An Analysis of Logistic Regression and eXtreme Gradient Boosting for Predicting Oxygen-Stressed Conditions in Cape Cod Bay

Katharine Baker, Madison Simmons, Rebecca Trailor, Julia Weppler
Boston College, Chestnut Hill, MA, USA
{bakerkx, simmonmr, trailorr, wepplerj}@bc.edu

Abstract

*Studies have shown that while there are highly specified models for classifying hypoxic events in localized regions, there is a lack of models that predict *pre-hypoxic* events. These events are dangerous in their own right, creating oxygen-stressed environments, and they also serve as precursors to hypoxic conditions, making them essential to monitor in forecasting the development of hypoxia. Furthermore, no models currently classify hypoxic or pre-hypoxic events in the New England watershed.*

Therefore, our goal is to develop a model for the Cape Cod Bay (CCB) region that classifies station data as hypoxic or at risk for hypoxia, enabling scientists to take timely action to prevent or mitigate hypoxia in the CCB. Additionally, our model highlights the complex relationships within environmental data by incorporating multiple variables from diverse datasets. Our approach uses Gradient Boosting methods with SMOTE, and we compare these results with a Logistic Regression model, which incorporates fewer parameters while still targeting oxygen-stressed conditions in addition to hypoxic states.

After conducting training and testing, we found that our XGBoost model with SMOTE and hyperparameter tuning achieved the highest recall at 0.913. Our Logistic Regression model also performed well with SMOTE and tuning, achieving a recall of 0.818. Both models outperformed their respective baselines and yielded results comparable to other methods in related research.

1. Introduction

Hypoxia, defined by low dissolved oxygen (DO) levels in aquatic systems, poses serious risks to marine ecosystems, fisheries, and tourism. In severe cases, hypoxia can cause mass die-offs of fish and invertebrates, disrupting food webs and economies. Even moderate oxygen stress (DO below 7 mg/L) can hinder growth and reproduction in marine life, making early detection both ecologically and economically important [4]. For instance, hypoxic events in Cape Cod Bay (CCB) during the summers of 2019 and 2020 caused benthic die-offs and significant lobster mortality [3, 4]. This emergence of hypoxia in CCB differs from regions like Chesapeake Bay or the Gulf of Mexico, which have long-standing seasonal "dead zones" and established forecast models [2]. Accurate early prediction is crucial for enabling timely interventions.

Building a predictive model for hypoxia in CCB has broader relevance. Forecast systems in well-studied estuaries often rely on detailed, site-specific inputs or complex coupled physical-biogeochemical models.
In contrast, our aim is to develop a simpler, data-driven model using a minimal set of widely monitored water quality indicators, avoiding overfitting to localized conditions and enhancing potential transferability to other temperate coastal regions experiencing emerging hypoxia. We focus on a streamlined feature set—including DO, temperature, and nutrients like ammonium, nitrogen, and phosphorus—that are routinely measured, to predict whether DO will fall below the 7mg/L threshold. Prior studies caution that models with too many input variables, while seemingly accurate, can become overly complex and prone to overfitting [12].

This study compares two approaches for classifying oxygen-stressed conditions in CCB: logistic regression (LR) as a simple, interpretable baseline, and eXtreme Gradient Boosting (XGBoost), an ensemble tree-based method known for modeling complex relationships [5, 12]. We use a decade of in-situ water quality data from the Center for Coastal Studies' buoys in CCB, which include DO, temperature, and dissolved nutrients [1]. Because hypoxic cases are underrepresented, we use the Synthetic Minority Over-sampling Technique (SMOTE) to balance the dataset and ensure adequate learning from low-oxygen events. Hyperparameters for XGBoost are optimized using randomized search cross-validation to avoid overfitting.

By evaluating both models' predictive performance, we assess whether a linear model is sufficient or if XGBoost provides a meaningful improvement in forecasting emerging hypoxia in CCB. Our goal is to build a generalizable and accessible predictive tool for communities with limited monitoring infrastructure, especially in coastal regions facing episodic hypoxia for the first time.

1.1. Related Work

Various machine learning (ML) algorithms—from simple regressions to deep learning—have been explored for forecasting DO levels and classifying hypoxic events. Ensemble tree-based methods have shown strong performance due to their ability to capture non-linear interactions among environmental variables. For example, Politikos et al. [13] used an ensemble of classifiers (XGBoost, extremely randomized trees, and random forest) to predict hypoxia in a Mediterranean lagoon, finding that tree-based models outperformed logistic regression after applying SMOTE to address class imbalance. Ahn et al. [5] similarly investigated ensemble models with Bayesian hyperparameter tuning to improve DO prediction.

These findings suggest that ensemble models—such as XGBoost, random forest, and stacked classifiers—often represent the current standard in data-driven hypoxia prediction [12]. Deep learning methods, such as recurrent neural networks and LSTMs, have also been employed, especially when incorporating temporal or spatial dynamics. One study achieved over 90% accuracy for 12-hour forecasts using an LSTM and meteorological inputs [11]. However, deep learning models typically require large datasets and can be challenging to interpret. In contrast, tree-based models offer a useful balance of performance and interpretability. For instance, Politikos et al. [13] identified variables such as pH, chlorophyll-a, salinity, temperature, and solar radiation as the most influential drivers of hypoxia in their boosted tree models.

Despite progress, common challenges remain. Class imbalance is a frequent issue: hypoxic conditions are typically rare relative to normoxic observations, especially in systems where hypoxia is emerging. This can bias models toward the majority class. SMOTE and other oversampling techniques are widely used to address this, enhancing model sensitivity to early hypoxia signals [9]. Our study adopts this approach by using SMOTE to ensure that DO ¡ 7 mg/L cases are adequately represented during training.

Another concern is feature overload. Many studies use a wide range of environmental variables—sometimes dozens—which can increase noise, reduce generalizability, and introduce overfitting [12]. Li et al. [10], for example, found that weakly correlated inputs like turbidity reduced model accuracy, while omitting key variables like light availability also degraded performance. Important predictors across studies include water temperature, nutrient levels, salinity, and chlorophyll. Our approach builds on this by using a small, high-impact set of inputs to reduce complexity without sacrificing performance.

These insights have shaped our study's design. By comparing LR to XGBoost, applying SMOTE, and using a focused feature set, we aim to balance simplicity and effectiveness. Our work draws on lessons from studies such as Politikos et al. [13] and Li et al. [10], and contributes a streamlined, generalizable methodology for hypoxia forecasting.

1.2. Methods

We began by loading our datasets in Google Colab, using the Pandas library to read and manipulate CSV files containing historical water quality data from Cape Cod Bay. Each dataset contained observations recorded at specific stations, with features including temperature, dissolved oxygen (DO), ammonium, nitrate/nitrite (NOx), phosphate, and salinity. To prepare the data, we removed any records missing DO values, as they are critical to our classification target. For other missing features, we used mean imputation to preserve the sample size while minimizing introduced bias. After cleaning, the dataset was partitioned into an 80% training set and a 20% testing set using `train_test_split` from `sklearn.model_selection`, ensuring stratification by class to preserve the hypoxic/non-hypoxic distribution.

Our classification task framed the DO threshold of 7 mg/L as the decision boundary between "oxygen-stressed" (hypoxic or near-hypoxic) and normal conditions. The target variable was binarized accordingly. We selected two supervised machine learning algorithms—Logistic Regression (LR) and eXtreme Gradient Boosting (XGBoost)—to evaluate the trade-offs between interpretability and performance. LR was chosen for its simplicity and transparency in capturing linear relationships between features and class probabilities. It models the log-odds of the target class via a sigmoid function and is visualized in Figure 1.

XGBoost, on the other hand, is a tree-based ensemble method that constructs decision trees sequentially, with each new tree minimizing the residual errors of the previous ensemble. The model begins by estimating initial probabilities using the log-odds of the positive class. It then iteratively fits trees to the gradients of the loss function, using similarity scores to optimize node splits. Its use of regularization (via parameters such as learning rate and maximum depth) makes it highly effective in preventing overfitting.

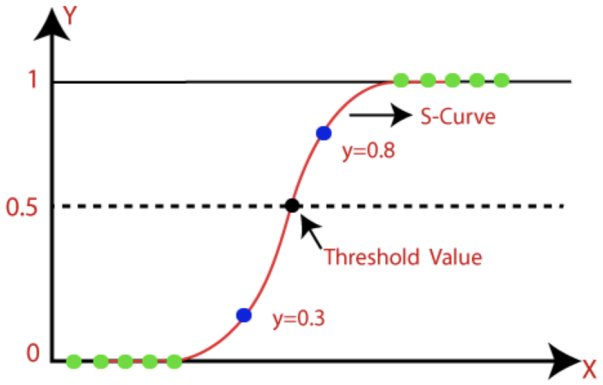

Figure 1. Sigmoid curve in logistic regression

Figure 2 provides a conceptual diagram of the boosting process, while Figure 3 shows an example of a tree trained on our dataset.

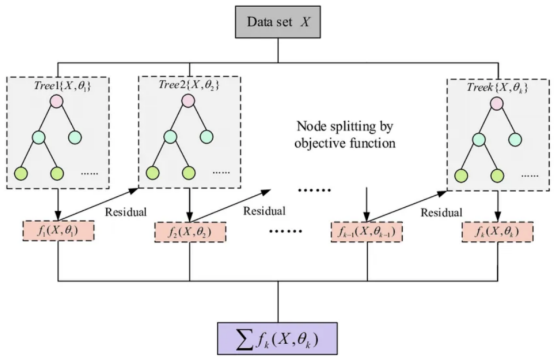

Figure 2. Conceptual illustration of gradient boosting

Due to the inherent class imbalance—fewer hypoxic events compared to normoxic observations—we implemented the Synthetic Minority Oversampling Technique (SMOTE) to generate synthetic samples for the minority class. SMOTE interpolates between existing minority class instances to create new, realistic examples. This helps prevent the model from being biased toward the majority class and improves sensitivity to low-oxygen conditions. We applied SMOTE using the `imblearn.over_sampling.SMOTE` module. It was performed only on the training data to avoid data leakage into the testing set.

To optimize the XGBoost model, we conducted hyperparameter tuning using `RandomizedSearchCV` from `sklearn.model_selection`. Key parameters tuned included the learning rate, number of estimators, maximum tree depth, subsample ratio, and column sampling rate. The objective was to maximize recall, as detecting hypoxic or oxygen-stressed conditions (true positives) is more critical than avoiding false positives in this context. Logistic regression was also tuned with L2 regularization strength to prevent overfitting.

Model evaluation was based on accuracy, precision, recall, and F1 score, with a particular emphasis on recall due to the ecological importance of detecting as many hypoxic events as possible. We used `classification_report` from `sklearn.metrics` to summarize performance. Additionally, we plotted confusion matrices to visualize model errors and confirm that SMOTE improved minority class detection. Finally, we extracted feature importance rankings from the trained XGBoost model using the `get_booster().get_score()` method, which calculates gain-based importance for each input variable, helping interpret the model's learned structure.

2. Experiments

The data we are using has been collected from Water Quality Monitoring Stations through the Center for Coastal Studies. This center is located in Provincetown, MA so it gathers data about the CCB. The map within Figure 4 shows the locations of all of the monitoring stations operated by the Center, however, our focus remained on the stations in the CCB.

The datasets we examined measured various features such as water temperature, salinity, dissolved oxygen levels, chlorophyll from various years, all of which are beneficial to the prediction of hypoxia or hypoxic conditions. We then created our dataframe with the help of the Pandas library by compiling each of our 24 csv files (one per each station) into one array.

To preprocess this data, we removed features such as id, internal_station_id, and collected_at as these features were insignificant to our model, not giving any information about hypoxic conditions. Additionally we removed any year rows that had incomplete information about dissolved oxygen levels because this feature is the primary indicator for hypoxia and this is how we will be evaluating our model. Further, we defined pre-hypoxic conditions to be when the dissolved oxygen is less than 7 mg/L. This will allow us to evaluate whether our model is correctly classifying pre-hypoxic events.

In terms of hyperparameter tuning, we used randomized search cross validation (RandomizedSearchCV) to determine the most optimal hyperparameter determinations from those randomly sampled, therefore more efficiently but without being too computationally expensive. Our model

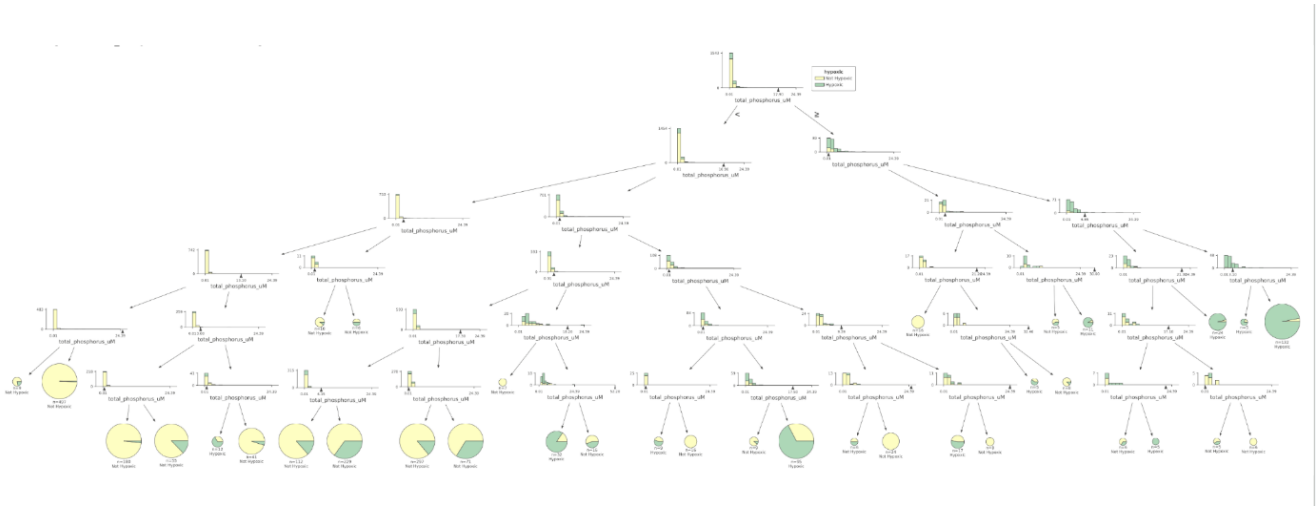

Figure 3. Visualization of a trained decision tree

Figure 4

Classification Report for XGBoost

	precision	recall	f1-score	support
0	0.90	0.91	0.91	357
1	0.77	0.75	0.76	140
accuracy			0.87	497
macro avg	0.83	0.83	0.83	497
weighted avg	0.86	0.87	0.86	497

Classification Report for Logistic Regression

	precision	recall	f1-score	support
0	0.90	0.80	0.84	357
1	0.59	0.76	0.67	140
accuracy			0.79	497
macro avg	0.75	0.78	0.76	497
weighted avg	0.81	0.79	0.79	497

Figure 5

performed five folds for each of the 30 candidates, totaling to 150 fits, with results including a maximum depth of 8, subsample of 0.8, and 500 estimators for the XGB model.

We implemented Logistic Regression and XGBoost as our baseline, comparative models. Following initial model implementation, we achieved the following results for LR and XGB shown in Figure 5. While the precision itself was quite high for both models, this was not an ideal performance metric for our project as it seemed the model was simply playing it safe due to the class imbalance of hypoxic versus non-hypoxic cases, and predicting many false negatives which could lead to researchers not detecting an oncoming hypoxic event. As such, we determined that recall would be a better metric to look to improve as it measures how many actual positive cases of hypoxic conditions were successfully detected. LR's recall was only 0.39, likely due to the class imbalance, while XGB performed slightly better at 0.71. However, there was still room for improvement and thus SMOTE was implemented.

After SMOTE was implemented, as shown in Figure 6 below, there is an improved recall of 0.75 for XGB and 0.76 for LR. Additionally, while it appears that the models here are performing equally well, when considering the other metrics including f1-score, XGB does seem to be better at making true predictions overall.

Another metric used for evaluating our results was area under the ROC curve, or AUROC, which indicates how well performing a model is at distinguishing between positive and negative classes. For the baseline XGB and LR performance, the AUROC was 0.901 and 0.816 respectively; after implementing SMOTE and RandomizedSearchCV, these values increased to 0.913 and 0.818

Classification Report for XGBoost

	precision	recall	f1-score	support
0	0.90	0.91	0.91	357
1	0.77	0.75	0.76	140
accuracy			0.87	497
macro avg	0.83	0.83	0.83	497
weighted avg	0.86	0.87	0.86	497

Classification Report for Logistic Regression

	precision	recall	f1-score	support
0	0.90	0.80	0.84	357
1	0.59	0.76	0.67	140
accuracy			0.79	497
macro avg	0.75	0.78	0.76	497
weighted avg	0.81	0.79	0.79	497

Figure 6

suggesting both models perform generally well. XGB does seem to pull ahead though as the more superior model for the scope of our project. The results for XGB can be seen below in Figure 7 and for LR in Figure 8.

In comparison to other literature results for AUROC, our tuned XGB model outperformed all other classification models we investigated, including Chen et al. [6] with 0.89, Erion et al. [8] with 0.86, and Li et al. [10] with 0.64. The only model we found with better performance was linear regression Elmoaqet et al. [7], 2014 with 0.93; these results too are shown below in Figure 7.

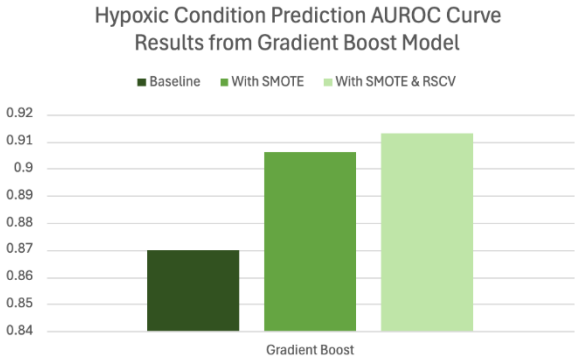

Figure 7

Therefore from these results, our XGB model which implemented optimized hyperparameters and SMOTE was the best performing classification model in terms of the AUROC metric. The Classification Report for XGBoost shown in Figure 7 represents the metrics for this model, with a macro average recall of 0.86, accuracy of 0.86, and F-1 score of 0.86.

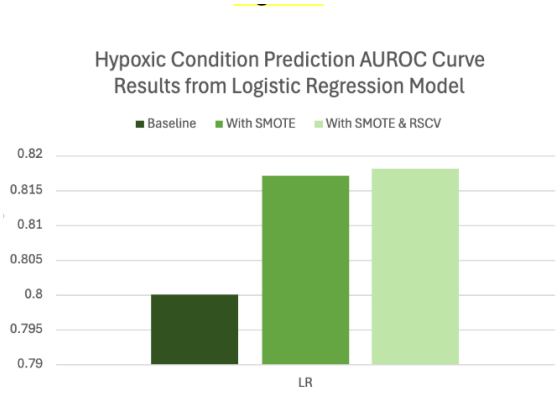

Figure 8

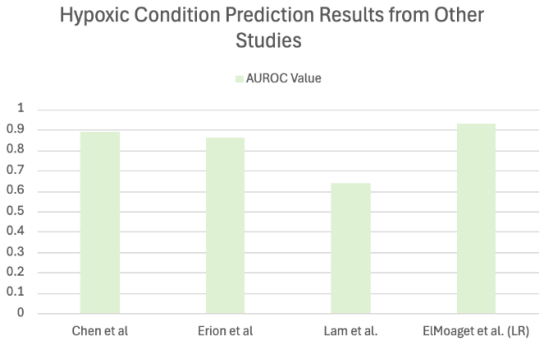

Figure 9

Given the context of this project within environmental forecasting and disaster intervention, it was important to determine which of the features used in our models were most impactful to model performance and classification. We applied SHAP (Shapley Additive exPlanations) value analysis to both our XGBoost classifier and baseline linear-regression model. SHAP values quantify, for each feature and each sample, how much that feature pushes the model output away from its baseline (the average prediction) in additive units—positive values drive the prediction higher, negative values drive it lower. The horizontal axis in each SHAP summary plot shows the magnitude of that push. In the XGBoost chart, it spans roughly –6 to +6 log-odds because tree splits and interaction effects cap individual contributions, while in the linear-regression chart it ranges from about –10 to +250 because each feature contributes via its coefficient multiplied by its raw value. In the XGBoost model, total dissolved nitrogen was the single strongest driver of high-risk predictions, with elevated ni-

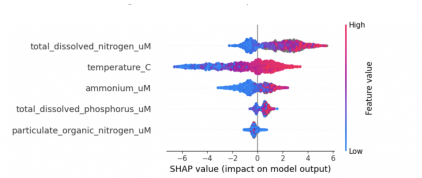

Figure 10

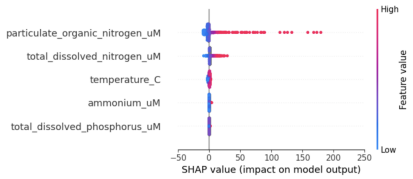

Figure 11

trogen concentrations consistently pushing outputs toward the stress class. Temperature was the next most influential feature, where warmer conditions increased predicted risk, while ammonium, total dissolved phosphorus, and particulate organic nitrogen played progressively smaller roles. In contrast, the linear model's SHAP plot was dominated by particulate organic nitrogen, as its large dynamic range and coefficient magnitude generated SHAP values an order of magnitude larger than all other predictors, effectively masking their contributions. The linear model's sensitivity to raw feature scale indicates the necessity of feature standardization or regularization (e.g., Ridge or Lasso) to ensure that no single water-quality parameter dictates model behavior.

3. Conclusion

After conducting our testing, we found that our models—developed using both XGBoost and Logistic Regression in conjunction with SMOTE for class balancing and hyperparameter tuning—produced AUROC values that were comparable to those reported in existing literature, including the study by Pigat et al. [12] While our models achieved similar predictive performance, a key distinction lies in their broader applicability and practical utility. Many of the models to which we compared our results were highly localized, trained on data from specific geographic areas with unique environmental conditions. As a result, these models often lack generalizability and may perform poorly when applied to different regions. Furthermore, those studies frequently relied on a large number of environmental and chemical features, many of which may not be readily available in resource-constrained settings or less-monitored watersheds. This poses a barrier to replication and implementation in underfunded regions that nevertheless experience significant water quality challenges.

In contrast, our model was intentionally designed to use a more parsimonious set of input features—those that are commonly available and can be measured at lower cost—without compromising performance. By minimizing the reliance on highly specific or hard-to-acquire data, we increase the accessibility and adaptability of hypoxic condition classification, thereby empowering a wider range of regions and stakeholders to monitor and respond to water quality issues. This is particularly critical for protecting vulnerable watersheds where early intervention can help prevent ecosystem degradation.

Additionally, our classification framework does not merely distinguish between hypoxic and non-hypoxic conditions; it also incorporates the detection of pre-hypoxic states. This distinction is crucial, as it enables environmental scientists and policymakers to anticipate the onset of hypoxia and intervene proactively rather than reactively. By identifying early warning signs, our approach supports more effective resource allocation, mitigation strategies, and long-term planning for ecosystem health.

Looking ahead, we hope to refine and tailor our model to specific watershed contexts by integrating localized data and collaborating with regional agencies. By doing so, we aim to enhance the model's precision while preserving its accessibility. Ultimately, our goal is to contribute to the ongoing effort to prevent the spread of hypoxia and support the restoration of already affected aquatic environments.

3.1. Contributions

Katharine Baker: Researching datasets and collecting data; building LR baseline model; implementing SMOTE; organize presentation slides; outline data collection; data preprocessing, XGBoost; milestone writeup work; Final Report: Abstract, Experiments, Conclusion, editing
Madison Simmons: Researching datasets and collecting data; fine-tuning LR model and analyzing/interpreting summary statistic results and their impact on next steps; milestone writeup work; XGBoost visualizations; Final Report: Methods, Figures, Experiments, editing
Rebecca Trailor: Researching datasets and collecting data; initial data preprocessing and XGBoost baseline; organizing materials in Github; milestone writeup work; XGBoost cross-validation implementation; Final Report: Methods, Conclusion, Figures, Resources, Github and LaTeXorganization and formatting
Julia Weppler: Project conception, project proposal, literature review for AUROC comparisons, background, data identification, RandomizedSearchCV for hyperparameter tuning; SHAP value analysis; milestone writeup work; Final Report: Introduction, Related Work

References

[1] Cape cod bay water quality monitoring program. Accessed: 2025-05-08. 1

[2] Chesapeake bay hypoxic volume. Accessed: 2025-05-08. 1

[3] Cape cod bay hypoxia. Accessed: 2025-05-08. 1

[4] In late summer 2019 and 2020, ... *Biogeosciences*, 19:3523–3539, 2022. Accessed: 2025-05-08. 1

[5] J. M. Ahn, J. Kim, and K. Kim. Ensemble machine learning of gradient boosting (xgboost, lightgbm, catboost) and attention-based cnn-lstm for harmful algal blooms forecasting. *Toxins*, 15(10):608, 2023. 1, 2

[6] S. Chen, R. Chen, C. Fu, D. Wang, Y. Li, and Y. Peng. Remote sensing estimation of chlorophyll-a in case-ii waters of coastal areas: Three-band model versus genetic algorithm–artificial neural networks model. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:3640–3652, 2021. 5

[7] H. Elmoaqet, D. M. Tilbury, and S. K. Ramachandran. Evaluating predictions of critical oxygen desaturation events. *Physiological Measurement*, 35(4):639–655, 2014. 5

[8] G. Erion, H. Chen, S. M. Lundberg, and S.-I. Lee. Anesthesiologist-level forecasting of hypoxemia with only spo data using deep learning, 2017. 5

[9] G. W. Jeon, Y. S. Lee, W. H. Hahn, and Y. H. Jun. A predictive model for perinatal brain injury using machine learning based on early birth data. *Children*, 11(11):1313, 2024. 2

[10] W. Li, J. Lv, Y. Wang, and X. Kong. Study on the impact of input parameters on seawater dissolved oxygen prediction models. *Journal of Marine Science and Engineering*, 13(3): 536, 2025. 2, 5

[11] S. Park, K. Kim, T. Hibino, and K. Kim. Machine learning-based prediction of seasonal hypoxia in eutrophic estuary using capacitive potentiometric sensor. *Marine Environmental Research*, 196:106445, 2024. 2

[12] L. Pigat, B. P. Geisler, S. Sheikhalishahi, J. Sander, M. Kaspar, M. Schmutz, S. O. Rohr, C. M. Wild, S. Goss, S. Zaghdoudi, and L. C. Hinske. Predicting hypoxia using machine learning: Systematic review. *JMIR Medical Informatics*, 12: e50642, 2024. 1, 2, 6

[13] D. Politikos, G. Petasis, and G. Katselis. Interpretable machine learning to forecast hypoxia in a lagoon. *Ecological Informatics*, 2021. 2