Classification of Calanoid vs. Cyclopoid Using FlowCam and Environmental Data

STA2453

Christina Feng

Winter 2025

1. Introduction

Zooplankton play a vital role in aquatic ecosystems by influencing food webs and nutrient cycles. Among the seven major zooplankton classes of interest to the government, Calanoid and Cyclopoid are particularly prominent. Developing a standardized method to classify these groups can significantly improve the monitoring of environmental changes and deepen our understanding of aquatic ecosystems.

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. Table 1 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

Table 1: Data Structure

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 separated into different groups based on their meanings. For example, features related to the shape of zooplankton include aspect ratio, circularity, elongation and perimeter. Features related to the texture include transparency, roughness and intensity. The meanings of the features can be found here https://q.utoronto.ca/courses/374291/files/35884661?module_item_id=6525304.

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.

- **Preprocessing**

The FlowCam data for each zooplankton sample is stored in separate CSV files, each with corresponding environmental data. Before analysis, these datasets must be merged. While most FlowCam CSV files have consistent columns, some contain additional features. To maintain uniformity, only common columns are retained during merging. Additionally, environmental data is missing for approximately 4.6% of the samples. Since this represents a small portion of the overall dataset, these records are excluded from further analysis.

Although some columns do not contain missing data, their values are not meaningful enough to be considered as potential features. This issue is particularly common in the environmental data, where certain columns consist almost entirely of zeros. For example, the density of larval smelt is recorded as zero across all samples. Initially it seemed useful to include this variable, as there could be a relationship between yellow perch density and zooplankton. However, meaningful conclusions cannot be drawn due to the lack of variation. In total, 6 features like this have been identified and are excluded from the analysis.

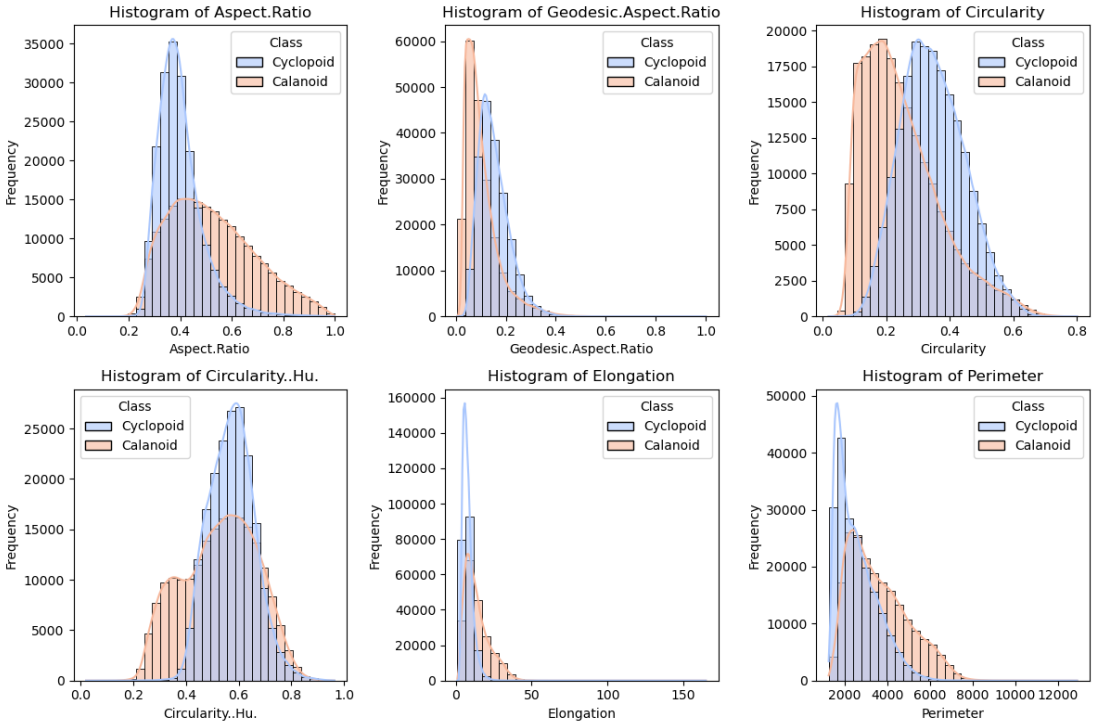

Figure 1: Histograms of some shape-related features

After data cleaning, the dataset consists of 394K records with 85 features (15 environmental and 70 image-based features). The labels for the zooplankton are replaced by Calanoid and Cyclopoid for better illustration.

- **Exploratory Analysis**

The dataset is almost balanced, with 49% Cyclopoid and 51% Calanoid. The two types of zooplankton differ in shape where Calanoids have long antennae and more extended body, while Cyclopoids are shorter and more oval. Our goal is to identify relevant FlowCam features to determine whether these morphological differences lead to a natural separation in the data.

Figure 1 shows histograms of selected features related to shapes. The distribution of the features look different for Calanoid and Cyclopoid and it is insightful to dive deeper. For example, aspect ratio, the ratio of the lengths of the axes of the Legendre ellipse of inertia of the particle, shows a notable difference between the two types. Cyclopoids tend to cluster around 0.4, while Calanoids have a heavier tail. A plankton with an aspect ratio of 0.4 is more likely to be Cyclopoid, whereas an aspect ratio greater than 0.8 strongly indicates a Calanoid. For circularity, which is computed from perimeter and the filled area, follows comparable distributions for both groups but with different means. Cyclopoids generally have higher circularity. However, due to significant overlap in the distributions, this feature alone is insufficient for reliable classification.

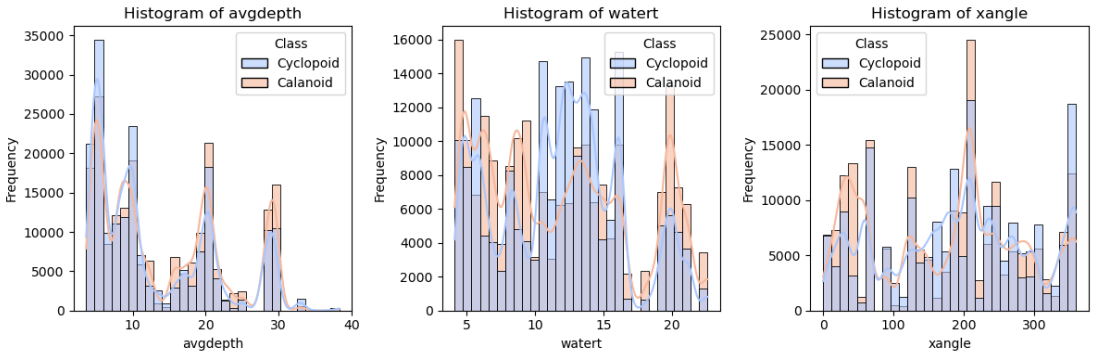

Figure 2: Histograms of some environmental features

For environmental features, Figure 2, the histogram of surface water temperature (watert), suggests that Calanoids are more commonly found in both colder and warmer waters, while Cyclopoids are more prevalent in water temperatures ranging from 10 to 15°C. Since samples from the same location share identical environmental factors, it is understandable that there is no clear separation between plankton classes based on environmental variables. Further investigation is needed to determine how environmental factors influence plankton classification.

Figure 3 is the correlation heatmap of meaningful features, highlighting several instances of near-perfect correlation. For example, Perimeter and Geodesic Length exhibit a correlation of 0.998, as Perimeter is derived from Geodesic Length. Similarly, Fiber Curl and Fiber Straightness show a strong negative correlation (-0.972) due to their inverse relationship in computation. This suggests that feature reduction may be useful to reduce model complexity.

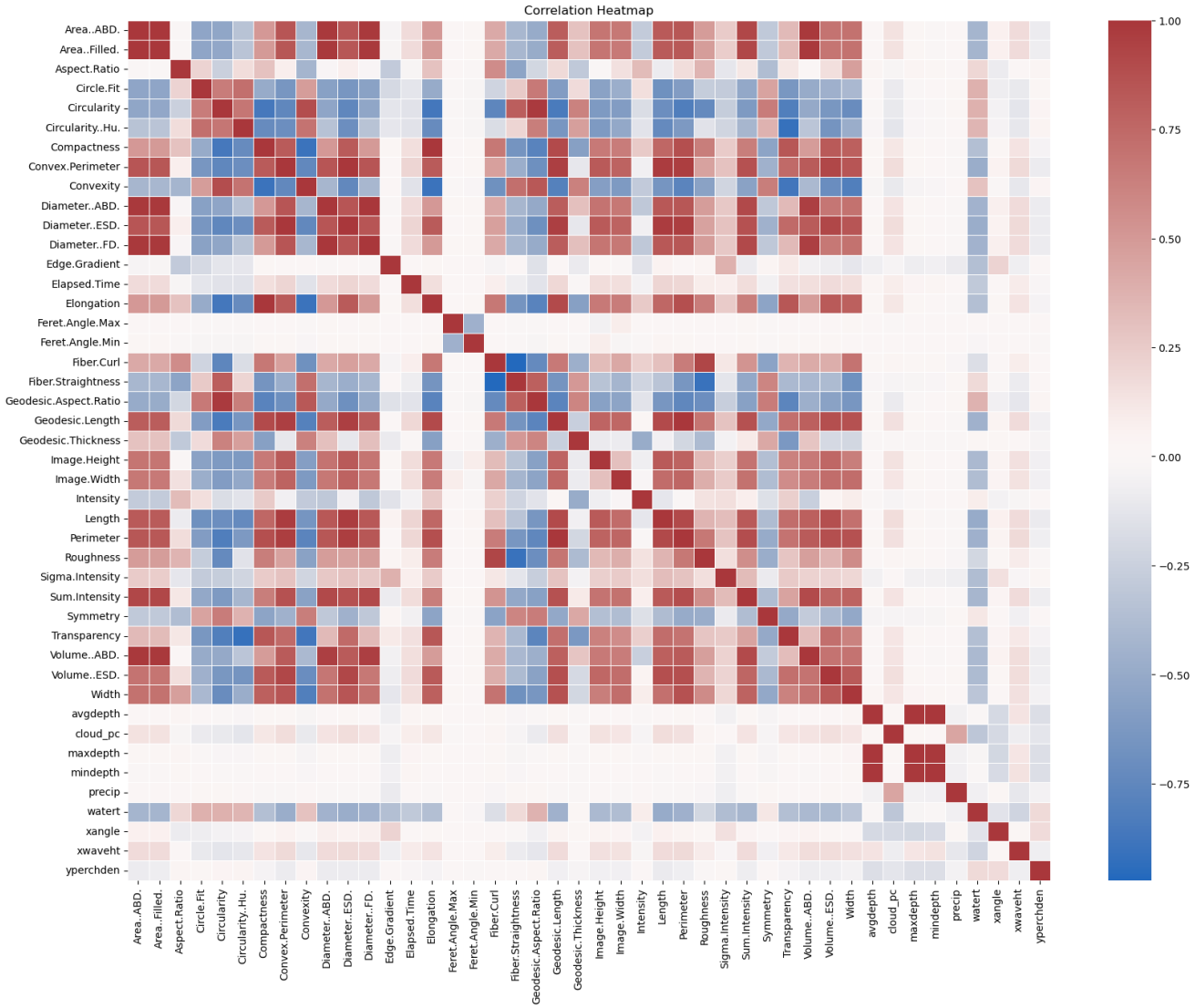

Figure 3: Correlation Heatmap

3. Methods

Random Forest and XGBoost offer a balanced combination of accuracy, interpretability, and robustness to noise and overfitting. Considering the limited computational resources, they are suitable for the classification of Calanoid and Cyclopoid compared to other machine learning models. For model evaluation, 5-fold stratified cross validation with an 80/20 training and test split is used.

The modeling process began with a full model that included all 70 image-based and 15 environmental features. To simplify the model and improve interpretability, feature reduction techniques were applied based on the definition of the features. Models were built on smaller subsets as a result: first using only environmental features, then combining selected top image-based and environmental features. The final model used a reduced set of features chosen to represent distinct morphological aspects such as

shape, texture, and size. The aim of exploring these smaller subsets was to develop a simpler, more interpretable model that still maintains strong classification performance.

4. Results

The full classification model, which incorporated all 70 image-based and 15 environmental features, achieved strong predictive performance. Both Random Forest and XGBoost classifiers had an overall accuracy of 93%, along with F1 scores and recall values of 93% for both Calanoid and Cyclopoid classes.

When the model was restricted to environmental features only, accuracy decreased to approximately 70%, with corresponding F1 scores and recall also at 69-70%. The most important environmental variables identified were water temperature and the density of larval yellow perch.

To evaluate the benefit of minimal feature sets, a simplified model was constructed using the top two image-based features, fiber curl and aspect ratio, combined with the two most important environmental variables. This model yielded improved performance, with XGBoost achieving an accuracy of 81%, an F1 score of 81%, and a recall of 83%. The Random Forest classifier produced similar results.

Finally, a refined model using 8 features was developed, selecting top-ranked variables that capture distinct morphological characteristics such as shape, texture, and size. This feature selection was motivated by the nature of the image-based features, which are designed to describe different structural aspects of zooplankton. One top-ranked feature was chosen from each morphological category, resulting in a total of six image-based features. Two additional environmental variables were included based on their relevance and contribution. For the performance, this model achieved 90% accuracy, F1 score, and recall across both classification methods. Further analysis showed that the inclusion of selected environmental features provided a modest improvement of approximately 2%.

Feature Set	Accuracy / F1 / Recall
Full model (all features)	93% – High precision and recall for both classes
Only environmental features	70% – Limited predictive power, low precision/recall
Top 2 image-based + top 2 environmental features	~80% – Solid mid-range performance
Selected best image-based features + top 2 environmental features	90% – Strong performance with reduced complexity

Table 2: Summary of model performance

Table 2 summarizes the performance of the models discussed. Since Random Forest and XGBoost performed similarly, only the results from Random Forest are shown. The findings highlight the importance of image-based features in classification. Even using

just the top two image-based features improves accuracy by around 10% compared to relying on environmental features alone.

5. Limitations

While the model performs well on the available data, several limitations should be considered. First, the classification relies heavily on features extracted from FlowCam. If the FlowCam system is unavailable, malfunctioning, or producing inconsistent data, the model's ability to classify zooplankton would be severely compromised. This dependency limits its practical use in settings without reliable access to such imaging equipment. Second, the environmental data used in the analysis were collected alongside the image data and may not be independent. This could restrict the interpretability and generalizability of the environmental features across different sampling contexts. Lastly, the data used in this project came from a single lake over a limited time period. As a result, the model may not perform as well when applied to different locations. These limit the generalizability of the model beyond the current project.

6. Conclusions

It is demonstrated that image-based features extracted from FlowCam are highly effective for classifying Calanoid and Cyclopoid. Environmental features alone are not sufficient for accurate classification but they still contribute useful context.

Through feature reduction, a simplified model using only 8 features was developed. This included 6 image-based features (fiber straightness, transparency, aspect ratio, convexity, geodesic thickness, symmetry) and 2 key environmental features (water temperature and density of larval yellow perch). The model maintained strong classification performance, achieving 90% accuracy, while offering improved interpretability and reduced complexity. These results demonstrate that a carefully selected subset of image-based and environmental features is sufficient for accurate classification of Calanoid and Cyclopoid. Overall, the objective of developing an effective and interpretable classification model was met.