Synaptiq
THE HUMANKIND OF AI
SYNAPTIQ |
SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Feasibility Study: Aquatic Insect Classification

Findings and Recommendations

Prepared by Synaptiq
Last Updated: May 29, 2023

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Table of Contents

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Overview

Aquatic insects are critical players in ecosystems and food webs within and surrounding freshwater rivers, lakes, and streams. They are the primary source of food for most freshwater fish and many birds. The diversity of aquatic insects is used by scientists and community organizations like the Deschutes River Alliance to monitor water quality and ecosystem health.

Equipping citizen scientists with tools to easily identify and count aquatic insects could provide a new source of valuable data for monitoring the health of these critical ecosystems. While there are several insect identification applications on the market today, most are geared towards terrestrial insects.

Our goal in this study is to:

1. Develop an image classification system with enough specificity to easy and accurately (90%+) identify of three key orders of aquatic insects in their nymph and adult life cycle stages: Trichoptera (Caddisflies), Ephemeroptera (Mayflies), and Plecoptera (Stoneflies); and
2. Use our learnings from this study to determine whether image classification at the family level is viable for aquatic insects and a worthy next step.

Methodology

The development of a new image classifier typically proceeds through the following steps:

1. Data Harvesting
2. Data Preparation
3. Model Training
4. Model Evaluation
5. Model Refinement

Data Harvesting

The most difficult step in building an accurate image classifier is obtaining enough high quality training data with accurate labels. Assembling a dataset from scratch can take weeks or months, and the process of labeling it adds even more effort.

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Therefore, the first step in building a new image classifier is to look for pre-existing labeled datasets. It did not take us long to find iNaturalist, a community of citizen scientists who have curated a database of over 130 million images of various organisms, with helpful labels for order/species and coordinates where each image was captured.

To collect our training dataset we queried the iNaturalist image database filtering by the orders Trichoptera, Ephemeroptera, and Plecoptera. We further narrowed the results using GPS coordinates [Lat = 0:60, Long = -40:-155] that correspond to North America. We achieved the image counts listed below:

Order name	Number of Images
Ephemeroptera (Mayflies)	21,609
Plecoptera (Stoneflies)	7,566
Trichoptera (Caddisflies)	25,177
Total:	**54,352**

Data Preparation

Fortunately, the data we harvested from iNaturalist was already labeled, saving an immense amount of time. The next step in preparation for training was to segment the images into different sets for training and evaluation.

When training a multi-class image classifier it is a best practice to train with equal numbers of examples of each class. Therefore, we randomly sampled 5,000 images from each order above as a starting point.

Next, we randomly split each batch of images into 80% training (4,000 per class) and 20% validation (1,000 per class). Then we reserved 10% of the validation set (100 images per class) to serve as the test set.

To increase the number of training images, we used standard augmentation techniques. This involves transforming the training set images in a variety of ways to create additional examples. Specifically, we applied random rotation, flipping, cropping, brightness and contrast adjustments, as well as histogram equalization to generate an additional 4,000 labeled images per class, effectively doubling the amount of training data.

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Model Training

Given the relatively small size of our dataset, we started with a pre-trained ResNet18 model which has 17 convolutional layers and one fully connected layer. The network was originally trained on the ImageNet dataset[1]. We froze the first 7 layers to preserve the lower-level features learned on ImageNet and adapted the higher-level features and the classifier (layer 18) to our task which has just 3 classes. The hope was that using the pre-trained network in a transfer learning scenario would lead to higher accuracy compared to, for example, training a smaller network from scratch.

Model Evaluation

We evaluated our model using accuracy as measured by the number of correct predictions divided by the total number of predictions. We also calculated precision and recall and plotted a confusion matrix to gain a more detailed feel for the model's behavior across classes.

Model Refinement

Even with the choice of an initial model (ResNet18 in this case) there are many hyperparameters that impact the final outcome, such as the learning rate, choice of optimizer, the number of layers to freeze, and whether/when to unfreeze them when doing transfer learning. We made sensible choices for each of them to keep the computational burden low, but building models while varying those parameters would probably yield a better final model.

Model Benchmarking

For the purposes of evaluating models outside of the training process, we created a benchmarking set of manually curated example images for each class which none of the models has seen before. In total, we assembled a benchmarking set of 113 images (roughly 40 images per class).

Findings and Results

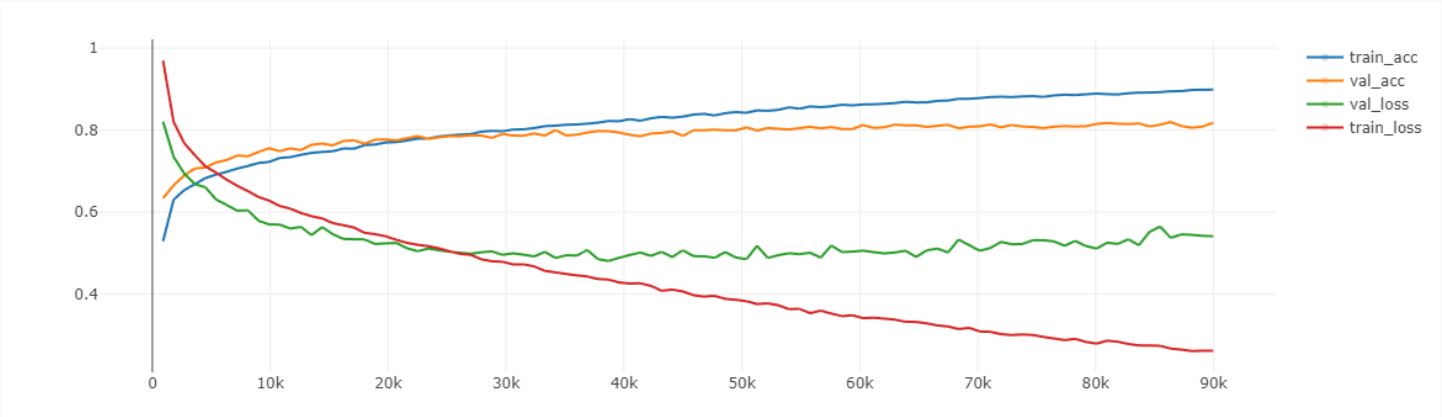
We conducted four total experiments in our effort to optimize the accuracy of the three class model.

[1] https://www.image-net.org/index.php

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

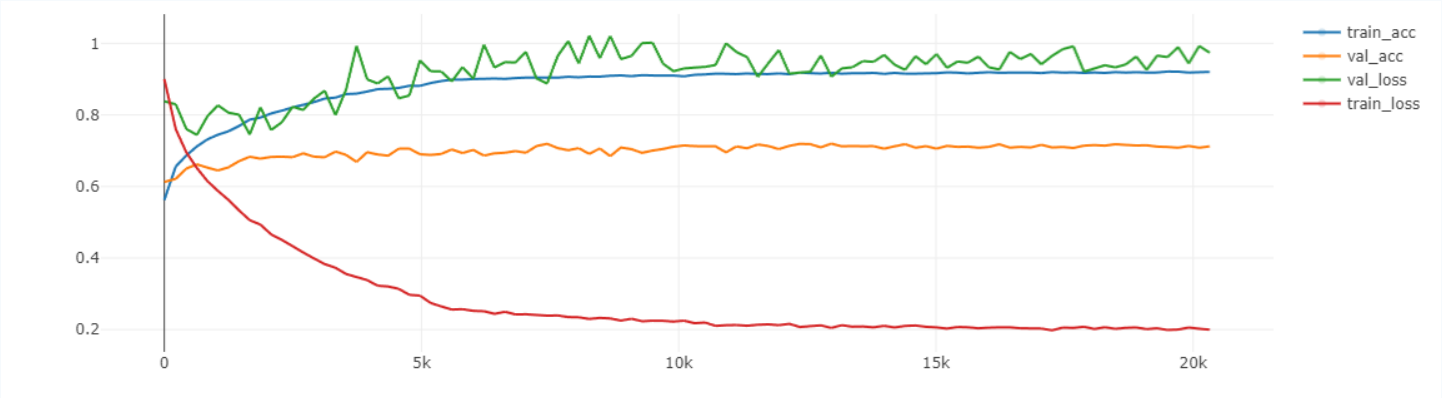
Experiment 1 - 5,000 images per class with pretrained weights

The training set accuracy was 89.9%, and the validation accuracy was reported at 81.8%. This model started to overfit on the training data as the validation accuracy stopped improving. While the training loss was decreasing and the validation loss was slightly increasing. The performance on the testing set was 80.0%.



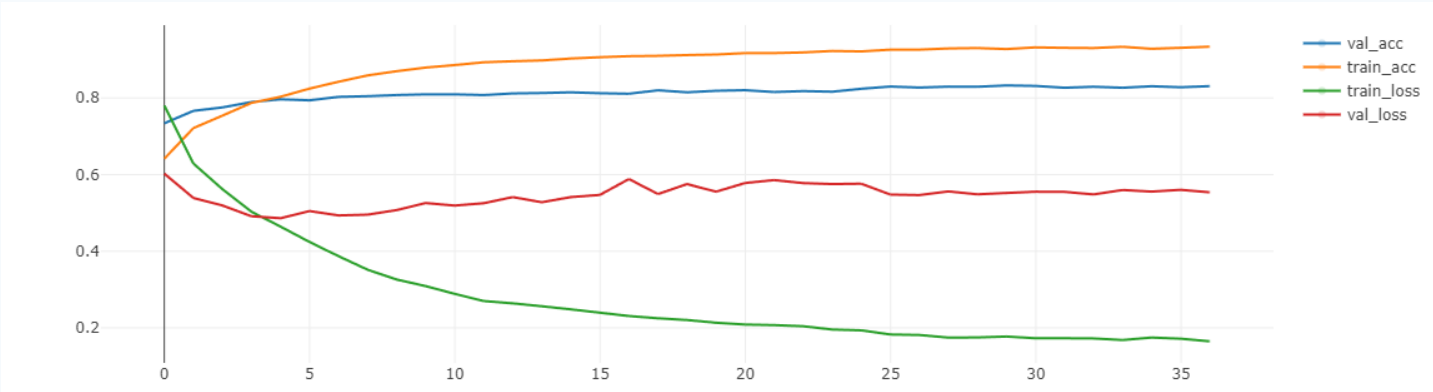
Experiment 2 - 7,000 images per class: SGD optimizer and added dropout(0.2)

For this experiment we increased the dataset size and tuned several parameters. We changed the Adam optimizer to SGD and added dropout of p=0.2. The train accuracy of this experiment was 92.02%, and the validation was 71.2% . The validation accuracy was the worst and the validation loss was the highest compared to the other experiments. The performance on the testing set was 78% with the validation loss being very high.



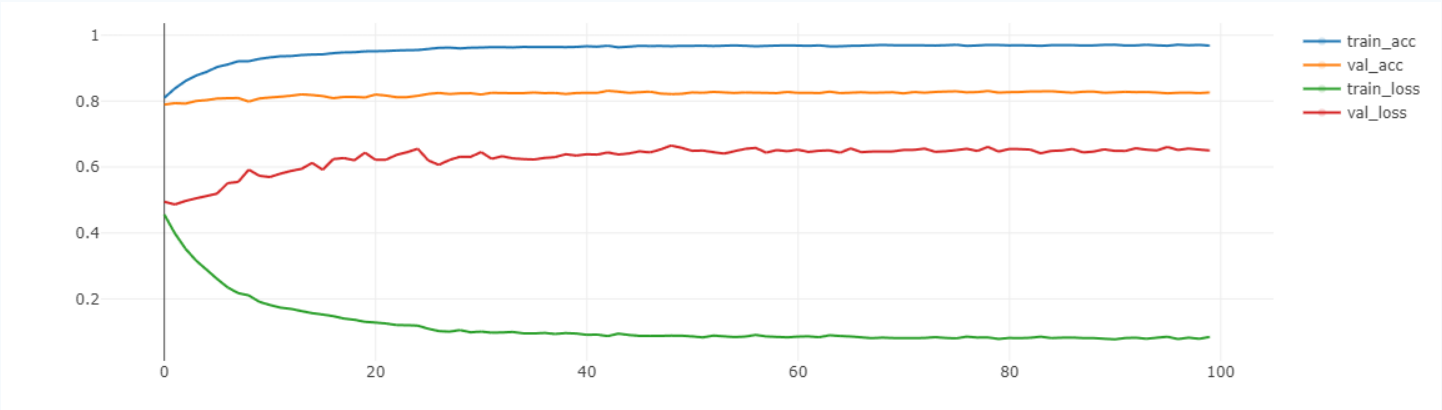
Experiment 3 - 7,000 images per class: Early stopping

In this experiment we used pretrained weights (for initialization) and stopped the training loop once the validation loss stopped decreasing. This resulted in training accuracy was 93.4%, and the validation accuracy was 83.1%. The performance on the testing set was 88.8%.



Experiment 4 - 7,000 images per class: Early stopping checkpoint and 7 frozen layers

The training accuracy of the model was 96.8% while the validation accuracy was 82.6%. We observed that after freezing seven layers, the validation loss starts increasing after a few epochs and the model does not show any further improvement. On the testing set, the model performed with an accuracy of 86.4% which was the best obtained.



Experiments Summary

Here are the results from all of our experiments:

	Accuracy		
	Training set	**Validation set**	**Testing set**
Experiment 1	89.9%	81.8%	80.0%
Experiment 2	92.0%	71.2%	78.0%
Experiment 3	93.4%	83.1%	88.8%.
Experiment 4	96.8%	82.6%	86.4%

Benchmark Performance

- The best accuracy on the benchmark set was **86.4%** using the "**Experiment 4 - 7,000 images per class: Early stopping checkpoint and 7 frozen layers**" model.
- Respectively the **"Experiment 3 - 7,000 images per class: Early stopping"** model achieve similar accuracy performance of **84.5%**

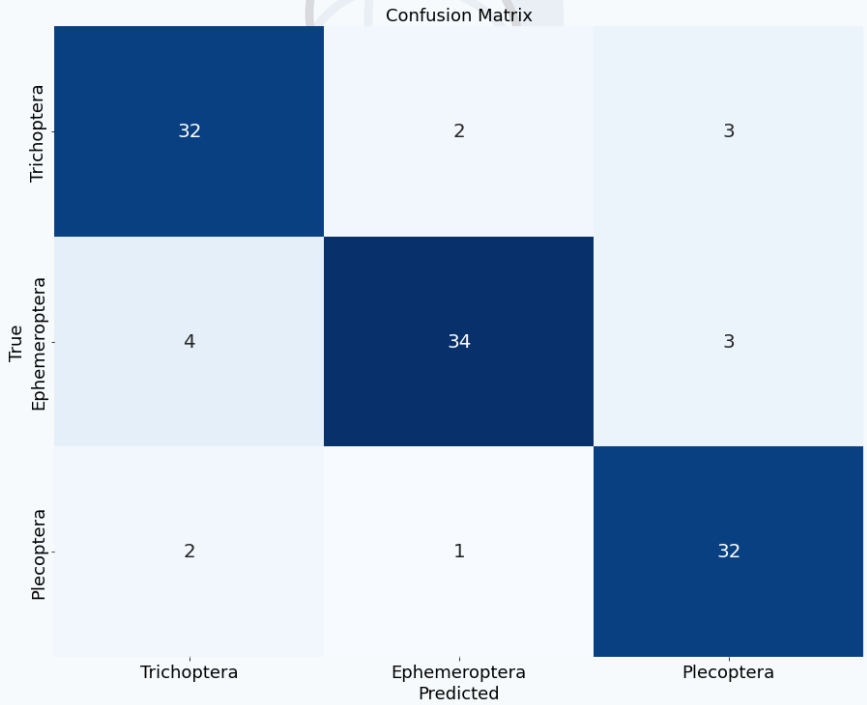
SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Confusion Matrix

A confusion matrix visualizes the performance evaluation of a classification model. It provides a summary of the predictions made by a model on a classification problem, comparing them to the actual ground truth labels. The rows of the confusion matrix represent the actual or true class labels, while the columns represent the predicted class labels. Each cell in the matrix indicates the number of samples that belong to a particular true class and were predicted as a particular predicted class. To calculate accuracy, precision, recall, and F1 score using the confusion matrix, with the following formulas:

- Accuracy = (TP + TN) / (TP + TN + FP + FN)
- Precision = TP / (TP + FP)
- Recall = TP / (TP + FN)
- F1 Score = 2 * (Precision * Recall) / (Precision + Recall)

Here is the confusion matrix for our benchmarking set:



Displaying the results in this way helps visualize the extent to which the model is correctly predicting the labels in the benchmarking set. When reading horizontally by row, you can easily see that the model most often makes the correct prediction, but that it sometimes “confuses”

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Ephemeroptera with either Trichoptera or Plecoptera. These are false negatives. Likewise, as you read vertically, you can see that the model is usually right, but provides false positives for up to 6 predictions.

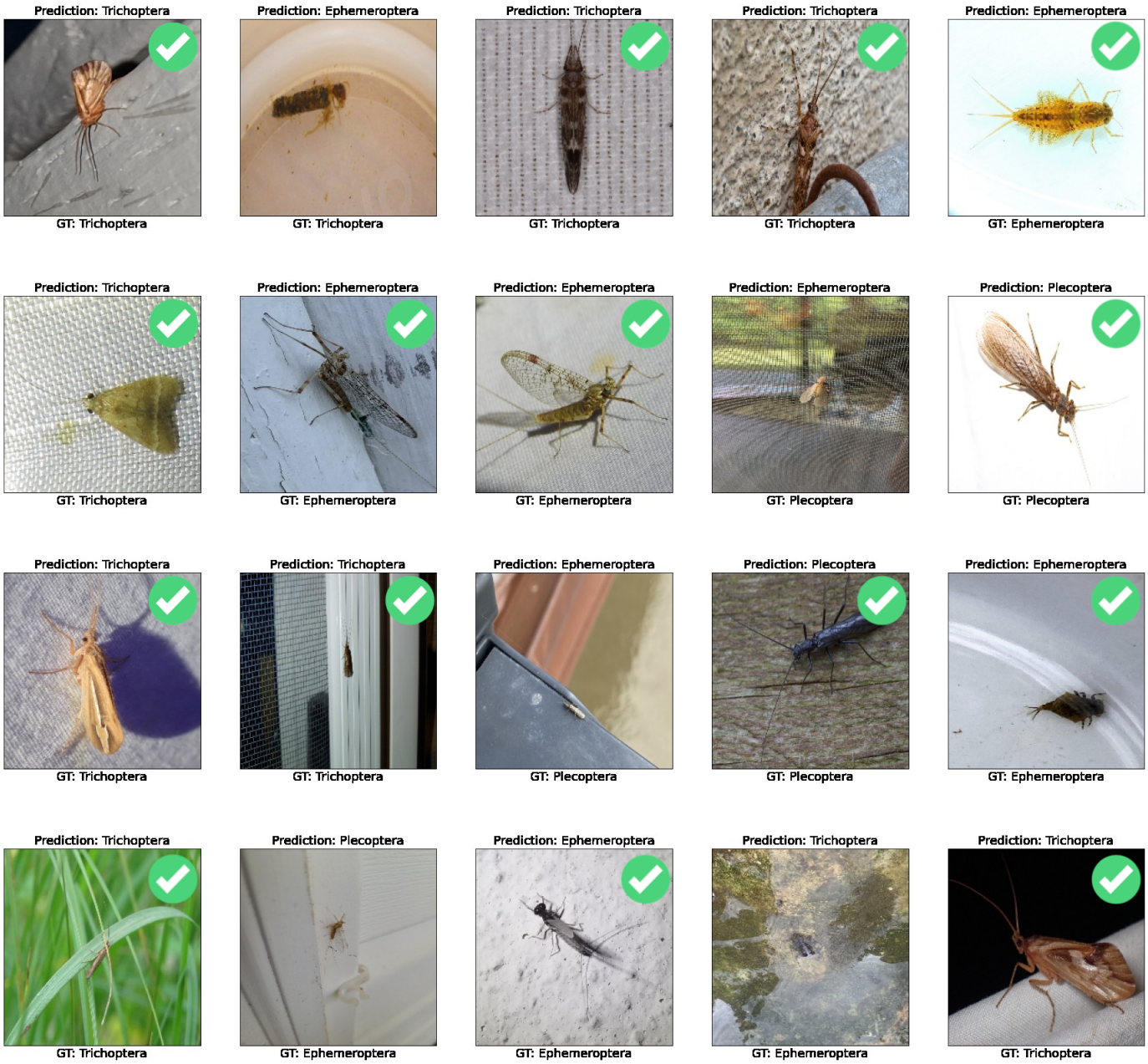
Here are the demonstrated results for each metric on the benchmarking set:

Metric	Score
Accuracy	86.40%
Precision	86.77%
Recall	86.69%
F1 Score	86.72%

These numbers are reasonable given the small size of the training data and the limited number of experiments. With more iterations, it should be possible to approach the high 90's on these metrics.

SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

To provide a sense for how the model performs, some representative predictions are presented below where GT means “Ground Truth”:



SYNAPTIQ.AI | +1 (484) 531 - 4167 | 10940 SW BARNES RD. #203, PORTLAND, OR 97225

Recommendations and Next Steps

While we did not achieve the ideal accuracy of 90%+ on our benchmark set, we felt there are still many avenues to explore in an effort to significantly improve model accuracy:

- Review the data set to ensure all labels are accurate
- Introduce true negatives into the training data
- Explore mechanisms to harvest or generate more data
- Experiment with different model architectures and hyperparameters
- Provide user guidance on consistent image capture

Likewise, we felt there were enough obvious visual differences in Trichoptera, Plecoptera and Ephemeroptera families that accurate image classification should be viable and worth pursuing.