Journal of Hazardous Materials Advances 23 (2026) 101348

Contents lists available at ScienceDirect

Journal of Hazardous Materials Advances

journal homepage: www.elsevier.com/locate/hazadv

Integrating advanced deep learning with ecological indicators for accurate prediction of harmful algal bloom occurrence and intensity

Md. Abdullah Al Mamun Hridoy [a], Petra Schneider [b,*], Chiara Bordin [c], Matteo Bodini [d], Abdullah Ibna Shawkat [e], Khairul Nizam Abdul Maulud [f], Md. Abdullah Al Mamun [g,h], Paolo Pastorino [i]

[a] Faculty of Fisheries, Sylhet Agricultural University, Sylhet, 3100, Bangladesh
[b] Department Water, Environment, Civil Engineering and Safety, Magdeburg-Stendal University of Applied Sciences, Breitscheidstr. 2, D-39114, Magdeburg, Germany
[c] Department of Computer Science, UiT The Arctic University of Norway, 6050, Langnes, 9037, Tromsø, Norway
[d] Department of Pathophysiology and Transplantation, University of Milan, Via Francesco Sforza 35, Milan, 20122, Italy
[e] Faculty of Agricultural Engineering and Technology, Sylhet Agricultural University, Sylhet, 3100, Bangladesh
[f] Department of Civil and Structural Engineering, Faculty of Engineering and Built Environment, Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
[g] Department of Fish Health Management, Laboratory of Fish Diseases Diagnosis and Pharmacology, Faculty of Fisheries, Sylhet Agricultural University, Sylhet, Bangladesh
[h] Faculty of Veterinary Medicine, Universitas Brawijaya, Malang, East Java, 65151, Indonesia
[i] Experimental Zooprophylactic Institute of Piedmont, Liguria and Aosta Valley Via Bologna, 148, 10154, Turin, Italy

ARTICLE INFO

Keywords:
Harmful algal blooms
Deep learning
FT-Transformer
Bloom Index
Eutrophication
Nutrient dynamics

ABSTRACT

Harmful algal blooms (HABs) are intensifying globally under climate warming and eutrophication, yet reliable forecasting remains challenging due to nonlinear and highly correlated environmental drivers. This study develops a dual-task deep learning framework for HAB occurrence classification and bloom-intensity prediction using FT-Transformer, Neural Oblivious Decision Ensembles (NODE), and SAINT. To ensure data reliability, the dataset was audited to distinguish observed and GAN-generated records, and synthetic-data fidelity was validated through statistical, correlation, and visual analyses. Potential target leakage associated with Bloom_Index was explicitly assessed, and a no-Bloom_Index ablation experiment was conducted to evaluate model robustness. Ecological analysis revealed that HAB events were associated with elevated sea surface temperature, chlorophyll anomaly, and nitrogen enrichment, together with reduced dissolved oxygen. FT-Transformer achieved the best classification performance (accuracy = 0.995 ± 0.0032, F1-score = 0.985 ± 0.0091, AUROC = 0.998 ± 0.0036), while NODE provided the highest regression accuracy for Bloom_Index prediction (MAE = 0.0476 ± 0.0172; R^2 = 0.9316 ± 0.0231). Ablation results confirmed that predictive performance remained robust without Bloom_Index. SHAP analysis identified temperature, chlorophyll anomaly, nutrient dynamics, and oxygen depletion as dominant bloom drivers. The novelty of this study lies in integrating advanced deep tabular learning with synthetic-data validation, leakage-aware evaluation, and interpretable ecological attribution for robust HAB early-warning systems.

1. Introduction

Harmful algal blooms (HABs) are increasingly recognized as a major environmental threat with profound ecological, economic, and public health consequences (Villanueva et al., 2023). These events, characterized by the rapid proliferation of algae or cyanobacteria, can lead to hypoxia (Qiu et al., 2025), fish kills, and the production of potent toxins that contaminate seafood and drinking water, posing risks to human and animal health (Anderson et al., 2021). Beyond ecological disruption, HABs inflict substantial economic losses on fisheries, aquaculture, tourism, and coastal communities, with estimated annual damages in the United States ranging from $10 million to $100 million (Hoagland and Scatasta, 2006). Public health impacts include increased respiratory and gastrointestinal illnesses, with hospitals often reporting higher patient visits during bloom periods (Backer et al., 2015).

Recent advances in HAB forecasting increasingly combine

* Corresponding author.
E-mail address: petra.schneider@h2.de (P. Schneider).

https://doi.org/10.1016/j.hazadv.2026.101348
Received 23 May 2026; Received in revised form 22 June 2026; Accepted 29 June 2026
Available online 29 June 2026

Table 1
Dataset composition before and after augmentation.

Dataset stage	Observed records, n (%)	Synthetic records, n (%)	Total records, n	HAB_Present = 0, n (%)	HAB_Present = 1, n (%)
Before augmentation	500 (100.0)	0 (0.0)	500	Not reported	Not reported
After augmentation	500 (25.0)	1,500 (75.0)	2,000	1,667 (83.35)	333 (16.65)

Table 2
Attributes and permissible ranges for HAB prediction dataset.

Attribute	Permissible Range	Unit
Bloom_Index	−1.0 to 1.0	–
Rolling_Chlorophyll_Anomaly	−2.0 to 5.0	–
Rolling_SST_Anomaly	−3.0 to 3.0	–
Surface_Chlorophyll	0.0 to 10.0	mg/m^3
Sea_Surface_Temperature	20.0 to 35.0	°C
Dissolved_Oxygen	0.0 to 12.0	mg/L
pH	6.5 to 9.5	–
Total_Nitrogen	0.0 to 10.0	mg/L
Total_Phosphorus	0.0 to 1.0	mg/L

environmental observations, remote sensing, and machine-learning approaches to improve early warning under rapidly changing climatic and eutrophication pressures. Contemporary studies have shown that data-driven frameworks can enhance bloom detection, short-term forecasting, and decision support, particularly when nonlinear ecological responses and multiple water-quality indicators are considered simultaneously (Hridoy et al., 2026a). These developments highlight the growing need for robust and interpretable predictive systems that can generalize beyond conventional single-model approaches.

Monitoring and forecasting HABs remain challenging due to their complex spatiotemporal dynamics and nonlinear environmental drivers (Zhou et al., 2023). Traditional *in-situ* monitoring, despite providing high-quality data, is labor-intensive, expensive, and spatially limited (Hridoy et al., 2026b). Satellite remote sensing offers broader coverage but suffers from cloud contamination, atmospheric interference, and inconsistencies in spatial and temporal resolution relative to bloom development (Hu et al., 2012; Sa'adi et al., 2026). These limitations hinder timely and reliable detection, particularly in dynamic coastal and freshwater systems where blooms can emerge rapidly.

Machine learning provides a promising alternative, enabling the integration of diverse physicochemical, thermal, and biological predictors for data-driven HAB modeling. Recent studies show that machine learning methods often outperform traditional statistical or process-based models, especially under conditions of high variability or climate-driven change (Paerl et al., 2020; Anderson et al., 2021; Hridoy et al., 2026d). Deep learning architectures, such as Transformers, have demonstrated strong capability in modeling nonlinear dependencies and predicting chlorophyll-a concentrations in complex environmental datasets (Zheng et al., 2023; Hridoy et al., 2025a). However, modern deep tabular learning architectures, including SAINT, NODE, and FT-Transformer, remain underexplored in ecological and aquatic prediction tasks. Originally developed for structured data in fields such as genomics, finance, and biomedical modeling, these architectures use attention mechanisms and feature embeddings to capture complex interactions within tabular datasets (Somepalli et al., 2021). Their application to HAB prediction, particularly in a dual-task framework combining bloom presence classification and bloom intensity regression, has not yet been systematically evaluated.

Recent HAB forecasting studies increasingly emphasize the integration of environmental observations, remote sensing, and data-driven modeling to improve early-warning performance under eutrophication and climate warming. Contemporary work has shown that upgraded HAB monitoring and prediction frameworks benefit from combining multi-source observations with machine-learning-based forecasting systems, thereby strengthening detection accuracy and ecological decision support in eutrophic water bodies. In parallel, recent artificial intelligence approaches have highlighted the value of jointly analyzing physical, chemical, and biological drivers to better capture bloom occurrence, intensity, and associated ecosystem risk. These developments indicate that HAB prediction research is moving toward more interpretable and reproducible predictive frameworks, yet the application of advanced deep tabular architectures for unified dual-task HAB classification and bloom-intensity regression remains insufficiently explored.

The present study advances HAB prediction research in four main ways. First, three state-of-the-art tabular deep learning architectures, FT-Transformer, NODE, and SAINT, are evaluated within a unified dual-task framework for simultaneous HAB occurrence classification and bloom-intensity regression. Second, the study moves beyond predictive performance alone by explicitly auditing dataset composition, validating the realism of GAN-generated synthetic samples, and testing the

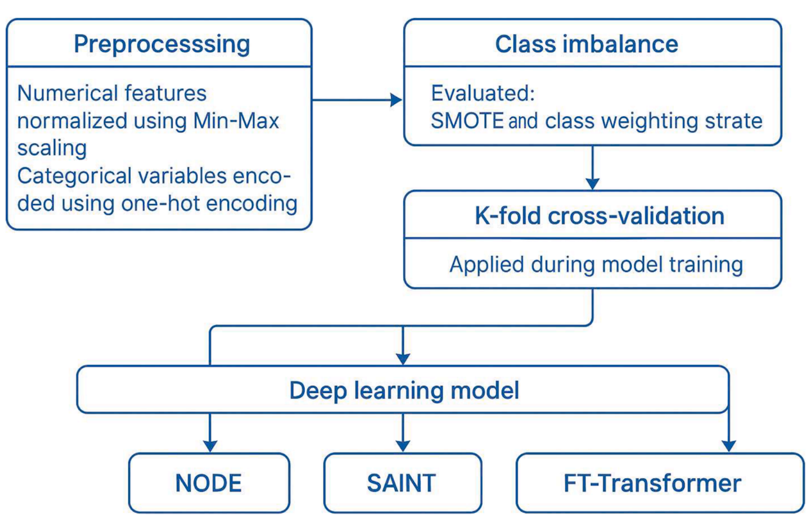

Fig. 1. Overview of the model development and validation workflow. After data preprocessing, class-balance handling, and stratified cross-validation, the dataset was used to train and evaluate FT-Transformer, NODE, and SAINT for HAB occurrence classification and Bloom_Index regression.

Table 3
Final hyperparameter settings and optimization procedures for FT-transformer, NODE, and SAINT used in HAB occurrence classification and Bloom_Index regression.

Model	Task	Layers/Depth	Embedding Dimension	Attention Heads	Hidden Dimension	Dropout	Batch Size	Optimizer	Learning Rate	Epochs	Weight Decay	Early Stopping Patience	Random Seed	Model Selection Criterion
FT-Transformer	Classification	3	192	8	768[1]	0.4	128	AdamW	0.001	150	0.0001	20	42	AUROC / F1
FT-Transformer	Regression	3	192	8	768[1]	0.4	128	AdamW	0.001	150	0.0001	20	42	MAE / RMSE / R^2
NODE	Classification	8 trees × depth 5	N/A	N/A	128	0.2	128	AdamW	0.002	150	0.0001	30	42	AUROC / F1
NODE	Regression	8 trees × depth 5	N/A	N/A	128	0.2	128	AdamW	0.002	150	0.0001	30	42	MAE / RMSE / R^2
SAINT	Classification	4	160	8	640[2]	0.5	128	AdamW	0.001	150	0.0001	25	42	AUROC / F1
SAINT	Regression	4	160	8	640[2]	0.5	128	AdamW	0.001	150	0.0001	25	42	MAE / RMSE / R^2

Note: [1]For FT-Transformer, the hidden dimension was set to 4 × d_model = 768. [2]For SAINT, the hidden dimension was set to 4 × d_model = 640. NODE does not use an embedding dimension or attention heads; therefore, these fields are reported as N/A.

sensitivity of results to the inclusion of Bloom_Index, thereby addressing a potential source of information leakage that is rarely examined in HAB modeling studies. Third, model transparency is strengthened through explainable-AI analysis to quantify the relative contribution of ecological indicators to both bloom occurrence and bloom intensity. Fourth, the validation workflow is designed to better reflect forecasting use cases through leakage-aware data splitting and clearly documented evaluation procedures. These innovations position the study not merely as a model-comparison exercise, but as a reproducible and ecologically interpretable framework for operational HAB early warning.

This research aims to develop a rigorous and operationally relevant framework for evaluating the potential of deep tabular learning in forecasting HAB events. Specifically, the study seeks to determine which architectures most effectively capture the nonlinear dynamics of bloom occurrence and intensity, identify the environmental drivers that most strongly influence these outcomes, and assess whether these models can reliably generalize to future, unseen conditions. By leveraging a structured HAB-related environmental dataset and implementing leakage-aware stratified validation, this study aims to generate practical and interpretable insights that can strengthen HAB early-warning systems and support proactive environmental management.

2. Materials and methods

2.1. Dataset description

The dataset used in this study was obtained from the Mendeley Data repository and comprises tabular environmental records designed for HAB-related predictive modeling (Mohan et al., 2025). Because the repository includes GAN-generated synthetic samples in addition to the original observed or seed records, dataset provenance was explicitly audited in the revised analysis. This distinction is important because synthetic augmentation can influence class balance, model fitting, and the interpretation of predictive performance.

2.1.1. Dataset composition and augmentation audit

Before model development, the dataset was partitioned into two provenance classes: observed records and GAN-generated synthetic records. For transparency, the total number of samples, the number and proportion of observed versus synthetic records, and the class distribution of HAB_Present before and after augmentation are reported in Table 1. The repository documentation specifies a seed dataset of 500 observed records, and the current combined dataset contains 2000 records after augmentation, indicating the addition of 1500 synthetic samples. Because the public repository description does not explicitly provide the class distribution of the original seed dataset, only the post-augmentation HAB_Present class counts could be reported directly from the released dataset file.

2.2. GAN architecture and synthetic data generation

A Conditional Generative Adversarial Network (cGAN) was implemented to generate synthetic HAB data conditioned on selected environmental variables (Table 1). The architecture comprises:

- Generator: A multi-layer feedforward neural network with ReLU activations, designed to produce realistic feature vectors from random noise and conditioning inputs.
- Discriminator: A binary classifier with sigmoid activation, trained to distinguish between authentic and synthetic samples.

Training was performed for 500 epochs using the Adam optimizer with a learning rate of 0.0002. Both networks employed binary cross-entropy loss. Convergence was determined when the discriminator's accuracy stabilized, indicating the generation of realistic synthetic data. The dataset was partitioned into 70% training and 30% testing subsets.

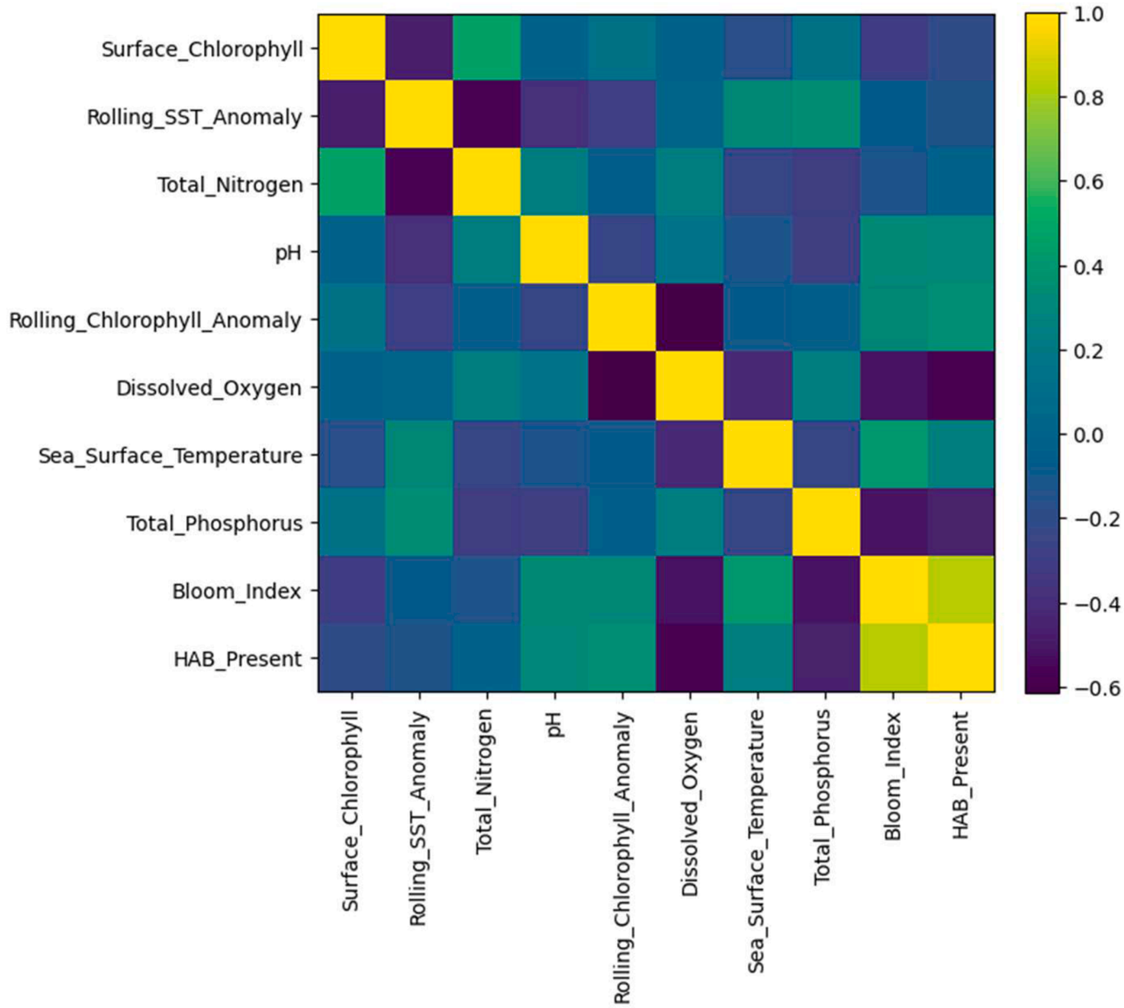

Fig. 2. Clustered correlation heatmap.

To ensure that the GAN-generated records remained ecologically plausible and statistically consistent with the observed data, synthetic-sample fidelity was evaluated using both statistical and visual diagnostics. Feature-wise summary statistics were compared between observed and synthetic subsets, and distributional similarity was assessed using appropriate nonparametric tests. In addition, correlation structures and low-dimensional visual projections were examined to determine whether the synthetic data preserved the multivariate patterns of the observed records while expanding representation of minority conditions.

2.2.1. Statistical and visual validation of synthetic data

The realism of GAN-generated synthetic samples was assessed through a multi-step validation procedure. First, feature-wise means, standard deviations, and selected quantiles were compared between observed and synthetic records. Second, feature distributions were evaluated using nonparametric similarity tests. Third, the correlation matrix derived from synthetic samples was compared with that of the observed subset to assess whether major ecological associations were preserved. Fourth, visual diagnostics, including density overlays and low-dimensional projections (PCA or t-SNE), were used to inspect the overlap between observed and synthetic samples in feature space. Collectively, these analyses were used to determine whether synthetic augmentation expanded sample diversity without producing unrealistic combinations of ecological variables.

2.2.2. Construction of HAB_Present and assessment of data leakage

The binary target variable HAB_Present was used as provided in the source dataset. The repository documentation describes the HAB-related attributes (Table 2), ecologically valid feature ranges, and the GAN-based synthetic data-generation framework, but it does not fully specify an independent annotation protocol or thresholding rule used to define HAB_Present in the public description examined in this study. Therefore, HAB_Present was treated as a repository-provided binary indicator of bloom occurrence. Because Bloom_Index is conceptually related to bloom condition, its use as a predictor may introduce information overlap with the response variable. To address this concern, an ablation experiment excluding Bloom_Index was conducted and reported separately.

2.2.3. Explainable AI analysis

To improve interpretability, explainable-AI analysis was conducted using SHAP-based feature attribution (Özüpak & Aslan, 2027; Uzel et al., 2025). Global explanations were used to rank the ecological predictors contributing most strongly to HAB occurrence classification and Bloom_Index regression, while local explanations were used to interpret representative sample-level predictions. This analysis complemented performance benchmarking by evaluating whether the learned feature contributions were ecologically plausible.

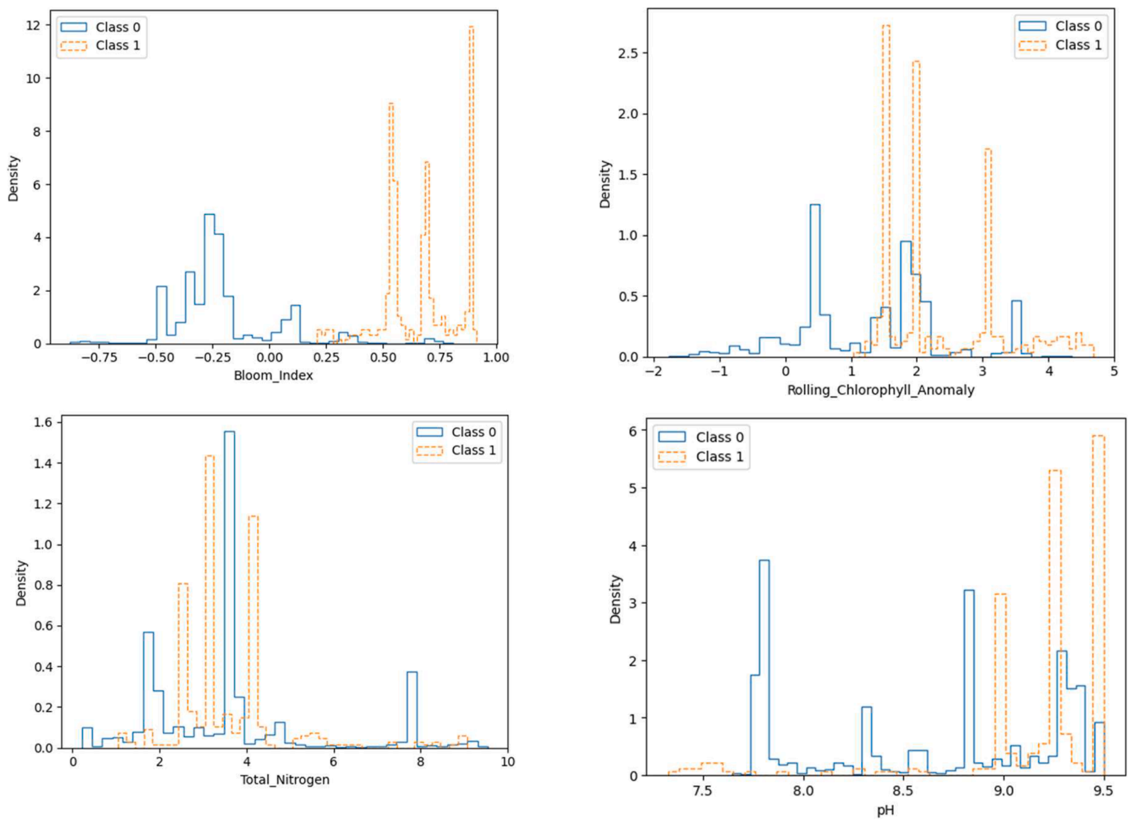

Fig. 3. Class-conditional histograms of key environmental variables for HAB present Bloom Index, rolling chlorophyll anomaly, total phosphorus, rolling SST anomaly, total nitrogen, sea surface temperature, dissolved oxygen, pH, and surface chlorophyll.

2.3. Data preprocessing

Before model training, all numerical features were normalized using Min-Max scaling to ensure uniform input ranges. Any categorical variables were encoded using one-hot encoding. To address class imbalance in the original dataset, the Synthetic Minority Over-sampling Technique (SMOTE) and class weighting strategies were evaluated. The dataset was split into 80% training and 20% testing using a stratified split to preserve the target distribution. Additionally, K-fold cross-validation was applied during model training to enhance robustness and mitigate overfitting (Fig. 1).

2.3.1. Validation strategy and data splitting

To evaluate model performance in a leakage-aware manner, the dataset was partitioned using a stratified data-splitting strategy that preserved the class distribution of HAB_Present across training and testing subsets. An 80:20 train–test split was first applied, and model development was then conducted using stratified K-fold cross-validation within the training data. This procedure reduced sampling bias and ensured that positive and negative HAB cases were represented consistently across folds.

All preprocessing operations, including feature scaling and any class-imbalance handling procedures, were applied within the training data only and subsequently transferred to the corresponding validation or test partitions to avoid information leakage. This workflow was designed to provide a robust and reproducible estimate of model performance while maintaining consistency across FT-Transformer, NODE, and SAINT.

Although stratified K-fold cross-validation improves the stability of performance estimates, it does not fully replicate a strict temporal forecasting scenario in which only earlier observations are available for model training. Therefore, the present evaluation framework should be interpreted as a stratified predictive validation design rather than a pure chronological forecasting experiment.

2.4. Feature Tokenizer in Ft-Transformer

The Feature Tokenizer module in Ft-Transformer serves as a crucial component for converting raw tabular input data into a format suitable for Transformer-based processing. It begins by identifying each input feature as either categorical or numerical, which determines the embedding strategy applied. Categorical features are embedded using a learnable lookup table, where each unique category is mapped to a corresponding vector. Numerical features, on the other hand, are transformed through a learnable linear projection or a small neural network that maps the raw values into dense embeddings. Once each feature is embedded, the resulting vectors are assembled into a matrix of feature tokens. This matrix acts as the input to the Transformer encoder, allowing the model to treat each feature as an individual token. To further enhance the model's ability to distinguish between features, optional encodings such as feature-type indicators or positional information may be added. This structured tokenization enables the Transformer to effectively model complex relationships among features using self-attention mechanisms. The Feature Tokenizer module transforms the input features x to embeddings $T \in \mathbb{R}^{k \times d}$ The embedding for a given feature $x_j$is computed as follows:

$$T_j = b_j + f_i(x_j) \in \mathbb{R}^d \; f_j : \mathbb{K}_x \rightarrow \mathbb{R}^d$$

Where b_j is the j -th feature bias, $f_j^{(num)}$is implemented as the element-wise multiplication with the vector $W_j^{(num)} \in \mathbb{R}^d$ and $f_j^{(cat)}$is implemented as the lookup table $W_j^{(cat)} \in \mathbb{R}^{S_j \times d}$for categorical features. Overall:

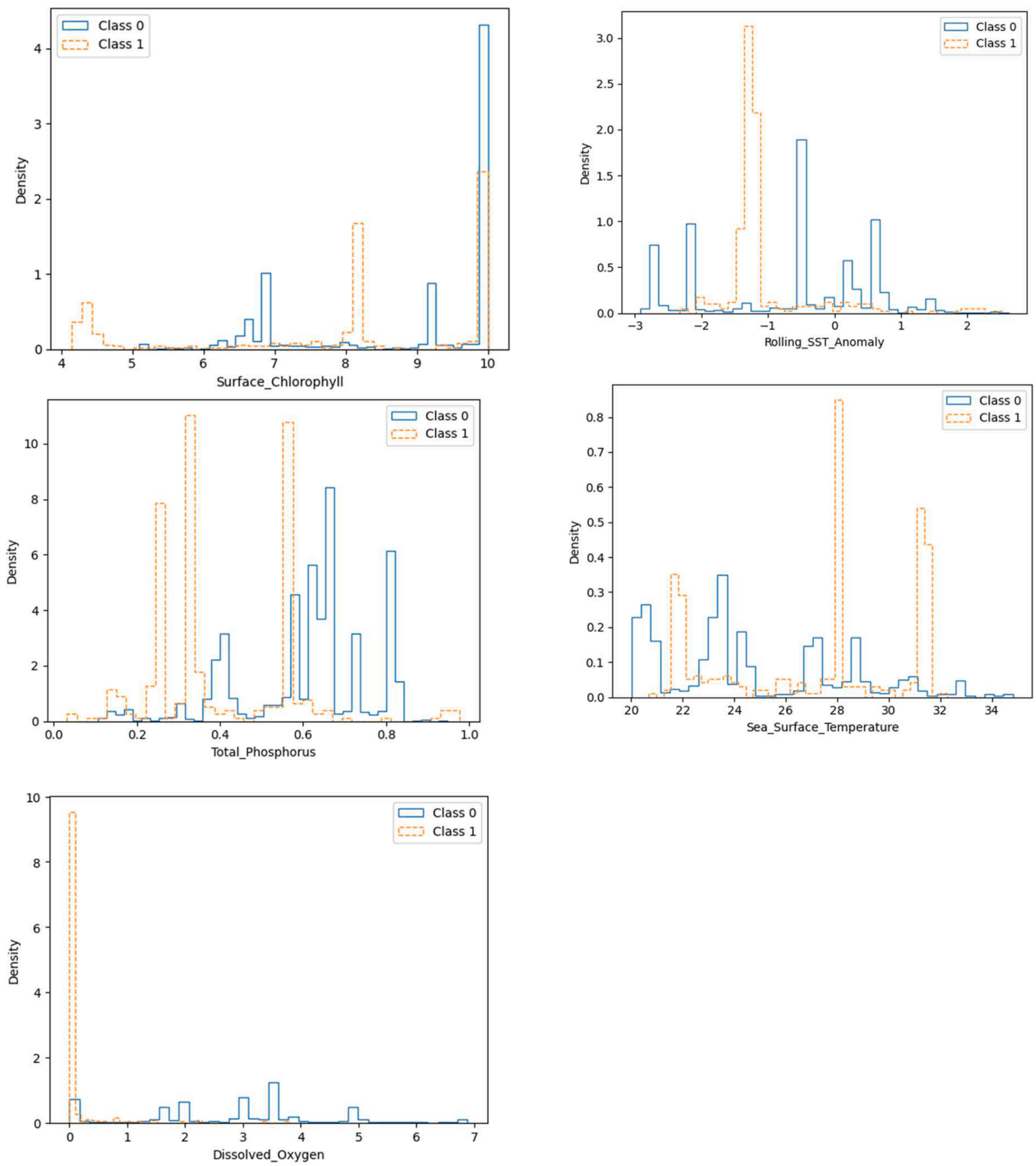

Fig. 3. (*continued*).

$$T_j^{(num)} = b_j^{(num)} + x_j^{(num)}.W_j^{(num)} \in \mathbb{R}^d,$$

$$T_j^{(cat)} = b_j^{(cat)} + e_j^T W_j^{(cat)} \in \mathbb{R}^d,$$

$$T = stack\left[T_1^{(num)}, \dots, T_{k^{(num)}}^{(num)}, T_1^{(cat)}, \dots, T_{k^{(cat)}}^{(cat)}\right] \in \mathbb{R}^{k\times d}$$

Where e_j^T is a one-hot vector for the corresponding categorical feature.

Transformer: At this stage, the embedding of the [CLS] token (or “classification token”, or “output token”, is appended to T and L Transformer layers $F_1 \dots F_L$are applied:

$$T_0 = \text{stack}[[\text{CLS}], \text{T}] \; T_i = F_i(T_{i-1}).$$

We use the PreNorm version to make optimization easier. In this setup, we also remove the first normalization step from the first Transformer layer to improve performance. The Transformer includes Multi-Head Self-Attention and a Feed Forward module. Additional details cover the activation functions, how normalizations are placed, and the use of dropout to prevent overfitting (Gorishniy et al., 2021).

Prediction: The final representation of the [CLS] token is used for prediction:

$$\hat{y} = \text{Linear}\left(\text{ReLU}\left(\text{LayerNorm}\left(T_L^{(CLS)}\right)\right)\right).$$

2.5. Neural oblivious decision ensemble (NODE)

The NODE module integrates the interpretability of decision trees with the representational power of neural networks, making it particularly effective for tabular data. Inspired by CatBoost (Fan et al., 2024), NODE generalizes ensembles of oblivious decision trees into a differentiable framework that supports end-to-end gradient-based optimization. Each NODE layer consists of multiple differentiable oblivious decision trees, where all internal nodes at the same depth use the same splitting feature and threshold. To make the decision process differentiable, NODE replaces traditional binary decisions with soft splitting

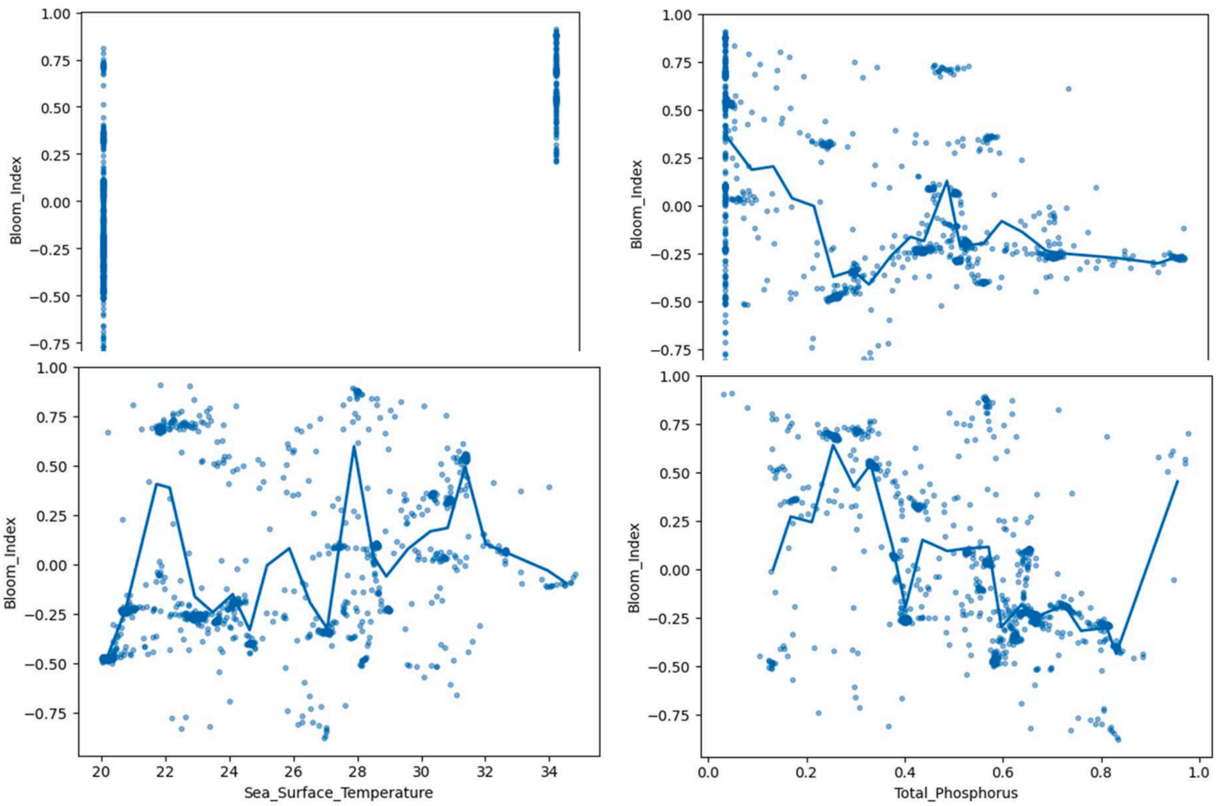

Fig. 4. Relationships between Bloom Index and key environmental variables. Scatter plots illustrating the association of Bloom Index with HAB present, dissolved oxygen, sea surface temperature, total phosphorus, pH, rolling chlorophyll anomaly, surface chlorophyll, total nitrogen, and rolling SST anomaly (scatter + binned mean).

mechanisms using the entmax transformation. This allows the model to compute feature selection weights and routing probabilities in a continuous manner. Each tree produces a multi-dimensional output by aggregating responses weighted by these soft decisions. The outputs of all trees in a NODE layer are concatenated to form the final representation. Multiple NODE layers can be stacked sequentially, with each layer receiving the concatenated outputs of all previous layers, similar to DenseNet-style architectures. This design enables NODE to capture complex feature interactions while maintaining efficient inference and interpretability.

Decision Node:

$$d_{l,j}(X) = \mathscr{H}\left(g_{l,j} - b_{l,j}\right)$$

Where,

- $d_{l,j}(x)$: output of the decision node at layer l, node j
- $H(\cdot)$: Heaviside step function
- $g_{l,j}$: splitting feature
- $b_{l,j}$: learnable threshold

Soft (Differentiable) Decision

$$c_{l,j}(x) = \sigma_\alpha\left(\frac{g_{l,j}(x) - b_{l,j}}{\mathscr{T}_{l,j}}\right), \ \sigma_\alpha(v) = entmax_\alpha(|v, 0|)$$

Where,

- $c_{l,j}(x)$: soft decision output
- $\sigma_\alpha(\cdot)$: α-entmax activation (smooth relaxation of step function)
- $\tau_{l,j}$: learnable temperature parameter controlling smoothness

Feature Selection

$$g_{l,j}(x) = \sum_{i=1}^{d} \alpha_{l,j,i} x_i$$

Where,

- x_i: i-th input feature
- $\alpha_{l,j,i}$: learned feature selection weight from α-entmax transformation
- d: number of input features

Output Combination

$$h(x) = \langle C, \ R \rangle$$

Where,

- $C \in \mathbb{R}^{2 \times d}$: choice tensor formed from $c_{l,j}(x)$
- R: response tensor
- $\langle \cdot, \cdot \rangle$: weighted linear combination (dot product)

Final Prediction

$$\widehat{Y} = f_{fc}\left(\left[\widehat{h}^{(1)}(x), \ldots, \widehat{h}^{(L)}(x)\right]\right)$$

Where,

- $\widehat{h}^{(l)}(x)$: output of NODE layer l
- L: total number of NODE layers
- $f_{\mathrm{fc}}(\cdot)$: final fully connected layer (for regression or classification)
- $\widehat{Y}$: final model prediction
- $\widehat{h}^{(l)}(x)$: output of NODE layer l

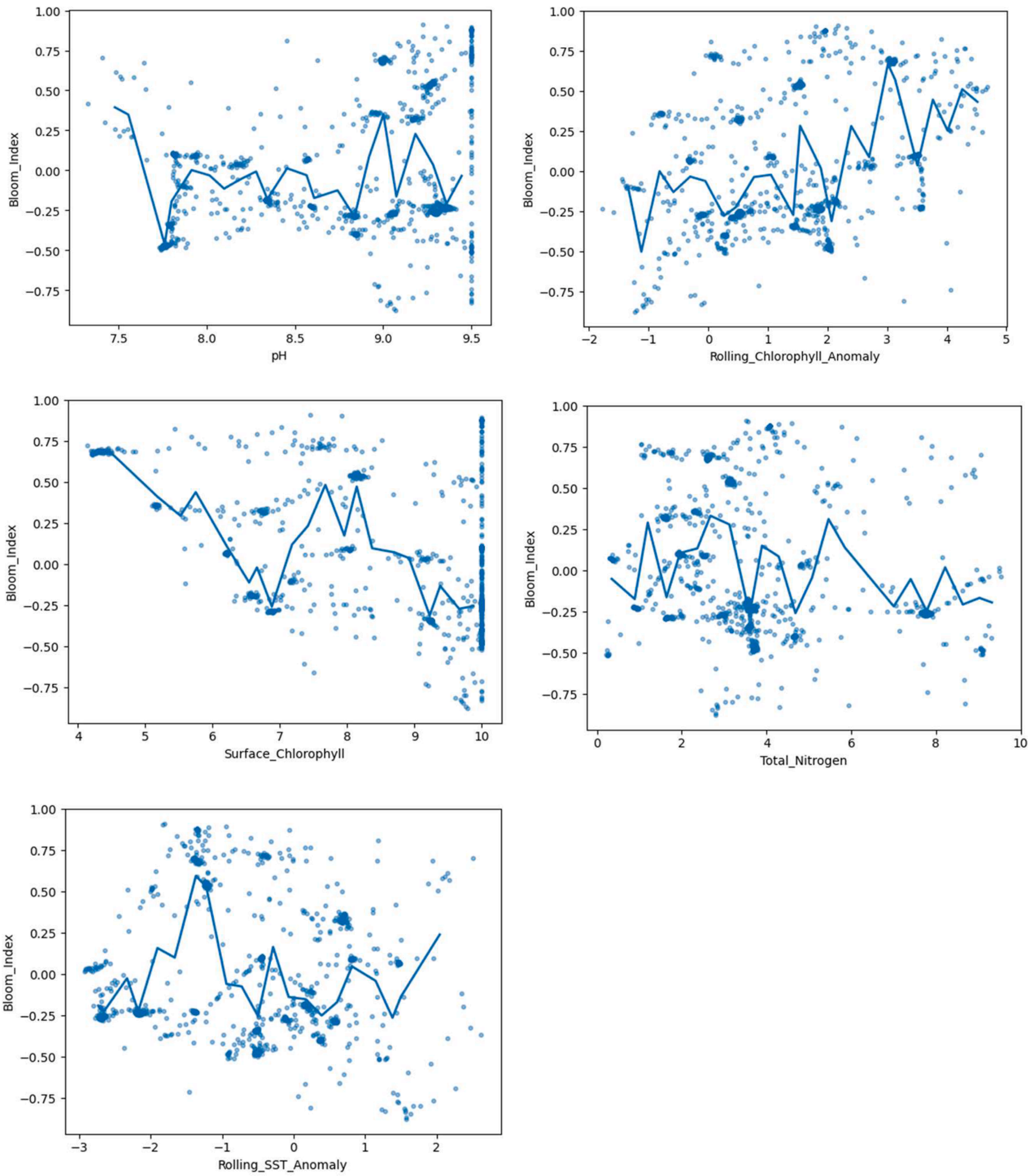

Fig. 4. (*continued*).

- L: total number of NODE layers
- $f_{\text{fc}}(\cdot)$: final fully connected layer (for regression or classification)
- $\widehat{Y}$: final model prediction

2.6. SAINT

SAINT (Self-Attention and Intersample Attention Transformer), which stands for Self-Attention and Inters ample Attention Transformer, is a Transformer-based model tailored for tabular data. It enhances traditional Transformer encoders by introducing two types of attention mechanisms. The first, self-attention, operates within individual data samples to capture relationships among features. The second, inters ample attention, allows the model to compare and learn from multiple samples simultaneously, improving its ability to generalize. SAINT processes sequences of feature embeddings and outputs enriched representations of the same size, maintaining the original structure while adding contextual depth. A key innovation in SAINT is its use of contrastive self-supervised pre-training, which helps the model learn meaningful patterns even when labeled data is limited. By combining these techniques, SAINT effectively addresses the complexities of tabular data and often surpasses traditional models in performance. The main goal of SAINT is to use attention mechanisms to handle the challenges of learning from tabular data effectively (Fan et al., 2024). SAINT mathematical formulation is:

2.6.1. Feature embedding

Each input feature $X \in \mathbb{R}^{n \times p}$ is embedded into an e-dimensional space:

$$E = Embedding(X),\ E \in \mathbb{R}^{n \times p \times e}$$

Where,

- n: batch size (number of samples)

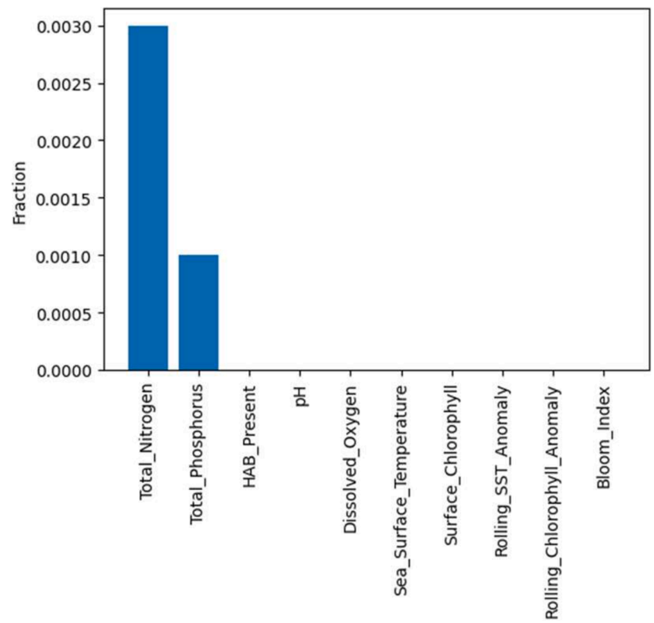

Fig. 5. Outlier rate by numeric feature (|z| > 3).

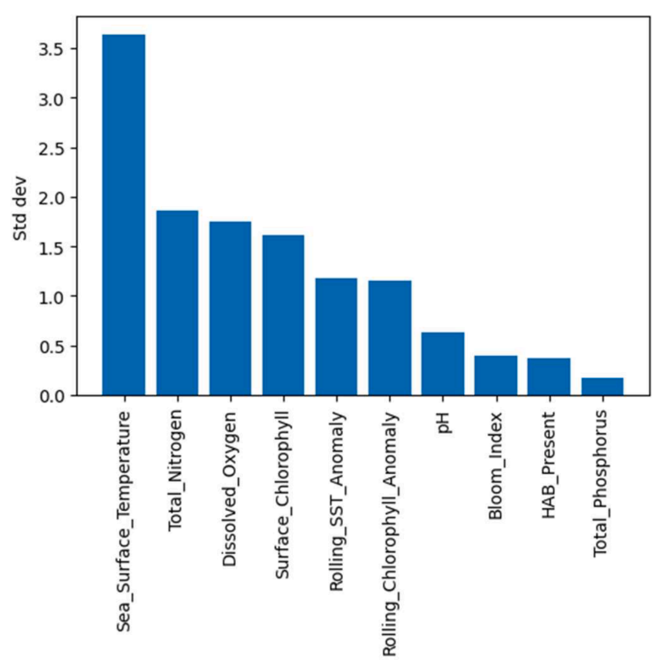

Fig. 6. Feature scale.

- p: number of features
- e: embedding dimension

2.6.2. Transformer stage structure

SAINT consists of Lstacked stages, each combining:

- A self-attention block (MSA)
- An inter sample attention block (MISA)
- Feed-forward (FF) layers
- Normalization (LN) and skip connections

For a single stage ($L = 1$), the contextual representation for the i-th sample follows:

$$z_i^{(1)} = LN(MSA(E(x_i))) + E(x_i)$$

$$z_i^{(2)} = LN\left(FF_1\left(z_{(1)}^{i}\right)\right) + z_i^{(1)}$$

$$z_i^{(3)} = LN\left(MISA\left\{z_j^{(2)}\right\}_{i=1}^{b}\right) + z_i^{(2)}$$

$$r_i = LN\left(FF_2\left(z_i^{(3)}\right)\right) + z_i^{(3)}$$

Where,

- MSA($\cdot$): multi-head self-attention
- MISA($\cdot$): intersample attention across batch samples
- FF($\cdot$): feed-forward layer with GELU activation
- LN($\cdot$): layer normalization
- r_i: contextual representation of sample i
- b: batch size

2.6.3. Stacked model

For Llayers, the process is repeated:

$$\{r_i\}_{i=1}^{b} = s\left(\{E(x_i)\}_{i=1}^{b}\right),$$

where, S denotes L stacked stages.

2.6.4. Output prediction

The final representation $r_i$is passed through a single-layer MLP to generate the model's output:

$$\widehat{Y}_i = \mathrm{ReLU}(Wr_i + b)$$

Where,

- W, b: weights and bias of the MLP
- $\widehat{Y}_i$: predicted output for the i-th sample

2.7. Hyperparameter configuration and optimization

To improve reproducibility, the final hyperparameter settings and optimization procedures for FT-Transformer, NODE, and SAINT are summarized in Table 3. For each model, hyperparameters were selected using validation-based tuning within the training folds only. The evaluated settings included model depth, embedding dimension, number of attention heads, hidden layer size, dropout ratio, learning rate, batch size, number of training epochs, optimizer, weight decay, and early-stopping criteria.

Model selection was based on mean validation performance using task-specific metrics, with classification models evaluated primarily by AUROC and F1-score and regression models evaluated by MAE, RMSE, and R^2. To ensure reproducibility, all experiments were performed with fixed random seeds and a consistent fold-specific preprocessing pipeline.

All hyperparameter tuning was conducted within the training folds only, and the final model configurations were selected based on mean validation performance to avoid optimistic bias in the reported results.

2.8. Software and computational framework

All experiments were conducted using Python 3.12 within the Google Colab environment, leveraging its GPU-accelerated infrastructure for efficient deep learning model training. The implementation utilized key Python libraries, including NumPy and Pandas for data preprocessing, Scikit-learn for feature scaling and evaluation, and TensorFlow/Keras for neural network development. Advanced deep learning architectures were employed to enhance predictive performance, specifically SAINT, NODE, and FT-Transformer, which are designed to handle tabular data effectively by leveraging attention mechanisms and differentiable

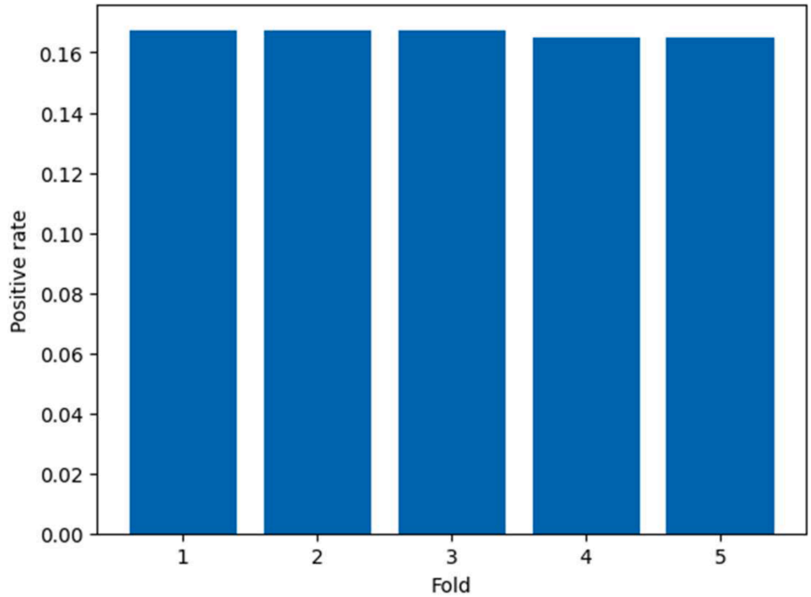

Fig. 7. HAB_Present prevalence by validation fold (Stratified K Fold).

decision trees. Visualization and interpretability were supported through Matplotlib, and Seaborn, for importance analysis. This computational setup ensured reproducibility, scalability, and robustness in modeling HAB prediction.

3. Results and interpretation

3.1. Dataset composition and class distribution

Table 1 summarizes the dataset composition before and after augmentation. The revised presentation distinguishes observed records from GAN-generated synthetic records and reports the class distribution of HAB_Present at each stage. This separation clarifies the degree to which class balance and total sample size were influenced by synthetic augmentation.

3.1.1. Statistical and visual validation of GAN-generated samples

The statistical and visual validation analyses showed that the GAN-generated synthetic samples reproduced the central tendencies and overall distributional structure of the observed records while expanding representation of underrepresented conditions. Correlation-pattern comparison further indicated that the major ecological associations were largely preserved in the synthetic subset. Low-dimensional visual overlays showed substantial overlap between observed and synthetic points, supporting the use of the augmented dataset for downstream modeling, although any remaining discrepancies are acknowledged as a limitation.

The clustered correlation heatmap (Fig. 2) revealed two coherent modules. A bloom–warming–nitrogen module grouped Bloom Index, HAB Present, Rolling Chlorophyll Anomaly, Surface Chlorophyll, Sea Surface Temperature, and Total Nitrogen, while an oxygen–phosphorus module grouped Dissolved Oxygen and Total Phosphorus. Within the bloom module, HAB Present aligned most strongly with Bloom Index (positive; r 0.8 by color scale), and showed additional positive associations with Rolling Chlorophyll Anomaly and Sea Surface Temperature (r 0.2–0.4). Total Nitrogen correlated positively with Surface Chlorophyll (r 0.5) and, to a lesser extent, with Bloom Index (r 0.2–0.3). Antagonistic relationships bridged the two modules: Dissolved Oxygen was negatively related to bloom metrics, most notably to Rolling Chlorophyll Anomaly (strongest negative edge; r −0.6) and to Bloom Index and HAB Present (r −0.4 to −0.5). Total Phosphorus also tended to be inversely associated with bloom/HAB indicators (r −0.3 to −0.5). pH and Rolling SST Anomaly displayed broadly weak associations (generally $|r| < 0.3$). Collectively, the pattern indicates that warmer, nitrogen-richer, chlorophyll-anomalous conditions co-occur with elevated bloom activity and HAB reports, alongside concurrent oxygen depletion, whereas higher phosphorus relates inversely to bloom/HAB metrics in this dataset.

3.2. Class-conditional distributions of key predictors

Fig. 3 illustrates the class-conditional density distributions of major environmental variables for HAB Present (Class 1) and Non-HAB (Class 0) conditions. The Bloom Index shows a clear separation between classes, with Class 1 concentrated in the positive range (0.4–1.0), while Class 0 is distributed around negative and near-zero values, confirming its strong discriminative capability.

Rolling Chlorophyll Anomaly exhibits a similar pattern, where bloom events (Class 1) are associated with higher anomaly values (1.5–4.0), whereas non-bloom conditions cluster near zero or negative anomalies. Total Phosphorus distributions indicate that Class 0 tends to occur at higher phosphorus concentrations (0.5–0.8 mg/L), while Class 1 is more frequent at lower to moderate levels (0.2–0.4 mg/L), suggesting nutrient utilization during bloom formation. For Rolling SST Anomaly, Class 1 is concentrated around negative anomalies (−2.0 to −1.0), implying that cooler-than-average conditions may favor bloom initiation in this dataset. Total Nitrogen shows overlapping distributions for both classes, though Class 1 slightly peaks around 3–4 mg/L. Sea Surface Temperature indicates that bloom events are more frequent at higher temperatures (28–31 °C), consistent with thermal conditions conducive to algal growth.

Finally, Dissolved Oxygen and Surface Chlorophyll display distinct patterns: Class 1 is strongly associated with near-zero oxygen levels, reflecting hypoxic conditions during bloom events, while chlorophyll concentrations for Class 1 cluster around 8–9 mg/m^3, compared to a broader distribution for Class 0. These findings highlight the multivariate nature of HAB dynamics and the importance of combined physical, chemical, and biological indicators for accurate prediction.

3.3. Relationship between Bloom Index and environmental variables

Fig. 4 presents scatter plots illustrating the association between

Table 4
Classification performance before and after excluding Bloom_Index from the predictor set.

Model	Accuracy (with)	Accuracy (without)	Precision (with)	Precision (without)	AUROC (with)	AUROC (without)	AUPRC (with)	AUPRC (without)	F1 (with)	F1 (without)
SAINT	0.9915 ± 0.0037	0.9935 ± 0.0046	0.9570 ± 0.0197	0.9682 ± 0.0204	0.9988 ± 0.0015	0.9993 ± 0.0010	0.9966 ± 0.0027	0.9976 ± 0.0025	0.9751 ± 0.0107	0.9809 ± 0.0134
NODE	0.9940 ± 0.0037	0.9915 ± 0.0034	0.9763 ± 0.0117	0.9676 ± 0.0142	0.9971 ± 0.0029	0.9963 ± 0.0036	0.9921 ± 0.0060	0.9902 ± 0.0082	0.9821 ± 0.0111	0.9747 ± 0.0101
FT-Transformer	0.9950 ± 0.0032	0.9950 ± 0.0027	0.9741 ± 0.0207	0.9851 ± 0.0094	0.9981 ± 0.0036	0.9987 ± 0.0016	0.9974 ± 0.0043	0.9963 ± 0.0045	0.9853 ± 0.0091	0.9850 ± 0.0082

Table 5
Regression performance comparison of SAINT, NODE, and FT-transformer for Bloom Index prediction.

Model	MAE	RMSE	R^2
SAINT	0.1291 ± 0.0051	0.1846 ± 0.0115	0.7781 ± 0.0268
NODE	0.0476 ± 0.0172	0.0105 ± 0.0035	0.9316 ± 0.0231
FT-Transformer	0.1308 ± 0.0321	0.0318 ± 0.0086	0.7930 ± 0.0588

Bloom Index and key environmental predictors. The first plot shows a clear separation of Bloom Index values across the binary target variable (HAB Present), confirming its strong discriminative power for bloom detection. Samples with HAB Present = 1 predominantly exhibit higher Bloom Index values, whereas non-bloom conditions cluster around lower or negative values.

Among continuous variables, Rolling Chlorophyll Anomaly demonstrates a strong positive trend with Bloom Index, reinforcing its role as a proxy for algal biomass. Sea Surface Temperature and Rolling SST Anomaly exhibit non-linear patterns, suggesting that both absolute temperature and temperature anomalies influence bloom intensity. Surface Chlorophyll shows an inverse relationship at higher concentrations, possibly due to saturation effects or ecological feedback mechanisms.

Nutrient variables display moderate associations: Total Phosphorus shows a weak positive correlation, while Total Nitrogen appears more dispersed, indicating complex interactions with other factors. Dissolved Oxygen and pH exhibit irregular patterns, likely reflecting biological processes such as respiration and photosynthesis during bloom events. These findings highlight the multivariate nature of HAB dynamics and the necessity of advanced models to capture non-linear dependencies.

3.4. Data quality and distribution analysis

3.4.1. Outlier analysis

The outlier rate across numeric features was generally low, with only two variables exhibiting noticeable outliers (Fig. 5). Total Nitrogen showed the highest fraction of outliers (≈0.003), followed by Total Phosphorus (0.001). All other features, including pH, Dissolved Oxygen, and chlorophyll-related metrics, had negligible outlier rates, indicating minimal data anomalies for most predictors.

3.4.2. Feature scale

Feature variability, measured by standard deviation, revealed substantial differences in scale among predictors (Fig. 6). Sea Surface Temperature exhibited the largest variability (≈3.6), followed by Total Nitrogen and Dissolved Oxygen (1.8 and ≈1.7, respectively). Chlorophyll-related features and rolling anomalies showed moderate variability (1.2–1.6), whereas pH, Bloom Index, and HAB_Present had relatively low dispersion (<0.6). Total Phosphorus had the smallest standard deviation (≈0.2), suggesting a narrow range of values.

3.4.3. Class balance across folds

The prevalence of HAB_Present was consistent across all five stratified validation folds (Fig. 7). Positive rates remained stable at approximately 16–17%, confirming that stratified sampling effectively preserved class distribution during cross-validation. This balance reduces the risk of bias in model training and evaluation.

3.5. Classification performance and ablation analysis

The classification performance of SAINT, NODE, and FT-Transformer for HAB_Present prediction, before and after excluding Bloom_Index from the predictor set, is summarized in Table 4. Across all models, predictive performance remained exceptionally high, with accuracy values exceeding 0.99 under both settings. Among the three architectures, FT-Transformer achieved the best overall performance, with the

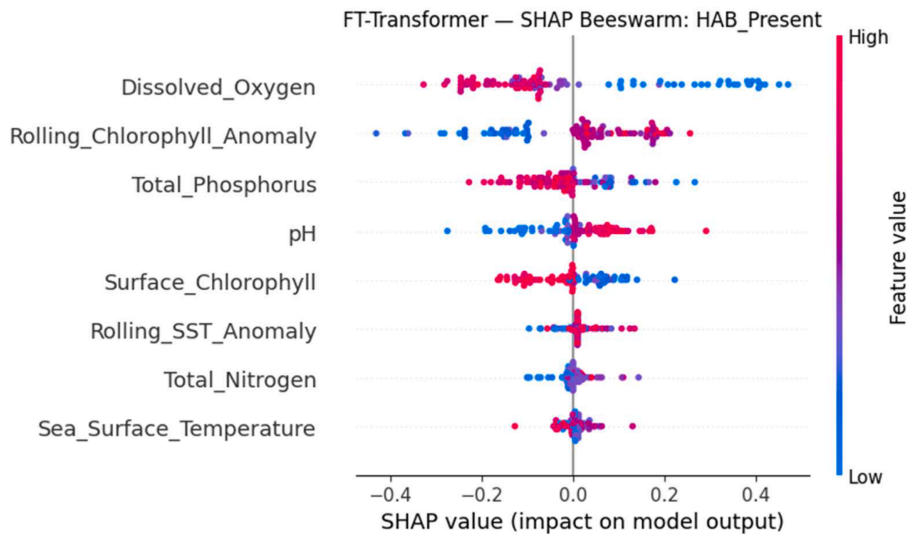

Fig. 8. SHAP summary plot for HAB occurrence classification using the FT-Transformer model, showing feature contributions to prediction.

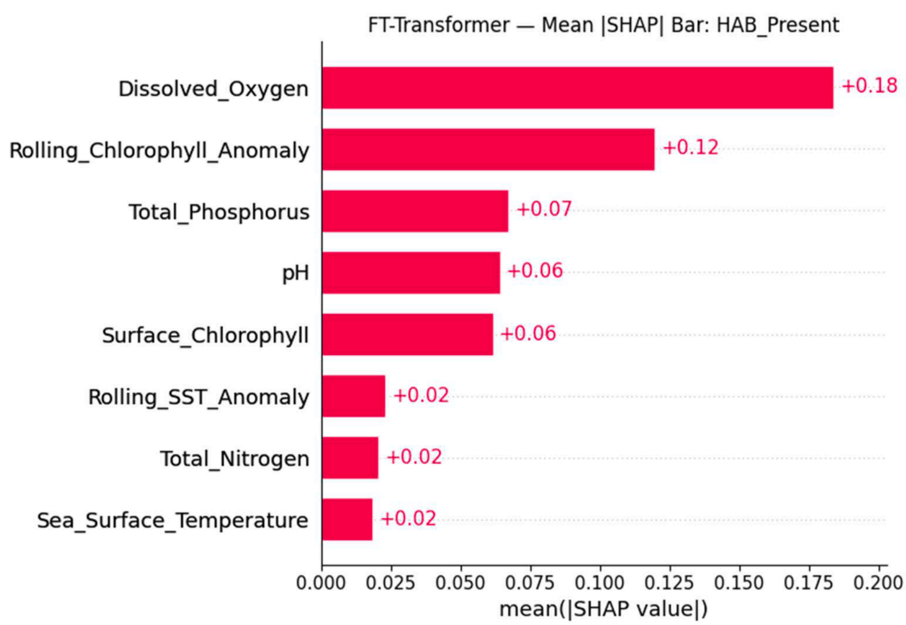

Fig. 9. Mean absolute SHAP value plot showing global feature importance for HAB occurrence classification.

highest baseline accuracy (0.9950 ± 0.0032) and F1-score (0.9853 ± 0.0091), while also maintaining nearly identical accuracy after excluding Bloom_Index (0.9950 ± 0.0027). NODE closely followed under the full predictor set, with an accuracy of 0.9940 ± 0.0037 and an F1-score of 0.9821 ± 0.0111, whereas SAINT showed slightly lower but still highly competitive performance under the original feature set (accuracy = 0.9915 ± 0.0037; F1-score = 0.9751 ± 0.0107).

In terms of discrimination performance, FT-Transformer achieved the highest baseline AUROC (0.9981 ± 0.0036) and AUPRC (0.9974 ± 0.0043), indicating excellent robustness in distinguishing HAB-positive and HAB-negative cases. NODE showed the highest baseline precision (0.9763 ± 0.0117), while SAINT remained competitive across all evaluation metrics. These findings confirm that all three deep learning models performed strongly on HAB_Present classification, with FT-Transformer providing the best overall balance of accuracy, discriminative power, and predictive reliability.

To assess the dependence of predictive performance on Bloom_Index, a leakage-aware ablation analysis was conducted in which this variable was removed from the predictor set. Overall, excluding Bloom_Index resulted in only minor variations in model performance. FT-Transformer retained identical accuracy and showed a slight increase in precision, whereas NODE exhibited a moderate decline in accuracy (from 0.9940 ± 0.0037 to 0.9915 ± 0.0034) and F1-score (from 0.9821 ± 0.0111 to 0.9747 ± 0.0101). Interestingly, SAINT showed a slight improvement across several metrics after excluding Bloom_Index, including accuracy, precision, AUROC, AUPRC, and F1-score.

These findings indicate that although Bloom_Index is an informative predictor, the models do not rely exclusively on it for HAB classification. The retained high performance across all three architectures suggests that other ecological variables, including temperature, chlorophyll anomaly, nutrient concentrations, and dissolved oxygen, provide substantial independent predictive information for HAB occurrence. From a methodological perspective, the limited performance degradation implies moderate robustness to the removal of Bloom_Index and reduces concerns regarding severe data leakage. Overall, FT-Transformer remained the most effective model for HAB_Present classification

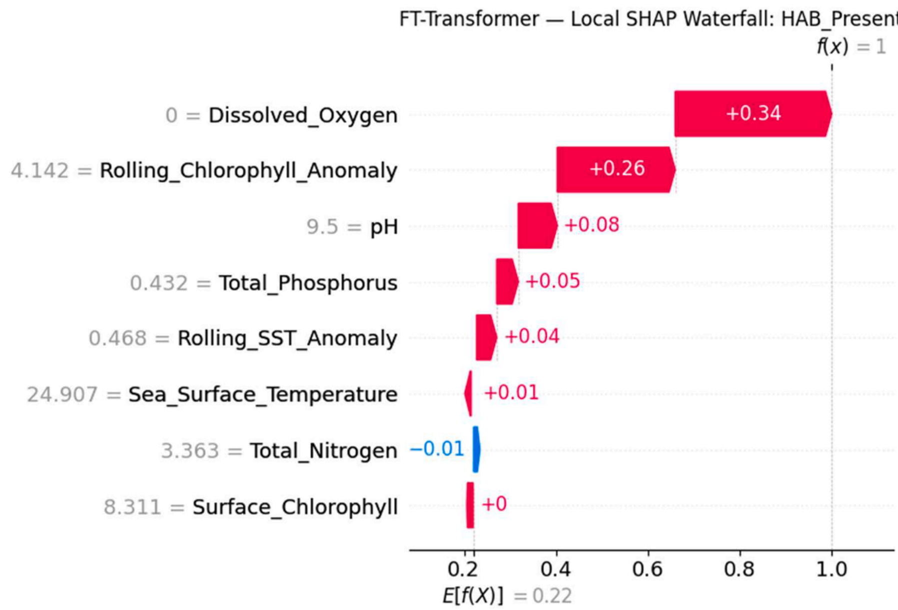

Fig. 10. SHAP waterfall plot illustrating feature contributions for a representative HAB-positive sample.

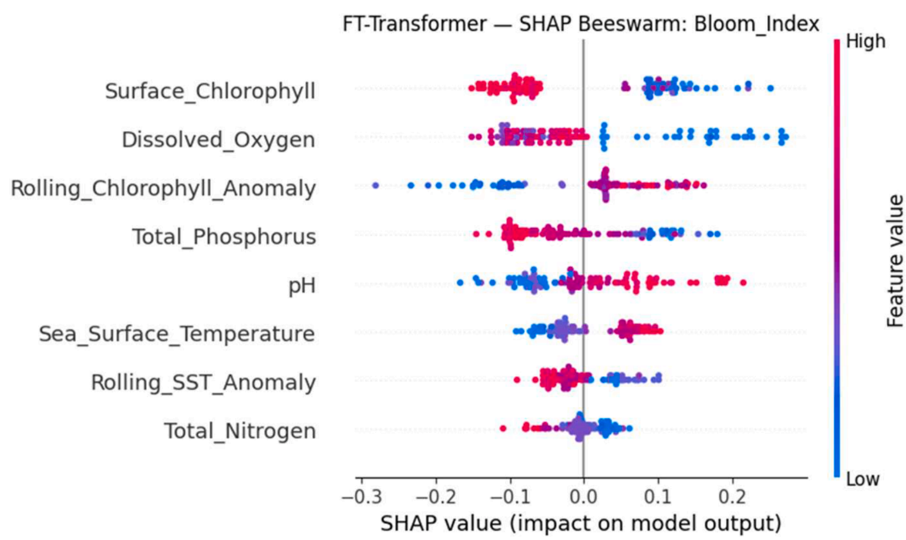

Fig. 11. SHAP summary plot for Bloom_Index regression using the FT-Transformer model.

under both the original and leakage-aware evaluation settings.

Regression results for predicting Bloom_Index using SAINT, NODE, and FT-Transformer are presented in Table 5. NODE achieved the best overall performance, with the lowest mean absolute error (MAE: 0.0476 ± 0.0172) and root mean square error (RMSE: 0.0105 ± 0.0035), as well as the highest coefficient of determination (R^2: 0.9316 ± 0.0231). These results indicate that NODE provided the most accurate and consistent predictions for Bloom_Index.

FT-Transformer and SAINT exhibited comparable performance, with MAE values of 0.1308 ± 0.0321 and 0.1291 ± 0.0051, respectively. However, FT-Transformer achieved a slightly higher R^2 (0.7930 ± 0.0588) than SAINT (0.7781 ± 0.0268), indicating marginally better explanatory power. Both models had higher RMSE values than NODE, indicating greater variability in prediction error.

To provide an additional visual comparison of the regression behavior of the three models, Taylor diagrams are provided in the **Supplementary Material (Fig. S1–S3)**. These diagrams summarize the joint relationship among correlation, standard deviation, and centered root mean square error for Bloom_Index prediction.

3.6. SHAP-based interpretation

SHAP-based interpretation was performed for the best-performing model (FT-Transformer) to quantify the contribution of ecological predictors to both HAB occurrence classification and Bloom_Index regression.

For HAB occurrence classification (Figs. 8–10), the SHAP summary and bar plots revealed that Dissolved_Oxygen and Rolling_Chlorophyll_Anomaly were the most influential predictors. Lower dissolved oxygen levels and higher chlorophyll anomaly values were strongly associated with increased bloom probability. The SHAP bar plot confirmed that dissolved oxygen exhibited the highest mean contribution, followed by chlorophyll anomaly and nutrient-related variables such as phosphorus.

For Bloom_Index regression (Figs. 11–13), the SHAP analysis showed a different dominance pattern. Surface_Chlorophyll emerged as the most

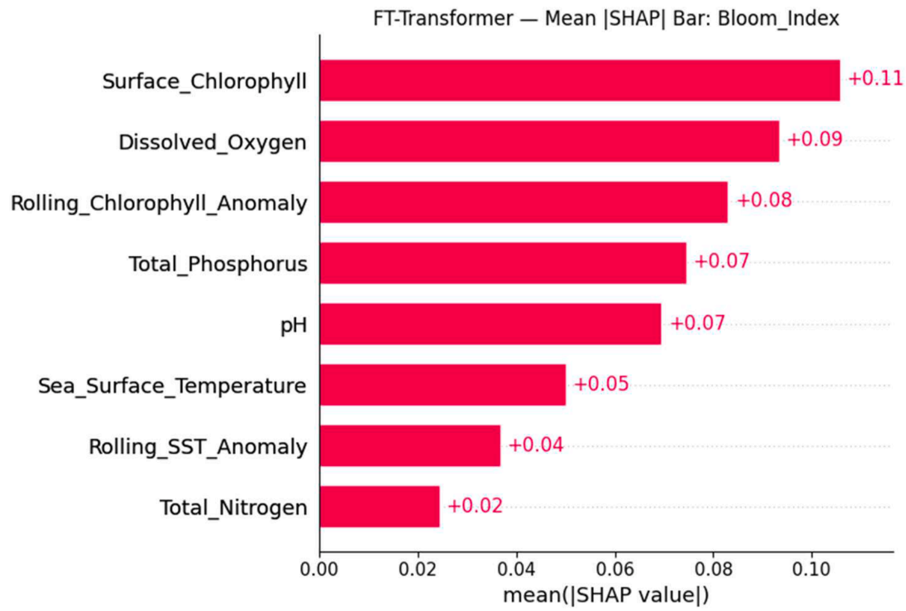

Fig. 12. Mean absolute SHAP value plot showing feature importance for Bloom_Index estimation.

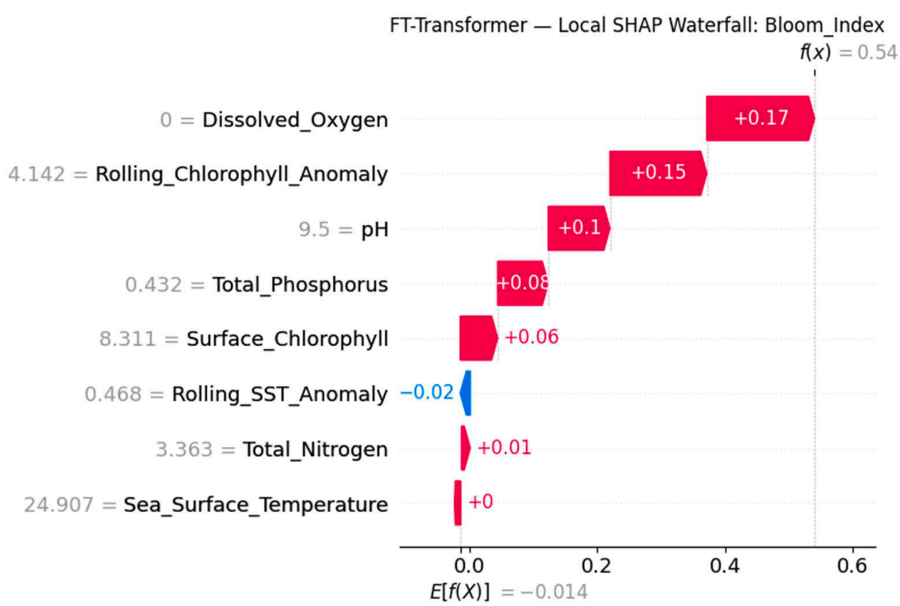

Fig. 13. SHAP waterfall plot illustrating feature contributions for a representative high Bloom_Index sample.

important predictor of bloom intensity, followed by Dissolved_Oxygen and Rolling_Chlorophyll_Anomaly. This indicates that while oxygen depletion plays a dominant role in determining bloom occurrence, chlorophyll concentration more directly influences bloom magnitude.

The local SHAP waterfall plots further illustrated the model's decision-making mechanism. For representative samples, the combined effects of reduced dissolved oxygen, elevated chlorophyll-related variables, and moderate nutrient concentrations contributed positively to both HAB occurrence and Bloom_Index estimation, whereas temperature-related variables exhibited comparatively smaller or mixed effects.

Overall, the SHAP analysis demonstrates that the model captures ecologically meaningful relationships, where oxygen depletion, phytoplankton biomass indicators, and nutrient dynamics jointly govern bloom formation and intensity. These findings are consistent with the correlation analysis and support the biological plausibility of the learned feature contributions.

4. Discussion

The clear formation of a *bloom–warming–nitrogen* module in the correlation structure underscores the tight coupling among thermal dynamics, nutrient enrichment, and phytoplankton proliferation. Similar interactive patterns have been observed in coastal and estuarine ecosystems worldwide, where elevated SST and nitrogen availability stimulate algal growth and favor HABs (Lan et al., 2024; Wei et al., 2023). The observed positive linkage between HAB occurrence, Bloom Index, and SST aligns with evidence that warming enhances stratification and nutrient recycling, creating favorable photic conditions for bloom development (Paerl, 2023; Yan et al., 2024). Conversely, the strong negative associations between DO and bloom metrics confirm the well-documented oxygen depletion during bloom decay phases (Testa et al., 2024; Spilling et al., 2014), further reinforcing eutrophication–hypoxia feedbacks.

Interestingly, phosphorus exhibited inverse correlations with bloom indicators, suggesting nutrient drawdown during bloom progression rather than phosphorus limitation per se. This contrasts with classical eutrophication models emphasizing phosphorus control (Maberly et al., 2020) but agrees with recent findings that nitrogen often dominates bloom dynamics in tropical and subtropical systems (Li et al., 2022). The weak relationship of pH and rolling SST anomaly indicates secondary roles of short-term fluctuations compared to cumulative thermal and nutrient effects. Overall, these findings confirm that HABs in this dataset are driven by co-occurring warming and nitrogen enrichment under oxygen-depleted regimes.

Among deep learning architectures, FT-Transformer achieved the best classification performance, closely followed by NODE and SAINT. The superior AUROC and AUPRC values indicate excellent discrimination between bloom and non-bloom states. Comparable high accuracies ($>$0.99) have rarely been reported in ecological prediction tasks, where data heterogeneity and spatiotemporal noise typically constrain performance (Amato et al., 2020). The success of transformer-based architectures suggests their capacity to capture complex, nonlinear feature dependencies and interactions across mixed-type environmental data. Similar outcomes were reported by Liu et al. (2025), who used transformer models for cyanobacterial bloom prediction in eutrophic lakes, achieving $>$0.98 accuracy, corroborating the present results.

For regression, NODE outperformed other models in predicting continuous Bloom Index, implying its superior ability to model hierarchical feature interactions without extensive tuning. While FT-Transformer and SAINT performed reasonably well, their higher RMSE values highlight sensitivity to small-sample variation or hyperparameter configuration. Previous studies have noted NODE's robustness for tabular ecological datasets due to its gradient-boosting-like optimization (Özdemir et al., 2023; Çınar et al., 2025). Thus, a hybrid ensemble combining NODE's regression precision with FT-Transformer's classification strength could further improve HAB prediction frameworks.

The ablation results further suggest that Bloom_Index, while strongly correlated with bloom occurrence, is not the sole driver of model performance, indicating that the models have learned meaningful ecological relationships beyond this bloom-proximate index. A notable finding from SHAP analysis is that dissolved oxygen primarily governs bloom occurrence, whereas chlorophyll-based indicators dominate bloom intensity estimation, highlighting distinct ecological controls across classification and regression tasks.

The integration of interpretable environmental indicators such as nitrogen enrichment, SST anomalies, and chlorophyll metrics with data-driven deep learning frameworks provides a scalable pathway for early warning systems. These findings can support near-real-time HAB risk assessment in coastal monitoring programs, particularly under climate-driven shifts in temperature and nutrient cycling (Hridoy et al., 2025b, 2025c, 2026c). Policymakers and environmental agencies could employ such models for adaptive management e.g., nutrient load regulation, aquaculture zoning, or targeted sampling during predicted high-risk periods. Furthermore, the inverse linkage between DO and bloom indicators reinforces the need to integrate oxygen monitoring into HAB surveillance systems to anticipate hypoxia-related fish kills and ecosystem stress.

Despite robust model performance, several limitations must be acknowledged. First, the dataset's spatial and temporal scope may restrict generalizability; environmental gradients beyond the sampled domain could produce different correlation structures. Second, the high classification accuracy may partly reflect limited class variability or potential overlap between predictor and response variables (e.g., chlorophyll indices embedded in Bloom Index). Third, deep learning models, while powerful, remain partially opaque in mechanistic interpretation. Model interpretability was strengthened in the revised analysis through SHAP-based feature attribution; however, additional explainability approaches such as integrated gradients and counterfactual analysis may further improve mechanistic understanding in future work. An additional limitation is that the public repository documentation does not fully describe the original seed-data annotation workflow and source metadata, which constrains independent verification of the pre-augmentation label distribution and full observational context.

Ensuring reproducibility was a central consideration in this study, particularly given the use of advanced deep tabular learning models and GAN-based data augmentation. All components of the workflow, including data preprocessing, synthetic data generation, model training, and evaluation, were designed to be fully transparent and modular, enabling independent replication (Hridoy et al., 2025d, 2026d; Alpsalaz et al., 2026). GAN-generated synthetic data were created from ecologically realistic distributions of key marine parameters (chlorophyll, SST anomalies, dissolved oxygen, TN, TP) to augment the seed dataset. This approach addresses class imbalance and the scarcity of rare bloom events while improving model generalization across heterogeneous marine conditions.

Machine learning models, including SAINT, NODE, FT-Transformer, and MLPs, were trained with documented hyperparameters and evaluated using a stratified train–test split together with stratified K-fold cross-validation to provide robust performance estimates on unseen samples. Fixed random seeds further supported reproducibility across repeated experimental runs. By providing fully documented, script-based workflows, this study enables other researchers and institutions to replicate the modeling pipeline, retrain models with local datasets (Salaeh et al., 2026), and evaluate forecasting performance in new regions (Pinthong et al., 2026). Beyond reproducibility, this framework supports operational applications: it facilitates the generation of synthetic scenarios, enhances early-warning capabilities for HAB events, and strengthens data-driven decision-making for fisheries management, ecosystem monitoring, and climate impact assessments.

Future research should extend these findings through spatiotemporal generalization and causal modeling. Incorporating satellite-derived dynamic variables (e.g., turbidity, mixed layer depth) and watershed nutrient flux data could enhance mechanistic understanding. Model explainability can be advanced using interpretable AI approaches to link predictor importance directly to ecological mechanisms. Transfer learning could enable application to other coastal systems with limited data. Finally, scenario-based modeling that combines climate projections with nutrient management simulations would inform policy strategies under future warming and eutrophication trajectories.

5. Conclusion

This study developed a dual-task deep learning framework for HAB occurrence classification and Bloom_Index regression by integrating ecological indicators with advanced tabular architectures, including FT-Transformer, NODE, and SAINT. Beyond benchmarking predictive performance, the revised framework contributes a dataset-composition audit distinguishing observed and GAN-generated synthetic records, statistical and visual validation of synthetic-data fidelity, an explicit leakage-aware ablation analysis excluding Bloom_Index, and SHAP-based interpretation of model behavior.

The results identified a bloom–warming–nitrogen module linking sea surface temperature, nitrogen enrichment, and chlorophyll anomalies with increased bloom activity, contrasted by an oxygen–phosphorus module inversely related to bloom intensity. FT-Transformer achieved the strongest classification performance, whereas NODE provided the most accurate Bloom_Index regression. Ablation analysis demonstrated that predictive performance remained robust even after excluding Bloom_Index, indicating that the models captured substantial independent information from the remaining ecological variables. SHAP analysis further showed that dissolved oxygen primarily governed bloom occurrence, whereas chlorophyll-based indicators dominated bloom-intensity estimation.

Overall, the findings demonstrate that coupling advanced deep tabular learning with synthetic-data validation, leakage-aware evaluation, and interpretable ecological attribution provides a robust framework for HAB prediction and early warning. Future work should extend this approach using broader independently documented datasets, additional environmental drivers, and spatiotemporal forecasting architectures to improve generalizability and operational deployment.

Code availability for reproducibility

https://github.com/MHridoy-123/Harmful-Algal-Bloom-Deep-learning-model_8032

Ethics and consent to participate declarations

Not applicable.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. Open Access (OA) funding was provided by the University of Applied Sciences Magdeburg-Stendal.

Software usage

While AI tools, including Microsoft Copilot (https://copilot.microsoft.com/), were used to assist with portions of the manuscript language, the scientific content, analysis, interpretation, and final approval of the manuscript were entirely the responsibility of the authors.

CRediT authorship contribution statement

Md. Abdullah Al Mamun Hridoy: Writing – original draft, Visualization, Validation, Software, Methodology, Data curation, Conceptualization. **Petra Schneider:** Writing – review & editing, Supervision, Funding acquisition. **Chiara Bordin:** Writing – review & editing, Supervision, Funding acquisition. **Matteo Bodini:** Writing – review & editing, Validation. **Abdullah Ibna Shawkat:** Investigation, Data curation. **Khairul Nizam Abdul Maulud:** Writing – review & editing, Validation. **Md. Abdullah Al Mamun:** Writing – review & editing, Validation. **Paolo Pastorino:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that there are no conflicts of interest. No funding, commercial, or institutional pressures influenced the design, execution, or reporting of this study. All affiliations listed were included solely for administrative and identification purposes.

Acknowledgment

The authors express sincere gratitude to the Mendeley Data repository and the original dataset contributors for making the HAB-related dataset publicly available, which greatly supported the development of this research. The authors also acknowledge the Google Colab platform and the open-source Python libraries used for data processing, model development, and visualization.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.hazadv.2026.101348.

Data availability

Data will be made available on request.

References

Alpsalaz, F., Özüpak, Y., Aslan, E., Uzel, H., 2026. Hybrid machine learning approach for enhanced fault detection and power estimation in photovoltaic systems. IET Renew. Power Gener. 20 (1), e70153.

Amato, F., Guignard, F., Robert, S., Kanevski, M., 2020. A novel framework for spatio-temporal prediction of environmental data using deep learning. Sci. Rep. 10 (1), 22243. https://doi.org/10.1038/s41598-020-79148-7.

... & Anderson, D.M., Fensin, E., Gobler, C.J., Hoeglund, A.E., Hubbard, K.A., Kulis, D.M., Trainer, V.L., 2021. Marine harmful algal blooms (HABs) in the United States: history, current status and future trends. Harmful Algae 102, 101975. https://doi.org/10.1016/j.hal.2021.101975.

Backer, L.C., Manassaram-Baptiste, D., LePrell, R., Bolton, B., 2015. Cyanobacteria and algae blooms: review of health and environmental data from the harmful algal bloom-related illness surveillance system (HABISS) 2007–2011. Toxins (Basel) 7 (4), 1048–1064.

Çınar, M., Aslan, E., Özüpak, Y., 2025. Comparison and optimization of machine learning methods for fault detection in district heating and cooling systems. Bull. Pol. Acad. Sci. Tech. Sci., e154063 https://doi.org/10.24425/bpasts.2025.154063.

Fan, Y., Waldmann, P., 2024. Tabular deep learning: a comparative study applied to multi-task genome-wide prediction. BMC Bioinform. 25 (1), 322.

Gorishniy, Y., Rubachev, I., Khrulkov, V., Babenko, A., 2021. Revisiting deep learning models for tabular data. Adv. Neural Inf. Process. Syst. 34, 18932–18943.

Hoagland, P., Scatasta, S., 2006. The economic effects of harmful algal blooms. Ecology of Harmful Algae. Springer, Berlin, Heidelberg, pp. 391–402.

... & Hridoy, M.A.A.M., Bodini, M., Mahin, M.R., Schneider, P., Pastorino, P., Bordin, C., Goliatt, L., 2026b. Operational early-warning forecasts of aquaculture water quality: Interpretable ML for TDS under walk-forward validation. Ecol. Inform. 96, 103804. https://doi.org/10.1016/j.ecoinf.2026.103804.

Hridoy, M.A.A.M., Bordin, C., Masood, A., Masood, K., 2025b. Predictive modelling of aquaculture water quality using IoT and advanced machine learning algorithms. Results Chem. 16, 102456. https://doi.org/10.1016/j.rechem.2025.102456.

Hridoy, M.A.A.M., Bordin, C., Pastorino, P., Maulud, K.N.A., Pathan, M.M., Baki, A.O., 2026d. Neural network-based prediction of water contamination dynamics in a rapidly developing tropical region. Case Stud. Chem. Environ. Eng. 13, 101342. https://doi.org/10.1016/j.cscee.2026.101342.

Hridoy, M.A.A.M., Ditthakit, P., Hosen, A., 2025c. Machine learning–driven prediction of heavy metal pollution for aquatic health surveillance and disease control. Machine Learning–Driven Prediction of Heavy Metal Pollution for Aquatic Health Surveillance and Disease Control. ISCIENCE-D-25-19889. https://doi.org/10.2139/ssrn.5930557.

... & Hridoy, M.A.A.M., Pastorino, P., Bordin, C., Bodini, M., Dhar, N., Schneider, P., Maulud, K.N.A, 2026c. Operational agricultural water management in river-based aquaculture: a machine-learning approach to predict dissolved oxygen in the Halda River, Bangladesh. Aquac. Eng. 113, 102693. https://doi.org/10.1016/j.aquaeng.2026.102693.

... & Hridoy, M.A.A.M., Sagor, P.S., Akter, P., Islam, M.H., Bordin, C., Islam, S.S., Ditthakit, P., 2026a. Decision optimization for inorganic contaminants in water systems using ecological informatics for sustainable management. Case Stud. Chem. Environ. Eng. 13, 101362. https://doi.org/10.1016/j.cscee.2026.101362.

Hridoy, M.A.A.M., Shawkat, A.I., Bordin, C., Acharjee, M.R., Masood, A., Baki, A.O., Al Mamun, M.A., 2025a. Advanced machine learning models for accurate water quality classification and WQI prediction: implications for aquatic disease risk management. Sci. Total Environ. 1008, 180965. https://doi.org/10.1016/j.scitotenv.2025.180965.

Hridoy, M. D., Bordin, C., Morshed, M. M., Hosen, M. T., Sagor, M. F. H., Shoeb, S.,... & Pastorino, P. (2025d) Machine-learning assessment of colloidal mechanisms regulating organic carbon and trace metals in the environment. 10.2139/ssrn.5955157.

Hu, C., Lee, Z., Franz, B., 2012. Chlorophyll aalgorithms for oligotrophic oceans: a novel approach based on three-band reflectance difference. J. Geophys. Res.: Oceans (C1), 117.

Lan, J., Liu, P., Hu, X., Zhu, S., 2024. Harmful algal blooms in eutrophic marine environments: causes, monitoring, and treatment. Water (Basel) 16 (17), 2525.

... & Li, C., Zhang, P., Zhu, G., Chen, C., Wang, Y., Zhu, M., Zheng, Q., 2022. Dynamics of nitrogen and phosphorus profile and its driving forces in a subtropical deep reservoir. Environ. Sci. Pollut. Res. 29 (19), 27738–27748.

Liu, Y., Yang, B., Xie, K., Sun, J., Zhu, S., 2025. Dongting Lake algal bloom forecasting: Robustness and accuracy analysis of deep learning models. J. Hazard. Mater. 485, 136804.

Maberly, S.C., Pitt, J.A., Davies, P.S., Carvalho, L., 2020. Nitrogen and phosphorus limitation and the management of small productive lakes. Inland Waters 10 (2), 159–172.

Mohan, D., Dayalani, A., Gupta, L., 2025. Generative Adversial Network (GAN) synthesized dataset on Harmful Algal Bloom (HAB). Mendeley Data V1. https://doi.org/10.17632/c3792mm5f6.1.

Özdemir, H., Baduna Koçyiğit, M., Akay, D., 2023. Flood susceptibility mapping with ensemble machine learning: a case of Eastern Mediterranean basin, Türkiye. Stoch. Environ. Res. Risk Assess. 37 (11), 4273–4290.

Özüpak, Y., Aslan, E., 2027. An enhanced interpretable ML approach for forecasting long-term CO_2 and N_2O emissions via optimized random forest modeling. Fuel 427, 139723. https://doi.org/10.1016/j.fuel.2026.139723.

Paerl, H.W., 2023. Climate change, phytoplankton, and HABs. Climate Change and Estuaries. CRC Press, pp. 315–334.

Paerl, H.W., Barnard, M.A., 2020. Mitigating the global expansion of harmful cyanobacterial blooms: moving targets in a human-and climatically-altered world. Harmful Algae 96, 101845.

Pinthong, S., Salaeh, N., Pham, Q.B., Hridoy, M.A.A.M., Bordin, C., Ditthakit, P., 2026. Coupled spiking neural network–extra trees framework for daily streamflow imputation with ERA5-land integration. Eng. Sci. 41. https://doi.org/10.30919/es2204.

... & Qiu, Y., Huang, J., Luo, J., Xiao, Q., Shen, M., Xiao, P., Duan, H., 2025. Monitoring, simulation and early warning of cyanobacterial harmful algal blooms: an upgraded framework for eutrophic lakes. Environ. Res. 264, 120296. https://doi.org/10.1016/j.envres.2024.120296.

... & Sa'adi, Z., Zainon Noor, Z., Zaidi, N.S., Mohamed, N.H., Abdul Razak, F., Foze, M.F., Kemarau, R.A., 2026. Youth-led spatio-temporal heatwave co-adaptation mapping through data-driven participatory dotmocracy approach. Int. J. Biometeorol. 70 (5), 157. https://doi.org/10.1007/s00484-026-03218-0.

Salaeh, N., Pinthong, S., Pham, Q.B., Hridoy, M.A.A.M., Bordin, C., Ditthakit, P., 2026. Adaptive parameter-switching GR2M model for streamflow simulation in tropical basins of Southern Thailand. Eng. Sci. 41, 2205. https://doi.org/10.30919/es2205.

Somepalli, G., Goldblum, M., Schwarzschild, A., Bruss, C.B., Goldstein, T., 2021. Saint: improved neural networks for tabular data via row attention and contrastive pre-training. ***arXiv preprint***. *arXiv:2106.01342*.

Spilling, K., Kremp, A., Klais, R., Olli, K., Tamminen, T., 2014. Spring bloom community change modifies carbon pathways and C: N: P: Chl a stoichiometry of coastal material fluxes. Biogeosciences 11 (24), 7275–7289.

... & Testa, J.M., Liu, W., Boynton, W.R., Breitburg, D., Friedrichs, C., Li, M., Brady, D.C., 2024. Physical and biological controls on short-term variations in dissolved oxygen in shallow waters of a large temperate estuary. Estuaries Coasts 47 (6), 1456–1474.

Uzel, H., Özüpak, Y., Alpsalaz, F., Aslan, E., Zaitsev, I., 2025. Acoustic-based fault diagnosis of electric motors using Mel spectrograms and convolutional neural networks. Sci. Rep. 16, 3379. https://doi.org/10.1038/s41598-025-33269-z.

... & Villanueva, P., Yang, J., Radmer, L., Liang, X., Leung, T., Ikuma, K., Lee, J., 2023. One-week-ahead prediction of cyanobacterial harmful algal blooms in Iowa lakes. Environ. Sci. Technol. 57 (49), 20636–20646. https://doi.org/10.1021/acs.est.3c07764.

Wei, W., Han, Y., Zhou, Y., 2023. Nutrients and sea surface temperature drive harmful algal blooms in China's coastal waters over the past decades. Environ. Res. Lett. 18 (9), 094068.

Yan, Z., Kamanmalek, S., Alamdari, N., Nikoo, M.R., 2024. Comprehensive insights into harmful algal blooms: a review of chemical, physical, biological, and climatological influencers with predictive modeling approaches. J. Environ. Eng. 150 (4), 03124002.

Zheng, G., Brown, C.W., DiGiacomo, P.M., 2023. Retrieval of oceanic chlorophyll concentration from GOES-R advanced baseline imager using deep learning. Remote Sens. Environ. 295, 113660.

Zhou, T., Li, Y., Jiang, B., Alatalo, J.M., Li, C., Ni, C., 2023. Tracking spatio-temporal dynamics of harmful algal blooms using long-term MODIS observations of Chaohu Lake in China from 2000 to 2021. Ecol. Indic. 146, 109842. https://doi.org/10.1016/j.ecolind.2022.109842.