Project Proposal: Classification of Calanoid vs. Cyclopoid Using FlowCam and Environmental Data

1. Research Question

Zooplankton are critical to ecosystems, influencing the food web and nutrient cycle. Among them, Calanoid and Cyclopoid are two predominant groups. Being able to classify them in a standardized approach can enhance the monitoring of environmental changes and the understanding of aquatic ecosystem.

This project aims to classify zooplankton into Calanoid and Cyclopoid using FlowCam image-based data and environmental factors. FlowCam is a device that captures the images microscopic particles in liquids. The particle properties are the imaged-based data that differentiates different types of zooplankton. The primary goal is to develop an efficient, interpretable, and computationally feasible model that can automate classification while leveraging image-based data and environmental features.

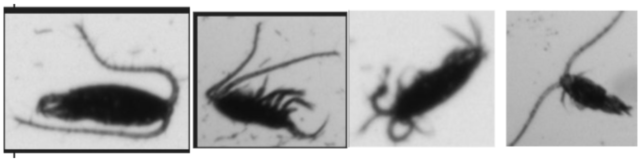
2. Data Description

The data used in this project is based on water samples collected from Lake Simcoe each spring in 2017, 2018 and 2019. Every data is labeled either Calanoid or Cyclopoid, and it has the corresponding imaged-based and environmental features. The following table shows what the structure of data look like in a nutshell.

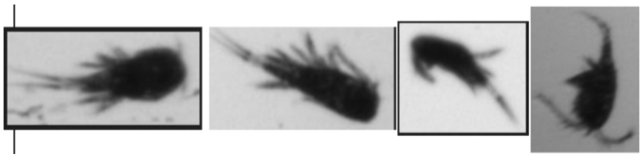
Particle.ID	Class	Image-based features (some examples)				Environmental features (some examples)		
		Aspect.Ratio	Compactness	Elongation	Transparency	distshore	CLOUD_PC	WaterT
1	Cyclopoid_1	0.3579	5.0161	13.6855	0.3854	187.9142	5	6.1
2	Calanoid_1	0.3326	4.0046	10.4855	0.3167	73.48868	5	6.4

The raw images of Calanoid and Cyclopoid are available as well but since FlowCam data comes with the important image-based features, the raw images are not directly used in this project. The images below are examples of Calanoid and Cyclopoid, which are just for facilitating the understanding of the project.

- Calanoid:



- Cyclopoid



The response variable is the zooplankton class which has been assigned manually. This is a binary classification problem, where Calanoid will be treated as 1 and Cyclopoid as 0 in the model.

The potential covariates are numerical values and include both image-based and environmental features. The imaged-based features are extracted from the images captured using FlowCam and are available for use. There are about 30 image-based features in total and below are some examples of the most important ones. The meanings of the features are referenced from https://q.utoronto.ca/courses/374291/files/35884661?module_item_id=6525304

- Aspect.Ratio: The ratio of the lengths of the axes of the Legendre ellipse of inertia of the particle.
- Compactness: A shape parameter derived from the perimeter and the (filled) area. The more convoluted the shape, the greater the value.
- Elongation: The inverse of Geodesic Aspect Ratio. 1 is the value for a circle or square; larger values are for elongated particles.
- Transparency: 0 is the value for a filled circle; values near 1 are for an elongated or irregular shape or a shape that has many interior holes.

The environmental features are measured at the time the water sample is collected. All samples taken at the same time and location share the same environmental features. There are about 15 environmental features in total and below are some examples of the most important ones.

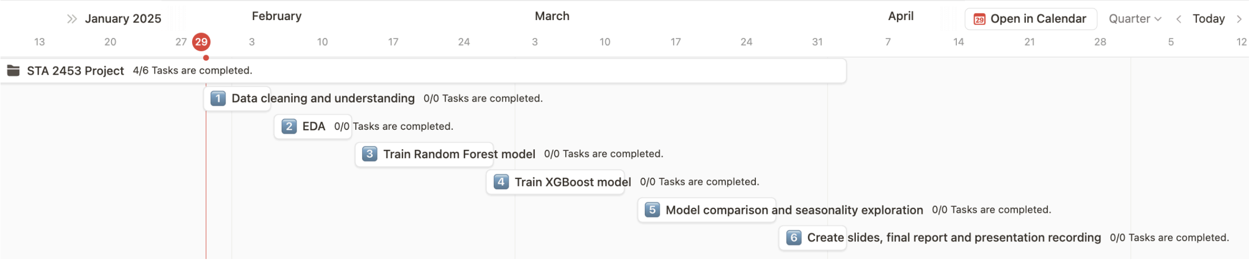
- distshore: Distance from shore (closest distance from shore).
- CLOUD_PC: Percent cloud cover
- WaterT: Surface water temperature

3. Methodology and Tools

The primary tool for this project will be Python. For exploratory data analysis, Matplotlib and Seaborn will be used for creating histograms, scatterplot, correlation heatmaps, etc. For data handling and processing, Pandas and NumPy will be the main libraries used, as they support data cleaning and transformation for large datasets.

Scikit-learn will be the primary library used for modeling. It provides machine learning models like Random Forest and XGBoost, which work well on large datasets, offer feature importance for interpretability, and are computationally efficient. Given the limited computational resources, these models in Scikit-learn are adequate for this project. Additionally, some statistical libraries may be used to investigate cyclical trends in species abundance.

4. Project Timeline



STA 2453 Project 4/6 Tasks are completed.

1 Data cleaning and understanding 0/0 Tasks are completed.

2 EDA 0/0 Tasks are completed.

3 Train Random Forest model 0/0 Tasks are completed.

4 Train XGBoost model 0/0 Tasks are completed.

5 Model comparison and seasonality exploration 0/0 Tasks are completed.

6 Create slides, final report and presentation recording 0/0 Tasks are completed.