EDA

February 14, 2025

I mainly focused on three aspects of the data.

The first aspect is the relationship between features in the images of plankton. Some features are similar, allowing me to group them. Grouping features based on their covariances can help build a more efficient model. It provides justification for removing redundant features or applying Principal Component Analysis (PCA). Alternatively, I can derive new composite features. For example, while not finalized yet, I might combine circularity and convexity.

Additionally, if we use linear models such as logistic regression or Support Vector Machines (SVM) in further studies, multicollinearity due to highly correlated features could hinder interpretation, making it difficult to determine which features are critical for classification. Similarly, in neural networks like Convolutional Neural Networks (CNNs), redundant information can lead to weight redundancy, as redundant features learn the same underlying pattern, causing inefficient learning. Moreover, redundant features can make the model overly complex, leading to overfitting.

The second aspect is the issue of unbalanced labels. Since the data documentation states that the dataset is unbalanced and we have specific classes to focus on, I investigated how severe the imbalance is.

The third aspect involves variations and unusual values.

Before conducting exploratory data analysis (EDA), I used data from SIMC_OverlapTiffsWithPP. Additionally, to improve understanding, I merged all the data into one dataset before proceeding with the EDA. Notably, 9 out of 252 files contained three extra columns: {'Biovolume..P..Spheroid.', 'Biovolume..Sphere.', 'Biovolume..Cylinder.'}. Since they were not present in most files, I removed them.

Before computing the covariance, I grouped features into three categories based on my knowledge:

Shape and Size Features: "Area..ABD.", "Area..Filled.", "Width", "Length", "Volume..ABD.", "Volume..ESD.", "Sphere.Volume", "Diameter..ABD.", "Diameter..ESD.", "Feret.Angle.Max", "Feret.Angle.Min" Structural and Shape Features: "Symmetry", "Circularity", "Convexity", "Aspect.Ratio", "Compactness", "Elongation", "Fiber.Curl", "Fiber.Straightness", "Roughness" Optical Features: "Transparency", "Sum.Intensity", "Intensity", "Sigma.Intensity", "Edge.Gradient"

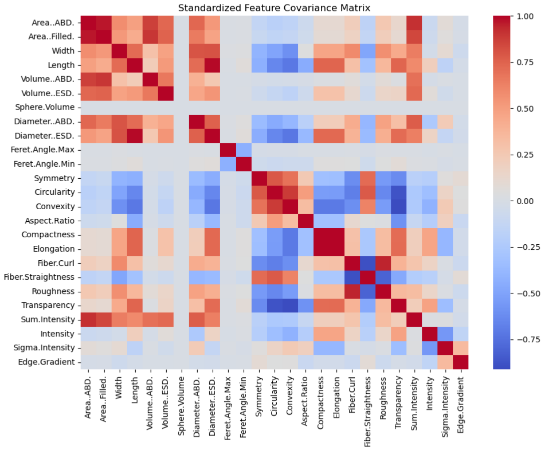
The definitions of these columns can be found in GlossaryParticleProperties.pdf.

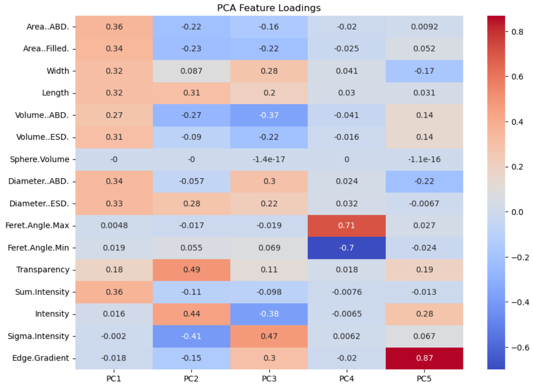
Next, I created a heatmap based on the covariance of each feature. Before drawing the heatmap, I standardized the data since the range of values varied significantly, and features had different scales.

As observed in the heatmap, my initial intuition regarding feature grouping was mostly correct. Three main clusters of highly correlated features (indicated in red) are visible. This is expected, as larger plankton areas correspond to larger width, length, and volume. Similarly, the structural and shape group exhibits strong internal relationships; for example, symmetry is inherently related to circularity.

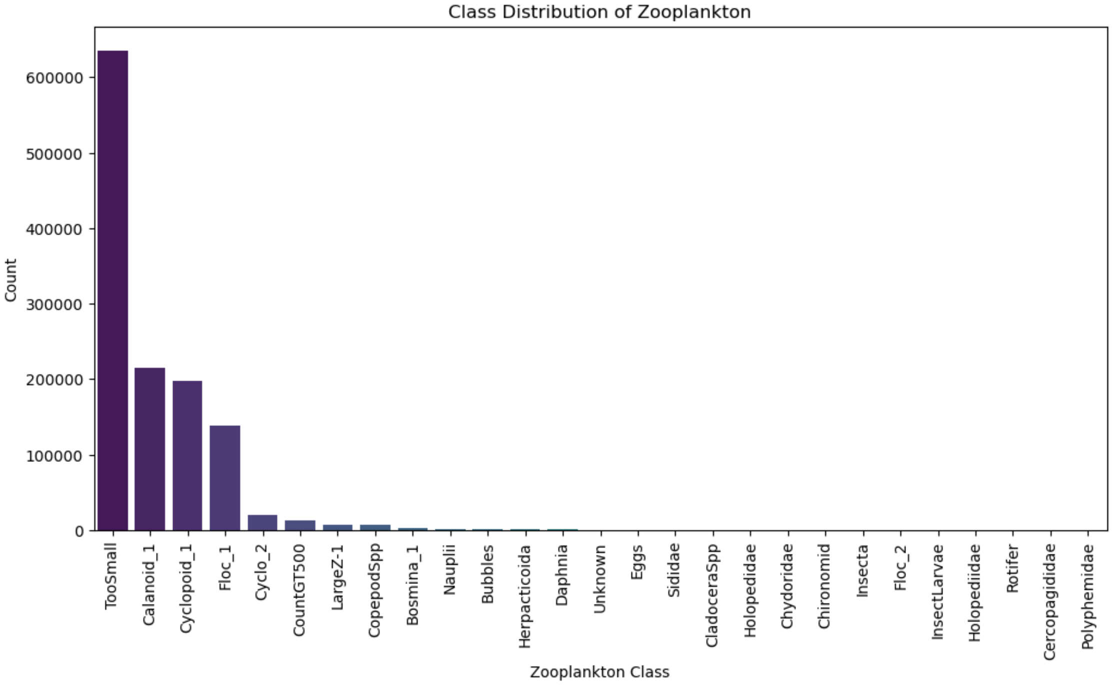
Interestingly, the shape and size group has a weak relationship with the structural group. This is reasonable since the shape and size of plankton are largely independent of their structural properties. Another intriguing finding is the relationship between dimension and size features with optical features. One possible explanation is that larger plankton block more light, appearing darker, leading to correlations between width, length, diameter, transparency, and sum intensity. Additionally, the formula for transparency (Transparency = 1 - (ABD Diameter / ESD Diameter)) further explains why dimension and size features are correlated with optical properties.

Based on this analysis, dimensionality reduction can be considered using feature groups. To evaluate potential dimensionality reduction, I examined a heatmap of PCA loadings. The results indicate that dimension and size features dominate PC1 along with Sum.Intensity, suggesting redundancy. Therefore, it may be beneficial to merge them into a single feature.





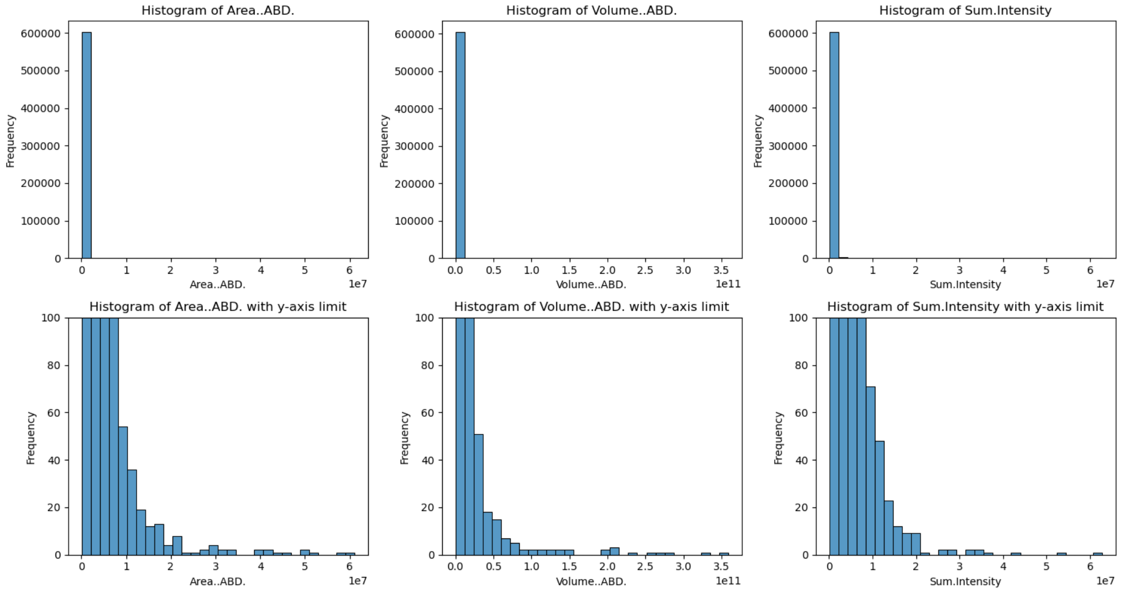
Next, I investigated the class imbalance. The dataset reveals that 51.25% of the data is labeled as “TooSmall”. Since this label does not provide meaningful information and exhibits high variability, it does not contribute to the classes we are focusing on. Given that more than half of the dataset is labeled as “TooSmall,” it could introduce significant noise into the model. Thus, I decided to remove these instances.



```
Percentage of 'TooSmall' class: 51.25%
```

Finally, to identify unusual values and variations, I generated bar plots and box plots to detect severe outliers after removing "TooSmall" data. I found that Sphere.Volume contained only zero values, so I removed it. Additionally, the range of area and volume was extremely large, spanning from 30,000 to 600,000, with large values primarily associated with CountGT500, Bubble, and Floc_1.

However, I cannot simply remove all extreme values, as they are essential for detecting the plankton we aim to classify. At the same time, correctly identifying other plankton is important to prevent misclassification. Removing extreme values might lead to biased model training. Therefore, I decided not to exclude these values and will focus on handling them appropriately in further analysis.



Top 10 largest values by 'Area..ABD.' and 'Volume..ABD.' and 'Sum.Intensity'

Area..ABD.	**Class**	**Volume..ABD.**	**Class**	**Sum.Intensity**	**Class**
6.1053e+07	CountGT500	3.5886e+11	CountGT500	6.2763e+07	CountGT500
5.8113e+07	Bubbles	3.3325e+11	Bubbles	5.2608e+07	Bubbles
5.1482e+07	Bubbles	2.7787e+11	Bubbles	4.3413e+07	Bubbles
5.0860e+07	Bubbles	2.7285e+11	Bubbles	3.6174e+07	CountGT500
4.9550e+07	CountGT500	2.6238e+11	CountGT500	3.4429e+07	CountGT500
4.6315e+07	Floc_1	2.3711e+11	Floc_1	3.4428e+07	Floc_1
4.3190e+07	Floc_1	2.1352e+11	Floc_1	3.2894e+07	Floc_1
4.2600e+07	Floc_1	2.0916e+11	LargeZ-1	3.2252e+07	LargeZ-1
4.2239e+07	CountGT500	2.0650e+11	CountGT500	2.7744e+07	Floc_1
4.0541e+07	Floc_1	1.9418e+11	Bubbles	2.7316e+07	Bubbles