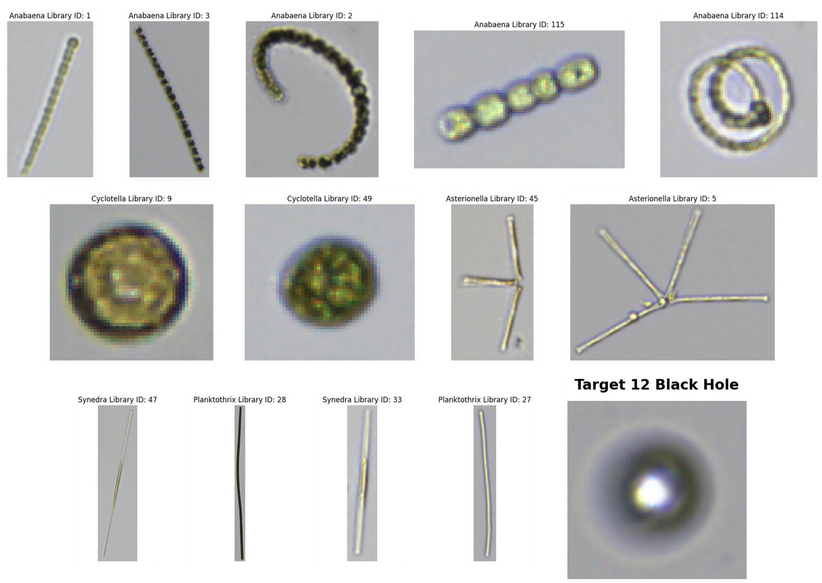
Algae Identification towards Automated Classification

Jihye Shin, Minju Kim, Ruoying Yuan, Luddy School of Informatics, Computing and Engineering, Indiana University Bloomington
Jill Minor, City of Bloomington

1. Motivation

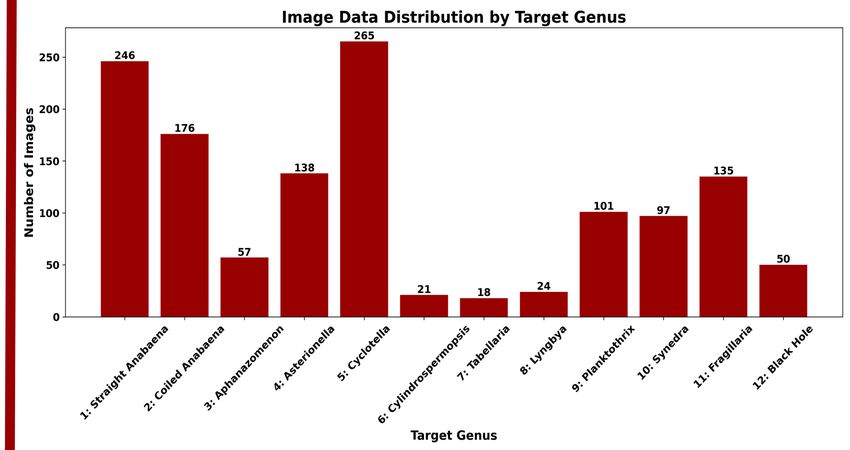
- **Bloomington's water, sourced from Monroe Lake**, experienced **poor quality** during the **hot months due to algae blooms**. This highlights the need for **improved algae monitoring and classification** to prevent future impacts on **water quality and public health**.
- Traditional tools like **FlowCam miss about 90% of algae images due to strict 'edge gradients' criteria.** This project applies computer vision to capture a broader range of algae types, enhancing performance.
- **Automating the algae classification** process reduces manual effort and errors, increases efficiency, and ensures quicker responses to water quality issues.

2. Challenges



- The presence of **numerous species with straight morphologies** similar to Straight Anabaena necessitates the development of specialized techniques to accurately differentiate these species.
- Some images contain **FlowCam lens artifacts** that appear as **black circles**, which would challenge the model's accuracy. We classified these images as **"Black Holes"** and curated them into **Target 12** for dataset clarity and integrity.

3. Data Preparation



- We **separated Anabaenas into straight and coiled** for better performance, as Anabaena is the biggest known culprit, also recorded in Indianapolis, for coinciding with taste and odor issues.
- To address the **imbalance** in the dataset, we applied data augmentation techniques to each taxa group in the training set.
- The **augmentation** process involved randomly applying transformations such as rotation, flipping, zooming, and distortion to the images.
- Each taxa group was augmented into a total of **400 images**, ensuring a more balanced representation for training the model.
- It is important to note that the augmentation was only applied to the training and validation set, preserving the integrity of the test sets.
- However, the necessity of data augmentation due to the original dataset's limitations would heighten the risk of model overfitting.

4. Methods

- Convolutional Neural Network (CNN) (1998)
- ResNet (2015)
- U-Net (2015)
- MobileNet (2017)

5. Results

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

- The goal of the model is to classify algae species from images, and it involves categorizing the images into 12 different species. In this task, the best accuracy and loss were achieved with ResNet and UNet models respectively.
- In the evaluation of the result (TPR and FPR), it was observed that the **performance for 'Straight Anabaena' was relatively lower** compared to 'Coiled Anabaena'. This discrepancy can be attributed to the presence of **other species with similar straight morphologies**, which complicates the accurate classification of 'straight anabaena.' Especially **Aphanazomenon** and **Planktothrix** as shown above.

6. Future Work

- Enhance model performance by **manually annotating additional 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.**