Algae Identification Towards Automated Classification

Minju Kim
MS in Data Science
mk159@iu.edu

Jihye Shin
PhD in Informatics
shinjih@iu.edu

Ruoying Yuan
PhD in Computer Science
ry1@iu.edu

Abstract

The critical need for advanced monitoring and classification of algae, particularly during the peak bloom periods in Monroe Lake, Bloomington, has driven the development of an automated classification system utilizing cutting-edge computer vision technologies. Traditional methods, such as FlowCam, often fail to capture a significant percentage of algae due to stringent measurement criteria, necessitating labor-intensive manual corrections. Our project addresses these challenges by implementing sophisticated machine learning models, including Convolutional Neural Networks (CNN), ResNet, U-net, and MobileNet, which significantly enhance the detection and classification accuracy of algae species. The study involved preprocessing a dataset of 3,060 images to isolate 1,287 images with relevant taxa, followed by rigorous augmentation to balance the dataset for effective model training. Comparative analysis of the models demonstrated that ResNet provided the highest accuracy and reliability in classifying algae, with promising implications for improving water quality management and public health protection. This work not only showcases the feasibility of applying artificial intelligence in environmental monitoring but also sets a precedent for future research in automated natural resource management.

I. Introduction

Increasing concerns about water quality in Monroe Lake, Bloomington—especially during the warmer months when algae blooms are prevalent—underscore the critical need for advancements in algae monitoring and classification. The majority of the area's residents rely on this water source, which is often compromised by the limitations of current detection methods. Predominantly used FlowCam devices, which classify algae based on a statistical method involving 60 different measurements including a narrow range of 'edge gradients,' often exclude about 90% of potential algae images. This exclusion is due to the stringent criteria not aligning with the natural variability of algae, as noted by user feedback. Such high rejection rates necessitate significant manual review and adjustment, undermining the efficiency of the classification process. This project addresses these limitations by employing computer vision technologies to improve the accuracy and efficiency of algae identification processes. By developing an automated classification system, we aim to minimize manual errors, enhance operational efficiency, and provide faster responses to water quality issues that critically impact public health.

II. Background and Related Work

Traditional approaches to algae classification often involve manual observation or semi-automated methods that rely on strict physical measurement criteria, which can significantly limit the scope of detection and recognition. Studies like those conducted by Ning et al. [1] have utilized conventional machine learning techniques such as Support Vector Machines and Random Forests to classify algae. These methods, however, struggle with the high variability and complex nature of algae appearances.

The adoption of advanced imaging technologies and computer vision in algae classification has been significantly explored in recent literature. Otálora et al. [2] and Abdullah et al. [3] have demonstrated the application of basic CNNs for microalgae classification, showcasing the potential of deep learning to surpass the limitations of traditional image processing techniques. Moreover, Xu et al. [4] highlighted how CNNs can achieve high accuracy in algae classification, even with datasets that are smaller and contain highly detailed images.

Recent studies by Chong et al. [5, 6] delve deeper into the identification and preprocessing challenges associated with microalgae. Their research outlines significant advancements in the digital image processing of isolated microalgae by incorporating complex classification algorithms, paving the way for more accurate and automated systems. These studies contribute essential insights into the evolving landscape of image-based algae classification, emphasizing the need for innovative solutions that can adapt to the intricate nature of such organisms.

The emergence of transformer-based models for image recognition, discussed by Dosovitskiy et al. [7] and

Touvron et al. [8], presents a new frontier for enhancing image-based classification systems. These models utilize mechanisms like self-attention to capture contextual relationships within images, which could be particularly beneficial for distinguishing between visually similar algae species.

Our project builds upon these advancements by implementing sophisticated neural network architectures such as ResNet and U-net, which have shown promising results in handling the diverse and complex structures of algae images captured from local water sources. The semi-automated approach using FlowCAM, as described by Mirasbekov et al. [9], aligns with our project's aim to automate algae classification, confirming the viability of combining imaging cytometry with artificial intelligence to improve accuracy and efficiency in environmental monitoring.

III. Methodology

A. Data Preprocessing

Our initial dataset extraction yielded 3,060 images from an established SQLite3 database, featuring 52 distinct algae taxa labels. Under the guidance of our mentor, Jill, we targeted key genera significant to the water quality of Monroe Lake, namely Diatoms (Asterionella, Aulocoseira, Fragillaria, Tabellaria, Cyclotella, Synedra) and Cyanobacteria (Dolichospermum aka anabaena, Cylindrospermopsis, Aphanazomenon, Planktothrix, Lyngbya). This focus refined the dataset to 1,287 images with 19 labels.

Original Label	Mapped Label
10X_FOV 100 Anabaena WILL CRASH	**Anabaena**
Example_Anabaena-coiled_10X_TR	
Example_Anabaena-straight_10X_TR	
Anabaena Curved 10X FOV 100	
Anabaena 10X FOV 100	
Example_Aphanazomenon_10X_TR	**Aphanazomenon**
Example_Asterionella_10X_TR	**Asterionella**
Asterionella 10X FOV100	
Example_Cyclotella_10X_TR	**Cyclotella**
Cyclotella 10X FOV100	
Example_Cylindrospermopsis_10X_TR	**Cylindrospermopsis**
Example_Tabellaria_10X_TR	**Tabellaria**
Example_Lyngbya_10X_TR	**Lyngbya**
Example_Planktothrix_10X_TR	**Planktothrix**
Example_Planktothrix-2_10X_TR	
Example_Synedra_10X_TR	**Synedra**
Synedra 10X FOV100	
Synedra	
Fragillaria 10X FOV100	**Fragillaria**

Table 1. Label Mapping for Algae Image Dataset

To enhance the practicality of the dataset for advanced machine learning applications, a meticulous process of data cleaning was undertaken. This included the removal of duplicate and non-essential entries, followed by a strategic consolidation of the 19 labels into a more focused set of 11 classes. A pivotal adjustment in this phase was the bifurcation of the 'Anabaena' category into 'Straight Anabaena' and 'Coiled Anabaena', recognizing the importance of morphological variations for precise classification.

Fig. 1. Anabaena's Two Distinct Morphologies

Also, our dataset's integrity was further fortified through the extraction of 50 'Black Holes' anomalies—lens artifacts from the imaging process. These anomalies, identified with the astute guidance of our mentor, Jill, were methodically segregated to obviate their potential to undermine the precision of our models. Consequently, this intervention necessitated the introduction of an additional 'Black Hole' class, culminating in a refined schema of 12 categories.

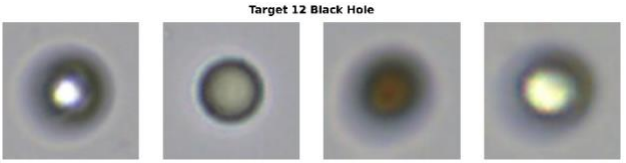

Fig. 2. Target 12 Black Hole

Algae Class	Numerical Label
Straight Anabaena	**1**
Coiled Anabaena	**2**
Aphanazomenon	**3**
Asterionella	**4**
Cyclotella	**5**
Cylindrospermopsis	**6**
Tabellaria	**7**
Lyngbya	**8**
Planktothrix	**9**
Synedra	**10**
Fragillaria	**11**
Black Hole	**12**

Table 2. Numerical Mapping for Algae Classes

The images are all 10x magnification under the FlowCam, but their shapes differ after segmentation by edge gradient. In preparation for the learning phase, we

standardized all images to a uniform resolution of 224x224 pixels using padding techniques. This precaution prevented any image distortion that could have compromised feature detection and classification accuracy. The padding color was chosen based on the predominant color at the border of each image to maintain visual consistency.

An examination of the class distribution for the 12 targeted genera revealed a marked imbalance, which is graphically depicted in the following figure:

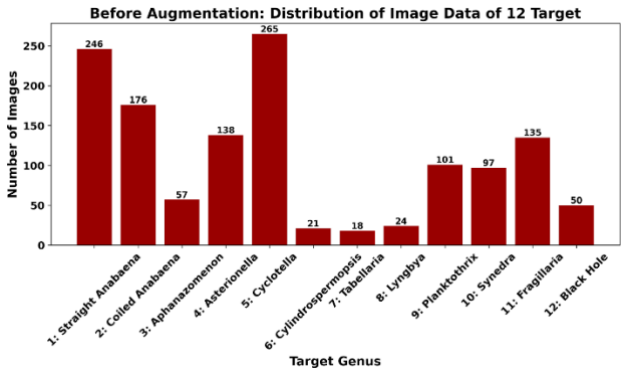

Fig. 3. Before Augmentation: Distribution of 12 Target

B. Data Augmentation

After splitting our dataset into training and test sets, to mitigate the impact of class imbalance identified in the initial data analysis, we executed an augmentation process solely on the training set. This was crucial to ensure the augmentation did not affect the test set, which remained untouched to preserve the validity of our model evaluation. We increased the representation of underrepresented classes, thereby balancing the dataset by expanding each class to 400 images. This balance was achieved through the random application of transformations such as rotation, flipping, zooming, and distortion to the images in the training set. By deliberately refraining from augmenting the test set, we maintained its integrity, which is essential for an unbiased assessment of the model's performance.

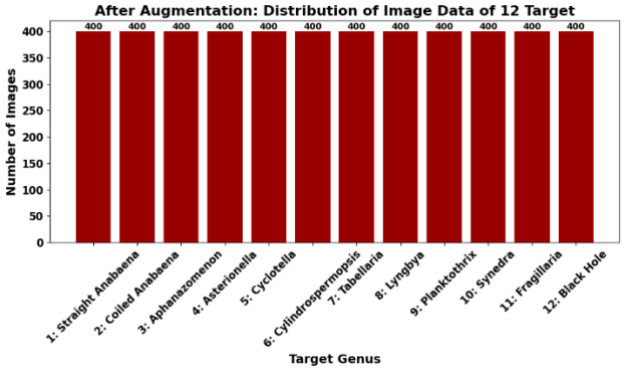

Fig. 4. After Augmentation: Distribution of 12 Target

C. Model Application

To ensure robust and accurate classification of algae species from microscopic images, we employed four widely recognized machine learning models that have been extensively validated for their performance in image recognition tasks. The choice of these models is supported by their documented success in similar applications, such as microalgae detection, as detailed by Ning et al. [1], Otálora et al. [2], and Abdullah et al. [3]. Each model brings unique strengths to the task, handling the diversity and complexity inherent in our algae image dataset.

1. Convolutional Neural Network (CNN)

Our Convolutional Neural Network (CNN) model, referenced from LeCun et al. [10], employs two primary convolutional layers, named conv1 and conv2, followed by max pooling to reduce dimensionality. The processed data is then flattened and directed through two dense layers, fc1 and fc2, which perform the final classification task.

2. ResNet

The ResNet architecture [11], designed with deeper networks in mind, initiates with a convolutional layer (conv1) and progresses through four subsequent stages of residual blocks (layer1 to layer4). These blocks are designed to facilitate training of deep networks by incorporating shortcut connections. The model concludes with an adaptive average pooling layer and a dense layer to yield the classification outputs.

3. U-Net

U-Net [12] operates on a dual-path system: an encoder that progressively downsamples the input, and a decoder that reconstructs the resolution through upsampling. The two paths work in tandem, with skip connections transferring contextual information from the encoder to the decoder. The final layer, prior to classification, applies adaptive pooling to standardize output dimensions.

4. MobileNet

MobileNet [13] leverages depthwise separable convolutions as a means of reducing computational load while maintaining efficacy. This model comprises a series of these specialized layers, punctuated by ReLU activation functions. The concluding layers include adaptive pooling and a dense layer that collectively facilitate the classification process.

Each models—CNN, ResNet, U-Net, and MobileNet— accepts input images of size (3, 224, 224) and

produces an output tensor of dimensions corresponding to the batch size and the 12 target classes, effectively mapping the input to predicted class probabilities.

IV. Results

In the evaluation of our results, we specifically reported the mean of the accuracy and loss metrics for all species, as well as the True Positive Rate (TPR) and False Positive Rate (FPR) for 'Straight Anabaena' and 'Coiled Anabaena'(Table 3). This detailed focus was necessary because Anabaena is a key species that significantly affects water quality. Accurately classifying Anabaena is crucial for monitoring and managing water quality effectively. By examining the TPR and FPR, we aimed to confirm whether the models were correctly classifying Anabaena, which is vital for addressing water quality concerns.

Method	All 12 Species		Straight Anabaena		Coiled Anabaena	
	Accuracy	**Loss**	**TPR**	**FPR**	**TPR**	**FPR**
CNN	83.333	0.666	0.481	0.2	0.83	0.0
ResNet	**94.928**	0.131	**0.96**	**0.17**	**0.94**	**0.0**
U-Net	94.203	**0.091**	0.92	**0.17**	0.94	**0.0**
MobileNet	76.087	0.573	0.44	0.42	0.72	0.07
Method	Aphanazomenon		Planktothrix		Black Hole	
	TPR	**FPR**	**TPR**	**FPR**	**TPR**	**FPR**
CNN	0.86	**0.14**	**0.9**	0.5	1.0	0.0
ResNet	0.93	0.17	0.8	**0.0**	1.0	0.0
U-Net	**1.0**	0.375	0.7	**0.0**	1.0	0.0
MobileNet	0.86	0.19	0.5	0.25	0.8	0.0

Table 3. Performance Evaluation of All Four Models

The results indicate a lower performance for 'Straight Anabaena' compared to 'Coiled Anabaena', which is largely due to the prevalence of other algae species with similar straight morphologies in the water. This similarity among straight-shaped species presents a classification challenge, leading to reduced accuracy in distinguishing 'Straight Anabaena' from other visually similar algae. This issue is not confined to Anabaena alone; the results also show lower performance metrics for other algae types with straight forms, compared to 'Coiled Anabaena'.

This suggests that the abundance of straight-shaped species in the water affects the classification accuracy. Recognizing and addressing this challenge is essential for improving the identification systems for species that critically influence water quality. Understanding these nuances is crucial for further improving the accuracy of algae classification systems, particularly for those species with critical impacts on ecosystem health and water quality.

Figure 5 (True Positive Rate (TPR) by Model and Species) and Figure 6 (False Positive Rate (FPR) by Model and Species) provide a visual summary of the TPR and FPR for each model, allowing us to quickly ascertain which models are more adept at accurately classifying each species. These visual representations are particularly useful in highlighting the challenges in distinguishing 'Straight Anabaena' from similar species due to their morphological resemblance.

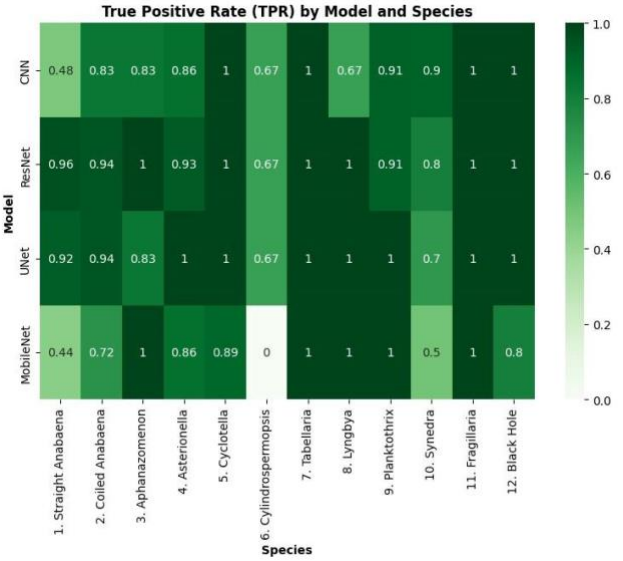

Fig. 5. True Positive Rate (TPR) by Model and Species

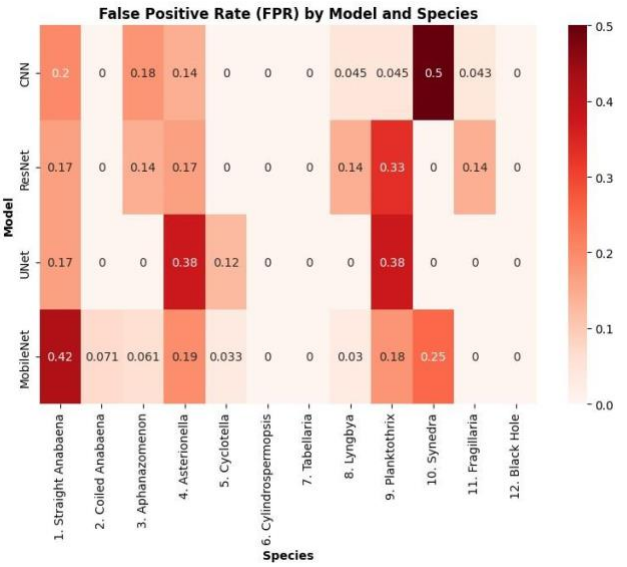

Fig. 6. False Positive Rate (FPR) by Model and Species

In the case of 'Straight Anabaena', the TPR is generally lower when compared to 'Coiled Anabaena'. This trend, which is vividly presented in Figure 4, points to the difficulty models face when differentiating between species with linear shapes. This issue is further compounded for other straight-shaped algae species, as depicted in Figure 5, where the FPR is notably higher, indicating a common challenge across models in classifying straight-shaped algae.

These visualizations underscore the challenge of accurately classifying species with similar morphological features and affirm the need for advanced techniques to improve model discrimination power. As we move to refine these systems, understanding the nuances captured in these figures is paramount for the enhancement of algae classification systems that impact water quality management.

V. Discussion

The comparative performance results, as summarized in Table 3 and depicted in Figures 5 and 6, indicate that ResNet performs the best among the four different deep learning methods. It achieves the highest accuracy and the highest true positive rate (TPR) for the target class while maintaining the lowest false positive rate (FPR) for the target class. Therefore, when we apply ResNet to the unlabeled dataset, it will correctly distinguish Anabaena from other similar-looking algae. We also listed some straight algae that look similar to straight Anabaena, and ResNet performs well in classifying those classes.

Figure 6 further informs our discussion by illustrating the FPR across different species. The consistent performance of ResNet in minimizing false positives emphasizes its potential for real-world application where the cost of misidentification can be high. Moreover, the visualization reveals that the challenge of distinguishing 'Straight Anabaena' is not solely limited to this species; other straight-shaped algae also present similar difficulties for classification models.

However, a detailed observation of Figure 5 indicates that the TPR for 'Straight Anabaena' is not as high as for 'Coiled Anabaena'. This discrepancy underscores the inherent complexities in classifying species with linear morphologies. Although ResNet outperforms other models, it still faces challenges in differentiating between species with straight forms, highlighting an area for potential improvement in future iterations of the model.

The insights gleaned from Figures 5 and 6 not only validate the quantitative data presented in Table 3 but also provide a nuanced understanding of each model's strengths and weaknesses. This understanding is crucial for further developments in the field of automated algae classification, where precision and accuracy are paramount. Recognizing these nuances and challenges is the first step towards enhancing the identification systems for species that significantly influence water quality and, by extension, ecosystem health.

VI. Conclusion

In this study, we developed a deep learning based algae classification system to overcome the limitations of traditional algae monitoring methods at Monroe Lake. By utilizing some computer vision technologies, our system transcends the constraints of the FlowCam devices, facilitating more accurate identification and classification of algae. Sophisticated machine learning models such as CNN, ResNet, U-net, and MobileNet were implemented, optimized through meticulous data preprocessing and augmentation. Among these, the ResNet model demonstrated superior accuracy and reliability. Notably, our models showed enhanced performance compared to the commonly used CNNs in most algae classification research, suggesting a substantial improvement over traditional approaches.

This project not only proves the feasibility of applying artificial intelligence in environmental monitoring but also sets a precedent for future research in the automation of natural resource management, offering significant implications for water quality management and public health protection. As part of future work, it will be beneficial to:

- Enhance model performance by acquiring more annotated data points, enabling more robust feature learning for deep learning models with extensive parameters.
- Develop an automated classification system by deploying the most effective model on unlabeled data, ensuring efficient and accurate algae classification, reducing manual errors, and enabling faster responses to water quality issues.
- Devise methods to improve performance on algae species with straight morphologies, which currently pose challenges due to their similar appearances, further refining the system's accuracy.

These steps will not only integrate a wider range of data for various algae species but also apply this technology to different aquatic environments globally, enhancing the efficiency of water quality management and algae monitoring worldwide. Additionally, the techniques used in this research could be adapted to other areas of environmental monitoring, contributing to the development of conservation strategies and sustainable management practices.

VII. Acknowledgements

We extend our heartfelt gratitude to Jill Minor, Data Analyst at the City of Bloomington, for her invaluable contribution to our project. Jill provided us with not only the critical data needed for our analysis but also her expertise and timely responses to our inquiries. Her insights were instrumental in our progress, and our success is a testament to her generous support.

Our sincere thanks also go to Professor David Crandall, who has not only imparted the knowledge necessary to undertake such work but also facilitated this collaboration with Jill Minor. This project was an opportunity to apply the theoretical foundations learned in his Computer Vision class to a real-world context, bridging the gap between academic study and practical application. His guidance has been a cornerstone of our learning journey, and this experience has been invaluable.

References

[1] Ning, Hongwei, Rui Li, and Teng Zhou. "Machine learning for microalgae detection and utilization." Frontiers in Marine Science 9 (2022): 947394.
[2] Otálora, P., et al. "An artificial intelligence approach for identification of microalgae cultures." New Biotechnology 77 (2023): 58-67.
[3] Abdullah, et al. "Computer vision based deep learning approach for the detection and classification of algae species using microscopic images." Water 14.14 (2022): 2219.
[4] Xu, Linquan, et al. "Accurate Classification of Algae Using Deep Convolutional Neural Network with a Small Database." ACS ES&T Water 2.11 (2022): 1921-1928.
[5] Chong, Jun Wei Roy, et al. "Microalgae identification: Future of image processing and digital algorithm." Bioresource technology 369 (2023): 128418.
[6] Chong, Jun Wei Roy, et al. "Trends in digital image processing of isolated microalgae by incorporating classification algorithm." Biotechnology advances (2023): 108095.
[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." ArXiv. /abs/2010.11929.
[8] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2020). "Training data-efficient image transformers & distillation through attention." ArXiv. /abs/2012.12877.
[9] Mirasbekov, Y., Zhumakhanova, A., Zhantuyakova, A., Sarkytbayev, K., Malashenkov, D. V., Baishulakova, A., ... & Barteneva, N. S. (2021). "Semi-automated classification of colonial Microcystis by FlowCAM imaging flow cytometry in mesocosm experiment reveals high heterogeneity during seasonal bloom." Scientific Reports, 11(1), 9377.
[10] LeCun, Yann, et al. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86.11 (1998): 2278-2324.
[11] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[12] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer International Publishing, 2015.
[13] Howard, Andrew G., et al. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

The Source Code

https://github.iu.edu/cs-b657-sp2024/mk159-ry1-shinjih_project