Harmful Algal Bloom (HAB) Prediction Using Ensemble Machine Learning Models and XAI Technique

Eddie Zhang[1,2] , Omer Mermer[2], Yusuf Sermet[2], İbrahim Demir[2, 3]

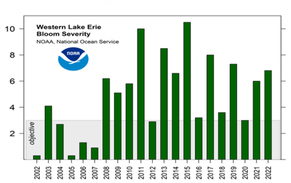
[1]The Harker School, [2]IIHR Hydroscience & Engineering, University of Iowa, [3]Electrical and Computer Engineering, University of Iowa

Introduction

- Harmful algal blooms (HABs) present in multiple bodies of water (Saxena, 2017)
- Result of various factors including climate change and agricultural water pollution (Saxena, 2017)
- High amounts of nutrients lead to higher concentrations of algae in water (Patel, 2017)
- Negative Effects on Environments
  - Deplete water of oxygen
  - Contaminate drinking water
  - Threat to biodiversity + animal life in surrounding areas

Algae Bloom Examples (Saxena, 2017; Molinari, 2024)

- Study Area: Lake Erie
  - Multiple harmful algae blooms in last years (Patel, 2017)
  - Important area for fishing and drinking water



Algal Blooms in Lake Eerie (Patel, 2017; Stumpf, 2024)

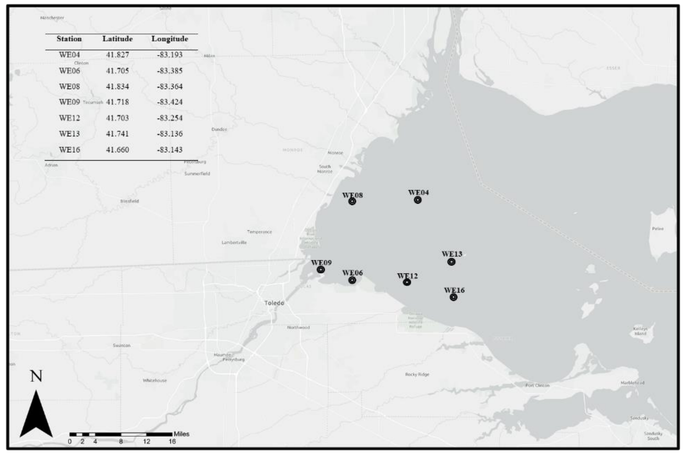
Background Information

- Water supply reservoir in South Korea predictions using RF and XGB (Jeong et. al, 2022)
- SHAP values on three different machine learning models to identify relationships between chlorophyll-a concentrations and various water quality factors (Shukla et. al, 2024)
- Use of multiple deep learning and linear models to predict chlorophyll-a values as an index of algae bloom prediction (Busari et. al, 2024)
- Lake Erie algae modeling and prediction using long short term memory networks based off of different features in water quality (Ai et. al, 2024)
- Use of remote sensing images and image processing techniques to detect and forecast algae blooms in Taihu Lake in China (Cao et. al, 2024)

Research Gap

- No **comprehensive and comparative study** of **ensemble** learning and linear machine learning techniques on **algae bloom data**
- Lack of understanding of the effects of using **stronger** and **weaker learners** in **ensemble regressors**
- Lack of **in depth analysis** on most important **features** to a **machine learning model** using **XAI**

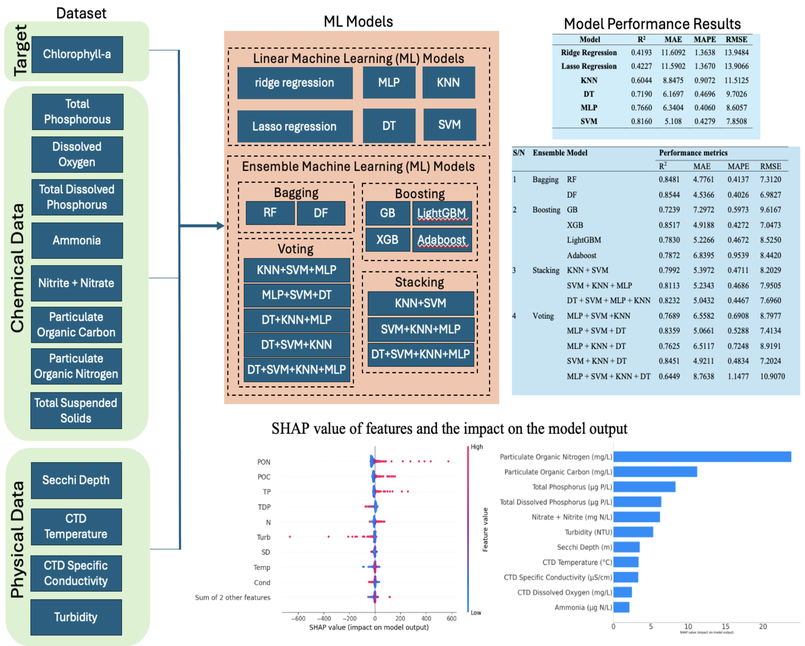
Methodology



Location Map of Seven Data Collection Stations

Variables	Min	Max	Mean	Median	Standard Deviation
Secchi Depth (m)	0.00	6.50	1.15	0.90	0.917
CTD Temperature (°C)	2.90	29.70	22.02	22.90	0.134
CTD Specific Conductivity (µS/cm)	19.90	583.30	297.77	280.40	2.314
CTD Dissolved Oxygen (mg/L)	4.20	13.04	7.58	7.60	0.039
Turbidity (NTU)	0.68	1148.00	19.77	9.30	55.561
Total Phosphorus (µgP/L)	4.00	2482.24	77.83	48.14	132.919
Total Dissolved Phosphorus (µgP/L)	0.16	273.58	20.42	9.27	28.248
Ammonia (µgN/L)	0.04	2108.70	33.61	12.25	87.502
Nitrate + Nitrite (mgN/L)	0.00	9.45	0.88	0.34	1.341
Particulate Organic Carbon (mg/L)	0.14	219.34	2.53	1.30	10.576
Particulate Organic Nitrogen (mg/L)	0.01	40.93	0.43	0.21	1.895
Total Suspended Solids (mg/L)	0.82	540.80	17.99	10.20	33.228
Chlorophyll-a (µg/L)	0.71	678.40	30.59	16.14	50.701

Characteristics of the Lake Erie Dataset

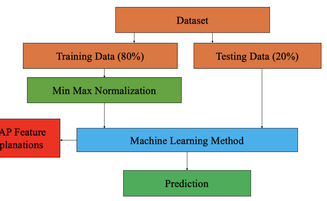


Overview of Methodology Including ML Models Tested SHAP Diagrams and Model Performances

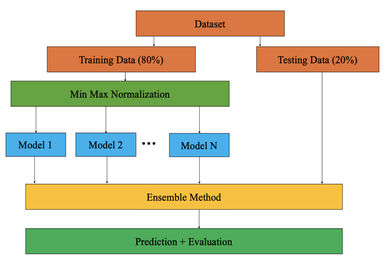
Model Performance Results

Model	R^2	MAE	MAPE	RMSE
Ridge Regression	0.4193	11.6092	1.3638	13.9484
Lasso Regression	0.4227	11.5902	1.3670	13.9066
KNN	0.6044	8.8475	0.9072	11.5125
DT	0.7190	6.1697	0.4696	9.7026
MLP	0.7660	6.3404	0.4060	8.6057
SVM	0.8160	5.108	0.4279	7.8508

S/N	Ensemble Model		Performance metrics			
			R^2	MAE	MAPE	RMSE
1	Bagging	RF	0.8481	4.7761	0.4137	7.3120
		DF	0.8544	4.5366	0.4026	6.9827
2	Boosting	GB	0.7239	7.2972	0.5973	9.6167
		XGB	0.8517	4.9188	0.4272	7.0473
		LightGBM	0.7830	5.2266	0.4672	8.5250
		Adaboost	0.7872	6.8395	0.9539	8.4420
3	Stacking	KNN + SVM	0.7992	5.3972	0.4711	8.2029
		SVM + KNN + MLP	0.8113	5.2343	0.4686	7.9505
		DT + SVM + MLP + KNN	0.8232	5.0432	0.4467	7.6960
4	Voting	MLP + SVM +KNN	0.7689	6.5582	0.6908	8.7977
		MLP + SVM + DT	0.8359	5.0661	0.5288	7.4134
		MLP + KNN + DT	0.7625	6.5117	0.7248	8.9191
		SVM + KNN + DT	0.8451	4.9211	0.4834	7.2024
		MLP + SVM + KNN + DT	0.6449	8.7638	1.1477	10.9070

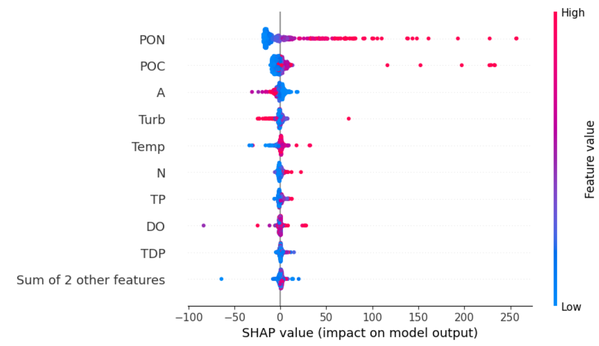


Architecture for Individual Models



Architecture for Ensembled Models Using the Voting and Stacking Classifiers

Results



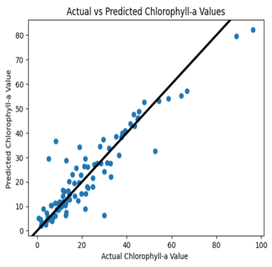
Beeswarm Plot for XGBoost Ensemble Model

ML Model	Input features (mean|SHAP|)				
Lasso	TP (25)	Turb (6.5)	A (2.0)	TDP (1.5)	POC (0.5)
Ridge	TP (35)	Turb (10)	A (2.5)	TDP (2.0)	POC (0.7)
KNN	N (6.0)	TDP (3.5)	DO (2.5)	Temp (2.0)	Cond (1.1)
DT	PON (20.0)	POC (6.0)	A (4.0)	Turb (2.5)	DO (2.0)
MLP	TP (15.0)	Turb (12.2)	A (10.0)	POC (5.3)	PON (2.0)
AdaBoost	PON (14.5)	POC (2.2)	A (1.8)	Temp (0.3)	N (0.1)
LightGBM	PON (17.7)	POC (7.5)	A (2.8)	Turb (2.7)	TP (2.5)
RF	PON (9.5)	POC (6.0)	A (2.5)	Turb (1.8)	Temp (1.5)
SVM	PON (24)	POC (12)	TP (7.5)	TDP (6.0)	N (5.9)
GB	PON (25)	POC (6)	TP (5)	Turb (5)	A (4.5)
XGBoost	PON (18)	POC (7)	A (3)	Turb (2.7)	Temp (2.5)
DF	PON (14)	POC (5)	A (2)	Turb (1.5)	N (1)

SHAP Values for Top 5 Features for Each Model

Model	Training runtime (s)
Lasso	1.0s
Ridge	1.0s
MLP	144.9s
RF	5.2s
LightGBM	7.0s
SVM	12.7s
Adaboost	5.2s
DT	1.0s
KNN	0.5s
DF	561.8s
GB	1.0s
XGBoost	1.0s

Training Time for Each Model



Actual vs Predicted Chlorophyll-a Values

Conclusions

- Most accurate **linear model** -> **SVM**, achieves **R^2 value** of **0.8160**
- **Ensemble models** more **accurate** than linear models with **DF** and **XGBoost** achieving **R^2** values of **0.851 and 0.854 respectively**
- **Fusion** of **weak learners** using **voting** and **stacking** regressors **improves accuracies**
- **Explainable AI** techniques and **decoding** the **black box** of machine learning models using **SHAP** reveals that the concentrations of **particulate organic carbon** and **particulate organic nitrogen** are **most important** factors in predictions
- Future Works: Developing a real-time monitoring system for algal blooms

Selected References

Ai, H., Zhang, K., Sun, J., & Zhang, H. (2023). Short-term lake erie algal bloom prediction by classification and regression models. *Water Research*, *232*, 119710. https://doi.org/10.1016/j.watres.2023.119710

Busari, I., Sahoo, D., Harmel, R. D., & Haggard, B. E. (2024). Prediction of chlorophyll-a as an index of harmful algal blooms using machine learning models. *Journal of Natural Resources and Agricultural Ecosystems*, *2*(2), 53-61. https://doi.org/10.13031/jnrae.15812

Cao, H., Han, L., & Li, L. (2022). A deep learning method for cyanobacterial harmful algae blooms prediction in taihu lake, china. *Harmful Algae*, *113*, 102189. https://doi.org/10.1016/j.hal.2022.102189

Jeong, B., Chapeta, M. R., Kim, M., Kim, J., Shin, J., & Cha, Y. (2022). Machine learning-based prediction of harmful algal blooms in water supply reservoirs. *Water Quality Research Journal*, *57*(4), 304-318. https://doi.org/10.2166/wqrj.2022.019

Molinari, C. (2024, January 12). *Chilean algae bloom continues to cause salmon mortalities, hitting Blumar with USD 8.5 million loss.* (n.d.). Retrieved July 18, 2024, from https://www.seafoodsource.com/news/premium/aquaculture/chilean-algae-bloom-continues-to-cause-salmon-mortalities-hitting-blumar-with-usd-8-5-million-loss

Patel, J. K., & Parshina-Kottas, Y. (2017, October 3). Miles of Algae Covering Lake Erie. *The New York Times*. https://www.nytimes.com/interactive/2017/10/03/science/earth/lake-erie.html

Saxena, R. (2017, April 29). *Toxic algae on the rise as our oceans warm*. Ars Technica. https://arstechnica.com/science/2017/04/harmful-algal-blooms-occur-more-often-now-that-oceans-are-warming/

Shukla, R. K., Boegman, L., & Kumar, P. (2024). Application of interpretable machine learning and causal discovery to understand chlorophyll-a variation in a large shallow lake. *SSRN*. https://doi.org/10.2139/ssrn.4821781

Stumpf, R. (2022, November 16). *2022 Lake Erie Algal Bloom More Severe than Predicted by Seasonal Forecast*. (2021, November 5). NCCOS Coastal Science Website. https://coastalscience.noaa.gov/news/2022-lake-erie-algal-bloom-more-severe-than-predicted-by-seasonal-forecast/

Acknowledgements

I would like to thank the SSTP program and the Belin Blank Center for giving me this opportunity. I would also like to thank Dr. Demir for mentoring me in his lab and allowing me to gain research experience. Lastly, I would also like to thank Dr. Mermer and Dr. Sermet for their support and mentorship on this project.