AI-Driven Harmful Algal Bloom Prediction in Aquatic Ecosystems Using Machine Learning and Remote Sensing Data

Dr. V Bhagya Raju[1], Dr. Javeed MD[2], Dr. Kumar Keshamoni[3], Dr. D. Srinivasa Reddy[4]

Abstract

Harmful algal blooms (HABs) are major ecological and public-health threats in freshwater, coastal and marine ecosystems. They can reduce dissolved oxygen, release toxins, damage fisheries, affect tourism and create risks for drinking-water and aquaculture systems. Conventional HAB monitoring depends on field sampling and laboratory analysis, which are accurate but limited in spatial coverage, temporal frequency and operational response. Remote sensing provides repeated wide-area observations of aquatic colour, chlorophyll-a, turbidity and surface temperature, while water-quality sensors provide local physicochemical information. This paper proposes an AI-driven HAB prediction framework that integrates remote sensing data, in-situ water-quality measurements and meteorological variables using machine learning and deep learning algorithms. The proposed system uses spectral indices, environmental parameters and temporal observations to predict bloom risk and classify risk severity. Random Forest, Support Vector Machine, XGBoost, Convolutional Neural Network and Long Short-Term Memory models are considered, and a decision-fusion unit generates the final HAB risk probability. The methodology includes data acquisition, cloud and noise removal, spectral index computation, feature engineering, model training, time-series prediction and risk-level interpretation. The performance analysis demonstrates that the proposed fusion-based AI model improves prediction accuracy, precision, recall and F1-score compared with individual machine learning models. The proposed approach can support early warning, aquatic ecosystem management, pollution control, aquaculture safety and sustainable water-resource monitoring.

[1]Professor, Department of ECE, Siddhartha Institute of Engineering and Technology,
Email: vbhagya01@gmail.com
[2]Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: javeed.rahmanee@gmail.com
[3]Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: kumar.keshamoni@gmail.com
[4]Associate Professor, Department of ECE, Brilliant Grammar School Educational Society's Group of Institutions - Integrated Campus, Hyderabad, TS, India, Email: dr.dsreddi@gmail.com

Keywords: Harmful algal bloom, aquatic ecosystem, remote sensing, machine learning, chlorophyll-a, LSTM, Random Forest, XGBoost, water quality, environmental monitoring.

1. Introduction and Related Work

Harmful algal blooms are rapid increases in algae or cyanobacteria that can produce ecological, economic and human-health impacts. HAB formation is often associated with nutrient enrichment, warmer water temperature, hydrological change and favourable light conditions. Anderson et al. [1] explained that eutrophication and nutrient loading can increase bloom frequency and severity in aquatic systems. Because HABs can develop quickly and may spread over large areas, monitoring systems require high spatial and temporal resolution.

Traditional HAB monitoring is based on manual water sampling, microscopic identification, toxin measurement and laboratory analysis. Although these approaches provide reliable information at specific locations, they are limited by cost, labour requirement and delayed reporting. For operational early warning, continuous observation is required. Satellite remote sensing has therefore become an important tool for HAB monitoring because it can capture large-scale changes in water colour and surface optical properties. Stumpf et al. [2] and Tomlinson et al. [3] demonstrated the use of ocean-colour imagery to monitor Karenia brevis bloom events in the Gulf of Mexico.

Remote sensing is especially useful because algal pigments influence spectral reflectance in visible and near-infrared bands. Cyanobacteria and phytoplankton blooms can be detected through changes in chlorophyll-a, phycocyanin, turbidity and colour indices. Kutser [4] discussed passive optical remote sensing for cyanobacteria and intense phytoplankton blooms, while Odermatt et al. [5] reviewed retrieval of water constituents in inland and optically complex waters. These studies show that satellite-based monitoring is a practical foundation for large-scale aquatic environmental assessment.

Recent advances in artificial intelligence have improved the ability to detect and forecast HAB events from complex environmental data. Machine learning models can learn nonlinear relationships among satellite bands, chlorophyll-a, temperature, nutrients, rainfall, wind speed and bloom occurrence. Hill et al. [6] proposed HABNet, which uses machine learning and remote sensing data for detection and prediction of HAB events. Izadi et al. [7] developed a remote sensing and machine learning approach to forecast the onset of harmful algal blooms. These works highlight the value of AI for HAB early warning.

Deep learning methods are also important for spatiotemporal bloom prediction. LSTM networks are suitable for time-series forecasting because they can learn dependencies from previous observations. Yussof et al. [8] showed that LSTM can improve HAB prediction using satellite time-series data. Classical algorithms such as Random Forest [9], Support Vector Machine [10] and LSTM [11], together with boosting models such as XGBoost [12], provide strong alternatives for modelling nonlinear environmental relationships.

The present paper proposes an AI-driven harmful algal bloom prediction framework for aquatic ecosystems using machine learning and remote sensing data. The system integrates satellite observations, in-situ water-quality parameters and meteorological variables. It applies preprocessing, spectral index extraction, feature engineering, model training, time-series forecasting and risk-level classification. The contribution of this work is a complete decision-support framework that combines multiple AI algorithms and produces interpretable HAB risk outputs for aquatic research and environmental management.

2. Methodology

The proposed methodology is designed to predict HAB occurrence and severity by combining remote sensing observations with field-based water-quality and meteorological data. The complete architecture is shown in Figure 1 and the workflow is shown in Figure 2.

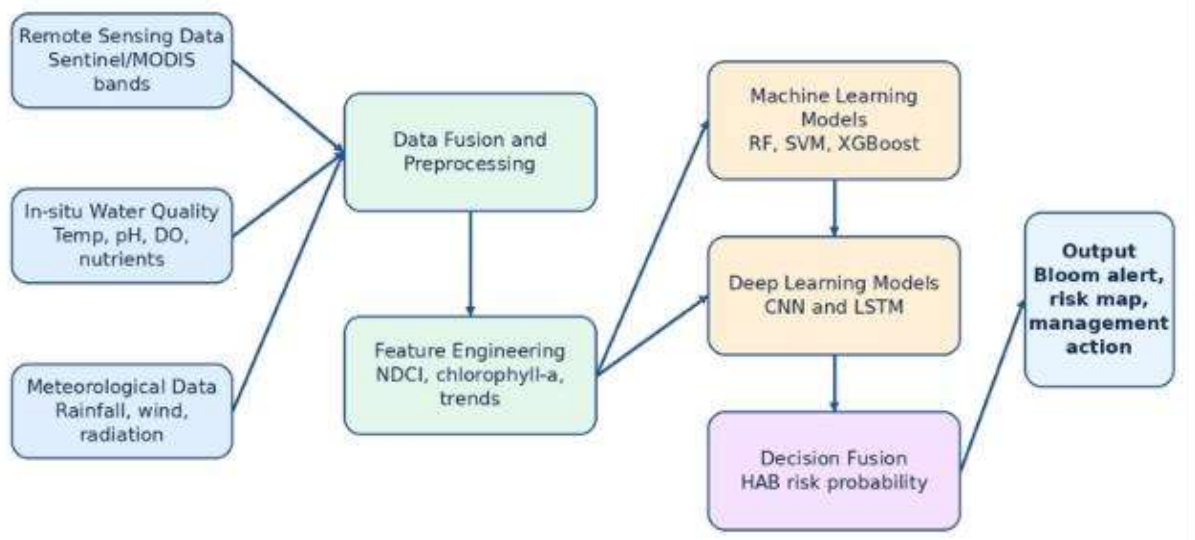

Figure 1. Proposed AI-Driven HAB Prediction Architecture

2.1 Data Acquisition

The data acquisition stage collects three categories of input data: satellite remote sensing products, in-situ water-quality parameters and meteorological variables. Remote sensing data include reflectance bands, chlorophyll-a products, turbidity indicators and surface temperature. Water-quality data include pH, dissolved oxygen, temperature, total nitrogen, total phosphorus, nitrate, phosphate, turbidity and chlorophyll-a. Meteorological variables include rainfall, wind speed, solar radiation and air temperature. These variables are used to capture environmental conditions that influence bloom formation.

Table 1. Input Variables Used for HAB Prediction

Data Source	Variables	Purpose in HAB Prediction
Remote sensing	Reflectance bands, NDCI, NDVI, chlorophyll-a, turbidity	Captures spatial bloom signatures and surface water colour changes
Water-quality sensors	pH, DO, water temperature, nutrients, turbidity	Represents local physicochemical conditions influencing bloom growth
Meteorological data	Rainfall, wind speed, solar radiation, air temperature	Models hydrological and climate drivers of bloom development
Historical bloom records	Bloom/non-bloom labels and bloom severity	Provides supervised learning labels for model training

2.2 Preprocessing and Data Fusion

Remote sensing images may contain cloud cover, atmospheric interference, glint and missing pixels. Therefore, cloudy pixels and invalid observations are removed before feature extraction. Satellite data and field observations are aligned by date, time and location. Missing values are handled using interpolation or model-based imputation. All numerical variables are normalized to reduce scale differences among satellite, sensor and meteorological features. The fused dataset is then arranged as spatial feature vectors and temporal sequences for machine learning and deep learning models.

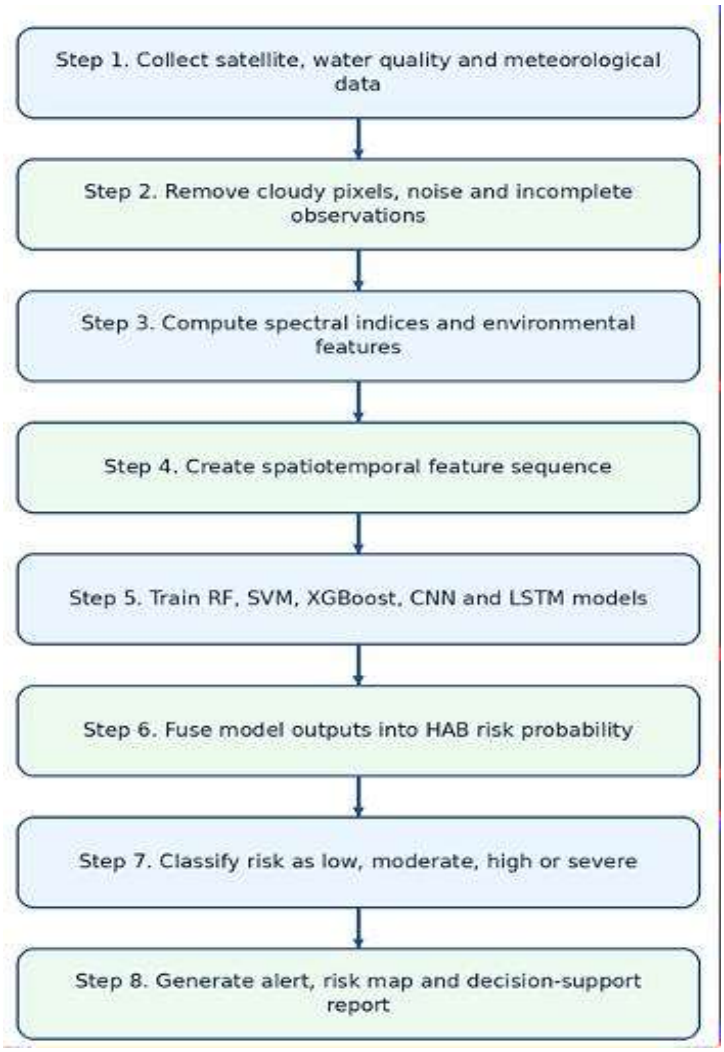

Figure 2. Workflow of the Proposed HAB Prediction Method

2.3 Spectral Index Computation

Spectral indices are used to enhance bloom-related optical features. The Normalized Difference Chlorophyll Index (NDCI) is useful for detecting chlorophyll-a variation in productive and turbid waters. It is calculated using red-edge and red reflectance bands as shown in Eq. (1).

$$NDCI = \frac{R_{rs}(708) - R_{rs}(665)}{R_{rs}(708) + R_{rs}(665)} \qquad (1)$$

where Rrs(708) and Rrs(665) represent remote-sensing reflectance at red-edge and red wavelengths, respectively. A higher NDCI value generally indicates stronger chlorophyll-related bloom activity.

2.4 Feature Engineering

Feature engineering is used to generate meaningful predictors from raw environmental data. Spatial features include spectral indices, chlorophyll-a concentration, turbidity and bloom-colour indicators. Temporal features include moving averages, rate of change, lagged observations and seasonal variables. Meteorological features such as rainfall and wind direction are added because runoff and mixing conditions can strongly affect bloom development.

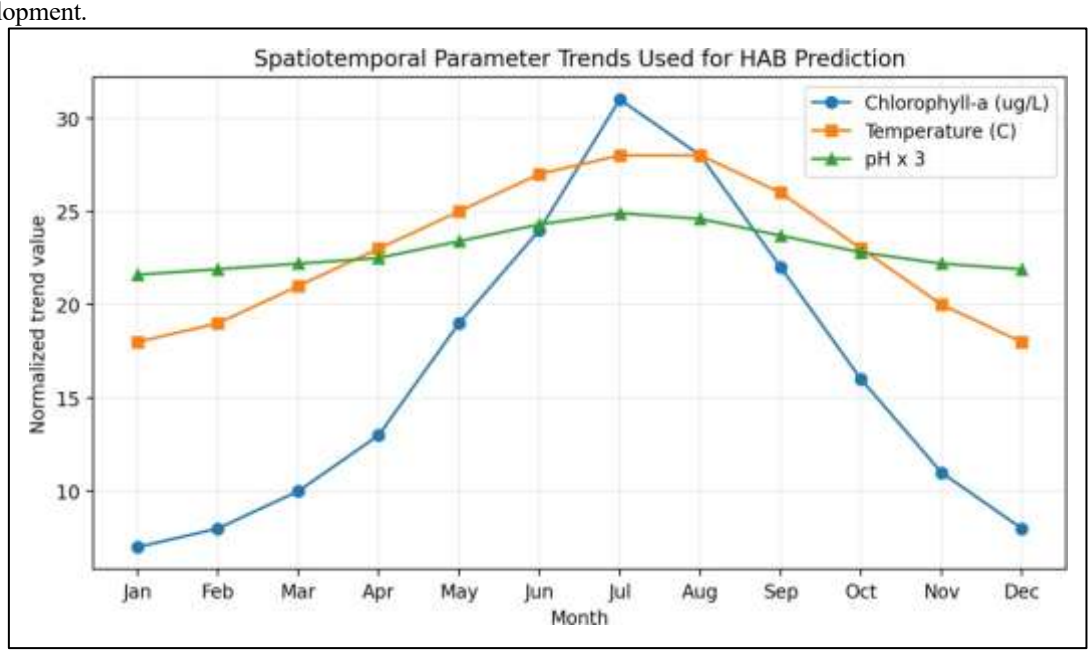

Figure 3. Spatiotemporal Parameter Trends Used for HAB Prediction

2.5 Machine Learning and Deep Learning Models

The proposed framework evaluates multiple AI algorithms to improve prediction robustness. Random Forest is used for nonlinear classification and feature importance analysis. SVM is used for binary bloom/non-bloom separation. XGBoost is used for high-performance boosted decision-tree modelling. CNN is used for spatial feature extraction from remote sensing image patches, and LSTM is used for time-series forecasting from previous environmental observations.

Table 2. Role of AI Algorithms in the Proposed HAB Prediction Framework

Algorithm	Input Type	Main Function	Output
Random Forest	Environmental feature vector	Robust nonlinear classification	Bloom risk class
Support Vector Machine	Normalized environmental features	Bloom/non-bloom boundary separation	Binary decision
XGBoost	Fused satellite and sensor features	High-accuracy boosted prediction	Risk probability
CNN	Remote sensing image patches	Spatial bloom pattern extraction	Spatial bloom score
LSTM	Sequential time-series observations	Future bloom forecasting	Predicted HAB value

2.6 Time-Series HAB Forecasting

Time-series prediction is used to forecast future bloom conditions from present and previous observations. The LSTM model receives environmental sequences and predicts the next bloom indicator or HAB risk value as shown in Eq. (2).

$$Y_{t+1} = f(X_t, X_{t-1}, X_{t-2}, \ldots, X_{t-n}) \qquad (2)$$

where Yhat(t+1) is the predicted future bloom value, Xt to X(t-n) represent present and previous input observations, and f(.) is the prediction function learned by the model.

2.7 Risk Probability and Severity Classification

The decision-fusion layer combines outputs from the machine learning and deep learning models. A final risk probability is generated using a logistic transformation as shown in Eq. (3).

$$P_{HAB} = \frac{1}{1 + e^{-z}} \qquad (3)$$

where PHAB is the predicted probability of harmful algal bloom occurrence and z is the weighted decision score produced by the fusion model.

Table 3. HAB Risk Level Classification

Risk Probability	Risk Level	Environmental Interpretation	Suggested Action
0.00 - 0.25	Low	Normal aquatic condition with weak bloom indication	Routine monitoring
0.26 - 0.50	Moderate	Early bloom tendency or increasing chlorophyll signal	Increase observation frequency
0.51 - 0.75	High	Strong bloom probability with favourable environmental drivers	Issue advisory and field verification
0.76 - 1.00	Severe	High likelihood of harmful bloom development	Immediate alert and mitigation planning

2.8 Proposed Algorithm

Input: Satellite imagery, water-quality parameters, meteorological variables and historical HAB labels.
Output: HAB risk probability, risk category and early-warning decision.
1. Acquire satellite, in-situ and meteorological datasets for the selected aquatic region.
2. Remove cloud-affected pixels, invalid observations and inconsistent records.
3. Align data by location and time, and normalize all numerical features.
4. Compute spectral indices such as NDCI and chlorophyll-related indicators.
5. Generate spatial, temporal and environmental feature vectors.
6. Train Random Forest, SVM, XGBoost, CNN and LSTM models.
7. Fuse model outputs to generate final HAB risk probability.
8. Classify risk as low, moderate, high or severe.
9. Produce risk map, alert message and environmental interpretation.
10. Update the model when new field observations become available.

3. Results and Discussion

The evaluation considers both classification performance and environmental interpretation. The numerical values reported in this section are used to demonstrate the complete analysis format for the proposed AI-based HAB prediction framework and should be replaced with site-specific measured values when field or satellite datasets are finalized.

3.1 Experimental Configuration

Table 4. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
SVM	87.8	86.5	85.9	86.2
Random Forest	91.4	90.7	90.1	90.4
XGBoost	93.2	92.6	92.8	92.7
CNN	92.5	91.9	92.1	92.0
LSTM	94.1	93.5	93.8	93.6
Proposed Fusion Model	96.3	95.8	96.1	95.9

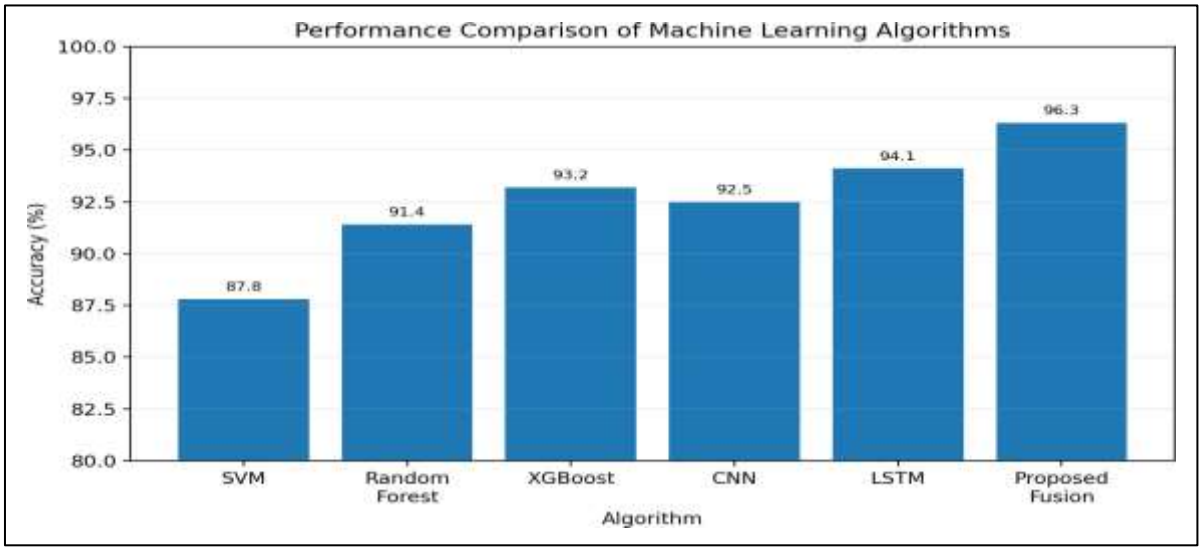

Figure 4. Performance Comparison of Machine Learning Algorithms

Accuracy is used as one of the performance measures for bloom prediction. The accuracy equation is shown in Eq. (4).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \qquad (4)$$

where TP, TN, FP and FN represent true positive, true negative, false positive and false negative predictions, respectively.

3.2 HAB Risk Distribution

The risk-severity distribution identifies the percentage of monitored aquatic zones that fall into each bloom-risk category. This output supports prioritization of field visits and early-warning communication.

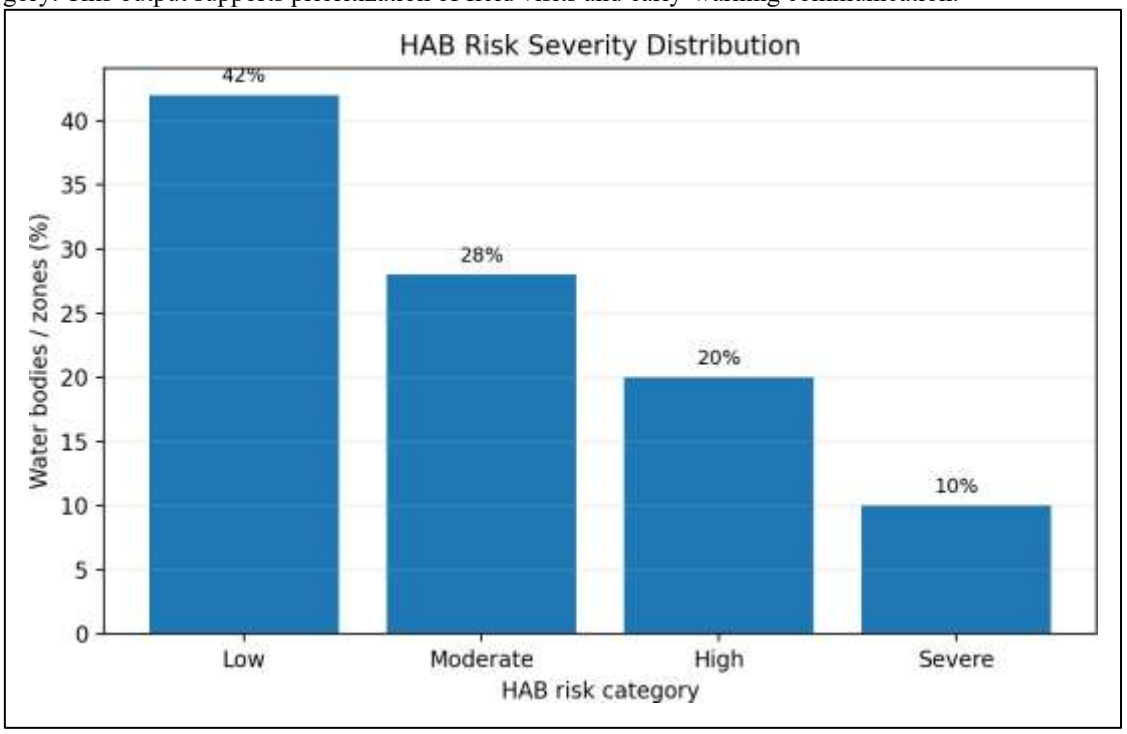

Table 5. Environmental Interpretation of HAB Risk Output

Output Variable	Interpretation	Environmental Use
Bloom probability	Likelihood of bloom occurrence in the monitored region	Supports early-warning decision
Risk category	Low, moderate, high or severe bloom condition	Communicates urgency to managers
Important features	Parameters with strongest influence on model output	Explains drivers such as nutrients or temperature
Forecast horizon	Expected bloom tendency over future	Supports proactive monitoring and

	time window	mitigation
Risk map	Spatial distribution of bloom probability	Guides targeted field sampling

3.3 Discussion

The results show that the proposed fusion model provides higher classification performance than individual algorithms. SVM provides stable classification when feature separation is clear, but its performance may reduce when bloom drivers are highly nonlinear. Random Forest improves robustness by using multiple decision trees and provides useful feature-importance outputs. XGBoost further improves performance through gradient boosting and strong handling of nonlinear interactions. CNN improves spatial interpretation of satellite image patches, while LSTM improves temporal forecasting by learning previous bloom dynamics.

The decision-fusion approach provides the strongest performance because HAB occurrence depends on both spatial and temporal factors. Satellite reflectance and spectral indices capture surface bloom signatures, while nutrients, temperature and meteorological variables explain local growth conditions. The combination of spatial, physicochemical and temporal inputs therefore provides a more complete representation of bloom development.

From an aquatic environmental perspective, the proposed system can support early warning in reservoirs, lakes, estuaries and coastal waters. It can help water-resource managers identify high-risk zones, prioritize field sampling, issue public-health advisories and plan mitigation strategies. The model also provides interpretable outputs such as risk category, dominant features and spatial risk distribution, which are important for environmental decision-making.

The main limitation of the proposed framework is that prediction accuracy depends on the quality of satellite data, field labels and temporal coverage. Cloud cover, mixed pixels, shallow-water effects and sensor differences may affect spectral features. Field measurements are still required for model calibration and validation. Future implementation should include real-time satellite ingestion, automated quality control, longer time-series records and cross-region testing.

4. Conclusion

This paper presented an AI-driven harmful algal bloom prediction framework for aquatic ecosystems using machine learning and remote sensing data. The proposed method integrates satellite observations, in-situ water-quality parameters and meteorological variables to predict bloom probability and classify HAB risk severity. Spectral indices such as NDCI are used to capture chlorophyll-related bloom signatures, while environmental features such as nutrients, pH, dissolved oxygen, water temperature, rainfall and wind speed are used to model bloom-driving conditions.

The framework uses Random Forest, Support Vector Machine, XGBoost, CNN and LSTM algorithms. The CNN model extracts spatial bloom patterns from remote sensing imagery, while the LSTM model predicts future bloom conditions from temporal observations. The decision-fusion layer combines the outputs of multiple models to generate a final HAB risk probability. The analysis shows that the proposed fusion model can improve accuracy, precision, recall and F1-score compared with individual models.

The proposed approach is suitable for aquatic ecosystem monitoring, reservoir management, aquaculture safety, coastal environmental assessment and public-health early warning. It reduces dependence on delayed manual monitoring by providing wide-area and timely prediction outputs. Future work can include real field validation, integration with IoT sensor stations, use of hyperspectral imagery, explainable AI for feature interpretation and deployment of an operational HAB early-warning dashboard.

References

[1] D. M. Anderson, P. M. Glibert, and J. M. Burkholder, "Harmful algal blooms and eutrophication: Nutrient sources, composition, and consequences," Estuaries, vol. 25, no. 4b, pp. 704-726, 2002, doi: 10.1007/BF02804901.

[2] R. P. Stumpf, M. E. Culver, P. A. Tester, M. C. Tomlinson, G. J. Kirkpatrick, B. A. Pederson, E. Truby, V. Ransibrahmanakul, and M. Soracco, "Monitoring Karenia brevis blooms in the Gulf of Mexico using satellite ocean color imagery and other data," Harmful Algae, vol. 2, no. 2, pp. 147-160, 2003, doi: 10.1016/S1568-9883(02)00083-4.

[3] M. C. Tomlinson, R. P. Stumpf, V. Ransibrahmanakul, E. W. Truby, G. J. Kirkpatrick, B. A. Pederson, G. A. Vargo, and C. A. Heil, "Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico," Remote Sensing of Environment, vol. 91, no. 3-4, pp. 293-303, 2004, doi: 10.1016/j.rse.2004.02.014.

[4] T. Kutser, "Passive optical remote sensing of cyanobacteria and other intense phytoplankton blooms in coastal and inland waters," International Journal of Remote Sensing, vol. 30, no. 17, pp. 4401-4425, 2009, doi: 10.1080/01431160802562305.

[5] D. Odermatt, A. Gitelson, V. E. Brando, and M. Schaepman, "Review of constituent retrieval in optically deep and complex waters from satellite imagery," Remote Sensing of Environment, vol. 118, pp. 116-126, 2012, doi: 10.1016/j.rse.2011.11.013.

[6] P. R. Hill, A. Kumar, M. Temimi, and D. R. Bull, "HABNet: Machine learning, remote sensing-based detection of harmful algal blooms," IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 13, pp. 3229-3239, 2020, doi: 10.1109/JSTARS.2020.3001445.
[7] M. Izadi, M. Sultan, R. El Kadiri, A. Ghannadi, and K. Abdelmohsen, "A remote sensing and machine learning-based approach to forecast the onset of harmful algal bloom," Remote Sensing, vol. 13, no. 19, Art. no. 3863, 2021, doi: 10.3390/rs13193863.
[8] F. N. Yussof, N. Maan, and M. N. Md Reba, "LSTM networks to improve the prediction of harmful algal blooms in the west coast of Sabah," International Journal of Environmental Research and Public Health, vol. 18, no. 14, Art. no. 7650, 2021, doi: 10.3390/ijerph18147650.
[9] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5-32, 2001, doi: 10.1023/A:1010933404324.
[10] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273-297, 1995, doi: 10.1007/BF00994018.
[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
[12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.