

Leveraging Artificial Intelligence for Cyanobacterial Bloom Prediction: A Hybrid Deep Learning and Generative Adversarial Network Framework for Accurate Forecasting and Proactive Management

Nadjette Dendani¹, Amel Saoudi², Nour Djihane Amara³, Nabiha Azizi⁴,
Julie Dugdale⁵, Rayenne Hadiby⁶

Abstract. This study presents an Artificial Intelligence-based system designed to predict cyanobacterial harmful algal blooms (CyanoHABs). The system utilizes Long Short-Term Memory (LSTM) networks to predict the timing of bloom occurrences and One-Dimensional Convolutional Neural Networks (1D-CNNs) to estimate cyanobacterial density. Additionally, Generative Adversarial Networks (GANs) are employed for data augmentation to enrich the database. The system's performance was validated using the Algerian Mexa database, achieving an R-squared (R^2) value of 98% and a root mean square error (RMSE) of 9% for cyanobacterial density prediction, and an R-squared value of 88% with a root mean square error of 31% for bloom timing prediction. These results highlight the system's robust predictive capabilities, enabling proactive monitoring and management of CyanoHABs to mitigate their adverse impacts on health and the environment.

Keywords: CyanoHABs, Cyanotoxins, Mexa Dam, 1D-CNN, GANs, LSTM, R^2 , RMSE

1. Introduction

Water is essential for all known forms of life, serving as a vital component for biological processes and a critical resource for ecosystems. Cyanobacteria are microscopic organisms commonly found in water bodies throughout the world, have become increasingly prevalent due to human activities, such as nutrient pollution and climate change, accelerating their

^{1,3,4} LabGed Laboratory, Computer Science Department, Badji Mokhtar University, Annaba, Algeria, e-mail: nadjet.dendani@univ-annaba.dz, nourjiheneamara@gmail.com, nabiha111@yahoo.fr

² Ecobiology Laboratory for Marine Environments and Coastal Areas, Biochemistry Department, Badji Mokhtar University, Annaba, Algeria, e-mail: amelsaoudi@yahoo.fr

⁵ Grenoble Informatics Laboratory, Grenoble Alps University, France, e-mail: julie.dugdale@univ-grenoble-alpes.fr

⁶ Public Health, Paris Saclay University, e-mail: rayene.hadiby@univsite-paris-saclay.fr

growth and spread [20]. These organisms produce cyanotoxins, harmful molecules contained within their cells, which are released into the water when cells die or rupture. Cyanotoxins can present serious health hazards to humans, animals, and aquatic ecosystems, leading to concerns about water quality and public health [28]. Different types of cyanotoxins exist, including Hepatotoxins, which damage liver cells; Neurotoxins, affecting the nervous system; and Dermatotoxins, causing skin irritation and allergic reactions upon contact [22]. Moreover, these proliferations pose a significant economic challenge for ecosystem managers, particularly in drinking water supply systems, which require both monitoring networks and appropriate treatment measures. These results in significant additional costs due to the large quantities of organic carbon generated during blooms which can also cause filters to clog.

Recently, AI has emerged as a promising solution to address the CyanoHAB problem and mitigate associated risks. By leveraging advanced machine learning (ML) and deep learning (DL) models, it is possible to achieve more accurate predictions and proactive management of these environmental threats. These models excel at handling complex, non-linear relationships within data, making them particularly suited for analyzing and forecasting CyanoHABs. Furthermore, ML and DL models have proven their value in decision-making across various fields, including economics, finance, social sciences, engineering, health sciences, and beyond [2]. Their application in the context of cyanobacteria and bloom events can lead to enhanced monitoring, early warning systems, and targeted interventions, ultimately contributing to better environmental management and public health outcomes.

Several ML and DL methods have been investigated in this domain, Mellios et al. (2020) [19] employed Decision Tree, K-Nearest Neighbors (KNN), Support-vector Machine (SVM) and Random Forest (RF) for Predicting Health Risk of Cyanobacterial Blooms in Northern European Lakes; after tuning and optimizing RF parameters, it achieved an impressive accuracy of 95.81%. S.Lee and D.Lee (2018) [18] utilized three DL models, including a multilayer Perceptron (MLP), a recurrent neural network (RNN) and a long short-term memory (LSTM) to predict chlorophyll-a concentrations. Their findings demonstrated that the LSTM model exhibited the highest prediction rate for harmful algal blooms (HABs).

Despite the promising results provided by existing studies in this area; they have not addressed and forecasted cyanobacterial density and timing bloom. These outputs are very important factors to solving CyanoHABs problem.

The primary contribution of our study is to address an environmental challenge by presenting a robust monitoring and forecasting system with the following key features:

- 1.Enhanced Data Diversity: Integrating an advanced data generation approach using Generative Adversarial Networks (GANs) and upsampling algorithms to ensure a varied and rich dataset.

- 2.Cyanobacterial Density Prediction: Exploring the application of a 1dimensional Convolutional Neural Networks (1D-CNNs) model to accurately predict cyanobacterial density.

- 3.Bloom Timing Forecast: Investigating the use of an adapted Long Short-Term Memory (LSTM) model for forecasting bloom timing, tailored to the size of the dataset.

This paper is structured as follows: an overview of the related work presented in Section 2. Section 3 provides a comprehensive background of the materials and methodologies utilized in conducting this work. Results and interpretations are described in Section 4. Finally, we concluded in section 5.

2. Related Work

The studies on cyanobacteria in eastern Algeria since the 1990s have revealed their presence and potential toxicity in various water bodies. The research conducted by Saoudi et al. (2015) [23] focuses on Mexa dam, identifying seven cyanobacteria genera, five of which are potentially toxic, and highlighting correlations between cyanobacteria density, chlorophyll-a and microcystins. The study emphasizes the impact of weather conditions on water parameters, suggesting increased blooms with global warming, posing risks to both environment and health.

A variety of studies have explored the application of ML and DL models in predicting CyanoHABs across different ecosystems and contexts. Mellios et al. (2020) [19] employed Decision Tree, K-Nearest Neighbors (KNN), Support-vector Machine (SVM) and Random Forest (RF), notably, after tuning and optimizing Random Forest parameters, it achieved an impressive accuracy of 95.81%. Similarly, Millie et al. (2014) [7] shed light on CyanoHABs dynamics in Lake Erie utilizing artificial neural networks (ANNs), their study demonstrated the predictive capabilities of ANNs for phytoplankton and Microcystis biomass, with accuracy values ranging from 0.87 to 0.97 for chlorophyll(a) concentrations and 0.71 to 0.90 for Microcystis biovolumes. Additionally, Tran and Hoang (2016) [26] utilized ANNs for forecasting the colonization growth of algae on building façades. The cross-validation process revealed that the ANN achieved a prediction performance with an R^2 of 0.91 and RMSE of 5.69. Park et al. (2021) [21] evaluated ANN and SVM as an early-warning system for cyanobacterial bloom outbreaks, identifying 6- and 7-day intervals as effective early-warning periods. Heggerud et al. (2024) [6] developed a Bayesian network that predicts future 2-week Cyanobacterial Blooms with an area under the curve (AUC) of 0.83. Additionally, Ahn et al. (2023) [13] integrated CNN and Transformer algorithms, comparing the results with those obtained using the Temporal Fusion Transformer (TFT) technique. Notably, the CNN-Transformer model exhibited superior performance with an R^2 value of 0.9, an MAE of 0.2, and an RMSE of 0.4. In their study, Cao et al. (2022) [10] developed a CNN-LSTM integrated model to predict CyanoHABs in Taihu Lake using 20 years of data obtained from MODIS images, The results demonstrated an R^2 value of 0.91 and a low mean relative error of 17.42%. Begliomini et al. (2023) [8] evaluated PRISMA hyperspectral data for estimating C-Phycocyanin concentrations in the Billings reservoir, Brazil, using three ML algorithms: RF, Extreme Gradient Boost (XGBoost), and SVM.

Furthermore, several studies have employed satellite imagery and remote sensing methodologies to forecast harmful algal blooms (HABs) in aquatic environments. For instance, Niu et al. (2024) [12] applied ML and DL techniques to forecast *Aureococcus anophagefferens* population density, achieving high accuracy with all models (RF, Support Vector Regression (SVR), Multilayer Perceptron (MLP), and CNN) having R^2 values exceeding 0.75, with RF surpassing 0.8. Xu et al. (2024) [27] used (Multi-Linear Regression) MLR, SVR, RF, and XGBoost to predict phytoplankton biomass. In addition, Kim et al. (2023) [11] utilized CNNs to model CyanoHABs of *Microcystis* sp. in South Korean rivers, achieving an accuracy and F1-score of 95.01% and 93.30%, respectively, for the testing dataset, with predictions exhibiting a mean R^2 of 0.55 for *Microcystis* abundances. Lee and Lee (2018) [18] utilized three DL models, including a MLP, a recurrent neural network

(RNN), and a long short-term memory (LSTM), to predict chlorophyll-a concentrations. Their findings demonstrated that the LSTM model exhibited the highest prediction rate for HABs. According to a study performed by Busari et al. (2024) [5], different sampling periods were employed to develop several LSTM models, which were augmented with a Monte Carlo dropout scheme for forecasting chlorophyll-a concentrations. The findings indicated that these models predicted chlorophyll-a concentrations 3 days and 7 days in advance using both daily and hourly water quality datasets. Lastly, Ai et al. (2023) [9] utilized a combination of riverine and meteorological features to predict HABs in Lake Erie. RF classification models achieved accuracies of up to 89.6%, while regression models achieved an R2 value of 0.69, and LSTM models predicted short-term trends with efficiencies ranging from 0.12 to 0.97.

While previous studies have contributed to the understanding of CyanoHABs, our research focuses on forecasting bloom timing and cyanobacterial density in Mexa Dam using advanced DL techniques such as 1D-CNN, GAN, and LSTM. Our findings highlight the exceptional predictive capabilities of these techniques.

3. Materials and Methods

In this section, we provide a comprehensive overview of the materials and methodologies utilized in conducting this research. The general architecture for our proposed intelligent CyanoHABs management system can be summarized in figure 1:

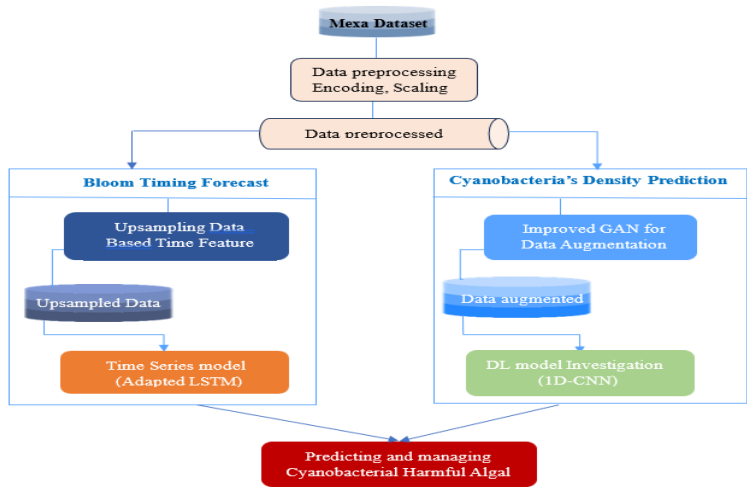


Figure 1. System Architecture.

Figure 1 illustrates a workflow for predicting and managing CyanoHABs. The process begins with preprocessing the Mexa dataset then diverges into two distinct prediction tasks. The first branch focuses on Bloom Timing Forecast. To address potential data imbalances, the data is upsampled based on time features. Subsequently, an adapted Long Short-Term Memory (LSTM) model is employed to forecast the timing of blooms.

The second branch tackles Cyanobacteria's Density Prediction. An improved Generative Adversarial Network (GAN) is utilized to augment the dataset, potentially enhancing model performance. A one-dimensional Convolutional Neural Network (1D-CNN) is then investigated for predicting cyanobacteria density. Both branches ultimately contribute to the overarching goal of predicting and managing CyanoHABs.

3.1. Site Description

The Mexa dam is situated in the commune of Bougous wilaya of El Tarf at a location known as the Gorge de Mexenne ($36^{\circ}45'14.31''\text{N}$ - $8^{\circ}23'33.68''\text{E}$). Figure 2 illustrates this location.

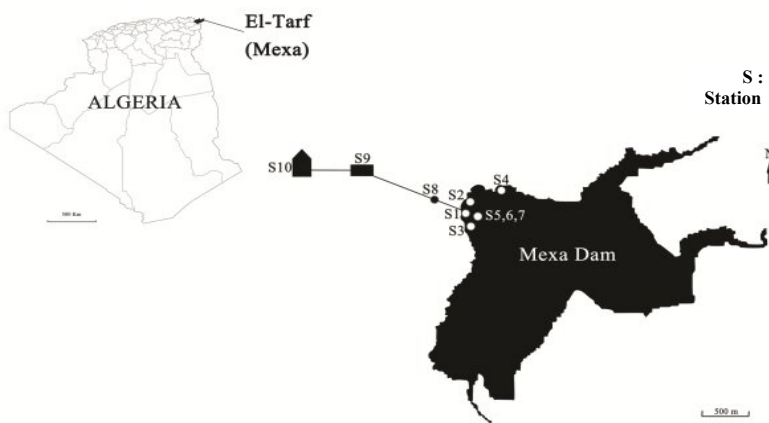


Figure 2. Mexa Dam Location

3.2. Data Description

Data was collected from ten stations, with samples being gathered monthly at each station over the 24 months from January to December during the years 2010 and 2011 [22], resulting in a total of 240 samples. The dataset comprises a total of 17 features, including the target variable (Cyanobacterial's density).

While the data used in this study was collected in 2010-2011, its empirical nature and specific collection context make it scientifically valuable. Sourced from in-situ sampling of the Mexa reservoir, the data offers high-precision, methodologically rigorous measurements that are difficult to reproduce. Such data is scarce and absent from public databases. Despite its age, it remains highly representative of the cyanobacteria dynamics in this ecosystem due to the minimal changes in the reservoir's environmental and hydrological conditions since data collection.

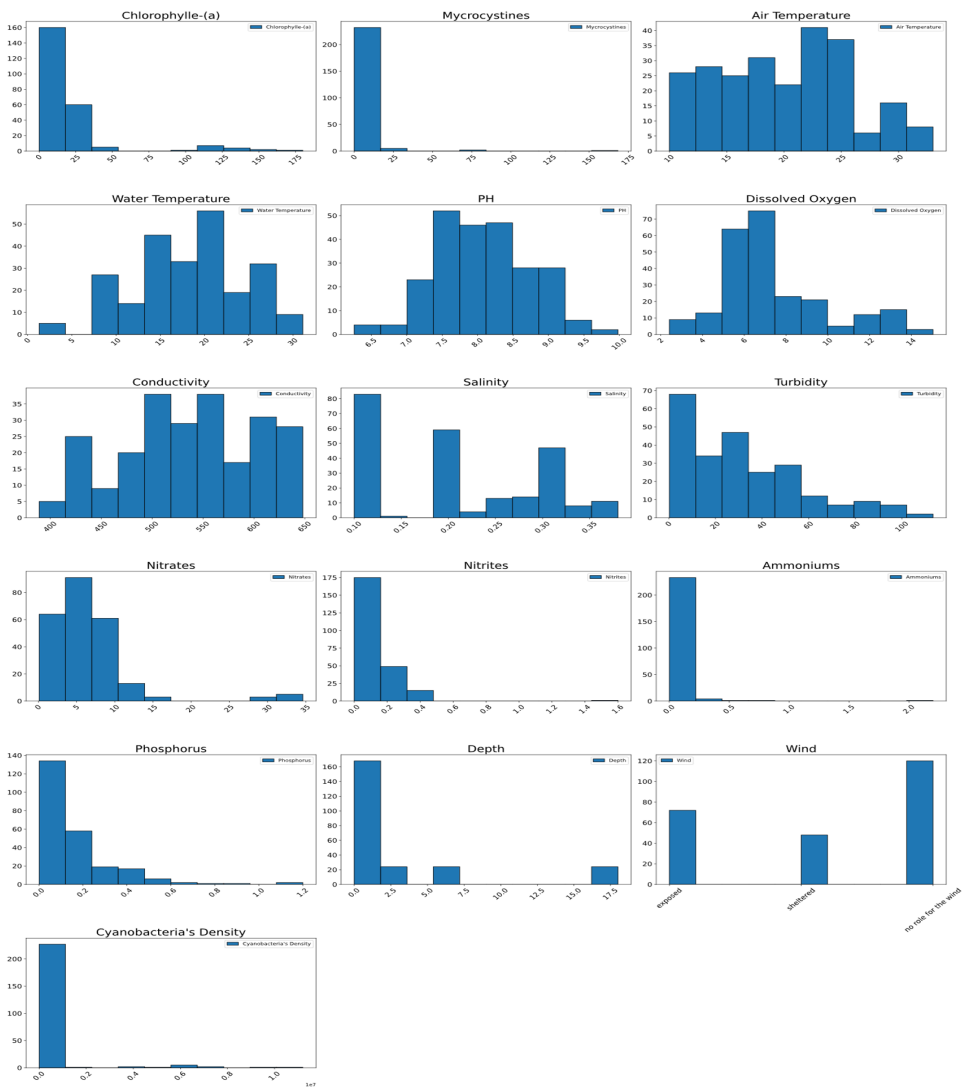


Figure 3. Data distribution of Mexa Dataset.

The Figure 3 below presents the data distribution of all features. In this dataset, there are five categories of features. These features encompass biotic, abiotic, Nutritional, physical, and meteorological parameters [23].

- Biotic: Cyanobacteria's density (Cell/ml), Chlorophyll (a) (µg/l), and Microcystins (µg/l).
- Abiotic: Water temperature (°C), pH, Dissolved oxygen (mg/l), Conductivity (µs/cm), Salinity (g/l), and Turbidity (NTU).
- Nutritional: Nitrates (mg/l), Nitrites (mg/l), Ammoniums (mg/l), and Phosphorus (mg/l).
- Physical: Depth (m), and Station's location.

- Meteorological: Air temperature (°C).

Right-skewed distributions are common, indicating that for most parameters, low values are more frequent with a few higher values. In contrast, normal distributions for parameters like PH and water temperature suggest a balanced central tendency. Additionally, bimodal and uniform distributions for parameters like salinity and conductivity indicate distinct groups or a wide spread of values. These distributions provide insights into the environmental and water quality parameters, useful for further analysis and decision-making in environmental studies.

Furthermore, preprocessing tasks are crucial for ensuring the quality and reliability of subsequent analyses. Handling missing values, scaling features, and transforming skewed distributions are essential steps to prepare the data for machine learning algorithms and statistical modeling.

3.3. Statistical Methods

Statistical methods are fundamental tools for extracting meaningful insights from data, playing a crucial role in various stages of data analysis. These methods encompass a range of techniques designed to prepare, analyze, and model data effectively.

3.3.1. Data Preprocessing

Several transformations and data preprocessing stages were applied to improve the quality and to improve the final decision of prediction [16]. The main adopted approaches are:

a) Data Encoding

Encoding is the process of converting categorical variables into a numerical form such as the 'Station's location'. It encompasses various techniques such as one-hot encoding, label encoding, and ordinal encoding

b) Data Scaling (Standardization)

Standardizing numerical data is applied to optimize performance of AI models, ensuring consistent scaling and preventing any single feature from dominating the model training process. Standardization involves transforming data to have a mean of 0 and a standard deviation of 1.

$$Z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Where:

Z_i : standardized value, x_i : data value, μ : mean, σ : standard deviation

c) Resampling Technique

Resampling is a statistical technique that involves the process of changing the frequency of a time series observation. There are two primary types of resampling [29]: upsampling and downsampling.

- Upsampling: Increasing the frequency of the data, such as converting daily data into hourly data, this is done by inserting new samples between existing ones.
- Downsampling: Decreasing the frequency of the data, such as converting hourly data into daily data.

In this study, we applied the time series models to both the original and upsampled data using the `resample` and `interpolate` functions in `pandas` taking Resampling Frequency: 'D' (Daily) resulting in 706 data points and Interpolation Method: 'linear' (to perform linear interpolation).

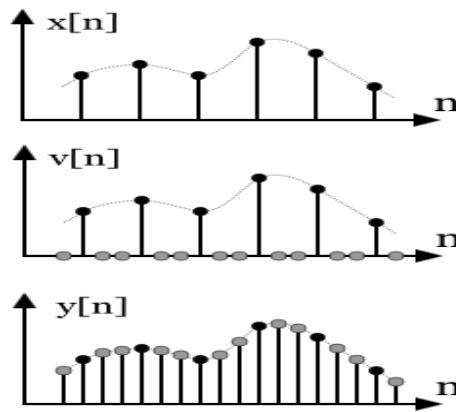


Figure 4. Upsampling and interpolating of original signal [24].

Figure 4 depicts the process of **upsampling and interpolation** of a discrete-time through three graphs labeled $x[n]$, $v[n]$, and $y[n]$.

- Original signal ($x[n]$): This is the initial signal, represented by a series of black dots connected by a smooth curve.
- Upsampling ($v[n]$): In this step, the original signal is upsampled by inserting zeros between the original samples. This effectively increases the sampling rate. The gray dots represent the inserted zeros.
- Interpolation ($y[n]$): Finally, a linear interpolation filter is applied to the upsampled signal. This filter smooths out the discontinuities caused by the inserted zeros and reconstructs a continuous signal. The resulting interpolated signal is represented by the gray dots and the curve connecting them.

3.3.2. Machine learning Models

To predict CyanoHAB occurrences, four distinct machine learning models were employed, each selected for their unique strengths in analyzing complex datasets [3]. Multiple Linear Regression (MLR) was utilized to establish relationships between water quality parameters and CyanoHAB events, offering valuable insights into environmental influences. Support Vector Regression (SVR) was enhanced through Grid Search Cross-Validation (GridSearchCV), a method that systematically explores hyperparameter combinations to optimize predictive accuracy including kernel type (linear, polynomial, RBF), regularization parameter C , kernel coefficient γ , and ϵ -insensitive tube size. The optimal configuration was selected based on 5-fold cross-validation to ensure robust performance. While SVR was our primary focus for hyperparameter tuning due to its sensitivity to parameter settings. Additionally, two tree-based approaches were implemented: Decision Tree Regression (DTR), which identifies threshold-based decision rules for prediction, and Random Forest Regression (RFR), which leverages the collective predictive capability of multiple decision trees to enhance model robustness and accuracy. To balance computational efficiency with performance, these models were implemented using default parameters as defined in the scikit-learn library. This combination of linear, kernel-based, and ensemble methods provided a robust framework for capturing diverse patterns and relationships within the dataset, enabling a comprehensive comparative analysis of CyanoHAB prediction approaches.

3.3.3. Deep Learning Models

Deep learning models, a subset of machine learning that leverages multi-layered neural networks, were employed in this study to explore complex patterns and relationships within the data. We investigated three distinct architectures: 1D Convolutional Neural Networks (1D-CNN) for feature extraction, Generative Adversarial Networks (GAN) for data augmentation, and Long Short-Term Memory (LSTM) networks for time series analysis. These models were chosen for their ability to capture intricate features, address data limitations, and effectively model temporal dependencies, respectively.

a) 1D Convolutional Neural Networks (1D-CNN)

1D-CNNs were employed to predict cyanobacterial density because of their proficiency in extracting local patterns and features from sequential data. Unlike traditional CNNs used for spatial data, 1D-CNNs are tailored for time-series analysis, making them suitable for identifying subtle changes in water quality parameters that precede bloom events. Their ability to handle variable-length input sequences and capture spatial-temporal dependencies was a critical factor in their selection.

we adopted an architecture inspired by successful models in the literature [17] and guided by our experimental trials. This architecture comprises two convolutional layers with 32 and 64 filters, respectively. This gradual increase in filter numbers allows the network to learn increasingly complex features at each layer. Each convolutional layer is followed by a max-

pooling layer (size=2) to downsample the data, reducing dimensionality and computational cost while preserving important features. The ReLU activation function was chosen for its computational efficiency and ability to introduce non-linearity into the network. The model is optimized using the Adam optimizer, which adapts learning rates for efficient convergence. The output of the convolutional layers is flattened and fed into two dense layers with 128 and 1 units, respectively. The 128-neuron hidden layer provides sufficient capacity for the network to learn complex relationships, while the final layer outputs a single value, suitable for the regression task. Figure 5 illustrates the 1D CNN architecture.

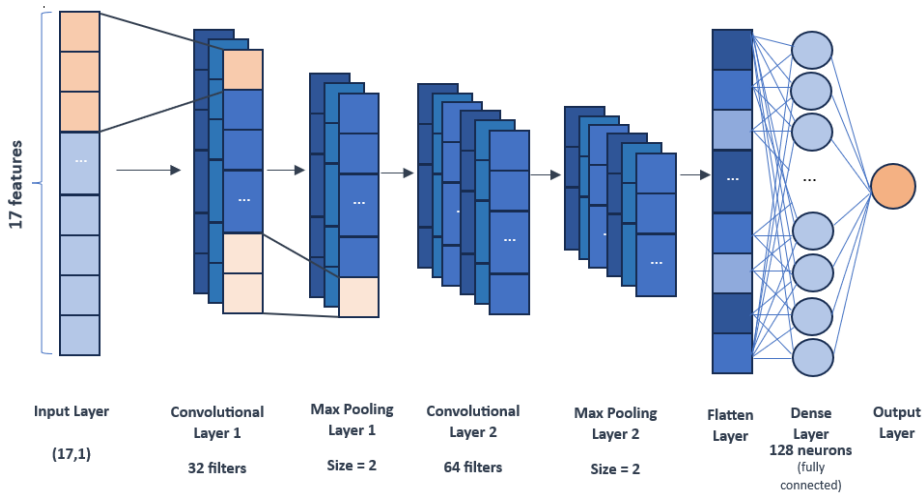


Figure 5. 1D-CNN Architecture.

b) Generative Adversarial Networks (GAN)

In this study, Generative Adversarial Networks (GANs) were employed to address data scarcity and imbalance in CyanoHAB events, which are infrequent compared to non-bloom conditions. The GAN model doubled the dataset by adding 240 synthetic samples, preserving the statistical properties of the original data to enhance model robustness and generalization. Additionally, GANs facilitated temporal data alignment, ensuring accurate representation of the temporal dynamics critical for predicting bloom timing and intensity. The GAN architecture, illustrated in Figure 6, consisted of a generator and a discriminator:

- The generator network included three dense layers: the first with 15 units and ReLU activation, the second with 30 units and ReLU activation, and the final layer with 18 units (matching the input data dimension) with a linear activation function.
- The discriminator network featured two dense layers with 25 and 50 units, respectively, both with ReLU activation, followed by a final layer with a sigmoid activation function to classify samples as real or fake [4].

Following data augmentation, a 1D-Convolutional Neural Network (1D-CNN) was applied to leverage the enriched dataset for improved predictive performance

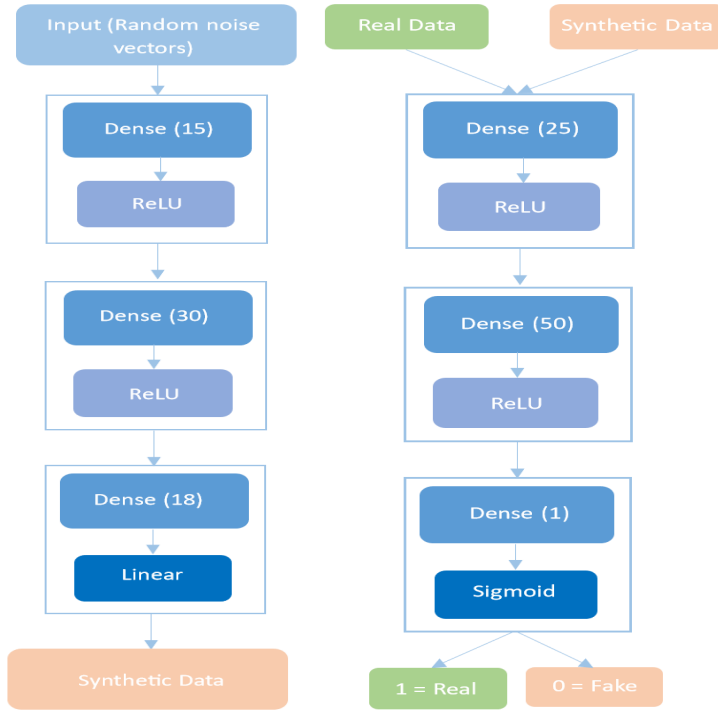


Figure 6. GAN architecture diagram with generator (left) and discriminator (right) network.

To improve the robustness of generative adversarial network GAN, several loss functions were analysed. The binary cross-entropy loss and the Adam optimizer were selected and are defined as follow:

$$BinaryCrossEntropy = -\frac{1}{N} \sum_{i=1}^N [y_i * \log(y_{pred}) + (1 - y_i) * \log(1 - y_{pred})] \quad (2)$$

y_i denotes the actual binary label (either 0 or 1), and y_{pred} represents the predicted probability of the positive class. The loss function consists of two terms: the first term, $y_i * \log(y_{pred})$, quantifies the loss when the true label is 1; the second term, $(1 - y_i) * \log(1 - y_{pred})$, quantifies the loss when the true label is 0. The logarithm function is employed to impose a higher penalty on larger errors, as the loss increases more sharply when the predicted probability diverges from the true label.

c) Long Short-Term Memory Networks (LSTM)

LSTM networks were chosen for predicting the timing of bloom occurrences due to their ability to capture long-term dependencies in time-series data. CyanoHABs are influenced by complex, temporally-dependent environmental factors such as temperature fluctuations, nutrient levels, and sunlight exposure. LSTMs are well-suited for such tasks as they can learn and remember patterns over extended periods, making them ideal for forecasting the onset of blooms. Two different architectures were employed (figures 7 and 8), on both original and upsampled data. The LSTM architecture (figure 7) consisted of two LSTM layers with 50 units each, followed by a dense layer with a single unit. The model was compiled with the Adam optimizer and a mean squared error loss function, with a batch size of 32 and trained for 86 epochs. While the improved architecture (figure 8) consisted of three LSTM layers, each with 50 units. Dropout regularization with a rate of 0.2 was applied after each LSTM layer to mitigate overfitting. The output layer was a dense layer with a single unit. The model was compiled with the Adam optimizer using mean squared error (MSE) loss function where a lower MSE indicates a smaller error and a more accurate estimator. The MSE is defined as follows:

$$MSE = \frac{\sum (y_i - y_i')^2}{n}$$

(3)

where y_i represents the actual values, y_i' represents the predicted values, and n is the number of observations.

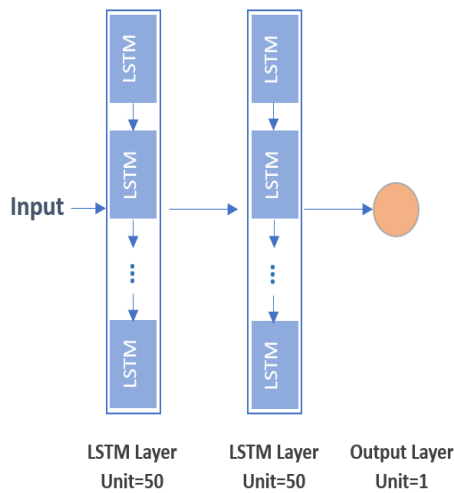


Figure 7. LSTM architecture.

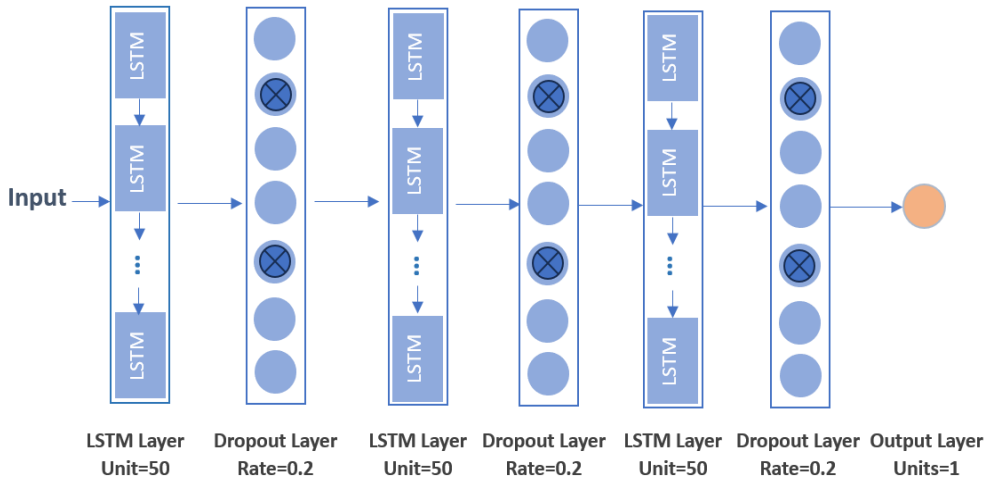


Figure 8. Improved LSTM architecture.

3.3.4. Performance Metrics

Performance metrics are quantitative measures used to evaluate the predictive performance of models, providing comprehensive insights into various aspects of model effectiveness [25], [15], [14]. The primary metrics selected for this study - R^2 , Adjusted R^2 , RMSE, and MAE- were chosen for their interpretability, widespread adoption in environmental monitoring and AI applications, and alignment with our objective: developing an AI-driven intelligent system for the early prediction of CyanoHABs. Adjusted R^2 accounts for model complexity, while RMSE and MAE offer complementary perspectives on error magnitude and distribution. This combination ensures a balanced assessment of both explanatory power and predictive accuracy, which is critical for developing a proactive decision-support tool. Their formulas are given as follows:

$$R^2 = 1 - \frac{\sum (y_i - y_i')^2}{\sum (y_i - \bar{y})^2} \quad (4)$$

$$RMSE = \sqrt{\frac{\sum (y_i - y_i')^2}{n}} \quad (5)$$

$$MAE = \frac{\sum |y_i - y_i'|}{n} \quad (6)$$

$$Adj. R^2 = 1 - \frac{(N-1)}{(N-P-1)} (1 - R^2)$$

(7)

Where:
yi represents the actual values of the dependent variable, and yi' denotes the predicted values.
n or N is the number of samples, and P represents the number of predictors.

4. Results and Interpretation

This section investigates the performance of various Machine Learning (ML) and Deep Learning (DL) models in predicting Cyanobacterial’s density [1] and timing bloom. We begin by evaluating the models with the original dataset. Subsequently, we assess the impact of data augmentation using GAN on model performance. Finally, we compare the performance of Long Short-Term Memory (LSTM) networks with and without data upsampling.

4.1. ML and DL Models with Original Data

The results of the ML and DL Models before data augmentation using GAN are summarized in the table below:

Table 1. Performance summary of ML and DL models with original data.

Model	R ²	MAE	RMSE	Adj. R ²
MLR	0.86	0.23	0.35	0.78
SVR	0.85	0.15	0.35	0.77
Optimized SVR	0.93	0.11	0.18	0.90
DTR	0.96	0.05	0.19	0.77
RFR	0.93	0.05	0.24	0.89
1D-CNN	0.96	0.09	0.14	0.93

Figure 9 provides a graphical representation of the performance metrics for various Machine Learning (ML) and Deep Learning (DL) models using the original dataset. This visualization helps compare the effectiveness of different models in predicting cyanobacterial density and bloom timing. We can assess which models offer the best predictive power and generalization capabilities

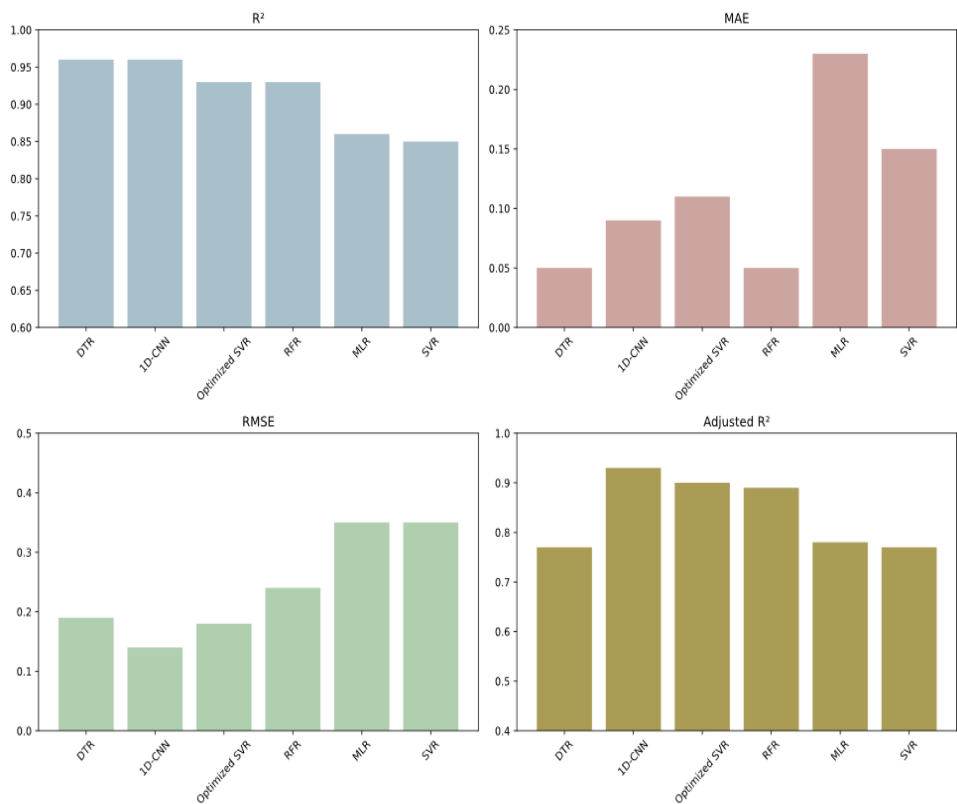


Figure 9. Graphical representation of performance metrics for ML and DL models with original data.

The results suggest that the 1D-CNN and (Decision Tree Regression) DTR models are top performers, with both achieving high R^2 values of 0.96, indicating their strong ability to explain the variance in the data. The 1D-CNN has the lowest RMSE, suggesting it handles large deviations particularly well, while DTR has the lowest MAE, reflecting minimal average errors. However, DTR’s lower adjusted R^2 compared to 1D-CNN suggests it might be prone to overfitting.

RFR also shows solid performance with a high R^2 of 0.93 and a balanced adjusted R^2 of 0.89, making it a reliable model with good generalization capabilities. The Optimized SVR improves significantly over SVR, with a notable increase in R^2 and reductions in both MAE and RMSE, indicating that hyperparameter optimization can significantly enhance model accuracy.

In contrast, MLR and SVR show comparatively lower performance, with R^2 values of 0.86 and 0.85 respectively. These models also exhibit higher errors and explain less variance, making them less favorable choices in this context. Overall, the 1D-CNN and RFR models offer the best combination of predictive power and generalization.

4.2. ML and DL Models with Augmented Data Using GANs

this section presents the results of ML and DL models using generated data from GAN model, as shown in the table below:

Table 2. Performance summary of ML and DL models with augmented data.

Model	R ²	MAE	RMSE	Adj. R ²
MLR	0.70	0.20	0.37	0.52
SVR	0.65	0.11	0.40	0.45
Optimized SVR	0.97	0.09	0.11	0.96
DTR	0.97	0.02	0.12	0.45
RFR	0.99	0.01	0.07	0.98
GAN-Enhanced 1D-CNN	0.98	0.06	0.09	0.97

Figure 10 also provides a graphical representation of the performance metrics for various Machine Learning (ML) and Deep Learning (DL) models but for this case using the augmented data.

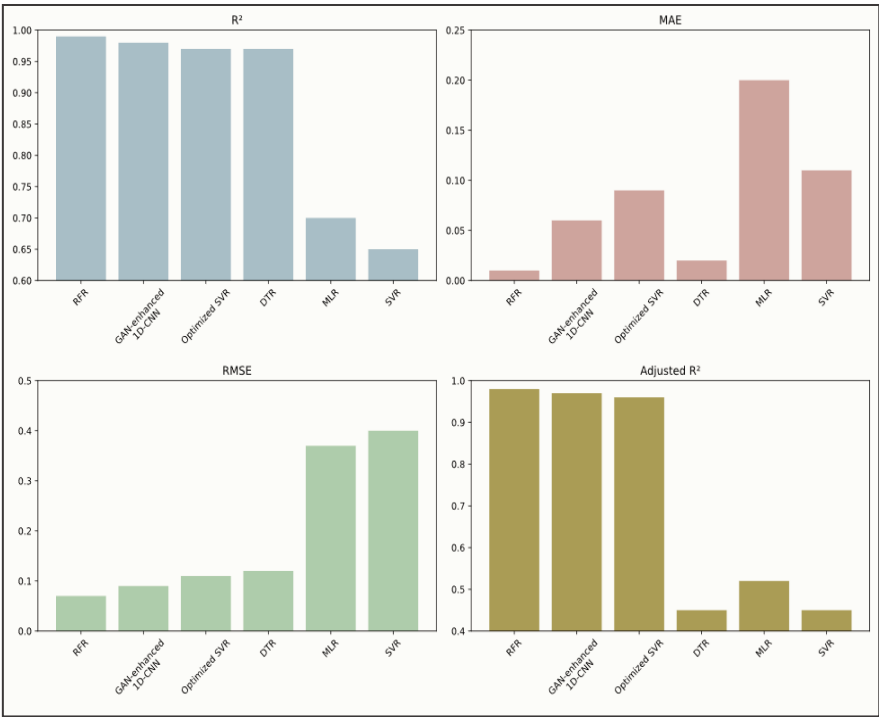


Figure10. Graphical representation of performance metrics of ML and DL models with augmented data.

After applying GAN-generated data augmentation and re-evaluating the models, the performance metrics have shifted notably:

RFR emerged as the top performer, with an R^2 of 0.99, indicating that it explains almost all the variance in the data. The extremely low MAE and RMSE values further confirm its superior accuracy and error management, making it the most reliable model in this context.

The GAN-enhanced 1D-CNN and Optimized SVR also demonstrated significant improvements. The 1D-CNN achieved an R^2 of 0.98, maintaining excellent predictive accuracy with low error metrics, while the Optimized SVR followed closely with an R^2 of 0.97, also showing strong performance in handling both average and larger deviations.

DTR maintained a high R^2 of 0.97, consistent with its performance before augmentation. However, its relatively low Adjusted R^2 of 0.45 now suggests potential overfitting, which could affect its generalizability.

In contrast, MLR and the SVR saw a decrease in their R^2 values, dropping to 0.70 and 0.65, respectively. These models now explain significantly less variance than the others, with higher RMSE values indicating larger errors. The augmentation process appears to have highlighted the limitations of these simpler models, as they struggle to maintain performance in the augmented data setting.

Overall, GAN-generated data augmentation has clearly enhanced the performance of more complex models like RFR, and 1D-CNN, while revealing the limitations of MLR and SVR.

Figure 11 shows the learning curves of the 1D-CNN model demonstrating a steady decrease in both training and validation loss over the 100 epochs, indicating effective learning from the data.

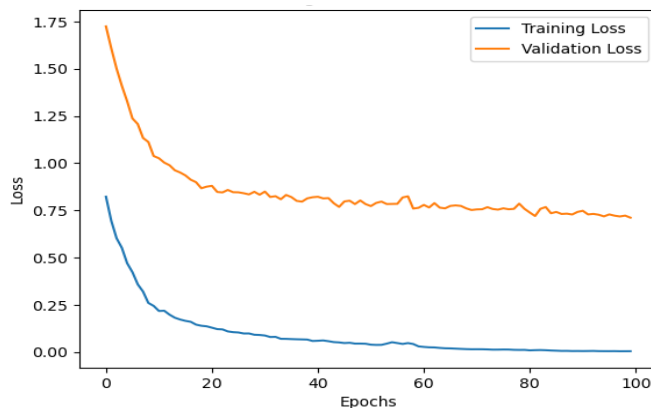


Figure 11. Learning curves for 1D-CNN model.

The learning curves of the GAN-enhanced 1D-CNN (Figure 12) showcase a remarkable reduction in both training and validation loss throughout the 100 epochs, surpassing the performance of the standard 1D-CNN. The convergence of the training and validation loss curves indicates effective learning with minimal overfitting, while the significantly lower

final loss values suggest that the diverse and enriched data generated by the GAN improves the model's performance, allowing it to better understand and predict the complex patterns within the dataset.

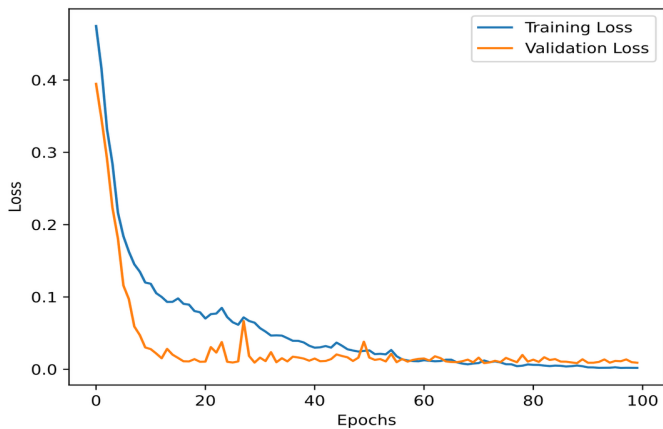


Figure 12. Learning curves for GAN-enhanced 1D-CNN model.

4.3. LSTM vs. LSTM with data upsampling

The results of the LSTM models, both before and after data upsampling for each architecture, are summarized in the table below:

Table 3. LSTM and LSTM with Upsampled data performance metrics.

Data	Metrics	LSTM Architecture	Improved LSTM Architecture
Original Data	R ²	0.51	0.29
	MAE	0.29	0.28
	RMSE	0.62	0.75
	Adj. R ²	0.14	-0.24
Upsampled Data	R ²	0.80	0.88
	MAE	0.32	0.24
	RMSE	0.41	0.31
	Adj. R ²	0.77	0.86

For the Original Data, the LSTM Architecture performs better overall with an R² of 0.51, indicating it explains about 51% of the variance, compared to 29% for the improved LSTM Architecture. However, both architectures struggle with relatively high RMSE values (0.62 for the First and 0.75 for the improved), indicating significant errors.

After applying Upsampling, both architectures show significant performance improvements. The R^2 values rise substantially to 0.80 for the LSTM Architecture and 0.88 for the improved LSTM Architecture, indicating that both models explain much more of the variance in the data post-augmentation. The MAE values increase slightly for the LSTM Architecture (0.32) but decrease for the Second Architecture (0.24), reflecting improved accuracy for the Second. RMSE values decrease for both architectures, showing reduced prediction errors. The Adjusted R^2 values also improve, with the LSTM Architecture reaching 0.77 and the improved LSTM Architecture achieving 0.86, indicating better model fits and generalizability after Upsampling.

Overall, the upsampling process significantly enhances the performance of both LSTM architectures in predicting cyanobacteria density 5 days in advance.

Figure 13 illustrates the actual and predicted values for both LSTM architectures using the original data. The graph on the left displays results for the LSTM Architecture, while the right graph shows those of the improved LSTM Architecture. The substantial deviations between the actual and predicted values in both graphs indicate that while each LSTM model captures some underlying patterns in the data, both architectures struggle to accurately model certain aspects. Notably, the "left finger" and "right finger" peaks, representing significant data points, show noticeable differences between the actual and predicted values.

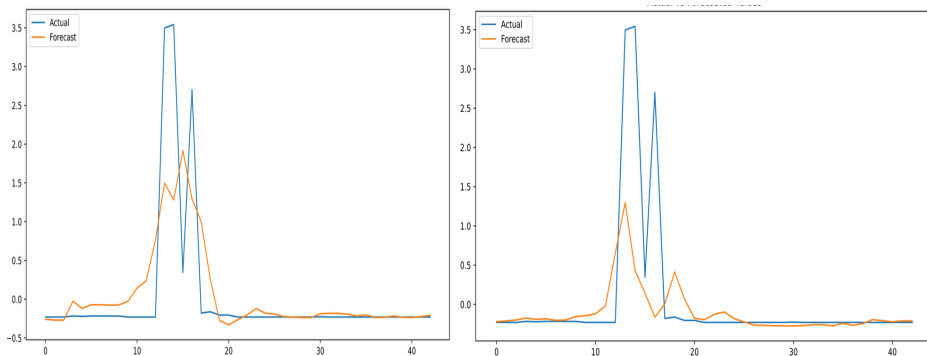


Figure 13. Actual and forecasted values for LSTM architectures with original data.

Figure 14 illustrates the actual and predicted values for both LSTM architectures using the upsampled data. The graph on the left displays results for the First Architecture, while the right graph shows the improved Architecture. The lines representing the actual and predicted values are much closer together compared to those in the original data. This indicates that the LSTM architectures with upsampled data achieve a higher level of accuracy in predictions, resulting in reduced deviations between the actual and forecasted values. The "left finger" and "right finger" peaks are more closely aligned in the upsampled data, demonstrating improved model performance.

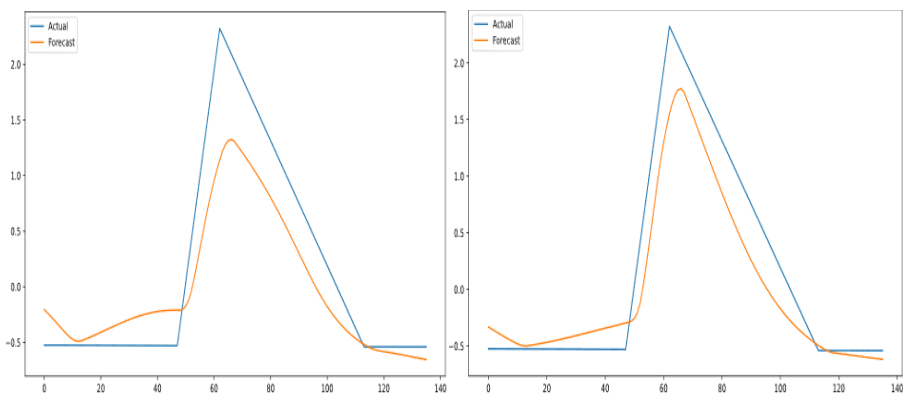


Figure 14. Actual and forecasted values for LSTM architectures with Upsampled data.

4.3.1. Bloom Detection Function

This function performs several tasks:

- It applies inverse standardization to the predictions to revert them to their original scale.
- It evaluates whether the predicted cyanobacteria density exceeds the threshold of 200,000 (Cell/ml) and classifies such instances as **Bloom**.

Figures 15 visualize cyanobacteria density predictions over time for the four architectures. The top two graphs display predictions for the architectures using the original data, while the bottom two graphs show predictions for the architectures with upsampled data. When the blue line (representing cyanobacteria density) surpasses the red dashed line, it indicates an algal bloom event. Negative values in the predictions suggest that the model anticipates low or absent cyanobacteria density during those time periods.

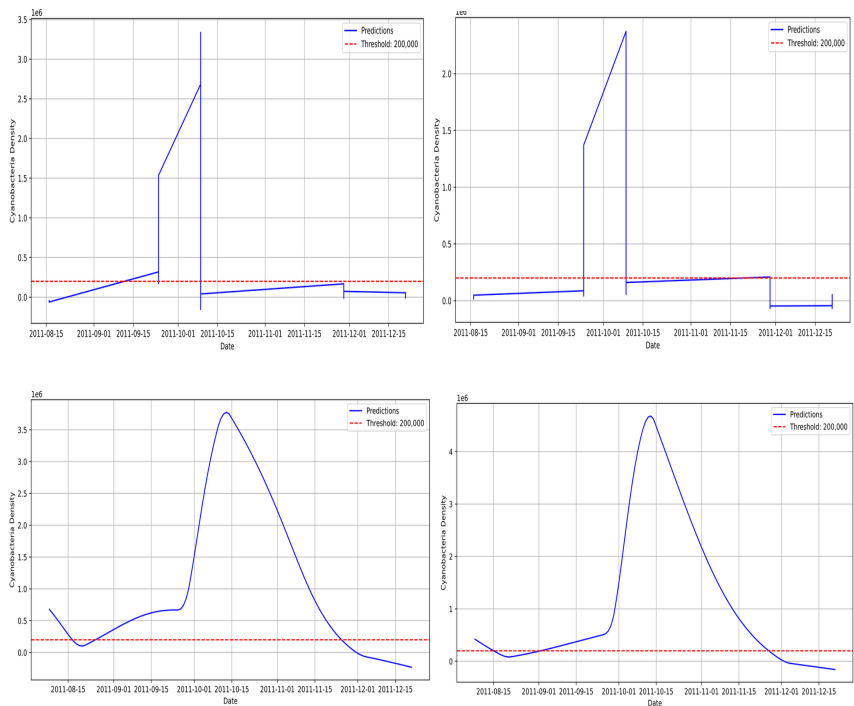


Figure 15. Detection of bloom events using LSTM predictions.

5. Conclusion

This research develops an intelligent AI-based system to predict CyanoHAB occurrences and severity using advanced deep learning architectures. Our approach combines Long Short-Term Memory (LSTM) networks for temporal bloom forecasting with 1D-Convolutional Neural Networks (1D-CNNs) for density prediction, both enhanced through Generative Adversarial Networks (GANs) for data augmentation to address class imbalance and limited sample size.

Methodologically, we implemented a comprehensive pipeline that includes:

- Temporal pattern recognition via LSTM for bloom timing prediction.
- Feature extraction from sequential data using 1D-CNN for density estimation
- Synthetic data generation through GANs to improve model robustness
- Comparative benchmarking against traditional regression models

Evaluated on the Algerian Mexa database, our system demonstrates strong predictive performance with:

- 98% R^2 and 9% RMSE for cyanobacterial density prediction
- 88% R^2 and 31% RMSE for bloom timing forecasting
- 5-day advance warning capability for proactive intervention

These results underscore the value of our hybrid deep learning approach for environmental monitoring, particularly in:

1. Improving prediction accuracy through model combination
2. Enhancing robustness via data augmentation
3. Providing actionable lead time for water management

The study contributes both technical advancements in AI for environmental science and practical solutions for CyanoHAB monitoring, with potential applications in water safety and ecosystem protection.

References

- [1] Amara, N. D., Dendani, N., Berrezek, A., Khemissa, H., Saoudi, A., and Azizi, N., GAN-Enhanced Deep Learning Approach for Forecasting the Potentially Toxic Cyanobacteria in Dams, in: Bennour, A., Bouridane, A., Almaadeed, S., Bouaziz, B., Edirisinghe, E. (eds), *Intelligent Systems and Pattern Recognition*. ISPR'2024. Communications in Computer and Information Science, vol. 2304, Springer, Cham, 2025, 42-52.
- [2] Awad, M. and Khanna, R., *Efficient Learning Machines*, Apress, 2015, 67-71.
- [3] Berrezek, A., Dendani, N., Amara, N. D., Azizi, N., and Saoudi, A., Ensemble Methods for Predicting Cyanobacteria's Potential Toxicity in Water Dams, in: Bennour, A., Bouridane, A., Almaadeed, S., Bouaziz, B., Edirisinghe, E. (eds), *Intelligent Systems and Pattern Recognition*. ISPR'2024. Communications in Computer and Information Science, vol. 2305, Springer, Cham, 2025, 161-174
- [4] Boukhamla, A., Bouziane, M. H., Laib, A., Azizi, N., Rouabhi, R., Merah, A., and Chaib, R., GANs Investigation for Multimodal Medical Data Interpretation: Basic Architectures and Overview, in: International Conference on Control, Automation and Diagnosis (ICCAD'2023), IEEE, 2023, 01-06.
- [5] Busari, I., Sahoo, D., and Jana, R., Prediction of Chlorophyll-a as an Indicator of Harmful Algal Blooms Using Deep Learning with Bayesian Approximation for Uncertainty Assessment, *Journal of Hydrology*, vol. 630, 2024.
- [6] Christopher, M. H., Jingjing, X., Hao, W., Mark, A. L., Ron, W. Z., Charlie, J. G. L., Rolf, D. V., and Pouria, R., Predicting Imminent Cyanobacterial Blooms in Lakes Using Incomplete Timely Data, *Water Resources Research*, vol. 60, p. 1, 2024.

-
- [7] David, F. M., Gary, R. W., Gary, L. F., Hunter, J. C., Ehsan, A., William, A. Y. I., Michael, J. S., and Robert, A. S., Using Artificial Intelligence for CyanoHAB Niche Modeling: Discovery and Visualization of Microcystis-Environmental Associations within Western Lake Erie, *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 71, p. 1, 2014.
- [8] Felipe, N. B., Claudio, C. B., Vitor, S. M., Evelyn, M. N., Rejane, S. P., Daniel, A. M., Thainara, M. L., Ryan, E. O., Nima, P., and Marta, C. L., Machine Learning for Cyanobacteria Mapping on Tropical Urban Reservoirs Using PRISMA Hyperspectral Data, *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, p. 1, 2023.
- [9] Haiping, A., Kai, Z., Jiachun, S., and Huichun, Z., Short-term Lake Erie Algal Bloom Prediction by Classification and Regression Models, *Water Research*, vol. 232, 2023.
- [10] Hongye, C., Ling, H., and Liangzhi, L., A Deep Learning Method for Cyanobacterial Harmful Algae Blooms Prediction in Taihu Lake, China, *Harmful Algae*, vol. 113, p. 1, 2022.
- [11] Hyo, G. K., Kyung, H. C., and Friedrich, R., Time-Series Modelling of Harmful Cyanobacteria Blooms by Convolutional Neural Networks and Wavelet Generated Time-Frequency Images of Environmental Driving Variables, *Water Research*, vol. 246, p. 1, 2023.
- [12] Jie, N., Yanqun, L., Mengyu, X., Linjian, O., Lei, C., Han, Q., and Songhui, L., Prediction of *Aureococcus anophagefferens* Using Machine Learning and Deep Learning, *Marine Pollution Bulletin*, vol. 200, p. 1, 2024.
- [13] Jung, M. A., Jungwook, K., Hongtae, K., and Kyunghyun, K., Harmful Cyanobacterial Blooms Forecasting Based on Improved CNN-Transformer and Temporal Fusion Transformer, *Environmental Technology & Innovation*, vol. 32, pp. 1-9, 2023.
- [14] Kambezidis, H., *Comprehensive Renewable Energy*, Elsevier, 2012.
- [15] Khan, A. Q., Awan, H. A., Rasul, M., Siddiqi, Z. A., and Pimanmas, A., Optimized Artificial Neural Network Model for Accurate Prediction of Compressive Strength of Normal and High Strength Concrete, *Cleaner Materials*, vol. 10, pp. 2-6, 2023.
- [16] Kotsiantis, S., Kanellopoulos, D., and Pintelas, P. E., Data Preprocessing for Supervised Learning, *International Journal of Computer Science*, vol. 1, no. 1, p. 111, 2006.
- [17] Lecun, Y., Bengio, Y., and Hinton, G., Deep Learning, *Nature*, vol. 521, no. 7553, 436-444, 2015.
- [18] Lee, S. and Lee, D., Improved Prediction of Harmful Algal Blooms in Four Major South Korea's Rivers using Deep Learning Models, *International Journal of Environmental Research and Public Health*, vol. 15, 2018.
- [19] Mellios, N., Moe, S. J., and Laspidou, C., Machine Learning Approaches for Predicting Health Risk of Cyanobacterial Blooms in Northern European Lakes, *Water*, vol. 12, no. 1191, p. 1, 2020.

- [20] Nowicka-Krawczyk, P., Żelazna-Wieczorek, J., Skrobek, I., Ziułkiewicz, M., Adamski, M., Kaminski, A., and Żmudzki, P., Persistent Cyanobacteria Blooms in Artificial Water Bodies: An Effect of Environmental Conditions or the Result of Anthropogenic Change, *International Journal of Environmental Research and Public Health*, vol. 19, no. 6990, p. 2, 2022.
- [21] Park, Y., Lee, H. K., Shin, J.-K., Chon, K., Kim, S., Cho, K. H., Kim, J. H., and Baek, S.-S., A Machine Learning Approach for Early Warning of Cyanobacterial Bloom Outbreaks in a Freshwater Reservoir, *Journal of Environmental Management*, vol. 288, p. 1, 2021.
- [22] Saoudi, A., Cyanobacteria and Cyanotoxins in the Water of the Mexa Dam (El-Tarf), Dept. Biochemistry. Ecobiology of Marine and Coastal Environments Laboratory, ANNABA, 2015-2016.
- [23] Saoudi, A., Choukri, B., Luc, B., Rachid, O., and Mourad, B., Environmental Parameters and Spatio- Temporal Dynamics of Cyanobacteria in the Reservoir of Mexa (Extreme North-East of Algeria), *Advances in Environmental Biology*, vol. 9, no. 11, p. 109, 2015.
- [24] Singh, V. K., Upsampling Interpolation of Discrete Signals, divilabs, 10 July 2014. [Online]. Available: <https://www.divilabs.com/2014/07/upsampling-interpolation-of-discrete.html>. [Accessed 04 June 2024].
- [25] Smith, G., Essential Statistics, Regression, and Econometrics, Elsevier Inc, 2012.
- [26] Thu-Hien, T. and Nhat-Duc, H., Predicting Colonization Growth of Algae on Mortar Surface with Artificial Neural Network, *Journal of Computing in Civil Engineering*, vol. 30, p. 1, 2016.
- [27] Xu, Y., Zhang, D., Lin, J., Peng, Q., Lei, X., Jin, T., Wang, J., and Yuan, R., Prediction of Phytoplankton Biomass and Identification of Key Influencing Factors Using Interpretable Machine Learning Models, *Ecological Indicators*, vol. 158, 1-6, 2024.
- [28] Zanchett, G. and Oliveira-Filho, E. C., Cyanobacteria and Cyanotoxins: From Impacts on Aquatic Ecosystems and Human Health to Anticarcinogenic Effects, Toxins, vol. 5, no. 10, p. 2, 2013.
- [29] Zemmam, N., Benzebouchi, N., Azizi, N., et al., Unbalanced Learning for Early Automatic Diagnosis of Diabetes Based on Enhanced Resampling Technique and Stacking Classifier, *International Journal of Intelligent Information Technologies (IJIT)*, vol. 18(1), p. 29, 2022.

Received 25.02.2025, Accepted 05.11.2025