# BDA 450 Report

Andrew Benfante

April 2023

## Abstract

Using RGB images of the underside of Arctic sea-ice, machine learning methods were used in training models to predict the concentration of chlorophyll and thus an indication of under-ice algal growth. In the interest of simplicity, linear models were trained using the RGB information of an images's dominant color and subsisting sensor data. Additionally, a simple, unrefined convolutional neural network was trained as a proof of concept for future modeling and as a means of providing contrast to the resultant accuracy of the linear methods. Accuracy of all models was found to be tolerable as an indicator of broad Chl *a* levels but further tuning of the CNN would be required for higher accuracy predictions that could substitute for traditional sensor data.

## 1 Introduction

### 1.1 Background

In 2014 and again in 2018, two similar studies were carried out using autonomous buoys equipped with a collection of instrument systems that were deployed in Arctic sea ice to study under-ice algae blooms and the effects of various environmental factors on seasonal growth cycles. The concentration of chlorophyll in the water column was measured using a fluorometer. Of the various environmental factors measured, the amount of available light was of chief interest to researchers[2][1]. In both studies, the amount of available light beneath the ice was found to be a limiting resource in regards to the rate and extent of algae growth. Algae growth was measured as concentration of Chl *a* fluorescence ($mg/m^3$).

In addition to quantitative sensor data, a digital RGB camera was deployed on each buoy's instrument tether $\sim$20 m below the surface pointing upwards towards the ice. In contrast to the hourly data collection for sensor readings, the majority of images were taken at roughly solar noon each day of the buoy's deployment and clearly display the instrument tether along with other common features including ice-transparency, algae-induced green hue, cracks in the ice

sheet, and filamentous algae attached to the tether. Images are 120 by 160 pixels.

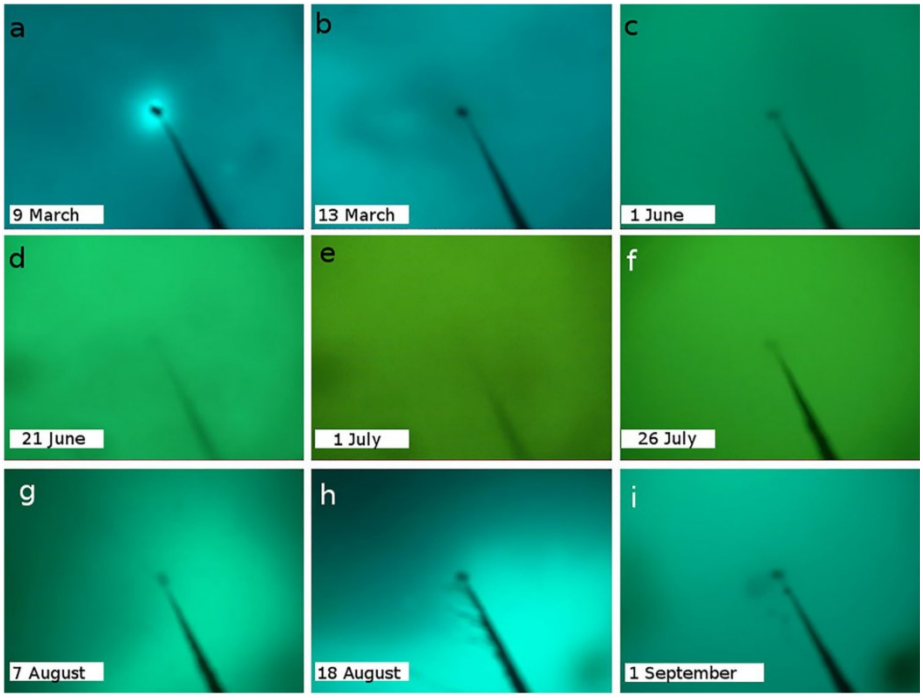


Figure 1: Time series of images[2].

## 1.2 Motivation

Deploying instruments and returning collected data in the Arctic circle is cost-prohibitive. Given the vivid color gradient across the time series of images collected in the studies, green hue corresponding to greater Chl *a* concentration, relatively cheap and low-maintenance digital image capture becomes a candidate for gathering data about algae growth using machine learning techniques. The significantly sized time-series sensor data and relatively small image sets already gathered can be used to train predictive models using outputs from minimal sensors in tandem with RGB data as well as using solely images to train computer-vision models. In the event that further experiments are conducted, new data can be used to continue training predicitive models, increasing their accuracy and effectiveness.

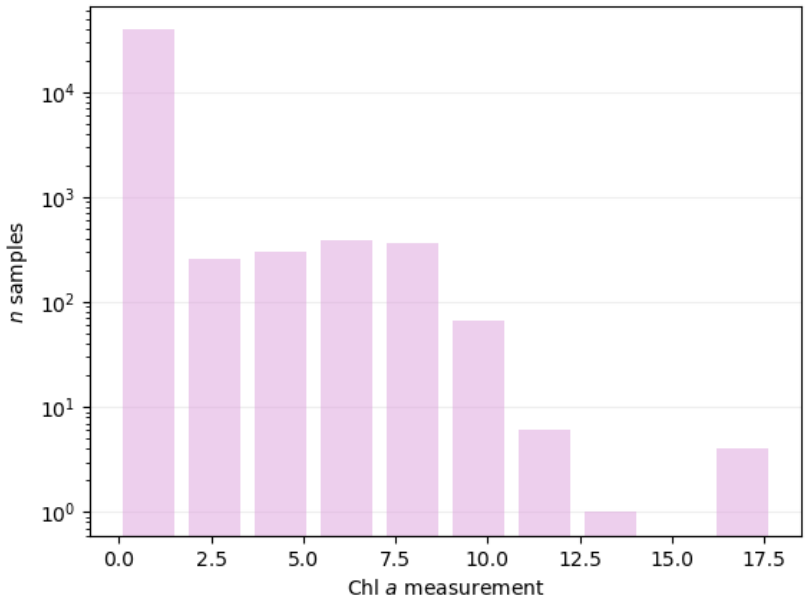


Figure 2: Distribution of raw Chl $a$ measurements

| Chl $a$ | $\sigma_{\text{Chl } a}$ | $n$ | $n_{\text{Avg}}$ |
|---|---|---|---|
| All | 1.223 | 40,686 | 1186 |
| $< 2$ | 0.238 | 39,365 | 1136 |
| $\geq 2$ | 2.084 | 1321 | 50 |

Table 1: Standard deviation for Chl $a$ over different ranges.

# 2 Methods

## 2.1 Note on Data Usage

Since the primary resource of interest was the collected RGB images, to accurately represent coinciding sensor data, the three observations nearest to time of image capture were averaged and used as an estimate for the data at time of image capture. In effect then, each image's corresponding sensor data accounted for the unweighted-average readings in a three hour window at which the time of image capture (predominantly solar noon) is roughly the median. This was also applied to the response variable in all models, Chl $a$. It should also be noted that in the original data, a significant portion of these measurements fell below a threshold of $2\,\text{mg/m}^3$ (Fig. 2). This discrepancy was preserved in the per-image averaged data (Table 1). This discrepancy was not explicitly accounted for in

any of the models.

## 2.2 Dominant Image Color as a Useful, Low-Dimension Predictor

A transformation of RGB pixel data to that image's two-most-dominant pixel colors was employed in an effort to reduce complexity of a prediction model and to reduce computational intensity during training. Examining image overlays for each image set (Fig. 3) as an approximation of each set's average image and dominant color is a low-overhead way to simplify the concept of image-based prediction for sets of mostly homogeneous images.

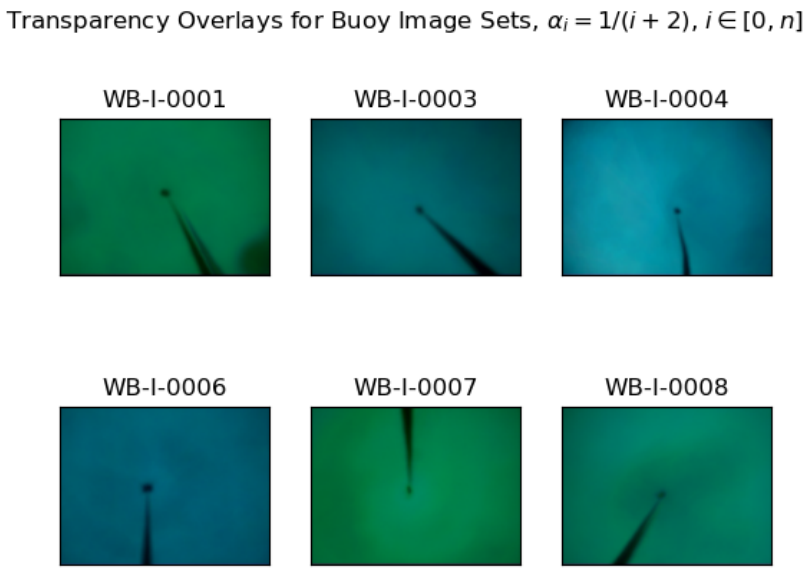


Figure 3: Transparency overlays incorporating all images in each individual buoy image set.

Aside from the angle of the buoys' instrument tethers and some vignetting, one of the primary differentiating factors between sets is the overall color of the images. Per the original studies and image collections, green color in the water is representative of higher concentration of chlorophyll. The differences between image-set overlays can be partially explained by the differing date ranges of each set i.e. images collected mostly during summer (peak algal bloom) will superimpose greener than images collected from Spring through Fall. Regardless, the color differences are apparent and are inspiration for further explo-

ration. Vignetting and ice translucency—while potentially containing useful information—would require a more robust learning method (e.g. Convolutional Neural Network) to harness in a prediction model. Again, in the original interest of reducing complexity, the potential of dominant colors was explored.

## 2.3 Quantization Method and Implementations

the `colorthief` library is based on a library originally implemented in JavaScript and uses modified median cut quantization (MMCQ) to extract RGB color palettes from RGB pixel arrays. This measurement differs from the average color of an image in that it is achieved by recursively sectioning off the 3D color space of an image (where each axis represents a color channel). The algorithm offers flexibility in terms of palette size and accounts for common errors and inaccuracies not addressed by the original median cut algorithm. Upon extracting a dominant color from each image, its color channels were found to be roughly normally distributed across the image set (Fig. 4).

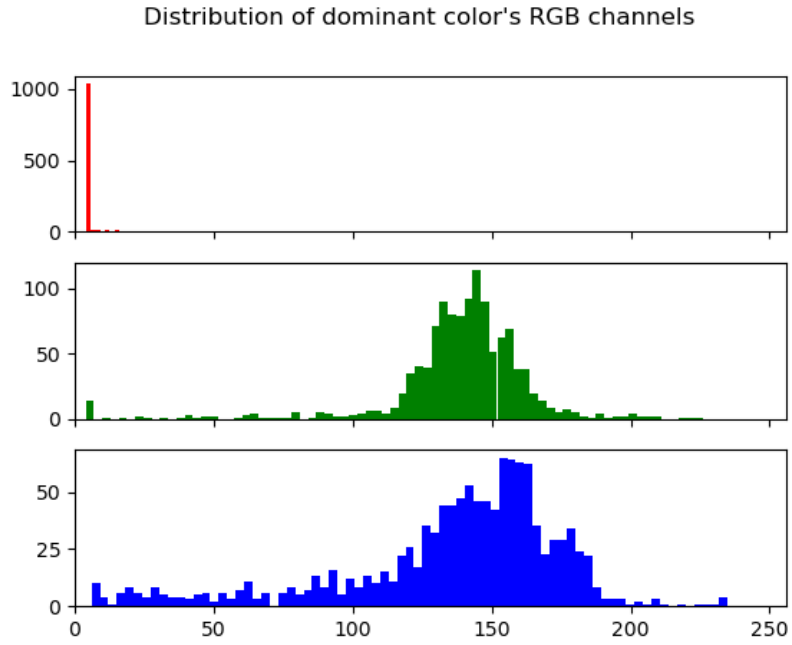


Figure 4: Dominant color channels by RGB value in [0, 255]. From Top to bottom, Red, Green, and Blue channels.

## 2.4 Regressing on Dominant Color and Common Factors

In the interest of training a predictive model to determine chlorophyll concentration using digital images from the buoy experiments, there is utility in harnessing the abundance of quantitative sensor data collected. It was a priority to choose reliable indicators of chlorophyll from the sensor data collected across both individual studies to bolster dominant color information. Naively, those sensor data common to all buoys were chosen and any identifying information concerning which readings came from which buoy were discarded. Across the two individual studies and the six buoys with available image data, 14 features appeared in all experiments (Table 2).

| Predictor | Description |
|---|---|
| PAR_1m | Available, full-spectrum light at 1 meter below the surface. |
| PAR_5dm | Available, full-spectrum light at a half meter below the surface. |
| air_temp | Surface air temperature. |
| chl_DOM_fluoro_460nm | Measure of dissolved organic material. |
| chl_backscatter_fluoro_532nm | 532nm-band irradiation Backscatter (higher readings equate to more particulate). |
| chl_fluorometer | Primary measurement of chlorophyll concentation. |
| latitude | Latitudinal position of the buoy at time of reading. |
| longitude | Longitudinal position of the buoy at time of reading. |
| month_dat_time_int | Number of seconds passed since 12:00 am, January 1st of that year. |
| pressure_db_20m | Pressure sensor (determines depth). |
| temp_10m | Temperature reading at 10 meters below the surface. |
| temp_25dm | Temperature reading at quarter meter below the surface. |
| temp_5m | Temperature reading at 5 meters below the surface. |
| temp_75d | Temperature reading at three-quarter meters below the surface. |

Table 2: Common factors used in linear modeling.

## 2.5 Regression Models

After extracting the dominant colors from the images and representing them as 3-channel RGB data and consolidating it with the sensor data, the data was fit

with three different linear models: an ordinary least squares (OLS) regression with normal error assumptions:

$$y_i = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + e_i,\ e \sim \text{i.i.d. } N(\mu, \sigma^2), \tag{1}$$

and then the same model penalized once with the $\ell_1$ norm (LASSO) and once with the $\ell_2$ norm (Ridge):

$$\min \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|; \tag{2}$$

$$\min \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2. \tag{3}$$

In both of the latter models, the tuning parameter $\lambda$ was chosen using a grid search scored on mean absolute error. All three models were trained using 10-fold cross validation (with 3 repetitions) using the scikit-learn library.

## 2.6 Convolutional Neural Network

In addition to the linear dominant-color models, a simple, untuned convolutional neural network (CNN) was constructed and trained using the TensorFlow framework. This was done in order to contrast the results of linear modeling methods as well as serve as proof of concept for future experiments. The network was comprised of two convolving layers each followed by a max pooling layer before a global average pooling layer, two hidden layers, and a dense output layer. The model was trained on down-sampled and randomly augmented images from the buoy image sets. The convolving layers and the hidden layers employed rectified linear units as activation functions and the output layer a raw linear classifier.

| **Layer (type)** | **Output Shape** | **Param #** |
|---|---|---|
| input_4 (InputLayer) | (None, 64, 64, 3) | 0 |
| conv2d_6 (Conv2D) | (None, 62, 62, 8) | 224 |
| max_pooling2d_6 (MaxPooling 2D) | (None, 31, 31, 8) | 0 |
| conv2d_7 (Conv2D) | (None, 29, 29, 16) | 1168 |
| max_pooling2d_7 (MaxPooling 2D) | (None, 14, 14, 16) | 0 |
| global_average_pooling2d_3 (Global-AveragePooling2D) | (None, 16) | 0 |
| dense_9 (Dense) | (None, 64) | 1088 |
| dense_10 (Dense) | (None, 64) | 4160 |
| dense_11 (Dense) | (None, 1) | 65 |

Table 3: Architecture of CNN.

| | |
|---|---|
| Total params | 6,705 |
| Trainable params | 6,705 |
| Non-trainable params | 0 |

Table 4: CNN parameters.

Data augmentation was implemented during image preprocessing to increase the virtual sample size that the model was trained on. Randomly applied transformations were horizontal flipping of the image, vertical flipping, and mild shearing. These transformations were chosen in part to avoid creating an association with Chl *a* measurement and the unmoving yet varied-between-sets location of the instrument tether in the frame for a given buoy image set. Transformations involving color and brightness were intentionally avoided given the significant relationship between visible light and algae growth as well as the saturated green hue of the water associated with peak Chl *a* levels.

# 3 Results

## 3.1 Linear Models

| **Predictor** | $t$ | $P > \|t\|$ |
|---|---|---|
| const | nan | nan |
| latitude | 5.946 | 0.000 |
| temp_25dm | nan | nan |
| air_temp | nan | nan |
| longitude | -2.673 | 0.008 |
| PAR_5dm | nan | nan |
| chl_backscatter_fluoro_532nm | nan | nan |
| pressure_db_20m | nan | nan |
| temp_10m | nan | nan |
| temp_5m | nan | nan |
| PAR_1m | -2.791 | 0.005 |
| chl_DOM_fluoro_460nm | nan | nan |
| temp_75dm | nan | nan |
| r1 | 20.672 | 0.000 |
| g1 | 9.325 | 0.000 |
| b1 | -19.188 | 0.000 |
| r2 | nan | nan |
| g2 | nan | nan |
| b2 | nan | nan |
| month_day_time_int | nan | nan |

Table 5: Significance of predictors in LASSO model.

Training linear models with various penalties resulted in similar performance metrics with the Ridge regression marginally outperforming LASSO and performing similarly to OLS (Table 6). Interestingly, the LASSO model selects the most intuitive parameters based on the research for its model: `latitude`, `longitude`, `PAR_1m`, `r1`, `g1`, and `b1` (Table 5).

| **Method** | **Test MAE** | **Test RMSE** | **Test $R^2$** | **Chl *a*** | $\sigma_{\textbf{Chl } a}$ |
|---|---|---|---|---|---|
| OLS | 0.403 | 0.671 | 0.685 | All | 1.223 |
| Ridge | 0.399 | 0.680 | 0.677 | $< 2$ | 0.238 |
| LASSO | 0.402 | 0.696 | 0.662 | $\geq 2$ | 2.084 |

Table 6: Performance metrics for linear models using test partition.

Looking at the accuracy graphs (Fig. 5 ), (Fig. 6), all of the linear models perform poorly at predicting higher concentrations of Chl *a* and lower concentrations are predominantly predicted with low precision. In all models, it was found that only the channels of the first dominant color in an image seem to be significant predictors of Chl *a*.

## 3.2 CNN

Compared to the linear models, the CNN demonstrated an improvement in performance (Table 7). Observing its accuracy graphs (Fig. 7), (Fig. 8), the model predicts higher concentrations with greater accuracy than the linear models although it still has difficulty predicting lower Chl *a* concentrations (measurements between 0 and 4). This along with the results of the linear models likely indicates nonlinear behavior of Chl *a*.

| **Metric** | **Training Data** | **Test Data** | **Chl *a*** | $\sigma_{\textbf{Chl } a}$ |
|---|---|---|---|---|
| RMSE | 0.285 | 0.444 | All | 1.223 |
| MAE | 0.254 | 0.292 | $< 2$ | 0.238 |
| $R^2$ | 0.812 | 0.690 | $\geq 2$ | 2.084 |

Table 7: Performance metrics for CNN

## 3.3 Accuracy Graphs

### 3.3.1

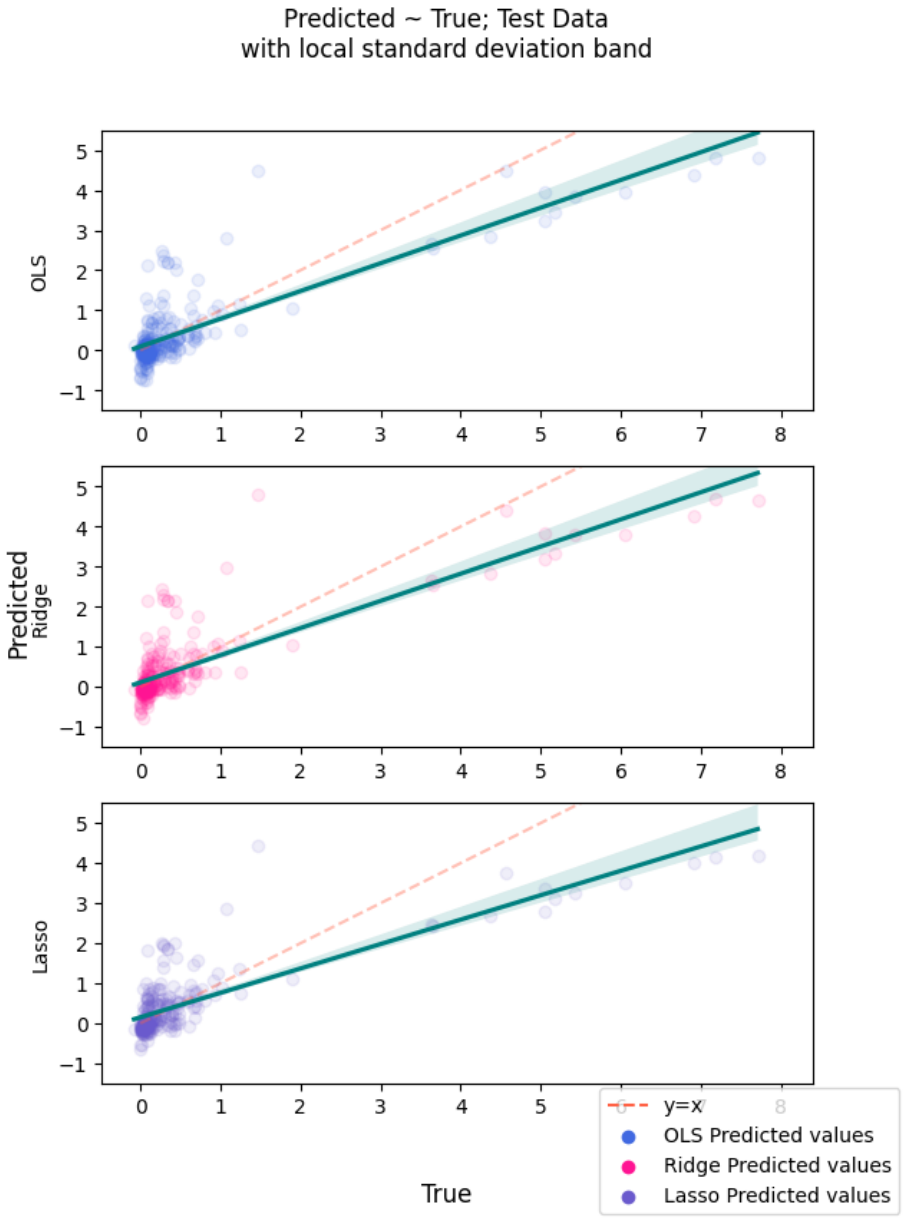


Figure 5: Linear model test performance: Predicted Chl $a$ $\sim$ True Chl $a$.

### 3.3.2

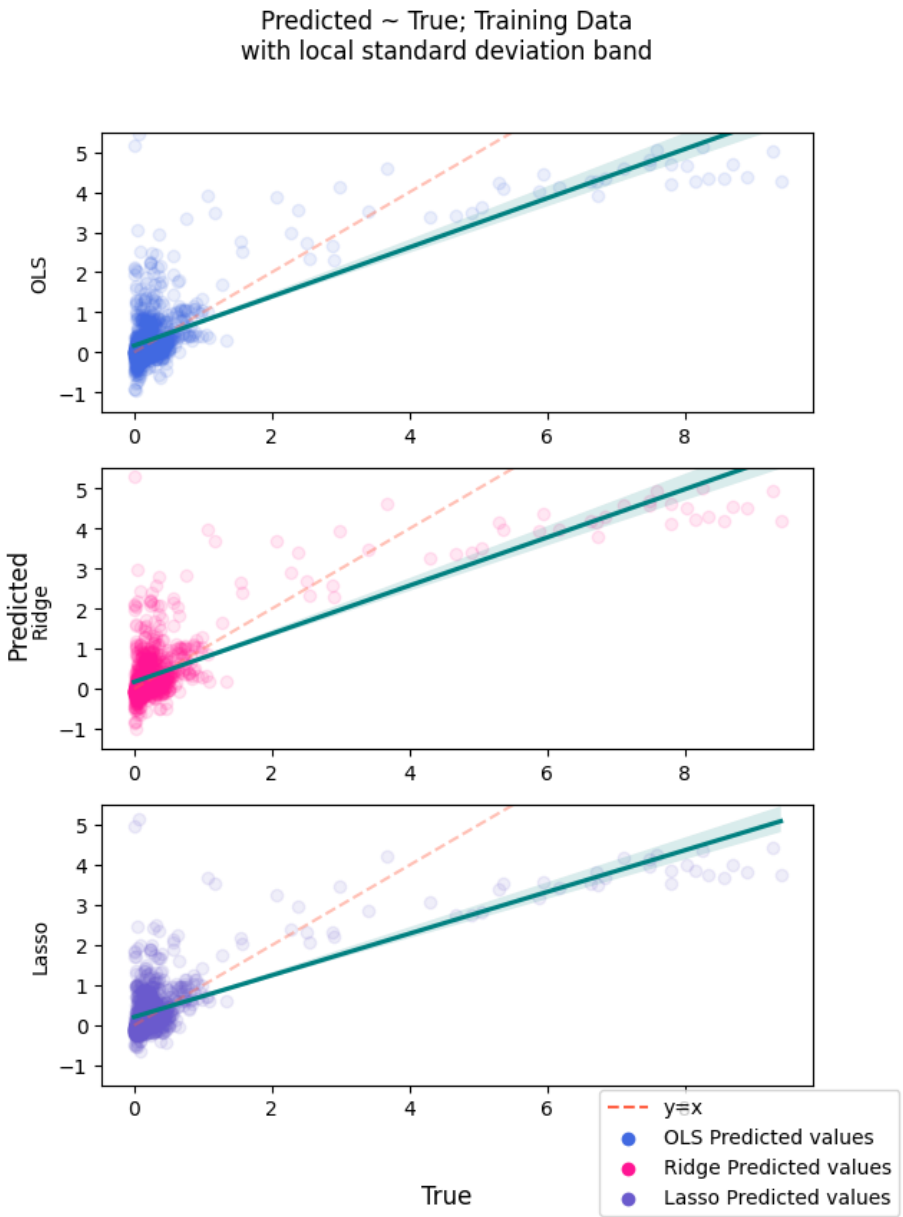


Figure 6: Linear model train performance: Predicted Chl $a$ ~ True Chl $a$.

#### 3.3.3

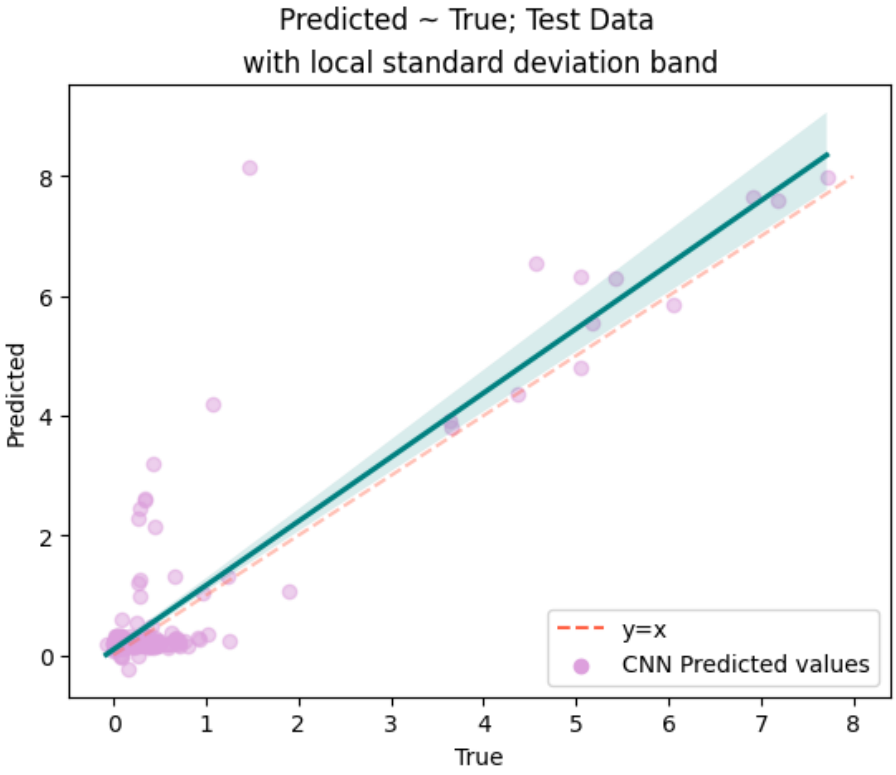


Figure 7: CNN test performance: Predicted Chl $a$ $\sim$ True Chl $a$.

#### 3.3.4

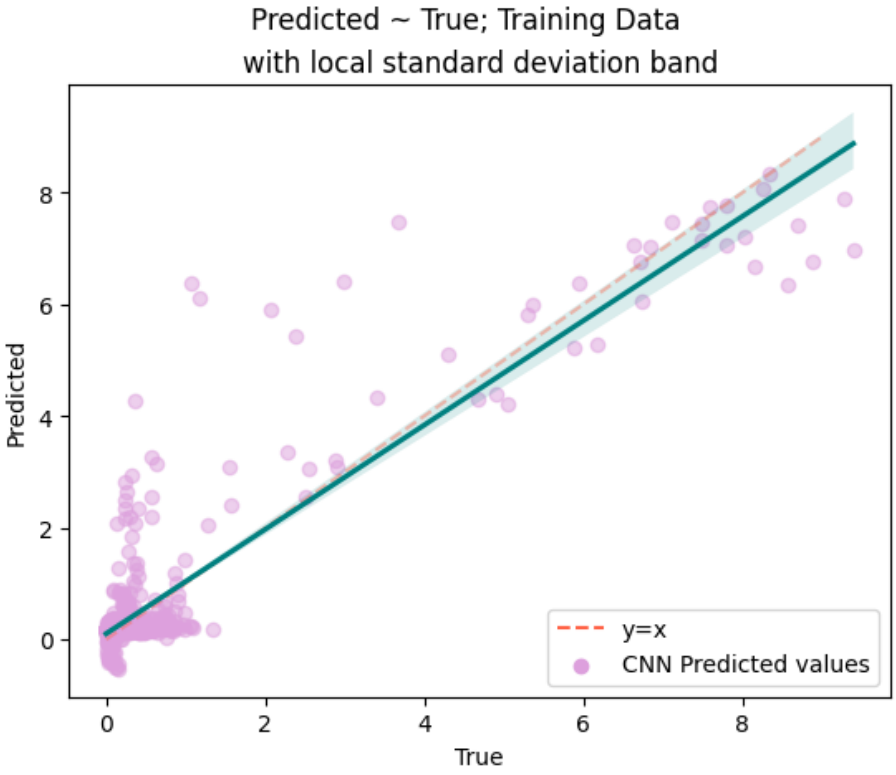


Figure 8: CNN train performance: Predicted Chl *a* ∼ True Chl *a*.

# 4 Discussion

## 4.1 Limitations

The sample size of the data was the primary limitation in model training. With roughly 1000 images, the risk of over fitting the data is high. Using the dominant color approach, the majority of the image information is lost. While this simplifies the model, it was found to be a limiting factor during training of the linear models. Conversely, employing data augmentation techniques during the training of the CNN proved to be beneficial to prediction accuracy. Regardless, an increased image training set would nonetheless improve the supervised learning process.

## 4.2 Conclusions

While there is some prediction accuracy, the dominant color approach appears to be ineffectual at predicting Chl *a*, at least for linear models and with a smaller sample size. Going forward, nonlinear models should be tried such as polynomial splines as well as models that explicitly account for the discrepency in amount of low-concentration samples vs. high-concentration samples. While the linear models trained here are not significantly accurate to justify replacing fluorometers, they could still serve a purpose as a fidelity check for diagnosing malfunctioning equipment or for experiments where categorical indication is preferred over precision of measurement. These models are not complex, highly legible and intuitive in terms of human understanding. This is always an advantage when working with and presenting experimental data.

The Convolutional Neural Net is imperfect and untuned. Going forward, experimenting with hyperparameter optimization and model architecture could prove rewarding as a low-cost alternative to more-sophisticated sensor equipment in future experiments. In this case, a higher number of autonomous RGB cameras could be deployed in lieu of or in addition to the instrument systems of the previous two experiments, broadening the geographic collection range, increasing the amount of useful data gathered, and enabling a rapid expansion of image training sets. In either case, understanding how to architect the model to effectively capture information concerning color, water transparency, as well as ice and snow thickness would likely result in high prediction accuracy based on the research in [2] and [1].

Another potential improvement could be found in more rigorous filtering of training images to remove darkened images, outlier images, and images taken during the fall and winter when algae growth is nonexistent or sporadic and atypical. Further standardization of image capture procedures would also ensure easy cleaning for future model training efforts.

# References

[1] Victoria J. Hill et al. "Contrasting Sea-Ice Algae Blooms in a Changing Arctic Documented by Autonomous Drifting Buoys". In: *Journal of Geophysical Research: Oceans* 127.7 (2022), e2021JC017848. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2021JC017848.

[2] Victoria J. Hill et al. "Light Availability and Phytoplankton Growth Beneath Arctic Sea Ice: Integrating Observations and Modeling". In: *Journal of Geophysical Research: Oceans* 123.5 (May 2018), 3651–3667.