PIVOT: CNN-Labeled Image Validation

Chandler Ault, Yamina Katariya, Yash Manne, Aditi Shrivastava, Alison Chase*, Claire Berschauer*

Abstract

There is an urgent need to accurately quantify oceanic phytoplankton to understand their role in the global Carbon Cycle. The prevailing method is a **complex, time-consuming, scattered** process of model development, image labeling, and model evaluation. In collaboration with researchers at the University of Washington Applied Physics Lab*, we developed a **unified, interactive tool to validate the Convolutional Neural Network (CNN) classification** of phytoplankton imagery.

Our in-house solution, the **Plankton Image Validation Optimization Toolkit (PIVOT)**, enables researchers to efficiently correct mislabeled images, improve the model, and track its performance. With **active learning**, PIVOT **strategically selects training images** to enhance model performance with minimal labeling while also incorporating a subset of random images for validation. Thus, we facilitate a **human-in-the-loop approach** to evaluate model performance while also accumulating a valuable dataset for subsequent use in semi-supervised learning.

Methods

Exploratory Analysis

The images utilized in this project were sourced from an Imaging FlowCytobot (IFCB) during four research cruises spanning 103 different dates and 5,299 sampling points. This effort resulted in a dataset comprising approximately 5,514,006 images of single-cell phytoplankton. Subsequently, a CNN was trained to classify these images into 10 groups: **Chlorophytes**, **Ciliates**, **Cryptophytes**, **Diatoms**, **Dictyo**, **Dinoflagellates**, **Euglenoids**, **Prymnesiophytes**, **Unidentifiable**, and **Other**.

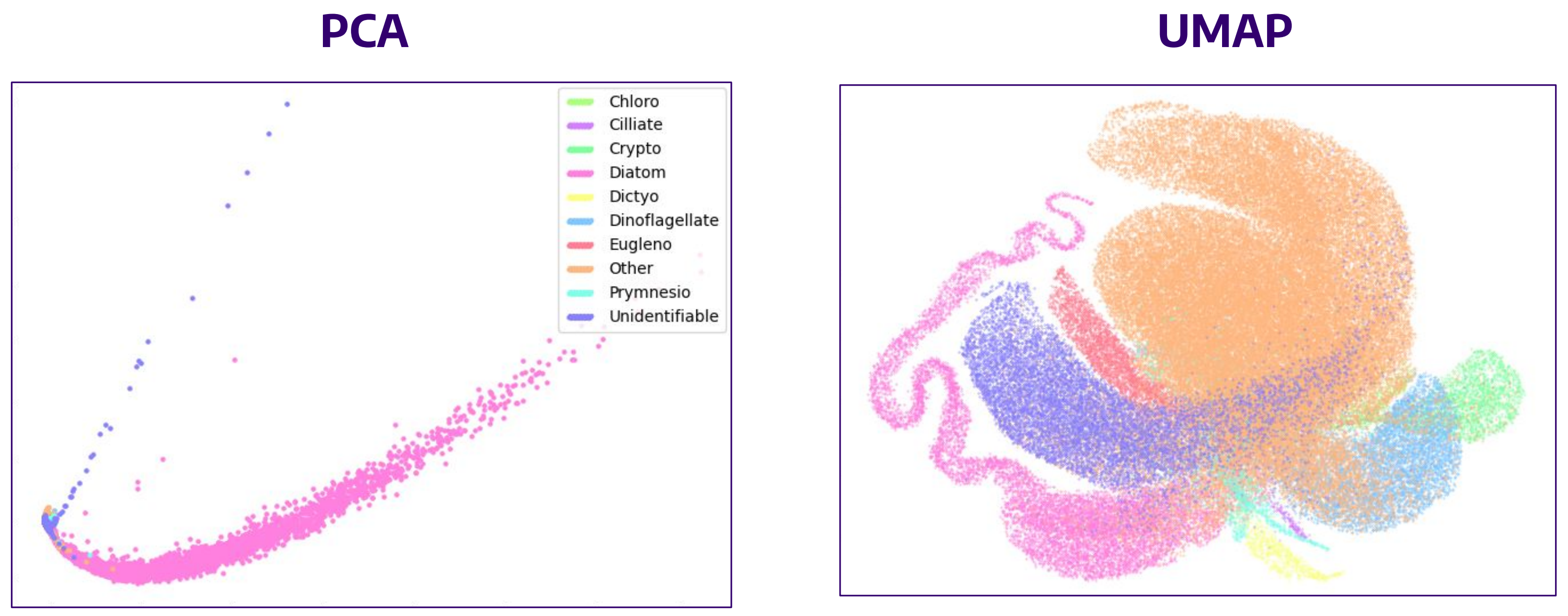
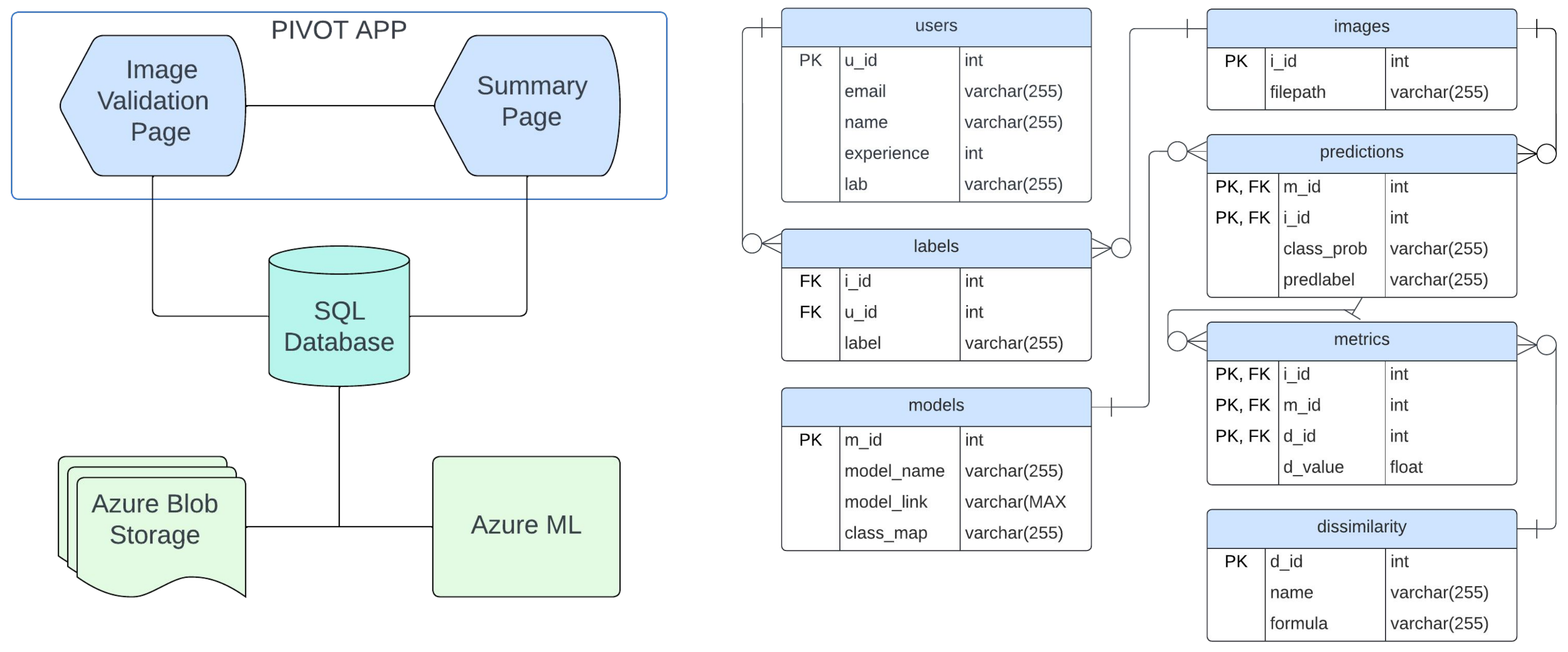


Figure 1: Skewed Distribution of Model Classes

In **Figure 1**, we see that most classes are fairly separable while the Unidentifiable and Other classes dominate the vector-space, potentially leading to erroneous misclassification. So, it might prove valuable to pre-filter these classes before proceeding with standard classification.

Application

We employ **Streamlit** to craft an interactive web tool meticulously tailored for the dynamic validation of CNN-based image classifiers. Streamlit **seamlessly integrates with Python**, enabling swift development of this application with a **focus on user-friendliness and accessibility** for researchers of varying technical proficiencies. Our web tool presents a intuitive interface, empowering researchers to **label images, review summary statistics, and refine the model effortlessly**, all while preserving user progress and offering supplemental resources as needed.



users		
PK	u_id	int
	email	varchar(255)
	name	varchar(255)
	experience	int
	lab	varchar(255)

labels		
FK	i_id	int
FK	u_id	int
	label	varchar(255)

models		
PK	m_id	int
	model_name	varchar(255)
	model_link	varchar(MAX
	class_map	varchar(255)

images		
PK	i_id	int
	filepath	varchar(255)

predictions		
PK, FK	m_id	int
PK, FK	i_id	int
	class_prob	varchar(255)
	predlabel	varchar(255)

metrics		
PK, FK	i_id	int
PK, FK	m_id	int
PK, FK	d_id	int
	d_value	float

dissimilarity		
PK	d_id	int
	name	varchar(255)
	formula	varchar(255)

Figure 2: System and Entity Relationship Diagrams

Data Storage

In our system architecture, data components are organized and stored within a **Azure SQL database** as seen in **Figure 2**. We developed utility functions and stored procedures that **dynamically retrieve data** based on user interactions and system events, such as model updates or new image availability. This database serves as the backbone of our application, enabling seamless data retrieval and manipulation for display on the website.

Model Pipeline

The deployment and hosting of the pre-trained CNN model are managed through **Azure Machine Learning**. We find that this service **prioritizes efficient deployment** and allows researchers to use the model without extensive setup. We conduct the entire process shown in **Figure 3** on the Azure virtual machine provided by the UTOPIA group at the eScience Institute.

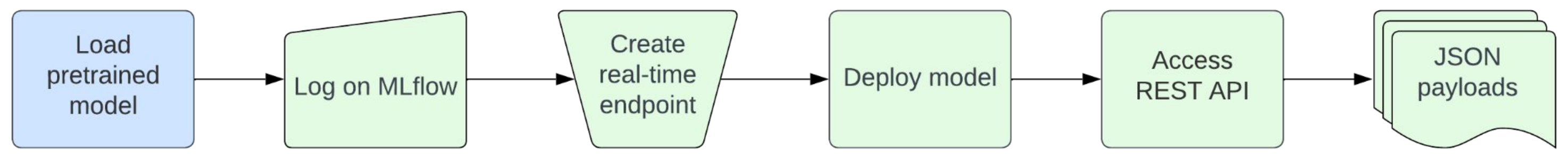


Figure 3: Azure Machine Learning Flow Diagram

Image Recommendation Optimization

We implement a framework to precompute an image importance score for each image per model. The **top N images** are then selected for validation for each user session. Additionally, a random selection of **R hold-out evaluation images** are also selected for labeling based on their weighted label score. These **two sets of N and R are combined** with a user-defined ratio to yield the final set of images to label for a user in a given session.

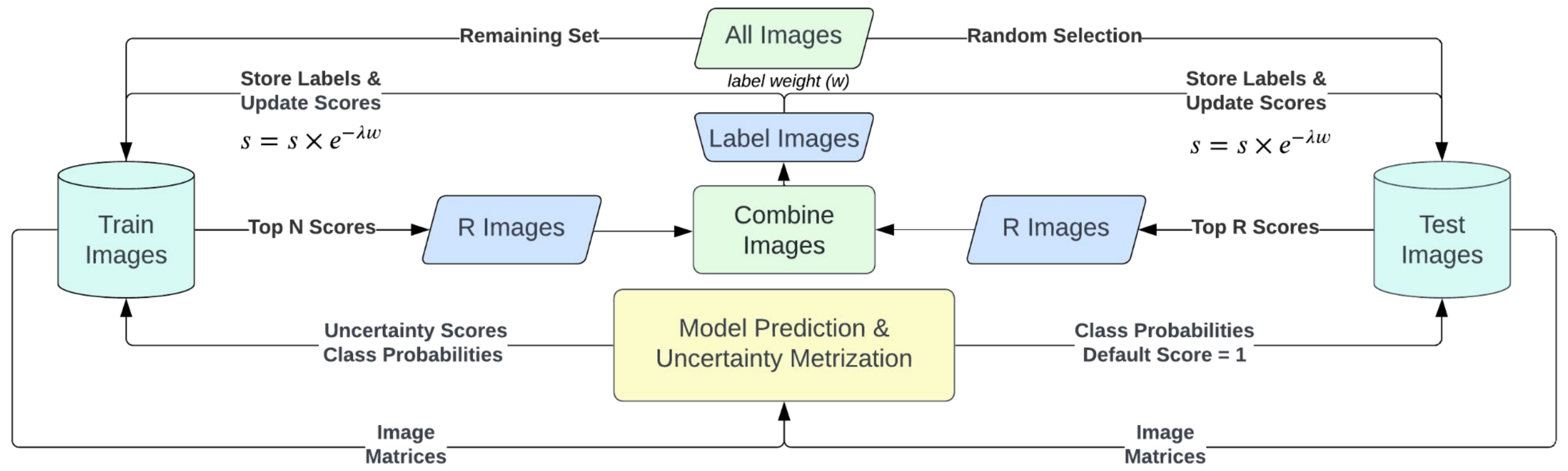

Remaining Set
All Images
Random Selection
Store Labels & Update Scores
label weight (w)
Store Labels & Update Scores
$s = s \times e^{-\lambda w}$
Label Images
$s = s \times e^{-\lambda w}$
Train Images
Top N Scores
R Images
Combine Images
R Images
Top R Scores
Test Images
Uncertainty Scores Class Probabilities
Model Prediction & Uncertainty Metrization
Class Probabilities Default Score = 1
Image Matrices
Image Matrices

Figure 4: Selection of Images to be Labeled

Results

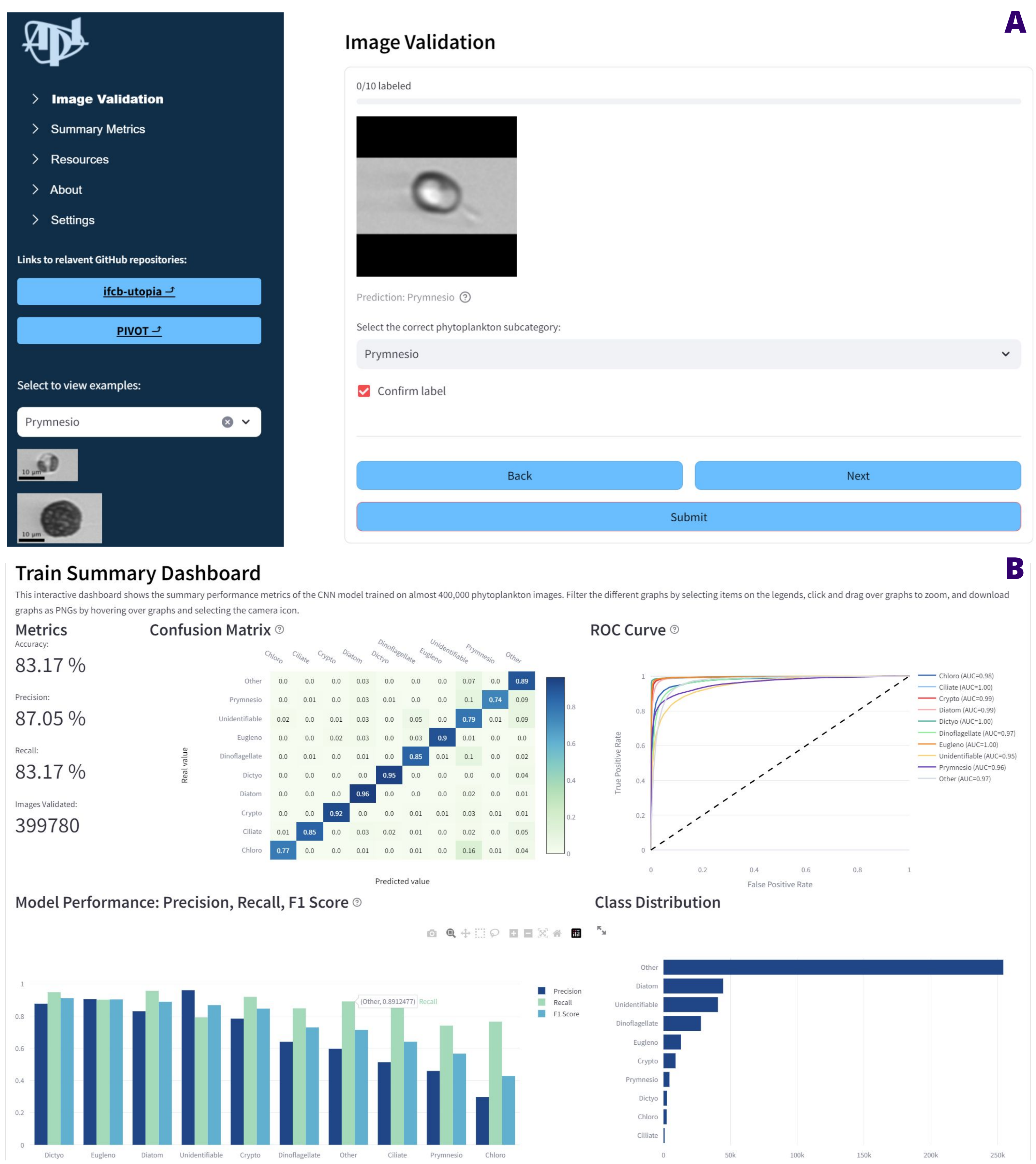


Image Validation

0/10 labeled

Prediction: Prymnesio

Select the correct phytoplankton subcategory: Prymnesio

☑ Confirm label

Links to relavent GitHub repositories:

Select to view examples: Prymnesio

Train Summary Dashboard

This interactive dashboard shows the summary performance metrics of the CNN model trained on almost 400,000 phytoplankton images. Filter the different graphs by selecting items on the legends, click and drag over graphs to zoom, and download graphs as PNGs by hovering over graphs and selecting the camera icon.

Metrics

Accuracy: 83.17 %

Precision: 87.05 %

Recall: 83.17 %

Images Validated: 399780

	Chloro	Ciliate	Crypto	Diatom	Dictyo	Dinoflagellate	Eugleno	Unidentifiable	Prymnesio	Other
Other	0.0	0.0	0.0	0.03	0.0	0.0	0.0	0.07	0.0	0.89
Prymnesio	0.0	0.01	0.0	0.03	0.01	0.0	0.0	0.1	0.74	0.09
Unidentifiable	0.02	0.0	0.01	0.03	0.0	0.05	0.0	0.79	0.01	0.09
Eugleno	0.0	0.0	0.02	0.03	0.0	0.03	0.9	0.01	0.0	0.0
Dinoflagellate	0.0	0.01	0.0	0.01	0.0	0.85	0.01	0.1	0.0	0.02
Dictyo	0.0	0.0	0.0	0.0	0.95	0.0	0.0	0.0	0.0	0.04
Diatom	0.0	0.0	0.0	0.96	0.0	0.0	0.0	0.02	0.0	0.01
Crypto	0.0	0.0	0.92	0.0	0.0	0.01	0.01	0.03	0.01	0.01
Ciliate	0.01	0.85	0.0	0.03	0.02	0.01	0.0	0.02	0.0	0.05
Chloro	0.77	0.0	0.0	0.01	0.0	0.01	0.0	0.16	0.01	0.04

Figure 5: Image Validation Tool (A) and Summary Metrics Dashboard: Train Set (B)

Conclusion

PIVOT addresses the absence of a robust platform for efficiently validating CNN models for phytoplankton image classification. It tackles the challenges of systematically processing and refining models used to label millions of single-celled autotroph images for oceanographic research. It serves as **a central hub for researchers** to validate ML model predictions and **strategically recommends images for maximum model improvement**.

Leveraging Streamlit and Azure's underlying compute, **we centralize data storage and model processing** while **ensuring future adaptability**. Demonstrably, PIVOT paves the way for connection to new models in future research endeavors. Still, work remains:

- Investigating additional Active Learning (AL) query strategies
- Optimizing storage of intermediate data required for AL querying
- Model Fine Tuning within in the AL loop