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

1. Data Preprocessing & Quality Check

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 yellow perch 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, 7 features like this have been identified and are excluded from the analysis.

After data cleaning, the dataset consists of 394K records with 78 features (8 environmental and 70 image-based features). The labels for the zooplankton are replaced by Calanoid and Cyclopoid for better illustration.

2. Univariate 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.

The FlowCam data contains more than the shape-related features. Similarly, we plot the histograms for features related to texture of the plankton in Figure 2. Transparency

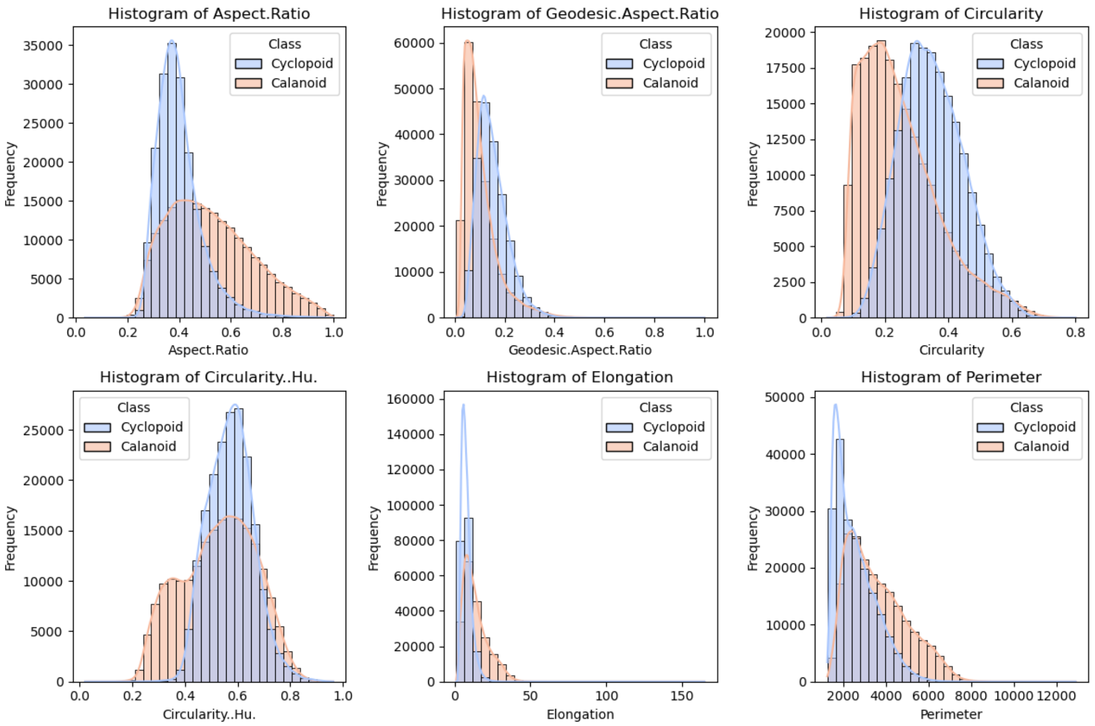

Figure 1: Histograms of some shape-related features

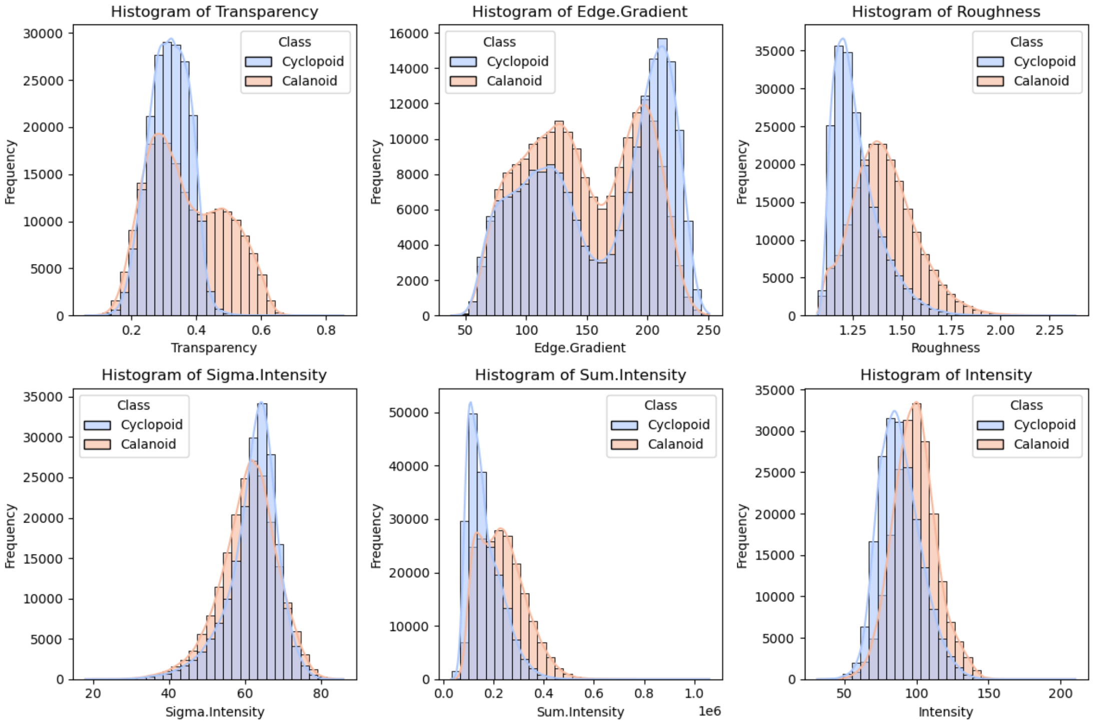

Figure 2: Histograms of some texture-related features

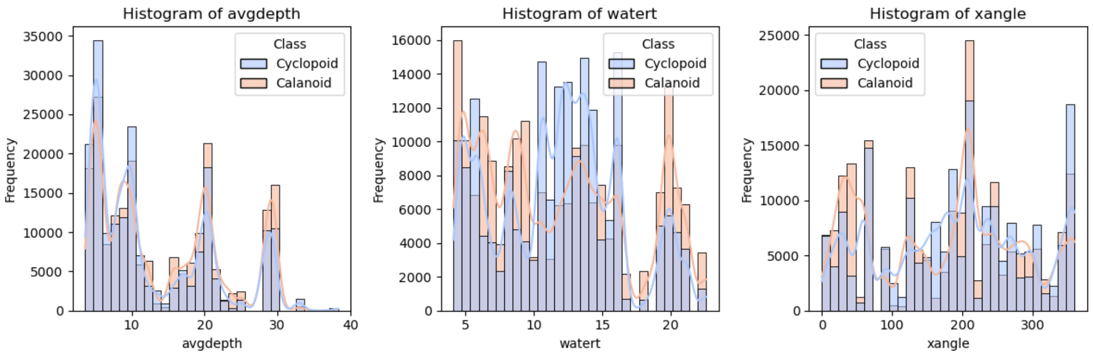

Figure 3: Histograms of some environmental features

and Roughness seem to be useful in classification but the final classifier needs to combine with other features.

For environmental features, 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.

3. Multivariate Analysis

Image-based features can be categorized into different groups based on their characteristics, such as shape or texture-related features. A closer examination of feature definitions reveals that some columns are derived from others, resulting in multicollinearity.

Figure 4 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.
There appears to be no strong correlation between environmental factors and image-based features. This suggests that separate models may be needed to explore the relationship between plankton types and environmental factors independently.

As multicollinearity may provide redundant information and reduce model generality, it is meaningful to perform feature selections.

4. Feature Engineering

Principal Component Analysis (PCA) is used to transform the data and retain the important information while reducing complexity. Figure 5, generated after proper

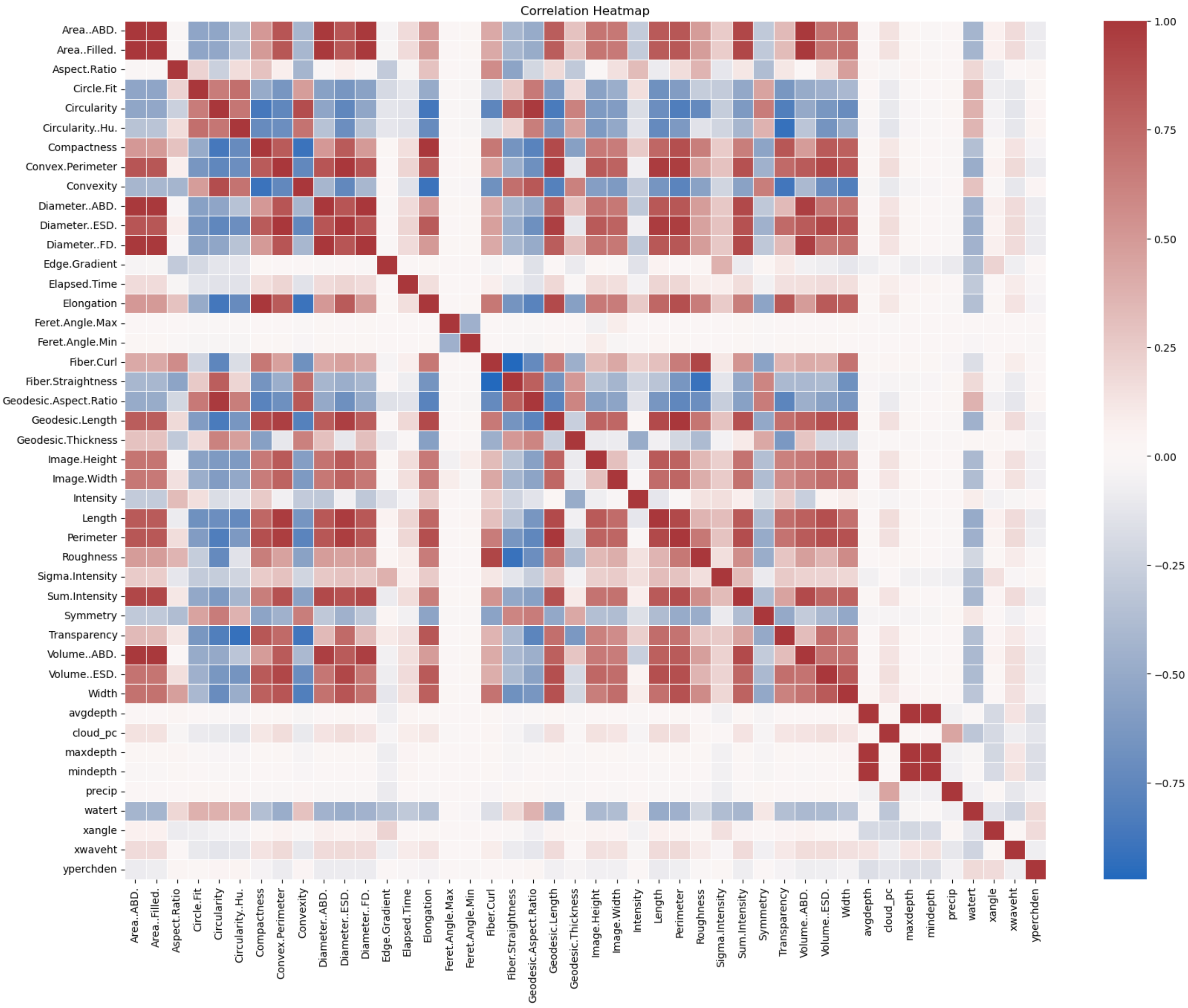

Figure 4: Correlation Heatmap

standardization, shows that 24 principal components explain 95% of the variance. Using this as the optimal number of components, the top five most influential features are Geodesic Length, Perimeter, Diameter ESD, Convex Perimeter, and Elongation. It's important to note that PCA transforms features into abstract components, making interpretation more challenging. Given that the goal of this project emphasizes interpretability, PCA may not be used into the final model development.

5. Conclusions & Next Steps

Initial data analysis suggests that some image-based features may be particularly useful for classifying the two types of zooplankton, while the impact of environmental features remains unclear. Many features exhibit high correlation, which must be addressed if multicollinearity affects the classifier model. PCA helps identify key features but considering the interpretability it may not be used in the future.

For next steps, I will train an initial classification model using Random Forest and XGBoost with hyperparameter tuning, compare their performance based on key metrics, and further explore the environmental data by analyzing seasonal patterns.

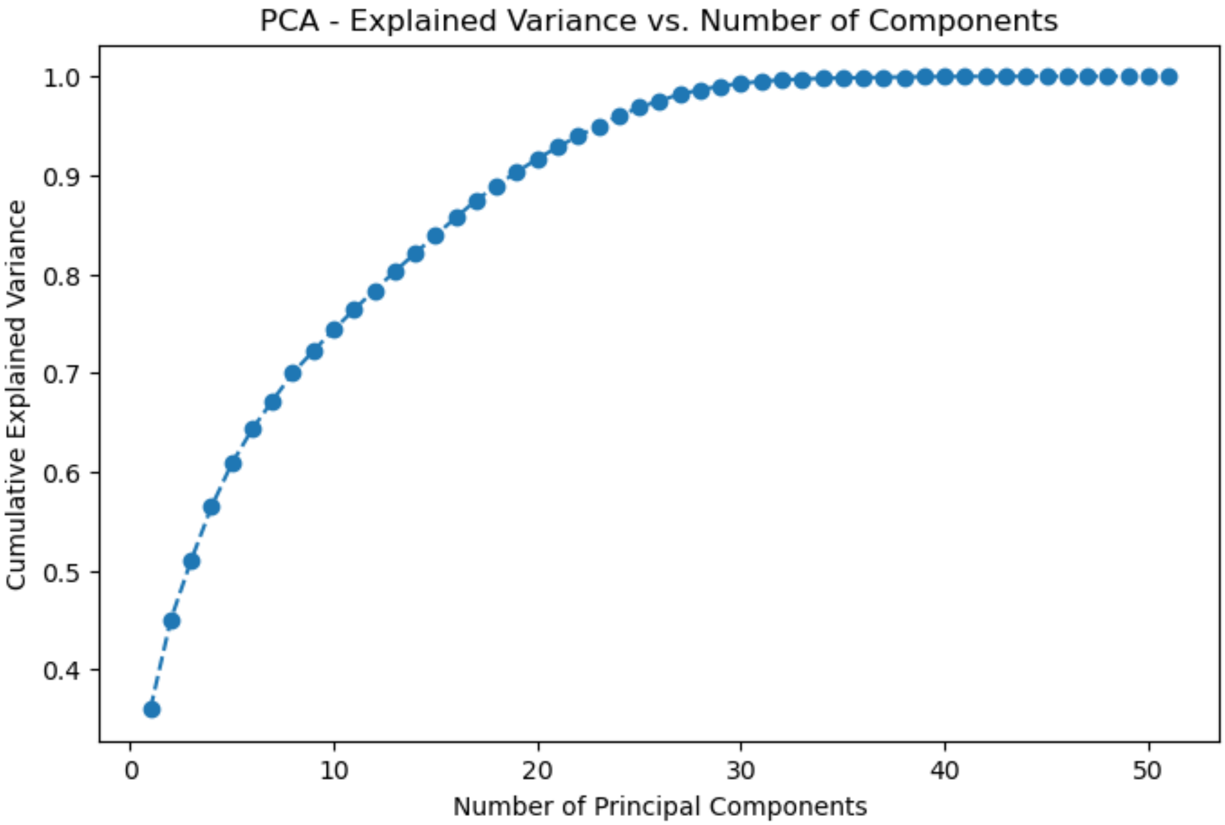

Figure 5: PCA Graph