Report

March 3, 2025

1 Introduction

This project aims to automate zooplankton classification using geometric and environmental features. Zooplankton are vital indicators of lake ecosystem health, serving as a crucial link in the food web by feeding on phytoplankton and becoming a food source for fish.

Fishing in Ontario's lakes has a significant economic impact. According to Ontario News (https://news.ontario.ca/en/release/1005496/ontario-boosting-its-fish-populations), more than 1.5 million anglers contribute $1.6 billion annually to the economy.

2 Data

I used data from the Ministry of Natural Resources and Forestry, Ontario. The dataset consists of .tif mosaics containing zooplankton images, along with geometric and environmental features. For example, each plankton image includes attributes such as transparency, symmetry, latitude, and longitude.

The classification model will use the "Class" column, focusing on classifying seven classes: Calanoid_1, Cyclopoid_1, Bosmina_1, Harpacticoida, Chironomid, Chydoridae, and Daphnia.

The data comes from two lakes in Ontario: Lake Huron and Lake Simcoe. However, the dataset is highly imbalanced, and some geometric features and labels may be missing.

3 Methods

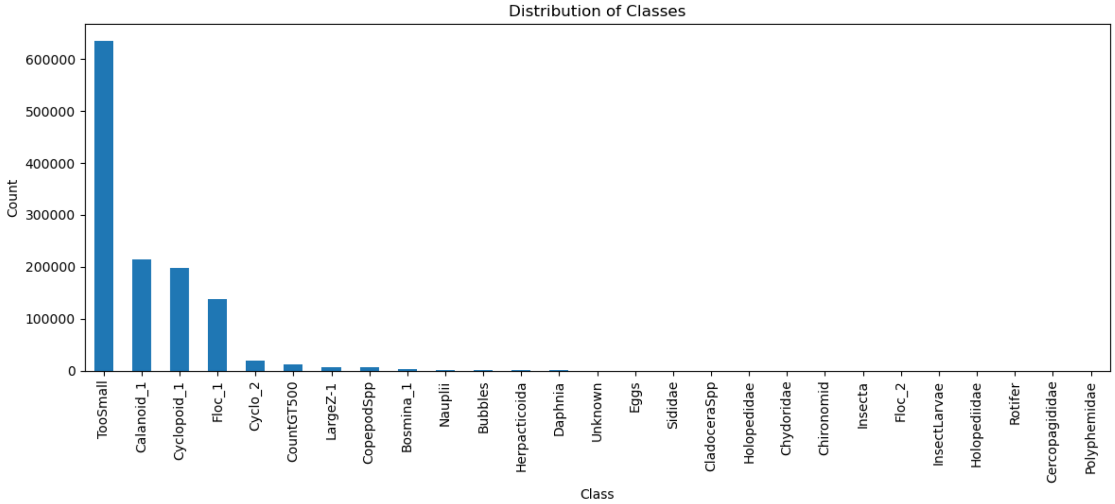
3.1 Data processing

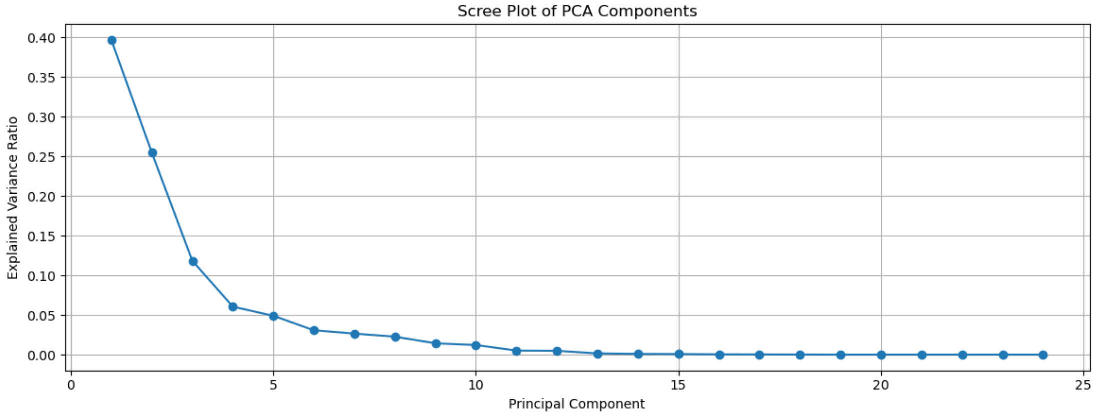
Before applying classification models, I merged all available data from Lake Simcoe for training and testing.

As shown in the graph below, the "TooSmall" class accounts for 51.25% of the dataset. Since this class provides no meaningful information about plankton (as these samples were too small to classify), I decided to exclude them to prevent introducing significant noise into the model.

Since many features are heavily right-skewed, I applied a log transformation to positively skewed data. Additionally, I performed standardization since most features have a large range.

The covariance matrix revealed some highly correlated features, which could lead to multicollinearity. To address this, I used Principal Component Analysis (PCA) with seven principal components. The number of components was determined using the elbow point from the scree plot, as shown in the graph below.





3.2 Models

I will test various statistical and machine learning models, including logistic regression, XGBoost, and random forest.

The goal is to classify the seven plankton classes of interest while also identifying "Others" for plankton outside this list. One approach to achieving this is hierarchical classification. This involves training a binary classifier to differentiate Known vs. Others, and if a sample is classified as Known, a second classifier predicts the specific plankton type. Therefore there will be three steps for models, which are train model for binary classification, train model for multiclass classification, and combine them. Eventually we split dataset into two. First we need to split the dataset into two, one will be used to train binary and multiclass model and the other will be used to test with combined model of binary and multiclass. To do this, we can avoid data double dipping. However, I decided to use same data for binary and multiclass training since we have fully independent test set for combined model testing.

3.2.1 Logistic Regression

Logistic regression requires several assumptions:

1. Linearly Separable Features: Zooplankton features are not linearly separable, but I will use logistic regression as a baseline due to its simplicity, interpretability, and ability to handle imbalanced classes.
2. No Multicollinearity: Since some features are highly correlated, I applied PCA as part of exploratory data analysis (EDA). The scree plot analysis suggested using seven principal components.
3. Standardized Features: Since feature ranges vary significantly, I standardized all features.

The binary classifier performed well, with high precision and recall for identifying Known plankton. However, it struggled with Others, indicating that some misclassifications occurred. The multi-class classification results reveal severe misclassification for some species (Chironomid, Chydoridae, Daphnia, and Herpacticoida), likely due to the extreme class imbalance. The hierarchical approach marginally improved the recall of Others and balanced class performance slightly. However, some species remain difficult to classify due to extreme class imbalance. Since logistic regression has clear limitations, I will explore XGBoost and random forest next to handle class imbalance and non-linear relationships better.

```
Binary Classifier Performance:
              precision    recall  f1-score   support

           0       0.96      0.65      0.77     29907
           1       0.86      0.99      0.92     66771

    accuracy                           0.88     96678
   macro avg       0.91      0.82      0.85     96678
weighted avg       0.89      0.88      0.87     96678


Multi Class Classifier Performance:
               precision    recall  f1-score   support

    Bosmina_1       0.32      0.89      0.47       457
   Calanoid_1       0.88      0.52      0.66     34469
   Chironomid       0.00      0.83      0.00         6
   Chydoridae       0.00      0.00      0.00         7
  Cyclopoid_1       0.77      0.57      0.66     31646
      Daphnia       0.02      0.67      0.03        88
Herpacticoida       0.01      0.61      0.02        98

     accuracy                           0.55     66771
    macro avg       0.29      0.59      0.26     66771
 weighted avg       0.82      0.55      0.65     66771


Final Hierarchical Classification Performance:
               precision    recall  f1-score   support

    Bosmina_1       0.22      0.89      0.36       592
   Calanoid_1       0.79      0.51      0.62     42869
   Chironomid       0.00      0.71      0.00         7
   Chydoridae       0.00      0.33      0.00         9
  Cyclopoid_1       0.67      0.57      0.62     39750
      Daphnia       0.01      0.53      0.02       118
Herpacticoida       0.01      0.62      0.01       118
        Other       0.96      0.65      0.77     37384

     accuracy                           0.58    120847
    macro avg       0.33      0.60      0.30    120847
 weighted avg       0.80      0.58      0.66    120847
```

3.2.2 XGBoost

XGBoost is well-suited for this dataset because it does not require strict distributional assumptions, can handle missing values, and performs well with imbalanced data.

3.2.3 Random Forest

Random forest is another suitable model since it can handle high-dimensional and correlated features without requiring assumptions about the data distribution.

4 Result

I expect logistic regression to have the lowest accuracy, while XGBoost and random forest should perform similarly, outperforming logistic regression.

5 Conclusion