A Multimodal Approach for Evaluating Algal Bloom Severity using Deep Learning

Fei Zhao, Chengcui Zhang, and Tianyang Wang
Department of Computer Science
The University of Alabama at Birmingham
Birmingham, USA
{larry5, czhang02, tw2}@uab.edu

***Abstract*—Harmful algal blooms (HABs) can have detrimental impacts on aquatic ecosystems, human health, and the economy. This paper presents a novel multimodal deep learning approach for assessing the severity levels of HABs, which will help to take necessary measures to mitigate the negative impacts. Unlike the other SOTA methods, the proposed method leverages three modalities: satellite image, elevation, and temperature data, to capture algal information. In particular, it utilizes an Attention-UNet-based encoder for satellite and elevation data, and a BiLSTM encoder for temperature data, to extract effective feature embeddings from respective modalities. In addition, we propose a geometric mean-based multimodal focal loss that modulates loss contributions of different modalities as a function of the confidence of different modalities. Our approach outperforms the SOTA unimodal and ensemble methods on tick-tick bloom (TTB) dataset, achieving a region-averaged root mean squared error (RA-RMSE) score of 0.8165.**

***Index Terms*—Satellite image, Multimodal, Deep learning, Algal bloom, Severity level**

I. Introduction

HABs pose a threat to aquatic ecosystems and human health due to their potential to cause oxygen depletion, fish kills, and release of harmful toxins. Accurately evaluating HABs is essential for environmental management and public health. Recently, deep learning techniques have been adopted to detect HABs using various modalities, e.g., satellite images. Although unimodality approaches show promise, they often struggle to capture the complex and dynamic nature of algal blooms, motivating multimodal approaches that can potentially better capture the complex interplay of multiple sources.

In this paper, we aim to use an unprecedentedly comprehensive set of modalities including satellite images, elevation, and temperature data, to assess HABs severity levels. Satellite images offer broad spatial coverage of water bodies; sequential temperature data reveal temporal temperature buildup; and elevation data is integrated with satellite images to enhance the model's ability to accurately capture water body areas, which improves the overall model performance. Two samples are presented in Fig. 1.

To the best of our knowledge, no prior studies have investigated a deep learning approach that combines satellite images, sequential temperature, and elevation data for HABs severity level classification. We provide a literature review focusing on HABs detection and evaluation in Section II. In Section III, we introduce our deep learning model and novel loss function. Section IV provides experimental results and ablation studies. Section V summarizes our findings and presents future research directions.

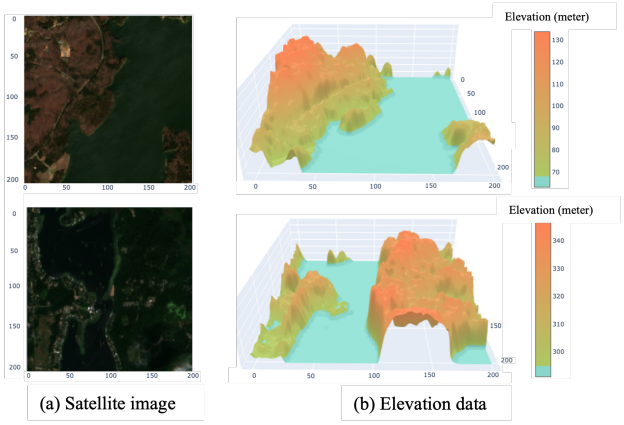

Fig. 1. Each row displays a satellite image (a) and elevation data (b) pair for a single sample. In the 3D visualization (b), light blue areas indicate potential water bodies.

II. Related Work

Several studies have adopted machine learning techniques for algal bloom detection and evaluation using various modalities. Classic methods, such as deep belief networks [1] and support vector machines [2], have relied on small, local datasets that may not be generalizable to other regions. Moreover, the data collection process for required historical water quality can be challenging. More recent studies have utilized optical imagery data, easily acquired from airborne or spaceborne, to address data collection issues [3]–[5]. For example, the authors in [3] proposed a unimodal tree-based LightGBM model that can predict HABs severity by using only RGB satellite images. However, satellite images may have a limited temporal resolution, making it challenging to capture the rapid dynamics of algal blooms that can change within hours or days. Also, algal growth is influenced by various factors, such as temperature and nutrient availability, which cannot be entirely captured by image data alone. The ensemble method [6], one of the best performers of the 2023 tick-tick bloom challenge, proposed to incorporate satellite images, elevation data, and other hand-crafted features, such as hard-coded location clusters of U.S., but not temperature data.

The hard-coded location clusters may not be generalizable to other non-US regions or scalable to larger datasets, and manual feature engineering can be labor-intensive, potentially missing relevant information. Additionally, ensemble methods often combine multiple models, leading to increased complexity and longer training and testing time, which can pose challenges when dealing with large-scale datasets or requiring real-time predictions.

Our proposed is a single-model based deep learning framework that combines an image-elevation stream and a temperature stream for HABs severity level classification. The image-elevation stream, which works with an Attention-UNet-encoder-decoder architecture, detects water bodies in satellite images with assistance of elevation data. The temperature stream uses a BiLSTM [7] model to extract temporal features from sequential temperature data. The high-level features generated from these two streams are complementary and can be combined to improve the overall model prediction accuracy. This approach represents a significant advance over previous unimodal models [3] and has outperformed the ensemble-based model [6]. Also, training on TTB large dataset enables our model to be more generalizable to unseen samples from different inland water bodies.

III. Methodology

A. Modality and Network Architecture

This study uses three modalities to predict HABs severity level: satellite images, elevation, and temperature. Each modality contains valuable information to address different aspects of the HABs assessment task. Therefore, we design a two-stream deep neural network architecture, shown in Fig. 2, consisting of an image-elevation stream and a temperature stream. The deep embeddings extracted from both branches are concatenated and sent through Subnet 3, consisting of 3 fully connected layers and 2 dropout layers, to predict HABs severity level.

Satellite Image and Elevation: The satellite image modality can provide valuable visual information about water bodies, such as color, texture, and spatial distribution, directly contributing to HABs severity level classification. However, whole-image CNN networks could extract unnecessary visual features (noises) beyond water bodies that may harm classification performance. To address this, we design an Attention-UNet-based encoder-decoder subnetwork (VGG16 backbone) [8], which takes a pair of satellite image and elevation data as input and generates a water mask. The attention mechanism [9] allows the model to selectively focus on important features (water body areas) in the input data. Here, elevation data may not directly contribute to HABs severity level prediction, but it can help to locate water body areas in the satellite image. Specifically, areas with relatively lower elevation values are more likely to be covered by water as shown in Fig. 1. TTB dataset contains samples with associated latitude, longitude, event date, and HABs severity level (1-5). We used 1000 meter buffer to extract both elevation data from Copernicus DEM set [10] and RGB satellite images with a shape of (204, 204, 3) from Sentinel set [10].

Temperature: The temperature is known to influence the growth of algae because it can affect the metabolic rate, growth rate, and reproductive rate of algae. Furthermore, the temperature modality captures the temporal patterns of algal blooms. By incorporating temperature data, we can better capture the dynamics of algal blooms over time and make more accurate predictions. For each sample, we extracted hourly temperature values for 14 days leading up to the event date. Therefore, for each sample, 336 Celsius values are retrieved from HRRR set [10] and form the temperature modality input feature. The processed temperature data is then fed into a BiLSTM encoder with 128 hidden states and 2 layers, which captures both past and future dependencies of algal blooms by processing the sequence data in both forward and backward directions.

B. Multimodal Focal Loss

Focal loss [11] is designed to reduce the loss contribution of samples that are correctly classified with high confidence. Its modulating factor is a function of the predicted probability of the correct class, where high probabilities receive lower weights, and low probabilities receive higher weights. Inspired by the work in [12], we propose a geometric mean-based multimodal focal loss. The intuition is that, similar to focal loss, for each modality's subnet, e.g., Subnet 1 of $M_{I,E}$ in Fig. 2, its loss contribution of samples, if correctly classified by its own modality's subnet or by the other modality's subnet with high confidence, e.g., Subnet 2 of M_T, should be reduced. Here *I, E, T* represent image, elevation, and temperature modalities, respectively. The original focal loss is defined as follows:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma} log(p_t) \tag{1}$$

where p_t is the predicted probability of the true class, α_t is the weighting factor for each class, and γ is the focusing parameter that controls the rate at which the loss decreases as the predicted probability increases. We modify focal loss and form a geometric mean-based multimodal focal loss (MFL):

$$MFL_{I,E}(p_t, q_t) = -\alpha_t(1 - q_t * \sqrt{p_t \cdot q_t})^{\gamma} log(p_t) \tag{2}$$

$$MFL_T(q_t, p_t) = -\alpha_t(1 - p_t * \sqrt{p_t \cdot q_t})^{\gamma} log(q_t) \tag{3}$$

The predicted probability of the true class for a sample by $M_{I,E}$ is represented as p_t while that of M_T is represented as q_t. The geometric mean of p_t and q_t is used to effectively manage situations where one probability is high while the other is low. The total loss of the proposed multimodal network is a weighted combination of *Dice loss* [13] and *Cross Entropy loss* (CE) for water segmentation, $CE_{I,E,T}$ *loss* for Subnet 3, $MFL_{I,E}$ loss for Subnet 1 and MFL_T loss for Subnet 2. Both modified multimodal focal losses and $CE_{I,E,T}$ are for the severity level prediction, but only the prediction from the combined Subnet 3 is used as the final prediction result.

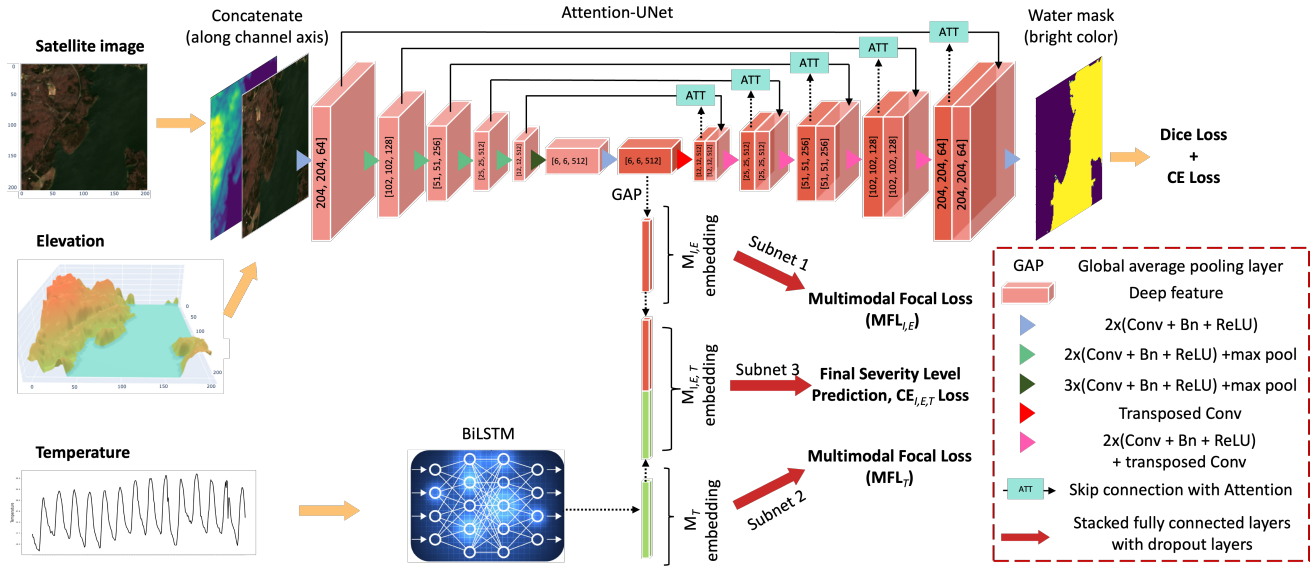

Fig. 2. The two-stream network architecture: the upper stream is an Attention-UNet-based encoder-decoder subnetwork for water segmentation, which takes RGB satellite images and the corresponding elevation data as input. The $M_{I,E}$ embedding extracted from the upper stream represents high-level features of the satellite image and elevation modalities. The bottom stream is a BiLSTM subnetwork that extracts high-level features of temperature modality. Those high-level features are concatenated to form a multimodal embedding for the downstream HABs severity level prediction task.

C. Experiments

Tick-tick bloom (TTB) dataset defined 4 geographic regions, including south, west, northeast, and midwest. Since missing modality problem is not within the research scope of this paper, 11,688 samples in TTB that have all three modalities are retrieved, including 6,730 samples from south, 2,502 samples from west, 1,581 samples from midwest, and 875 samples from northeast. Ground truth severity values are from 1 (no algal bloom) to 5, accounting for 42.1%, 19.9%, 17.4%, 20.2%, and 0.4% of the total samples. The dataset was split into training, validation, and test sets by the ratio of 8:1:1, while maintaining the same region and severity class distributions in each dataset. TTB's performance metric is RA-RMSE. RMSE is the square root of the mean of squared differences between estimated and observed values. Eq. 4 shows the error metric (a lower value is better).

$$\text{RA-RMSE} = \frac{1}{M}\sum_{i=1}^{M}\sqrt{\frac{1}{N_i}\sum_{j=1}^{N_i}(y_{i,j}-\hat{y}_{i,j})^2} \tag{4}$$

where M represents the number of regions, and N_i represents the number of samples in the i-th region. $y_{i,j}$ is the actual severity level of the j-th sample in the i-th region, and $\hat{y}_{i,j}$ is the predicted severity level for that sample.

Our model is trained with a learning rate of 0.0001 and a batch size of 14. Learning rate decay is applied. All models are trained for a maximum of 100 epochs with early stopping based on the validation accuracy. Image and elevation data are augmented by horizontal/vertical flips and 0˜30° rotations.

IV. Results and Ablation Study

Table I presents the results of our model, along with several SOTA methods for comparison. M represents Modality, and the subscripts *I, E, T, Others* stand for satellite image, elevation, temperature, and other modalities (e.g., geographic cluster index [6]), respectively. The subscript *Augment* represents a data augmentation strategy applied during training. It is worth noting that the dataset is imperfect, since there is a significant portion (8.8%) of image samples that do not show any water body but have at least one nearby, for which we search for its nearby water body. Due to the time-consuming data retrieval process, we use a coarse-grid search and take a step size of 1000-meter in each of the four directions (N/W/E/S), and stop as soon as a non-trivial water body is found or none is found after taking one step of search in each direction. According to our observation, in most cases taking only one step would find a non-trivial water body. If a water body is found, we use the new location to retrieve all modalities for that sample. The term "Single" in Table I refers to HABs severity level classification task, while "Multi" denotes a combination of the HABs severity level classification and water segmentation tasks. Our approach achieves a lower RA-RMSE compared to SOTA methods [3] and [6], demonstrating its superiority. We also tested the effect of the hard-coded location cluster index as adopted in [6] by incorporating it into Subnet 3 of $M_{I,E,T,Augment}$ model, which yields an even lower RA-RMSE of 0.7979. However, as aforementioned, its generalizability to samples from non-US regions is highly questionable.

A. Ablation Study of Elevation Data

We also examined the impact of raw elevation data versus processed (min-max normalized) elevation data in our multimodal approach. We use $M_{I,E}$ model to compare these two types of input. The performances is shown in Table II. The processed elevation data-based model outperforms the raw data-based model, as raw elevation data has more

TABLE I
RESULTS

Modality	Model	Task	RA-RMSE
M_I	LightGBM$_{TTB}$ [3]	Single	1.4078
$M_{I,E,Others}$	Ensemble$_{SOTA}$ [6]	Single	0.8762
M_T	BiLSTM	Single	1.0421
M_I	CNN (VGG16)	Single	0.9475
M_I	UNet	Multi	0.9144
$M_{I,E}$	UNet	Multi	0.8622
$M_{I,E,T}$	UNet + BiLSTM	Multi	0.8310
$M_{I,E,T,Augment}$	UNet + BiLSTM	Multi	**0.8165**

heterogeneity. For instance, in Fig. 1, the water areas of the first row sample have elevation values $\sim$ 70 meters, while the second row's water areas reach 300 meters. The second row's water areas have higher elevation values than the first row sample's maximum. The region-based min-max normalization allows the model to focus on areas with relatively lower elevation values in an elevation map. The result of using elevation data only for prediction is ignored because it is far worse than the others.

TABLE II
ABLATION STUDY OF ELEVATION DATA

Data Type	Modality	RA-RMSE
Raw Elevation	$M_{I,E}$	0.8751
Processed Elevation	$M_{I,E}$	**0.8622**

B. Ablation Study of Temperature Data

We further evaluated the impact of different temporal windows on our temperature-based model performance. Two temporal windows are tested:1) 7-days and 2) 14-days. Also, for the temperature data each day, two types of operation are applied: 1) 24 raw temperature data each day and 2) one mean value each day, which is the highest mean value of all 8-hour consecutive windows of a day. The motivation is to try to focus on the warmest window of a day, with the 8-hour representing the shortest daytime in U.S., except Alaska. We used $M_{I,E,T}$ model for this experiment, with results in Table III. The 14-day models consistently outperformed the 7-day models, and raw temperature-based models surpassed processed ones, suggesting that a longer temporal window and raw temperature data improve the model's accuracy. We also tested using all the 8 temperature values in the 8-hour window per day, or one temperature reading at a fixed time per day as features, but the performance is not as good as the ones presented.

TABLE III
ABLATION STUDY OF TEMPERATURE DATA

Data Type	Modality	RA-RMSE
7 days (Processed)	$M_{I,E,T}$	0.8563
14 days (Processed)	$M_{I,E,T}$	0.8373
7 days (Raw)	$M_{I,E,T}$	0.8479
14 days (Raw)	$M_{I,E,T}$	**0.8310**

C. Ablation Study of Loss

We also evaluate the impact of our proposed loss function using $M_{I,E,T,Augment}$ model, comparing it to a regular loss function CE. The results are presented in Table IV.

TABLE IV
ABLATION STUDY OF LOSS

Loss Function	Modality	RA-RMSE
Regular Loss (CE)	$M_{I,E,T,Augment}$	0.8577
Proposed Loss	$M_{I,E,T,Augment}$	**0.8165**

V. CONCLUSIONS & FUTURE WORK

We propose the first comprehensive multimodal deep learning approach to predict algal bloom severity levels, incorporating satellite image, elevation, and temperature modalities. This single-model (vs ensemble) based approach achieved superior performance compared to SOTA unimodal and ensemble methods, with the lowest RA-RMSE value. Ablation studies highlighted the significance of processed elevation data, raw temperature data, and the proposed loss function. Factors that negatively affect the accuracy include cloud occlusion and/or missing water bodies in the image, etc. Future work could include adding new modalities such as wind speed or investigating alternative fusion strategies and deep learning architectures for improved performance and interpretability, such as transformer-based networks.

REFERENCES

[1] F. Zhang et al., "Deep-learning-based approach for prediction of algal blooms," *Sustainability*, vol. 8, no. 10, pp. 1060, 2016.

[2] J. H. Kim et al., "Improving the performance of machine learning models for early warning of harmful algal blooms using an adaptive synthetic sampling method," *Water Research*, vol. 207, pp. 117821, 2021.

[3] Drivendata-how to predict harmful algal blooms using lightgbm and satellite imagery, https://drivendata.co/blog/tick-tick-bloom-benchmark.

[4] J. Shin et al., "Convolutional neural network model for discrimination of harmful algal bloom (hab) from non-habs using sentinel-3 olci imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 191, pp. 250–262, 2022.

[5] X. Wang et al., "Algaemask: an instance segmentation network for floating algae detection," *Journal of Marine Science and Engineering*, vol. 10, no. 8, pp. 1099, 2022.

[6] Drivendata-tick tick bloom: harmful algal bloom detection challenge, https://www.drivendata.org/competitions/143/tick-tick-bloom/page/649.

[7] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[8] O. Ronneberger et al., "U-net: convolutional networks for biomedical image segmentation," in *Proc. of Medical Image Computing and Computer-Assisted Intervention–MICCAI, Part III 18*. Springer, 2015, pp. 234–241.

[9] O. Oktay et al., "Attention u-net: learning where to look for the pancreas," *arXiv:1804.03999*, 2018.

[10] Drivendata-harmful algal bloom detection challenge: problem description, https://www.drivendata.org/competitions/143/tick-tick-bloom/page/650/#climate-data.

[11] T. Y. Lin et al., "Focal loss for dense object detection," in *Proc. of the IEEE Intl. Conf. on Computer Vision*, 2017, pp. 2980–2988.

[12] A. George and S. Marcel, "Cross modal focal loss for rgbd face anti-spoofing," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, 2021, pp. 7882–7891.

[13] C. H. Sudre et al., "Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations," in *Proc. of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, held in conjunction with MICCAI*, 2017, pp. 240–248.