Deep Learning for HABs Prediction with Multimodal Fusion

Fei Zhao
The University of Alabama at Birmingham
Birmingham, AL, USA
larry5@uab.edu

Chengcui Zhang
The University of Alabama at Birmingham
Birmingham, AL, USA
czhang02@uab.edu

ABSTRACT

Harmful Algal Blooms (HABs) present significant environmental and public health threats. Recent machine learning-based HABs monitoring methods often rely solely on unimodal data, e.g., satellite imagery, overlooking crucial environmental factors such as temperature. Moreover, existing multi-modal approaches grapple with real-time applicability and generalizability challenges due to the use of ensemble methodologies and hard-coded geolocation clusters. Addressing these gaps, this paper presents a novel deep learning model using a single-model-based multi-task framework. This framework is designed to segment water bodies and predict HABs severity levels concurrently, enabling the model to focus on areas of interest, thereby enhancing prediction accuracy. Our model integrates multimodal inputs, i.e., satellite imagery, elevation data, temperature readings, and geolocation details, via a dual-branch architecture: the Satellite-Elevation (SE) branch and the Temperature-Geolocation (TG) branch. Satellite and elevation data in the SE branch, being spatially coherent, assist in water area detection and feature extraction. Meanwhile, the TG branch, using sequential temperature data and geolocation information, captures temporal algal growth patterns and adjusts for temperature variations influenced by regional climatic differences, ensuring the model's adaptability across different geographic regions. Additionally, we propose a geometric multimodal focal loss to further enhance representation learning. On the Tick-Tick Bloom (TTB) dataset, our approach outperforms the SOTA methods by 15.65%.

CCS CONCEPTS

• **Information systems** → **Geographic information systems**; • **Computing methodologies** → **Computer vision**; **Neural networks**; • **Applied computing** → **Environmental sciences**.

KEYWORDS

Geolocation, Computer Vision, Deep Learning, Harmful Algal Blooms

ACM Reference Format:
Fei Zhao and Chengcui Zhang. 2023. Deep Learning for HABs Prediction with Multimodal Fusion. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3589132.3628370

SIGSPATIAL'23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11…$15.00
https://doi.org/10.1145/3589132.3628370

1 INTRODUCTION

HABs are a growing environmental and public health concern. These events can lead to severe consequences, such as oxygen depletion in water bodies, massive fish kills, and the release of harmful toxins contaminating drinking water. Given HABs' rapid onset and impacts, timely interventions are crucial. Central to early intervention efforts is the precise prediction of HABs occurrence.

Traditional methods of monitoring HABs, such as manual sampling and Sonde Buoys, lack scalability for vast regions and are labor-intensive. Recently, machine learning techniques have emerged, predominantly focusing on unimodal data, e.g., satellite imagery [1]. However, this singular focus often misses crucial environmental factors such as temperature variations, compromising prediction accuracy. The shift to multi-modal data has seen methods like the ensemble approach in [2], which leveraged satellite images and elevation data but overlooked temperature data. Another ensemble in [3] combined KNN and GBDT models, utilizing geolocation and temperature data. While promising, both of them depend on hard-coded geolocation clusters and ensemble models, facing challenges in model generalizability and real-time applicability.

To address these challenges, we propose the first work integrating four modalities using deep learning: satellite imagery, elevation data, temperature data, and geolocation data. Recognizing the distinct characteristics of these modalities, we strategically split them into two branches. The SE branch captures spatial features from satellite and elevation data, while the TG branch focuses on temporal temperature patterns and regional variations. Experimental results show that all the four modalities significantly contribute to HABs severity prediction. This architecture is further enhanced by our proposed geometric multimodal focal loss, which supervises individual branches, ensuring optimal representation learning. This design enhances the robustness of our model, allowing for a comprehensive understanding of HABs' intricate dynamics.

2 METHODOLOGY

2.1 Model Architecture

Satellite-Elevation (SE) Branch: Satellite imagery provides detailed visual insights into water bodies, capturing essential attributes such as color and texture. Elevation data, while not a direct predictor of HABs severity, complements satellite imagery by enhancing water body segmentation. By emphasizing areas of relatively lower elevation, it pinpoints potential water regions, reinforcing the model's focus of learning. For this reason, we combine them in the SE branch using an early fusion strategy for efficiency. We concatenate satellite imagery and elevation data along the channel axis to form the SE branch input, as shown in Fig. 1. This branch utilizes a UNet architecture with the Swin-Transformer [7] as the encoder. The Swin-Transformer is capable of handling long-range dependencies in spatial data, effectively detecting water bodies and

integrating elevation data. This produces the spatial representation: the embedding $M_{S,E}$, shown in Fig. 1.

Temperature-Geolocation (TG) Branch: Temperature plays a key role in algal growth dynamics. Our model incorporates hourly temperature values from the 14 days leading up to the event date, offering a detailed temporal profile. Yet, the impact of temperature on algal growth can vary with geographic locations due to regional climatic differences and other environmental factors. Thus, to contextualize the temperature data, geolocation data, represented as latitude and longitude coordinates, is integrated. In the TG branch, geolocation data is transformed into a 24-bit binary code using Geohash [4]. After being mapped into a geo-feature space by a linear layer, this high-level geo-feature is combined with the high-level features generated by the temperature BiLSTM-Transformer [6]. The combined features are then processed through another Transformer module [5] to derive a robust temporal representation of the HABs dynamics: the embedding $M_{T,G}$, shown in Fig. 1.

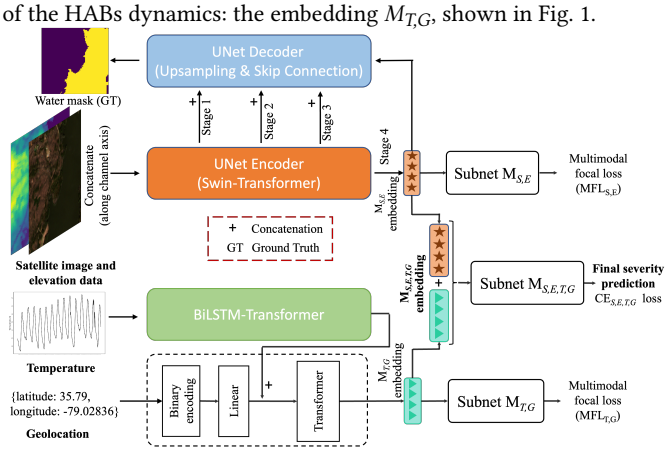

Figure 1: End-to-End architecture of the proposed model

2.2 Multi-modal Focal Loss

Inspired by Focal loss, our proposed multi-modal focal loss dynamically adjusts the loss contribution from each branch according to its prediction confidence. Specifically, if one branch can confidently classify a sample correctly, the contribution from the other branch to the overall loss will be moderated.

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

$$MFL_{T,G}(q_t, p_t) = -\alpha_t (1 - p_t * \sqrt{p_t \cdot q_t})^{\gamma} log(q_t) \quad (2)$$

In the above equations, p_t and q_t represent the predicted probabilities of the true class for a sample via Subnet $M_{S,E}$ and Subnet $M_{T,G}$. The scaling factor α_t balances the contribution of positive and negative samples, while γ serves as a focusing parameter, modulating the rate at which easy samples are down-weighted.

To generate the final prediction of HABs severity levels, the embeddings from the SE branch and the TG branch are concatenated and fed into Subnet $M_{S,E,T,G}$. The three subnets, shown in Fig. 1, consist of stacked fully connected layers with dropout. The overall loss combines Dice loss, Cross Entropy loss (CE), and multi-modal focal losses. While both multi-modal focal losses and $CE_{S,E,T,G}$ target severity prediction, the final result comes from Subnet $M_{S,E,T,G}$ only.

3 EXPERIMENTS AND RESULTS

The TTB dataset underpins our HABs severity assessment. Each sample in this dataset comprises a satellite image, elevation data, sequential temperature readings, a pair of latitude and longitude coordinates, a HABs occurrence date, and the corresponding ground truth severity level. Both the satellite image and its associated elevation data span a 2km x 2km geographic area. The temperature data comprises 336 consecutive hourly readings leading up to the HABs event. The TTB dataset is distributed across four U.S. regions: south (6,730 samples), west (2,502 samples), northeast (875 samples), and midwest (1,581 samples). Severity levels are categorized from 1 (no algal bloom) to 5. They constitute 42.1%, 19.9%, 17.4%, 20.2%, and 0.4% of the total, respectively. For evaluation purposes, the dataset was partitioned into training, validation, and test sets at a ratio of 8:1:1. This split was carefully done while maintaining the same region and severity class distributions in each of these subsets.

Table 1: Results

Modality	Model	RA-RMSE
M_S	LightGBM$_{SOTA}$[1]	1.4078
$M_{S,T,Others}$	Ensemble$_{SOTA}$[3]	1.0690
$M_{S,E,Others}$	Ensemble$_{SOTA}$[2]	**0.8762**
M_S	Swin-Trans-UNet	0.8901
$M_{S,E}$	Swin-Trans-UNet	0.8186
$M_{S,E,T}$	Swin-Trans-UNet, BiLSTM-Trans	0.7898
$M_{S,E,T,G}$	**Swin-Trans-UNet, BiLSTM-Trans**	**0.7391**

The results of our proposed model in comparison to SOTA methods [1–3] are shown in Table 1. In this table, 'M' denotes Modality, where the subscripts *S, E, T, G, Others* represent satellite image, elevation, temperature, geolocation, and additional modalities (such as hand-crafted geographic cluster index [2, 3]), respectively. We adopt the same Region-Averaged Root Mean Squared Error (RA-RMSE) from [1–3] as the metric (**a lower value is better**). Our model exhibits an impressive performance by achieving the lowest RA-RMSE of 0.7391, which is 15.65% lower than SOTA methods. The ablation study shows that the incorporation of additional modalities consistently enhances the performance. This highlights the inherent advantage of a multimodal approach over unimodal methods, emphasizing the importance of leveraging diverse data sources for a holistic understanding and prediction of HABs severity.

4 CONCLUSION

We introduce a novel deep learning model that integrates multiple modalities for HABs prediction. Our model's performance on the TTB dataset underscores its potential as a leading tool in the fight against HABs. As we move forward, several avenues beckon further exploration. Incorporating additional modalities (e.g., wind, humidity), further refinement of our multimodal fusion loss, and the exploration of other deep learning architectures could further enhance the model's predictive capabilities.

REFERENCES

[1] DrivenData. 2023. *How to predict harmful algal blooms using LightGBM and satellite imagery.* https://drivendata.co/blog/tick-tick-bloom-benchmark
[2] DrivenData. 2023. *Tick Tick Bloom Challenge.* https://www.drivendata.org/competitions/143/tick-tick-bloom/page/649
[3] DrivenData. 2023. *Tick Tick Bloom Challenge: Sheep.* https://www.drivendata.org/competitions/143/tick-tick-bloom/leaderboard
[4] J.B. Zhang et al. 2021. POI-Transformers: POI Entity Matching through POI Embeddings by Incorporating Semantic and Geographic Information. (2021).
[5] Vaswani et al. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).
[6] Z.H. Huang et al. 2020. TRANS-BLSTM: Transformer with bidirectional LSTM for language understanding. *arXiv preprint arXiv:2003.07000* (2020).
[7] Z. Liu et al. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 9992–10002.