of predictive features including macro-climatic drivers like the El Niño Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD), turbidity, pH, and coral genera susceptibility. Addressing this gap is crucial for building adaptive management tools suited to India's unique reef systems.

This paper argues that a rigorously evaluated suite of temporally aware and multi- variate machine learning models, especially LSTM networks and ensemble approaches like XGBoost, LightGBM, and Random Forest, when fed finely grained, site-specific environmental and climatic data, can substantially enhance the accuracy of coral bleaching forecasts in Indian reef ecosystems. The paper proceeds with a discussion on the significance of the problem, literature review, methodology, results, and conclusion.

Background and Significance

Coral reefs are among the most biologically diverse ecosystems on Earth, supporting approximately 25% of all marine species despite occupying less than 1% of the ocean floor [13]. In India, coral reefs in regions like the Andaman & Nicobar Islands, Lakshadweep, Gulf of Mannar, and Gulf of Kachchh provide essential ecosystem services protecting coastlines, supporting fisheries, and sustaining livelihoods [12]. However, these ecosystems are increasingly threatened by recurrent coral bleaching events caused by environmental stressors.

Mass bleaching events in 1998, 2010, and 2016 have highlighted the growing vulnerability of Indian reefs to elevated sea surface temperatures (SST), and changing salinity patterns [1][17]. Coral bleaching primarily results from the expulsion of symbiotic zooxanthellae due to thermal stress, but it is also influenced by factors like freshwater influx from monsoonal rains, turbidity, ocean acidification, and climatic oscillations such as the El Niño-Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD) [6][14].

Traditional prediction models are mostly dependent on SST thresholds, which often overlook the interactions among several environmental variables. These models are especially inadequate for India's reefs, which differ in species composition and local oceanographic conditions [19]. This lack of region-specific predictive capacity is a critical gap in current marine management strategies.

Our study addresses this gap by proposing machine learning models capable of integrating multivariate datasets such as SST, salinity, pH, fCO2, species susceptibility, and climate indices to provide a more accurate, India-specific prediction of coral bleaching events. This approach has the potential to improve early warning systems, inform conservation policies, and support regional marine biodiversity resilience.

Literature Review

Positioning within Marine Ecology and Climate Science

The interdisciplinary nature of coral bleaching prediction places this research at the nexus of Climate Science and Marine Biology, with technical integration from Environmental Data Science and Machine Learning. As coral reefs are key indicators of climate-induced stress in marine environments, understanding the drivers and dynamics of bleaching aligns with the broader goals of climate impact research [5]. At the same time, research on coral ecosystems and species-specific resilience is essential to conservation science and marine biology.

Origins of Coral Bleaching Research and Key Contributions

Over the last two decades, the climate science research community has placed a strong emphasis on quantifying the impact of global warming on marine systems, particularly through metrics such as Sea Surface Temperature (SST) anomalies and Degree Heating Weeks (DHW). These metrics are commonly employed in climate models derived from satellite data and aid in tracking accumulated thermal stress. Marine biologists, on the other hand, have investigated coral physiology, species susceptibility, and the ecological impacts of bleaching events across diverse reef systems [6].

Recent years have seen a growing convergence between these fields, as traditional statistical models and field-based assessments struggle to cope with the multivariate complexity and spatial variability of bleaching events. This has created a niche for machine learning models, which offer tools capable of processing and learning from large, heterogeneous datasets that include climatic, oceanographic, and biological variables.

Within India, this type of ML research is still nascent. Building on remote sensing-derived environmental and climate data, machine learning models, especially LSTM and ensemble methods like XGBoost, LightGBM, and Random Forest, can turn passive monitoring into precise, site-specific coral bleaching forecasts for India's reefs. This research is situated within this emerging subfield of environmental AI, aiming to enhance India-specific coral bleaching prediction by integrating environmental stressors, climate indices (e.g., ENSO, IOD), and biodiversity metrics using advanced ML techniques such as Random Forest, XGBoost, and LSTM-based models.

The study of coral bleaching initially emerged within the domain of marine biology in the mid-20th century, following early observations of coral discoloration and mortality events in response to elevated temperatures. As ecological concerns grew in the 1980s and 1990s, researchers began documenting mass bleaching events on a global scale. Key studies established the foundational understanding that coral bleaching is closely related to thermal stress and climate variability, particularly during El Nino years [4][2]. Deeper research into coral physiology, reef biodiversity, and stress resilience mechanisms was prompted by these early findings, which established the foundation of coral reef ecology and made bleaching a quantifiable stress response.

Simultaneously, climate science evolved as a structured discipline in the late 20th century, developing tools for monitoring large-scale ocean-atmosphere phenomena like the El Nino-Southern Oscillation (ENSO) and Indian Ocean Dipole (IOD). Eventually, it discovered these macro-climatic indices had a substantial correlation with bleaching events, offering worldwide models for stress forecasting [11]. With satellite remote sensing introduced in the 1980s, agencies like NOAA began producing sea surface temperature (SST) datasets, enabling the development of global coral bleaching alert systems. These were among the earliest attempts to quantitatively model bleaching risks, albeit with limited resolution and predictive capacity.

By the early 2000s, researchers began integrating insights from oceanography, climatology, and ecology to form a more interdisciplinary approach to coral reef monitoring. The emergence of environmental data science and the rising accessibility of satellite and in-situ oceanographic data laid the groundwork for computational models. Statistical and rule-based models were first used to interpret SST trends and bleaching thresholds. However, these conventional methods were constrained by their dependence on singular variables like SST and lacked the ability to capture nonlinear interactions between multiple stressors.

In this context, machine learning has emerged as a powerful evolution of predictive modeling within the discipline. As coral reef degradation accelerates due to climate change, ML offers the potential to combine climatic, oceanographic, and biological datasets for real-time, high-resolution predictions. The integration of environmental AI into marine science is a recent but rapidly growing development, offering novel ways to understand, forecast, and mitigate coral bleaching events.

Historical Progression

The evolution of coral bleaching prediction methods mirrors a broader shift from empirical observation and rule-based thresholds to data-integrated, machine learning-driven modeling. This transformation reflects an increasing awareness of the complex, nonlinear, and region-specific nature of environmental stressors affecting coral reefs.

Pre-AI Era Observational Science and SST-Threshold models

The first mass coral bleaching events were recorded in the early 1980s, coinciding with anomalous warming during strong El Nino periods. Studies by Glynn [4] and Brown [2] established thermal stress, particularly sustained sea surface temperature (SST) elevations, as the dominant trigger for coral bleaching. These findings catalyzed the development of Degree Heating Weeks (DHW), a metric that aggregates thermal stress over time above a calculated baseline.

By the late 1990s, institutions like NOAA's Coral Reef Watch began using satellite- derived SST anomaly data to issue global bleaching alerts, significantly advancing early warning capabilities [8]. These models were widely adopted due to their simplicity and wide-scale applicability. However, their reliance on temperature alone introduced major limitations.

McClanahan et al. [10] demonstrated that SST-only models could either overestimate or underestimate bleaching severity in specific reef locations, depending on factors such as recovery history and unaccounted environmental variables. For example, two reefs experiencing similar SST anomalies could experience vastly different bleaching responses due to differences in water turbidity or species resilience. This exposed a key shortcoming of threshold-based models: their failure to capture ecological and environmental nuances.

Additionally, these early models did not integrate climatic oscillations like the Indian Ocean Dipole (IOD) or ENSO, despite their significant influence on Indian Ocean temperatures and monsoon variability [11].

Post-AI Era: Emergence of Machine Learning and Data-Driven Forecasting

To address the shortcomings of the previously discussed models, researchers turned to machine learning. Beginning in the 2010s, models like Random Forest, Support Vector Machines (SVMs), and Gaussian Process Regression were applied to eco- logical datasets containing variables such as SST, DHW, salinity, pH, turbidity, and chlorophyll-a. These models demonstrated better performance than traditional regression models, especially in predicting bleaching across multiple locations [19][9]. One of the biggest advantages of machine learning is its flexibility in the face of real-world data challenges. ML algorithms can gracefully handle gaps or inconsistencies in the input, tease out complex, nonlinear relationships that traditional methods often miss, and even highlight which environmental factors matter most. On top of that, they're well suited to skewed datasets, in our case, where true bleaching events are relatively rare, so we don't lose predictive power just because positive cases are few and far between.

Udomchaipitak et al. [19] found that Random Forest outperformed other ML methods, achieving up to 88% accuracy when using remote sensing variables. However, even high-performing models often relied on global datasets, raising concerns about their applicability to region-specific reef systems such as those in India [1].

Furthermore, while ML models improved prediction, they often lacked temporal awareness, treating each observation as independent, without accounting for how stress accumulates over time.

Deep Learning and Time-Series Forecasting

To overcome this, deep learning models, particularly Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, have gained attention. These models are tailored for sequential data and can detect temporal dependencies in SST, salinity, and other environmental parameters, ideal for forecasting bleaching risk based on evolving climate trends. Despite their promise, LSTMs remain under- utilized in coral reef science, especially in the Indian context where localized, monthly reef-level data is still being standardized.

Meanwhile, Convolutional Neural Networks (CNNs) such as YOLOv8 have been successfully applied to underwater imagery to automatically detect and classify bleached coral colonies [15]. These image-based systems are highly effective for post- event analysis, offering fine-scale detection and monitoring capabilities. However, they do not predict bleaching in advance and are heavily dependent on annotated imagery, which is often lacking for Indian reef systems.

Contemporary Debate

Contemporary research on coral bleaching prediction has largely coalesced around two dominant machine learning frameworks: multivariate ensemble models and image-based deep learning classifiers. These methodologies represent significant progress over traditional SST-threshold models, but each presents clear limitations, particularly when applied to the complex and understudied ecological context of Indian coral reefs.

Model Type A: Multivariate Ensemble Models

Multivariate ensemble models such as Random Forest (RF) and XGBoost are widely used for environmental forecasting due to their ability to model nonlinear interactions, handle missing data, and perform robustly on imbalanced datasets. These models typically integrate multiple oceanographic and climatic variables, sea surface temperature (SST), degree heating weeks (DHW), salinity, pH, fCO2, turbidity, and macro-climatic indices like the El Nino-Southern Oscillation (ENSO) and the Indian Ocean Dipole (IOD), to predict bleaching probability.

Udomchaipitak et al. [19] demonstrated that Random Forest achieved a predictive accuracy of 88.24% in modeling bleaching outcomes using satellite-derived geospatial variables. Similarly, Madireddy et al. [9] reported high performance from Random Forest and Gradient Boosting methods across global coral reef datasets.

Despite these strengths, ensemble models remain temporally static, relying on aggregated snapshots of environmental data rather than modeling the accumulation of stress over time. Moreover, these models are frequently trained on global or inter- polated datasets, which often lack ecological resolution and specificity when applied to Indian reef systems. Coral reefs in India, particularly in regions like the Gulf of Mannar, Lakshadweep, the Andaman & Nicobar Islands, and the Gulf of Kachchh, are influenced by regional phenomena such as monsoonal salinity variation, localized sedimentation, and reef-specific species susceptibility. These dynamics are poorly captured by generalized, global-scale ensemble models.

Model Type B: Image-Based Convolutional Neural Networks

A second major modeling approach involves Convolutional Neural Networks (CNNs) for underwater image classification, particularly for detecting bleached coral colonies. YOLOv8 (You Only Look Once, version 8) is a high-speed object detection algorithm capable of classifying images in real-time with high spatial precision.

Patel et al. [15] utilized YOLOv8 to automate the detection of bleached coral from annotated underwater imagery, demonstrating high precision and recall in classifying bleached versus healthy colonies. These models offer several advantages: they reduce dependence on time-intensive field surveys, can be deployed in real-time monitoring scenarios, and deliver location-specific visual diagnostics with minimal post-processing requirements.

However, CNN-based systems are retrospective, designed for post-event classification rather than forecasting. They do not incorporate environmental drivers of bleaching such as SST anomalies or salinity changes and are therefore incapable of modeling causal or time-based relationships. Furthermore, the effectiveness of such models is constrained by the availability of extensive, labeled image datasets, which are scarce or underdeveloped for Indian reef regions.

Bridging Gaps in Current Research

While significant advancements have been made in coral bleaching prediction through the adoption of multivariate ensemble models and image-based deep learning systems, there remains a clear methodological and regional gap in the literature. Most existing models are either designed to classify bleaching events retrospectively or are built on global-scale datasets that do not capture the localized, dynamic, and temporally driven nature of bleaching in Indian coral reef ecosystems.

One of the biggest advantages of machine learning is its flexibility in the face of real-world data challenges. ML algorithms can gracefully handle gaps or inconsistencies in the input, tease out complex, nonlinear relationships that traditional methods often miss, and even highlight which environmental factors matter most. On top of that, they're well suited to skewed datasets in our case, where true bleaching events are relatively rare so we don't lose predictive power just because positive cases are few and far between.

Multivariate models such as Random Forest and XGBoost have shown promising predictive performance, particularly in studies where diverse environmental variables such as SST, DHW, salinity, pH, and turbidity are considered [19][9]. However, these models are generally temporally static, lacking mechanisms to understand how environmental stress accumulates and interacts over time. They also tend to generalize across reef systems, underrepresenting the biophysical heterogeneity of coral reef environments in the Indian Ocean region, where monsoonal variability, species-specific susceptibility, and localized anthropogenic stressors play a significant role in bleaching outcomes [17].

On the other hand, deep learning models like YOLOv8 are capable of high-accuracy detection of bleaching through image classification, yet they are fundamentally non- predictive and rely on large, annotated datasets, which are unavailable for many Indian reef systems [15]. Furthermore, they do not engage with oceanographic or climatological data, thereby lacking explanatory or forecasting power.

Additionally, while Long Short-Term Memory (LSTM) networks and other recur- rent neural network (RNN) architectures have proven effective in time-series modeling in fields like hydrology, air quality, and even more man-centric traffic forecasting [7][18], there is a notable absence of their application in coral bleaching research, especially in regional contexts such as India. The temporal dimension of stress, which is central to the biology of coral bleaching, remains underexplored in current models.

This absence of temporally aware, region-specific, and ecologically integrated machine learning models highlights a critical gap in the coral bleaching literature, one that must be addressed to improve both the accuracy and applicability of early warning systems for vulnerable coral reef ecosystems in India.

Research Question

This leads to the central research question: How can the evaluation of temporally aware and multivariate machine learning models, such as LSTM, XGBoost, LightGBM, and Random Forest, be used to improve the prediction of coral bleaching events in Indian reef ecosystems using localized environmental and climatic data?

Methodology

Proposed Argument

This research argues that combining temporally structured deep learning models with multivariate ensemble methods can significantly improve the predictive accuracy of coral bleaching forecasts, particularly in underrepresented regions like India. The approach emphasizes data localization, temporal modeling, and model interpretability, in contrast to the generalized or post-hoc classification frameworks currently dominant in the field.

By leveraging a comprehensive dataset of monthly environmental variables (1995-2020) across four major Indian reef systems and applying a range of predictive algorithms, this study fills a key methodological gap. The model ensemble not only allows comparison across different algorithmic classes but also provides insight into how temporal memory, tree-based decision logic, and boosting frameworks handle ecological forecasting.

Addressing the Gap in Current Research

This proposal addresses the gap in coral bleaching prediction literature by employing a mixed-model framework capable of both temporal pattern recognition and multivariate environmental factor analysis. The combination of LSTM, XGBoost, LightGBM, and Random Forest allows for comparative analysis between different model families: time-series deep learning, boosting-based decision trees, and ensemble bagging models. Each model is applied to the same region-specific dataset, thereby controlling for data variation while highlighting algorithmic performance differences in a real-world ecological context. The methodological gap identified in the literature lies in the absence of temporally structured models applied specifically to Indian coral reef systems, which are influenced by region-specific stressors such as monsoonal salinity changes, sedimentation, and localized species composition. While previous models have demonstrated strong performance on global datasets, they often ignore local stress dynamics, general-specific bleaching susceptibility, and the progressive buildup of stress over time. Existing models either generalize across reef systems or lack the capacity to incorporate cumulative environmental stress. This study fills that gap by testing both temporally aware (LSTM) and multivariate ensemble (XGBoost, LightGBM, Random Forest) models using harmonized datasets curated from Indian reef regions.

Uniqueness

The uniqueness of this study lies in its comparative approach using both temporal and non-temporal models across the same dataset, enabling benchmarking across machine learning paradigms. Unlike prior studies that depend heavily on satellite-only data or visual post-classification, this study integrates environmental variables with historical bleaching records and coral genera distributions. Moreover, it applies Long Short-Term Memory (LSTM) networks, rarely used in coral bleaching research, to learn from historical sequences of environmental stress. By doing so, the research introduces temporal depth into bleaching prediction. This is complemented by boosting-based models (XGBoost and LightGBM) and a bagging-based baseline (Random Forest), ensuring performance is evaluated not only on accuracy but also model complexity and interpretability.

Data Collection and Data Analysis

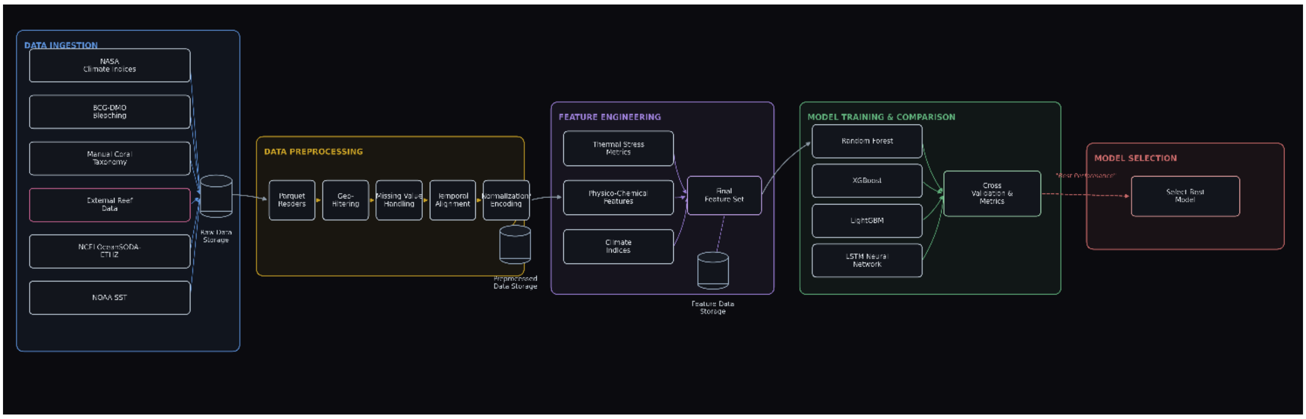

Figure 1: Research Methodology Flowchart

An end-to-end pipeline ingests and preprocesses multi-source environmental, climatic, and biological data (1995-2020), engineers thermal-stress, physico-chemical, and climate features, and then trains and compares Random Forest, XGBoost, LightGBM, and LSTM models using SMOTE-balanced, chronologically split time series to select the best predictor of site-specific coral bleaching in India.

Data was compiled from multiple validated sources. SST and DHW metrics were sourced from NOAA Coral Reef Watch. Ocean chemistry data, including salinity, pH, and fCO2, were drawn from OceanSODA-ETHZ (via NCEI). ENSO and IOD indices were sourced from NASA. Coral genera distributions were collected from the National Biodiversity Authority, while bleaching records were obtained from BCO-DMO and validated against the synthesis by Thangadurai et al. [17].

Preprocessing included standardizing date formats, spatial tagging by reef zone, and conversion of biological data into one-hot encoded features to reflect genera composition. Environmental variables were aggregated to monthly statistics using mean, maximum, and minimum values. All data was converted into a monthly time-series format from 1995 to 2020 and geospatially mapped to four Indian reef zones: the Andaman & Nicobar Islands, Lakshadweep, Gulf of Mannar, and Gulf of Kachchh. Furthermore, redundant features were removed using a correlation matrix (Figure 2: Correlation Matrix) for feature reduction and significant features were identified using feature selection (Figure 3: Time Series Plot).

Figure 2: Correlation Matrix: Feature to feature mapping highlighting variable correlations

A binary bleaching label was constructed based on recorded events, with 1 indicating bleaching and 0 indicating absence of bleaching. The final dataset consisted of 1,201 rows of monthly observations and contained environmental and biological data aligned on a monthly basis across four Indian reef systems over a 25-year period. Each record included SST metrics (mean, min, max, DHW), salinity, pH, fCO2, ENSO, IOD, and genera presence. The binary bleaching outcome was used as the dependent variable. This dataset was used uniformly across all models to ensure experimental consistency.

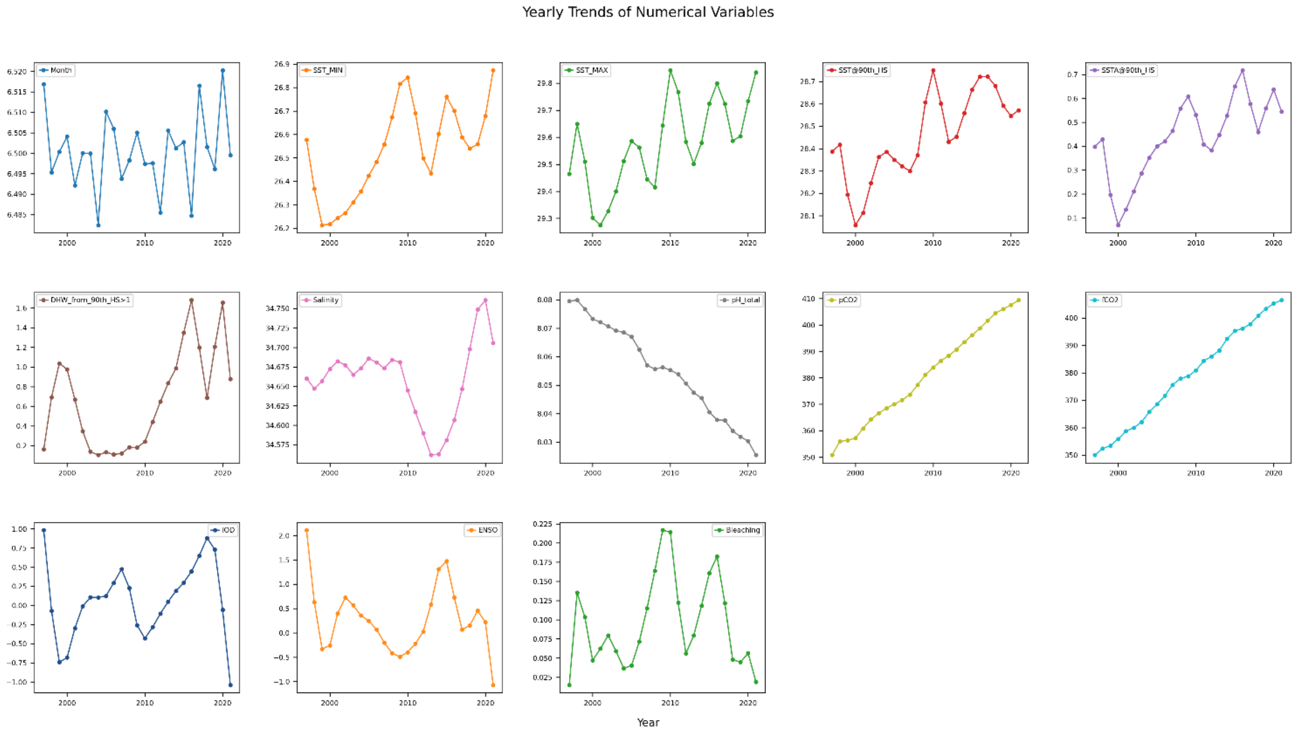

Figure 3: Time Series Plot: Depiction of feature changes over time

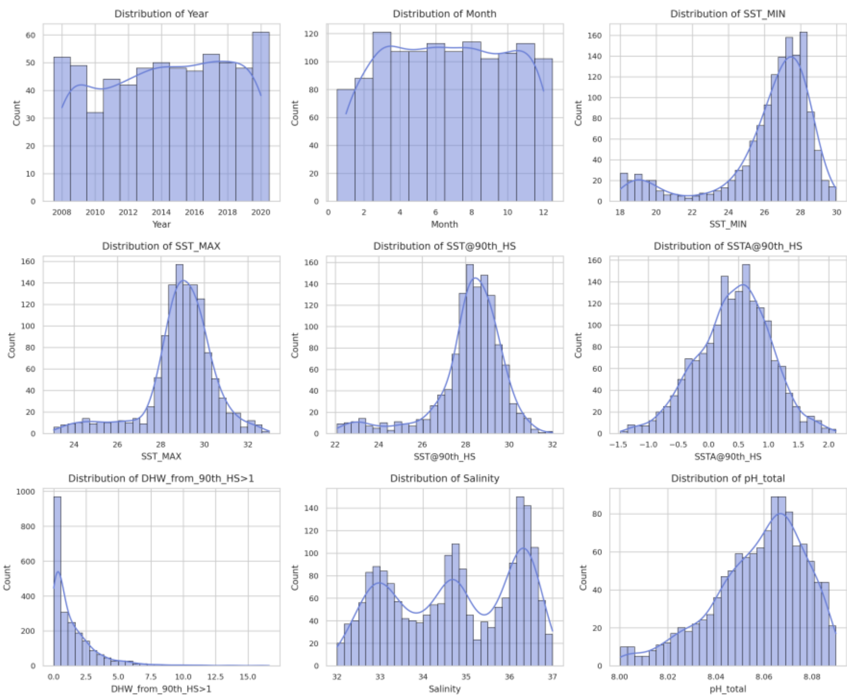

Figure 4: Factor Distribution Plot: Individual variables and their respective distributions

An 80:20 chronological train-test split was used to preserve temporal integrity. To address class imbalance, where non-bleaching months far outnumbered bleaching months, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to the training set. SMOTE generates synthetic minority samples based on k-nearest neighbor interpolation, improving generalization in imbalanced binary classification tasks [3].

Four models were then trained on this dataset, with each model being trained on the same features to allow direct comparison. Random Forest was configured with depth constraints and a fixed number of trees. XGBoost and LightGBM were tuned using grid search and early stopping. LSTM was trained using sliding-window time sequences and employed dropout and batch normalization for regularization. The model with the best performance, as determined through evaluation metrics, was then used as the final model.

Results and Findings

Model	Class	Accuracy	Precision	Recall	F1-Score
LSTM	0	0.90	0.99	0.90	0.95
	1		0.32	0.91	0.47
LightGBM	0	0.97	0.98	0.98	0.98
	1		0.60	0.60	0.60
XGBoost	0	0.96	0.99	0.97	0.98
	1		0.50	0.70	0.58
Random Forest	0	0.93	0.98	0.94	0.96
	1		0.35	0.64	0.45

Table 1: Evaluation Metrics for LSTM: LightGBM XGBoost and Random Forest Models

The models were assessed on a temporally held-out test set using several classification metrics: accuracy, precision, recall, F1-score, and ROC-AUC. Given the class imbalance, where bleaching events were infrequent, particular emphasis was placed on recall and F1-score for the positive (bleaching) class.

The LSTM model exhibited the highest recall for the minority class (0.909), indicating its strength in correctly identifying actual bleaching events.

However, it also recorded a relatively low precision (0.312) for the same class, reflecting a tendency to overpredict bleaching. Despite this, its F1-score for the bleaching class (0.465) was the highest among all models. The macro-averaged F1-score was 0.706 and ROC-AUC was also strong at 0.9636, suggesting that the model performed well overall in distinguishing between bleaching and non-bleaching cases.

The LightGBM model showed more balanced performance. Its recall for the bleaching class was 0.60, with a precision of 0.60 and an F1-score of 0.60, indicating consistent classification ability. Its overall accuracy was 0.79, with a macro F1-score of 0.79. While it did not outperform LSTM in terms of recall, it avoided the overprediction trend and achieved high class-wise precision and stability.

XGBoost performed similarly to LightGBM, with a recall of 0.70, a precision of 0.50, and an F1-score of 0.58 for the bleaching class. It achieved a macro F1-score of 0.78 and a weighted F1-score of 0.96. This performance suggests that while the model had a strong ability to identify bleaching, it still produced more false positives than LightGBM.

The Random Forest model, while achieving high accuracy (0.79) and strong preci- sion for the non-bleaching class (0.98), performed the weakest on the minority class. The recall for the bleaching class was 0.64, but the precision dropped to 0.35, result- ing in an F1-score of only 0.45. The macro F1-score (0.71) indicates that while the model had interpretability advantages, it struggled with identifying bleaching cases effectively.

These results indicate that LSTM is the most sensitive model in capturing true bleaching events due to its temporal memory, but it also has a higher false-positive rate. In contrast, LightGBM and XGBoost offer more balanced performance with moderate recall and better precision. Random Forest, while interpretable and stable for the majority class, failed to generalize effectively to the minority class.

Importantly, these findings support the core research question. By integrating both temporally aware and multivariate models, the study demonstrates that it is possible to improve coral bleaching prediction using localized data. The LSTM model's superior recall highlights the value of temporal modeling in capturing the cumulative effects of environmental stress. Meanwhile, the performance of LightGBM and XGBoost con- firms the utility of gradient-boosted decision trees in extracting interactions across multiple environmental variables.

The implications of these findings are significant. In ecological forecasting, missing a bleaching event (false negative) is typically more consequential than a false alarm (false positive). Hence, the recall and F1-score for the bleaching class must be prioritized. The models developed here, particularly LSTM, demonstrate strong potential for use in early warning systems and reef management interventions, especially when supplemented with real-time data and further fine-tuning.

Conclusion

This study proposed that combining temporally aware and multivariate machine learning models could improve the accuracy and relevance of coral bleaching prediction in Indian reef ecosystems. Addressing the research question: How can the evaluation of temporally aware and multivariate machine learning models be used to improve the prediction of coral bleaching events in Indian reef ecosystems using localized environmental and climatic data? The research evaluated four models: Long Short-Term Memory (LSTM), XGBoost, LightGBM, and Random Forest.

The central problem identified in the literature was the lack of models capable of integrating both time-dependent and multivariate environmental data, particularly those calibrated to the region-specific stressors affecting Indian reefs. Most existing approaches either relied on single-variable SST thresholds or performed post-event classification without the ability to forecast bleaching onset. The findings demonstrate that the LSTM model achieved the highest recall (0.909) for bleaching events, reflecting its ability to capture sequential environmental patterns leading up to stress. However, this came at the cost of precision, resulting in a higher rate of false positives. LightGBM and XGBoost exhibited more balanced performance, with LightGBM offering the best precision-recall tradeoff and macro-averaged F1- score. Random Forest, while interpretable, underperformed the minority class and was therefore less suitable for risk-sensitive prediction.

These results show that temporally aware and ensemble models, when trained on localized multivariate datasets, can significantly enhance predictive capabilities and support reef conservation planning. The use of recall and F1-score as priority metrics ensures ecological relevance by reducing the likelihood of missed bleaching events.

Limitations of the study include the reliance on interpolated biological data, the use of monthly resolution, and the absence of image-based integration. Future work should explore hybrid models that combine satellite imagery, real-time monitoring, and domain-specific biological data to build operational early-warning systems.

Data Availability Statement

The datasets analyzed during the current study are derived from publicly available sources. Sea Surface Temperature (SST) and Degree Heating Weeks (DHW) data were obtained from NOAA Coral Reef Watch. Ocean chemistry variables (salinity, pH, fCO) were sourced from OceanSODA-ETHZ via NCEI. ENSO and IOD indices were collected from NASA climate databases. Coral genera distributions were acquired from the National Biodiversity Authority, and bleaching records were sourced from BCO-DMO and Thangadurai et al. [17].

All preprocessing, harmonization, and modeling scripts used to build the dataset and train the machine learning models are available from the corresponding authors upon reasonable request. Due to licensing constraints, redistribution of raw third-party datasets may require direct access from the respective data providers.

References

1. Arora, M., et al. (2019). Bleaching stress on Indian coral reef regions during mass coral bleaching years using NOAA OISST data. Current Science. https://doi.org/10.18520/cs/v117/i2/242-250
2. Brown, B. E. (1997). Coral bleaching: Causes and consequences. Coral Reefs, 16(1), S129–S138. https://doi.org/10.1007/s003380050249
3. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953
4. Glynn, P. W. (1993). Coral reef bleaching: Ecological perspectives. Coral Reefs, 12(1), 1–17. https://doi.org/10.1007/BF00303779
5. Hoegh-Guldberg, O., et al. (2007). Coral reefs under rapid climate change and ocean acidification. Science, 318(5857), 1737–1742. https://doi.org/10.1126/ science.1152509
6. Hughes, T. P., Kerry, J. T., & A´ lvarez-Noriega, M., et al. (2017). Global warming and recurrent mass bleaching of corals. Nature, 543(7645), 373–377. https://doi.org/10.1038/nature21707
7. Kratzert, F., Klotz, D., Herrnegger, M., Hochreiter, S., & Nearing, G. S. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089–5110. https://doi.org/10.5194/hess-23-5089-2019
8. Liu, G., Strong, A. E., & Skirving, W. (2011). Remote sensing of sea surface tem- peratures during 2002 Barrier Reef coral bleaching. Eos, Transactions American Geophysical Union, 87(15), 137–144. https://doi.org/10.1029/2003EO150001.
9. Madireddy, I., Bosch, E., & McCalla, S. (2023). Using machine learning to develop a global coral bleaching predictor. Journal of Emerging Investigators, JEI 22.056. https://doi.org/10.59720/22-056
10. McClanahan, T. R., et al. (2007). Modelling susceptibility of coral reefs to envi- ronmental stress using remote sensing data and GIS models. Ecological Modelling, 212(2–3), 180–199. https://doi.org/10.1016/j.ecolmodel.2007.10.033
11. McPhaden, M. J., Zebiak, S. E., & Glantz, M. H. (2006). ENSO as an integrating concept in earth science. Science, 314(5806), 1740–1745. https://doi.org/10.1126/ science.1132588
12. Ministry of Environment, Forest and Climate Change (MoEFCC). (2014). India's Fifth National Report to the Convention on Biological Diversity. https://www.cbd.int/doc/world/in/in-nr-05-en.pdf
13. NOAA Coral Reef Conservation Program. (2023). What are Coral Reefs? https://coralreef.noaa.gov/
14. NOAA Coral Reef Watch. (2022). Coral Bleaching Data and Products. https://coralreefwatch.noaa.gov/
15. Patel, J., et al. (2024). Automated detection of coral bleaching using YOLOv8. The Academic, 2(12), 938–953. https://doi.org/10.5281/zenodo.14737335
16. Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Infor- mative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone. 0118432

17. Thangadurai, T., et al. (2024). History of recurrent short and long-term coral bleaching events in Indian coral reefs: a systematic review of contrasting bleaching patterns, lessons learned, and future directions. bioRxiv. https://doi.org/10.1101/ 2024.05.18.594244

18. Tian, Y., et al. (2018). LSTM-based traffic flow prediction with missing data. Neurocomputing, 318, 297-305. https://doi.org/10.1016/j.neucom.2018.08.067

19. Udomchaipitak, T., et al. (2022). Forecast coral bleaching by machine learning of remotely sensed geospatial data. International Journal of Design & Nature and Ecodynamics, 17(3), 423–431. https://doi.org/10.18280/ijdne.170313