DS - GA 1017 RDS Final Report

Fiona Chow and Jennah Gosciak

Spring 2023

Background

The purpose of this automated decision making system (ADS) is to use satellite imagery to detect and classify the severity of cyanobacteria blooms in small, inland water bodies. This will help water quality managers better allocate resources for in situ sampling and make more informed decisions around public health warnings for drinking water and recreation. The primary goal of this ADS is to achieve accuracy.

The data and ADS come from a data science competition hosted by NASA and DrivenData. DrivenData is a company that hosts social impact data science competitions like those found on Kaggle. We chose to audit the solution of the second placed winner for the "Tick Tick Bloom: Harmful Algal Bloom Detection Challenge" that completed on Feb 17, 2023. The data and code implementing the ADS can be found on the winner's GitHub page with documentation. DrivenData requires that all winning solutions are open source under The MIT License.

We selected this particular ADS because we were curious to see if there were subpopulations for which the ADS was less accurate in prediction. Algal blooms are an environmental justice issue. Exposure to high levels of cyanobacteria has been linked to cancer, birth defects, and even death (Gorham et al., 2020; Schaider et al., 2019). Prior research suggests that in the U.S. there are significant racial and socioeconomic disparities in access to clean drinking water (Schaider et al., 2019). Because this ADS could help water quality managers better allocate resources for in situ sampling and make informed decisions around public health warnings for drinking water and recreation, the algorithm must be accurate not only for the overall population but also for sensitive subpopulations.

Input and Output

Description

The competition provides 23,570 rows of training and test data. Since the test data do not have labels, we conduct our audit using the training data (n=17,060). Each row in the training data is a unique in situ sample collected for a given date and location going back to 2013. We sample from the training data to speed up the runtime of training and hyperparameter tuning. First, we restrict our data to instances recorded after 2016. From the filtered data, we sample so as to ensure proportional levels in each region to the original data while using a 72%/28% train/test split that is similar to what was done during the competition. To illustrate this, we present the number of sampled training observations by year in figure 1. Since the competition test data used samples with unseen latitude and longitude (i.e. "spatial holdouts"), we made sure there was no spatial overlap in our training and test data. Also similar to the ADS implementation, we then split the new training data (the 72%) further into a training and validation set. We tune the hyperparameters on the validation set using the same approach to tuning as in the competition. The author of the ADS uses custom functions for hyperparameter tuning. The function takes in a defined number of rows for selecting the validation set and performs cross-validation across 10 splits. We report performance on the test set that was unseen during training and validation.

Auxiliary Data. We are using American Community Survey (ACS) 2015-2019 5-year estimates at the census tract level for our audit. We chose to use the 2015-2019 estimates, as there have been documented problems and delays associated with subsequent surveys due to COVID-19 (Wines and Cramer). There are approximately 74,000 census tracts for all 50 U.S. states, Washington D.C., and Puerto Rico. However, only

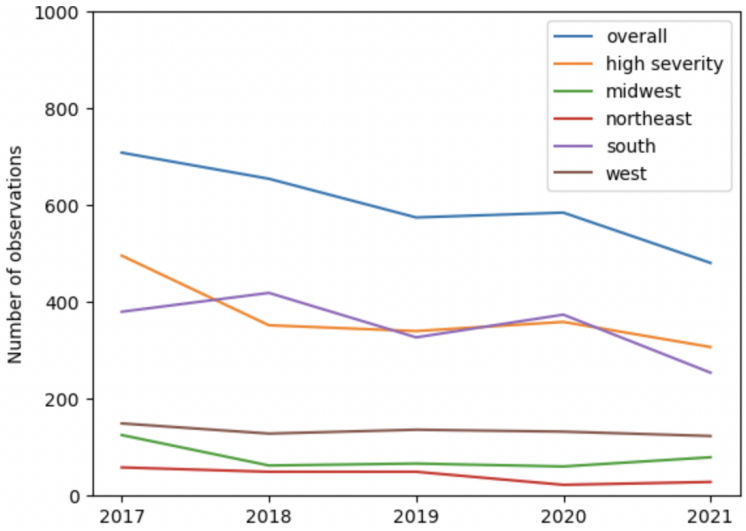

Figure 1: Number of training data observations by year

1,660 census tracts are uniquely represented in the training data. This occurs because some in situ samples appear to come clustered in the same geographic areas. For example, we observed that 5,308 samples come from only a handful of census tracts in Chatham County, North Carolina. The features we have collected using census data include: race, ethnicity, median household income, and poverty rate.

Input

All the input datasets can be linked using a unique string identifier, *uid*g. *uid* identifies the date and location (latitude and longitude) for each sample. We present detailed information on each of the input datasets used in the ADS. Note, as we explain above, we use a sampled version of these datasets in our actual audit to reduce the runtime of training and hyperparameter tuning and this section will detail information of this sampled dataset.

- **Train labels:** These data only cover the training records and provide information on the outcome for testing and validation purposes. We plot the value distributions in figure A1. We do not look at correlations as the relationship between severity and density is clearly defined.

Feature	Type	Description
region	string	unique values: northeast, south, west, midwest
severity	integer	unique integer values: 1-5, correspond to density levels
density	float	cyanobacteria cells per mL
uid	string	unique identifier

Table 1: Description of features in the labeled training data

- **Metadata:** The metadata provide information on the context of the sample. The columns include:

Feature	Type	Description
latitude	float	latitude where the sample was collected
longitude	float	longitude where the sample was collected
cluster	integer	generated from latitude and longitude to account for spatial variation
date	string	data the sample was collected
split	string	denotes "train" or "test" data in the original sample
uid	string	unique identifier

Table 2: Description of features in the metadata

- **Elevation:** Elevation data for the ADS come from the Copernicus Digital Elevation Model (DEM) with 30-meter resolution. The DEM is a digital surface model that can provide information on buildings, infrastructure, and vegetation. There are 3,000 entries of elevation data and no missing data in any of the columns. The columns, some of which are constructs that the author of the ADS generated, are:

Feature	Type	Description
latitude	float	
longitude	float	
box	integer	the distance (in meters) from the given latitude and longitude coordinates
elevation	float	
mine	float	minimum elevation
maxe	float	maximum elevation
dife	float	difference in elevation
avge	float	average elevation
stde	float	standard deviation in the elevation
DateTime	string	date of the data download
uid	string	unique identifier

Table 3: Description of features in the elevation data

Each input feature of the elevation data has the following value distribution:

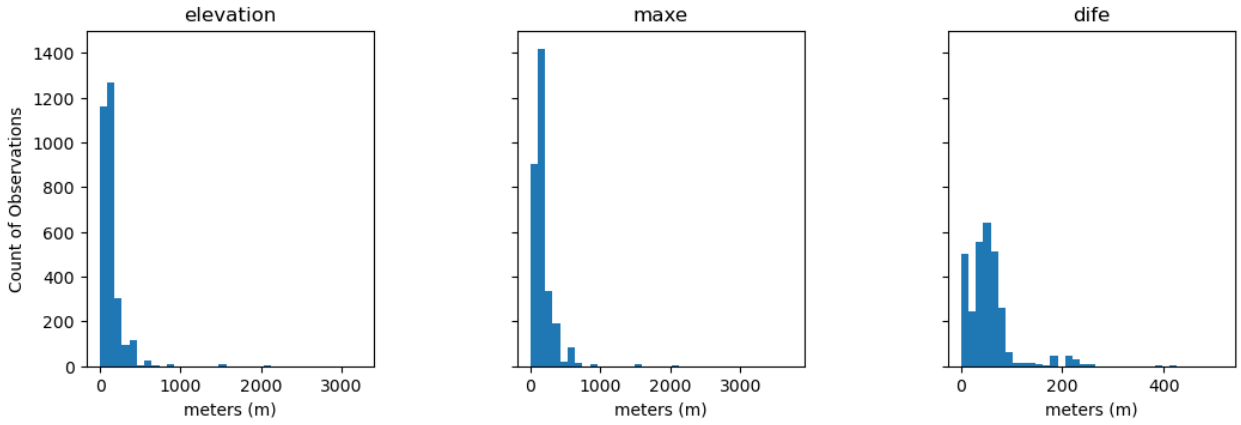

Figure 2: Histograms of input features from elevation data

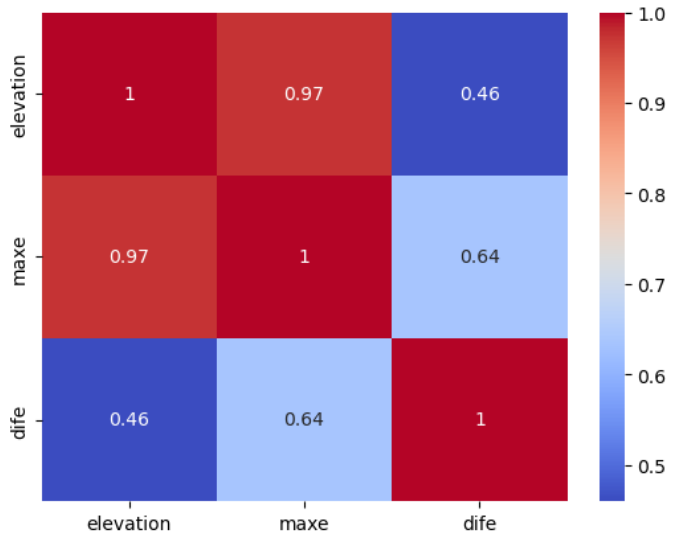

Figure 3: Correlation of input features from elevation data

In figure 2, elevation data range from 0 to 4,000 meters. The median value for elevation, maximum elevation and difference in elevation is 115, 154 and 50 meters respectively. In figure 3, the correlation matrix shows that there is a strong relationship between maximum elevation and elevation data ($r = 0.97$), which is expected as both are based on elevation in a pre-specified fixed bounding box.

- **Climate:** Data come from the National Oceanic and Atmospheric Administration (NOAA) and provide information on temperature, wind, and precipitation. The author of the ADS acknowledged that he has not used climate data, and so we are not presenting any additional information on this dataset.
- **Satellite:** Sentinel-2 Level-2A satellite imagery come from the European Commission in partnership with the European Space Agency (ESA) and it provides information on the spectral bands – red, blue and green– at 1000 and 2500 meter radius from latitude and longitude and at ten-day intervals. The water areas were identified by k-means image segmentation and the satellite images were selected based on low cloud cover of less than five percent so as to increase the likelihood of capturing algal bloom detection in water surfaces. The columns, some of which are constructs that the author created, are:

Feature	Type	Description
imtype	object	
prop_lake_1000	float	estimate of water area at 1000 meters from latitude/longitude
r_1000	float	estimate of red inside water area at 1000 meters
g_1000	float	estimate of green inside water area at 1000 meters
b_1000	float	estimate of blue inside water area at 1000 meters
prop_lake_2500	float	estimate of water area at 2500 meters from latitude/longitude
r_2500	float	estimate of red inside water area at 2500 meters
g_2500	float	estimate of green inside water area at 2500 meters
b_2500	float	indicates estimate of blue inside water area at 2500 meters

Table 4: Description of features in the satellite data

Missing values were encoded with value '-1'. There may be missing data because satellite images could be limited due to cloud cover. In this case, there are 1,695 entries of satellite images and no missing data in any of the columns.

In figure 4, the red, green and blue (R/G/B) pixel values range from 0 (representing the minimum possible color intensity) to 255 (representing the maximum possible color intensity) and is the actual

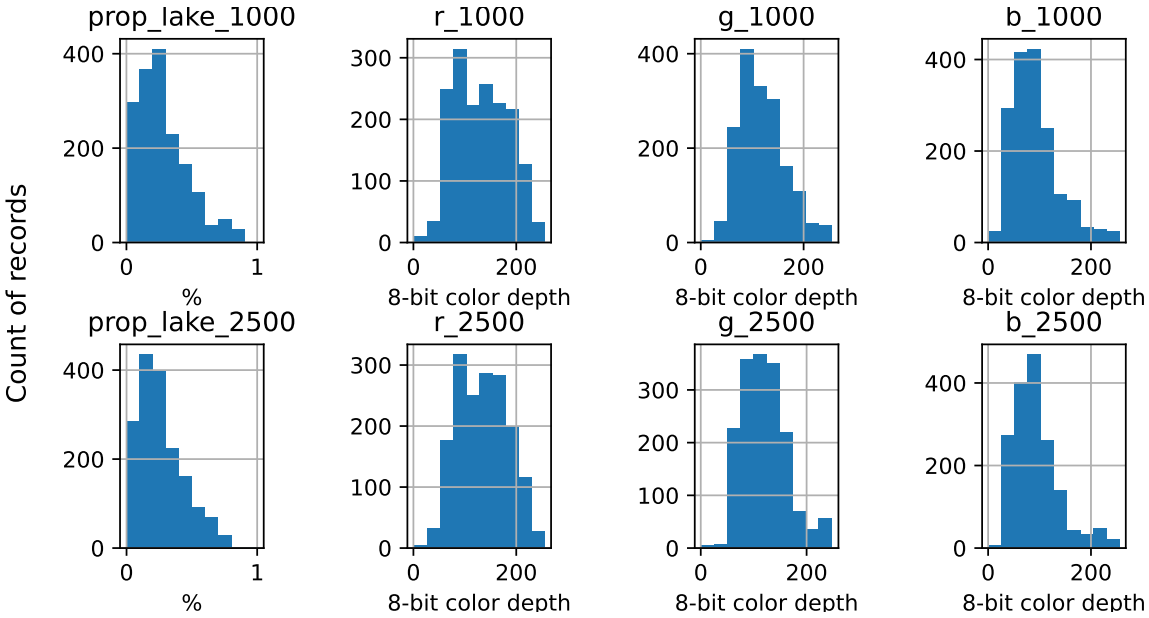

Figure 4: Histogram of input features from satellite images

normalized range from which the Sentinel-2 images are created with. The proportion of image that is classified as lake values range from 0 (representing the entire image is not lake) to 1 (representing the entire image is lake).

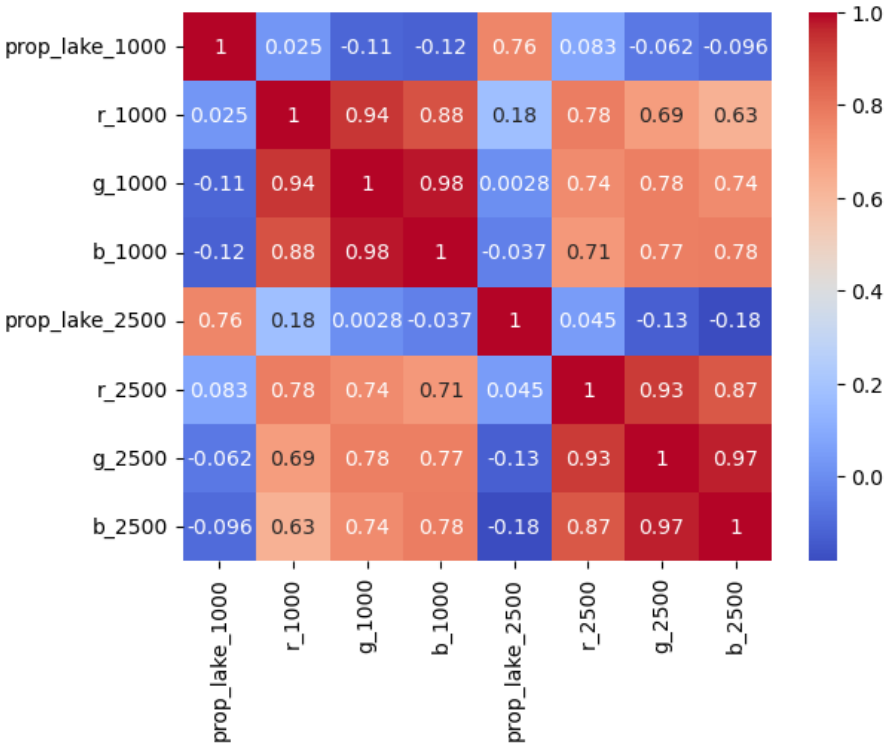

Figure 5: Correlation of input features from satellite images

In figure 5, the correlation matrix shows that there is a strong relationship between blue and green spectral bands ($r = 0.98$) and red and green spectral bands ($r = 0.94$) at their respective distance from latitude and longitude. In addition, there is a strong relationship between the spectral bands of the same color at different distances from the latitude and longitude ($r = 0.8$).

Output

The output of the system formally is a severity level that takes integer values 1 through 5. According to DrivenData, the severity is based on cyanobacteria density. Cyanobacteria density is another column in the training data, which ranges from 0 to 804,667,500 cells per mL. As table 5 illustrates, density values are non-overlapping for distinct severity levels and higher density values are associated with higher severity levels. For example, severity level 5 encompasses the the highest density values, greater than 10 million cells per mL. According to the World Health Organization (WHO), moderate and high risk health exposures occur at density levels $\geq 20,000$ cells per mL (WHO, 2003). Using this classification approach, we construct a binary outcome label for a high risk health exposure. The binarized label aligns with severity levels 2-5, as table A1 indicates. Slightly less than half (≈ 44 percent) of the training data are low-risk. DrivenData specifies that the submission is based on severity levels and not the raw, underlying densities.

Severity	Density range (cells/mL)
1	$< 20,000$
2	$20,000 - < 100,000$
3	$100,000 - < 1,000,000$
4	$1,000,000 - < 10,000,000$
5	$\geq 10,000,000$

Table 5: Formal severity level ranges

	Severity	Percent of Total
Low	1	38
High	4	22
	2	20
	3	20
	5	< 1

Table 6: Binarized severity based on estimated risk

Implementation and Validation

Data cleaning / Pre-processing

In this section, we document several decision that the ADS winner made, which may have had consequences on both accuracy and fairness performance. Because the ADS does not use scikit-learn, we are not able to run ml-inspect and instead our inspection of the ADS is manual.

During pre-processing, the winner of this ADS did not normalize the elevation data and the satellite data already came normalized.

One decision the winner of this ADS made was to create an ad-hoc cluster variable (represented by the ordinal variable 'cluster') as there is substantial spatial variation in the target variable. For example, the south region had the lowest average severity levels at 1.57 while the west region was 3.74. By creating this ad-hoc variable, he was able to better model the patterns in the different regions based on more granular spatial groupings.

High Level Implementation

The solution uses an ensemble of three different boosted tree models – XGBoost, CatBoost and LightBoost – and features such as region, date, location cluster, elevation and Sentinel-2 satellite images – red, blue and green spectral bands at 1,000 and 2,500 meters from latitude and longitude.

To optimize the tree models, the solution employed a process of hyper-tuning nine different parameters. Two of these parameters took integer values and the remaining seven were categorical. The integer-valued parameters were the number of boosted trees and maximum tree depth. The categorical parameters included the type of elevation data, type of XY coordinates, type of slope data, type of region data, a boolean

indicating whether to use sample weights when fitting the model, a boolean indicating whether to treat categorical features as numeric, and the type of satellite data to use.

ADS Validation

This ADS is validated by the region-averaged root mean squared error (RMSE). The smaller the error value, the more accurate the model is. Region-averaged RMSE is calculated by taking the square root of the mean of squared differences between estimated and observed severity values for each region and averaging across regions.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(Y_i - \hat{Y_i})^2}$$

$$RegionAveragedRMSE = \frac{RMSE_{Midwest} + RMSE_{West} + RMSE_{South} + RMSE_{Northeast}}{4}$$

- Y_i = the true severity level for the i^{th} record
- $\hat{Y_i}$ = the predicted severity level for the i^{th} record
- N = the total number of records

Given this performance metric, the author of the ADS used regression models to generate a continuous outcome. He then rounded the predictions to the nearest integer value. For the ensemble models, he averaged across the predictions before rounding, and then rounded the final result. The ADS met its goal of achieving high performance as evidenced by winning second place in the competition with a sufficiently low region-averaged RMSE of 0.7616. Using the sampled dataset, our implementation of the ADS scored a similar region-averaged RMSE on the test set of 0.7275.

Outcomes

We study the performance of the ADS across four subpopulations. Since we are using a binary version of the outcome variable, we use metrics commonly studied in binary classification problems, which we outline below. To further assess the validity of our findings, we compare our results to the performance of the ADS across subpopulations using RMSE averaged over the four regions.

Subpopulations

Based on the ACS data we have, we identify sensitive groups using the following construct definitions. We evaluate the performance of the ADS across these sensitive groups.

- **Above average poverty rate:** We have ACS data at both the census tract and state level. We create indicator variables that denote whether the poverty rate in a given census tract is above the statewide average.
- **Above average shares of racial and ethnic subgroups:** We create a series of indicator variables, corresponding to racial and ethnic subgroups that denote whether a given census tract has an above average population share of each subgroup. We generate these indicator variables for each subgroup in comparison to the statewide average. We also create an indicator that aggregates all non-white racial subgroups. For this audit, we focus on the indicator for non-white racial subgroups to reduce the total number of subgroups that we discuss. However, an expansion of this audit should ultimately consider the performance of the ADS across all racial subgroups as there may be important heterogeneity.
- **Low Income Community:** this is a designation from Internal Revenue Code §45D(e). Broadly, it refers to any census tract where the poverty rate is greater than or equal to 20 percent or the median family income is less than 80 percent of either the statewide median family income or the metropolitan area median family income–whichever is greater. We propose a modified version of this definition, given the complexity associated with recreating the first approach, that only compares the median family income to the statewide median.

Accuracy

Since we turned the ADS into a binary classification problem, we assess performance across conventional metrics like recall, precision, and accuracy. Binary classification makes sense in this audit because we are most interested in whether the ADS fails to identify high-risk groups; the precise level of severity is less significant.

We evaluate the performance of the ADS overall and across the various subpopulations outlined above. We prioritize recall and false negative rates (FNR) in our assessment of the ADS as there are greater risks for failing to detect algal blooms: not issuing a public health warning, exposure to toxic bacteria, and serious health and safety consequence.

Table 7 presents the results from both the baseline classifier and the overall highest-performing ensemble classifier across various performance metrics that we care about. Table 8 shows the performance of the ADS model across subgroups. We look at the performance of the model in census tracts that are *above* the statewide average for a given subgroup definition and *below*.

	Baseline	Overall
Accuracy	0.69	0.81
Precision	0.69	0.82
Recall	1	0.93
False Negative Rate	0	0.07
False Positive Rate	1	0.48

Table 7: Evaluation of Performance Overall in Comparison to the Baseline Classifier

From table 8, we see that in general, accuracy is high overall and across subgroups. As table 6 indicates, the base rate for predicting high severity is 62%. This means a baseline classifier that predicts high severity (prediction = 1) in all cases will achieve 62% accuracy. Accuracy overall and for subgroups is higher than the base rate in all cases. In addition, both recall and precision are high; recall, which we care about the most, is ($> 89\%$) overall and across subgroups. The relatively high rates of accuracy, recall, and precision suggest that the ADS has societal value and utility. Residents who live near toxic lakes and reservoirs will benefit from high recall while overburdened governmental agencies will benefit from the accuracy of the ADS.

		Subgroups in reference to the statewide average			
		Poverty rate	Low income	Non-White	Hispanic/Latino
Accuracy	Above	0.85	0.84	0.78	0.91
	Below	0.80	0.78	0.81	0.77
Precision	Above	0.89	0.85	0.81	0.92
	Below	0.80	0.79	0.82	0.77
Recall	Above	0.93	0.96	0.89	0.98
	Below	0.93	0.90	0.94	0.91
False negative rate	Above	0.07	0.04	0.11	0.02
	Below	0.07	0.10	0.06	0.09
False positive rate	Above	0.52	0.55	0.44	0.77
	Below	0.47	0.44	0.48	0.48

Table 8: Evaluation of Performance across Subgroups

The ADS does make errors, but the errors are frequently less consequential. For example, the FNRs are low overall and across subgroups. The highest FNRs are 11% for census tracts with non-white population rates above the statewide average. False positive rates are higher (≈ 0.5) overall and across subgroups, which is

less concerning. From the perspective of government agencies and water management organizations, the false positive rates are likely wasteful and increase business expenses. But from the perspective of the general public, they are not a significant concern.

Lastly, table 8 allows us to consider fairness, which we discuss further in the next section. Overall, the performance of the model does not vary significantly across subgroups. Systematically, the ADS is slightly lower-performing for census tracts with above average non-white population percentages and highest performing for census tracts with above average Hispanic/Latino population percentages. Interestingly, the ADS performs better for both high poverty and low income census tracts; both accuracy and precision are higher than for the performance of the model overall.

Fairness

According to the Aequitas Fairness Tree (Ghani, accessed April 17, 2023), the best fairness metric for this ADS use case is recall parity among the different subpopulations. Water quality managers can only allocate limited resources for in situ sampling because in situ sampling is labor intensive and expensive (Granger et al., 2018). Consequently, we should attempt to ensure the results of the ADS is distributed in a representative way. Since the false negative rate (FNR) is equal to 1−recall, we also look at false negative rate disparity among the sensitive subpopulations. Ultimately, we want to ensure that the ADS is not biased in errors, i.e. failing to provide assistance given group membership in a protected class.

We use the following fairness metrics from Saleiro et al. (2018): false negative rate disparity, recall disparity, and predicted prevalence disparity. g_i indicates membership in a protected group i, g_{ref} indicates the membership in the reference group, $\hat{Y}$ is the predicted value of the outcome, and Y is the true value of the outcome.

$$FNR_{disparity_{g_j}} = \frac{FNR_{g_j}}{FNR_{g_{ref}}} = \frac{Pr(\hat{Y} = 0|Y = 1, G = g_i)}{Pr(\hat{Y} = 0|Y = 1, G = g_{ref})}$$

$$Recall_{disparity_{g_j}} = \frac{Recall_{g_j}}{Recall_{g_{ref}}} = \frac{Pr(\hat{Y} = 1|Y = 1, G = g_i)}{Pr(\hat{Y} = 1|Y = 1, G = g_{ref})}$$

$$PPrev_{disparity_{g_j}} = \frac{SelectionRate_{g_i}}{SelectionRate_{g_{ref}}} = \frac{Pr(\hat{Y} = 1|G = g_i)}{Pr(\hat{Y} = 1|G = g_{ref})}$$

We consider these metrics to be fair using the 80% rule:

$$\tau \leq DisparityMeasure_{g_j} \leq \frac{1}{\tau} \Rightarrow 0.8 \leq DisparityMeasure_{g_j} \leq 1.25$$

We also consider additional fairness measures from FairLearn, including FNR difference, FPR difference, demographic parity difference, FNR ratio, FPR ratio, demographic parity ratio, and equalized odds ratio, which we define and present in table A4 in the appendix. Qualitatively, the differences are small overall (all are less than 0.15) and most ratios are between 0.8 and 1.25. The results do not lead us to different conclusions, compared to table 9.
The disparity metrics that we present in table 9 are equivalent equal to the ratio measures that FairLearn uses in the case where there are only two groups and the reference group is the group with the maximum value. For example, demographic parity ratio is defined as:

$$\frac{minimum_{g_i} P(\hat{Y} = 1|G = g_i)}{maximum_{g_i} P(\hat{Y} = 1|G = g_i)}$$

.

From table 9, we see that while there is recall parity across all subgroups, there are some FNR differences between subgroups (from table 8). There is a 5% higher FNR in the non-White subgroup which means that

the non-White subgroup has a higher likelihood of not receiving assistance when severity levels are high. Looking at non-White FNR disparity, where

$$FNR_{disparity_{non-white}} = \frac{FNR_{non-white}}{FNR_{white}} = \frac{0.11}{0.06} = 1.83$$

(numbers from table 8 – FNR and Non-White), the FNR disparity for the non-White subgroup is 1.83 and falls outside of the rule-of-thumb for fairness of $0.8 \leq disparity \leq 1.25$.

		Poverty rate	Low Income	Non-White	Hispanic/Latino
	FNR	1.12	0.38	1.83	0.26
Disparity	Recall	0.99	1.07	0.95	1.07
	Predicted prevalence	1.09	1.18	0.93	1.19

Table 9: Comparison of Fairness Metrics across Subgroups

Interpretability

In terms of interpretability, the model has a relatively low level of transparency due to its ensemble structure, which combines many boosted trees and incorporates numerous features, resulting in a highly non-linear model. Additionally, though we hoped to use LIME or SHAP to better understand the features used in the models, we ultimately found that the functions in both Python packages do not work well with the type of complex ensemble structured used here. Though the packages are model-agnostic, it is challenging to apply them to a pipeline that does not utilize built-in predict functions.

Instead, we relied on the 'feature importances' methods from the catboost, lightgbm and xgboost libraries. The results of the feature importances that we present here are similar to the feature importances that the author of the ADS shared with the winning competition model. As can be seen from table 10, longitude, cluster, elevation and month were the most important features in predicting an accurate outcome across all models. In other words, the primary focus of the ADS solution is on fitting a curve based on temporal and spatial variations, which may not generalize well to situations where the underlying spatial and temporal patterns are substantially different from those captured by the model. In such cases, the performance of the model may be suboptimal and additional data sources or modeling approaches may be required to achieve accurate predictions.

The feature importance of longitude is of some concern, as it could indicate leakage. Though we carefully sampled the test data to prevent any spatial overlap with the training data, during training and cross-validation the model may be memorizing patterns and information correlated with longitude that may not be available when the ADS is deployed. The creator of the ADS seems to have considered this and accounted for the probability that a sample is in the test data during cross-validation. However, based on the feature importances, we recommend further investigation of leakage, particularly with the successful implementation of tools like LIME and SHAP.

Catboost Variable	Feature Importance
longitude	0.246
cluster	0.243
elevation	0.178
stde	0.099
month_date	0.071
region	0.07
days_date	0.03
weekday_date	0.01
g_2500	0.008
prop_lake_500	0.007
g_500	0.007
prop_lake_2500	0.007
b_2500	0.006
b_500	0.006
imtype	0.004
r_2500	0.004
r_500	0.003

Lightboost Variable	Feature Importance
longitude	0.275
maxe	0.208
stde	0.177
month_date	0.101
days_date	0.094
prop_lake_500	0.052
cluster	0.044
r_500	0.023
b_500	0.013
weekday_date	0.008
region	0.003
imtype	0.003
g_500	0.0

XGBoost Variable	Feature Importance
longitude	0.566
cluster	0.078
month_date	0.071
latitude	0.068
stde	0.039
elevation	0.038
r_1000	0.029
weekday_date	0.026
g_1000	0.026
prop_lake_1000	0.023
b_1000	0.018
days_date	0.017

Table 10: Feature Importance of each tree model

Stability

To assess stability, we evaluated the performance of the ADS over 10 random samples of the original test data. We focused on the performance of a single model, trained on one dataset, to reduce runtime issues we encountered with hyperparameter tuning. It's important to evaluate the stability of the model over random test set samples, as the data used in deployment might look very different. Additionally, these findings will indicate whether researchers might be able to deploy this model in a range of disparate contexts.

As figure A2 demonstrates, overall the model is relatively stable across train/test splits. Only the false positive rate has some variation and a larger interquartile range. While a more thorough analysis might use external datasets and different contexts (e.g., large bodies of water as opposed to inland lakes and reservoirs– the focus of this contest), we exploit the fact that we filtered the data to records post-2016 and examine the performance of the ADS on unseen pre-2016 data. Pre-2016 data are more likely to differ from the training data than the held-out test set that we are using. As figure A3 shows, the ADS performs similarly well in terms of stability though there are some differences in the level of its performance. For example, accuracy and precision are both lower.

In figure A4, we examine the fairness metrics of the model across different samples of the test data. Just as we observed some evidence of disparate impact for samples in areas with above average non-White population percentages, we see that the fairness metrics fluctuate the most for this subgroup. The values for false negative rate disparity–one of the fairness metrics we prioritize–range from one to five. This means that in some test set samples, the false negative rate is five times larger for samples in census tracts with above average non-White population percentages. There is also more variation for the percent non-White and percent Hispanic or Latino subgroups when considering the predicted prevalence disparity. However, because we are less concerned about selection rates in the context of this ADS, the variation in predicted prevalence disparity is not a concern.

Summary

Data

The data are appropriate for this ADS, given that the central problem involves visible changes in environmental conditions. It is costly and time-consuming to manually sample all bodies of water in the U.S. and relying on manual sampling will likely lead to delays in starting mitigation and alerting the public. Computer vision techniques are already established methods for identifying cyanobacteria blooms, particularly in large bodies of water like oceans (Ye et al., 2021; Baek et al., 2021). However, there has been less attention paid to the challenge of identifying cyanobacteria blooms in small inland lakes and reservoirs. There are also seasonal and temporal patterns, so leveraging time-series data can be especially helpful in predicting new algal bloom cases. While detailed data on the water quality and surrounding environment would certainly improve the performance of the model, we do find that the winning model improves accuracy and precision while maintaining high recall relative to a baseline classifier.

Robustness, Accuracy, and Fairness

The ADS performs well. Accuracy is $\approx 81\%$, which is an improvement over a simple baseline classifier. In the test data, the base rate of high-severity samples is only $\approx 69\%$. The ADS also performs well on the performance metrics we care about, given the assistive nature of the intervention and the adverse consequences for failing to identify a positive case. We identified only one fairness concern with respect to non-White subgroups. Otherwise, the ADS does not appear to be systematically biased against sensitive subgroups like race, income, and ethnicity. FNR disparities are low for both low-income and Hispanic/Latino subgroups. Though not ideal, these disparity values suggest the model may perform better for these groups. Both performance metrics and fairness metrics remain moderately stable across different samples of unseen test data with the highest variation for the non-White subgroup as well.

State and local governments, water management agencies, and other environmental organizations will all benefit from the deployment of this ADS. The high accuracy and precision of the model, especially in comparison to the baseline classifier, means that inspections of inland lakes and reservoirs will occur more efficiently and with fewer costs. For nearby residents and those most affected by the outputs of the ADS, the high recall and low FNR–measures that are constant across various test set samples–should inspire confidence. These stakeholders are at high-risk if the ADS fails to detect an algal bloom and no inspection or warning is issued. The high recall may come at the expense of precision. Those who manage these bodies of water may advocate for higher precision regardless of the implications on recall to save time and resources, but fairness and equity considerations mean that we should continue to prioritize high recall and low FNRs in the deployment of the model.

Deployment

Overall, the ADS exhibited high levels of performance and fairness across various subpopulations. Although there were some differences in FNR for non-White subgroups, the magnitude of these differences was minimal. While it is possible for the ADS to overlook certain areas that require water sampling, having a human in

the loop could help address this issue by utilizing their expertise or alternative sources of information. This, in turn, can help reduce the disparity in FNR. Hence, we are comfortable in deploying this ADS in the public sector with some human oversight.

Recommendations

In terms of data collection, one area of improvement could be to collect more samples from the Northeast region, as it had the smallest proportion of training observations (see table 1) and the highest RMSE (see table A3). This could help to ensure that the model is trained on a representative sample and could therefore make more accurate and fair predictions. In terms of data processing, one area of improvement could be in the creation of the ad-hoc cluster variable. While this step is beneficial for accounting for spatial variations, there might be more systematic and statistically rigorous ways to account for such variations. Methods such as spatial autocorrelation analysis or spatial regression could be explored. In terms of analysis methodology, incorporating more domain knowledge into the model could be beneficial. This could be in the form of engineered features based on domain expertise.

Overall, our findings suggest the ADS is accurate, fair, and robust. However, we recommend that researchers involved in the project continue to audit the ADS and investigate the source of the FNR disparities. Perhaps there are issues with the underlying data or technical bias was introduced in the model pipeline. Additionally, future work should evaluate performance across disaggregated racial groups to assess whether one racial group is driving the disparity. Researchers should consider intersectional disparities as well. We briefly examined the relationship between poverty and race, for example, and did not observe any disparities. We recommend that a more formal, comprehensive investigation assess the performance across all combinations of racial, ethnic, and socioeconomic characteristics.

Relevant Links

- Audit GitHub Repository
- Original Second Place Repository
- Competition Website

Appendix

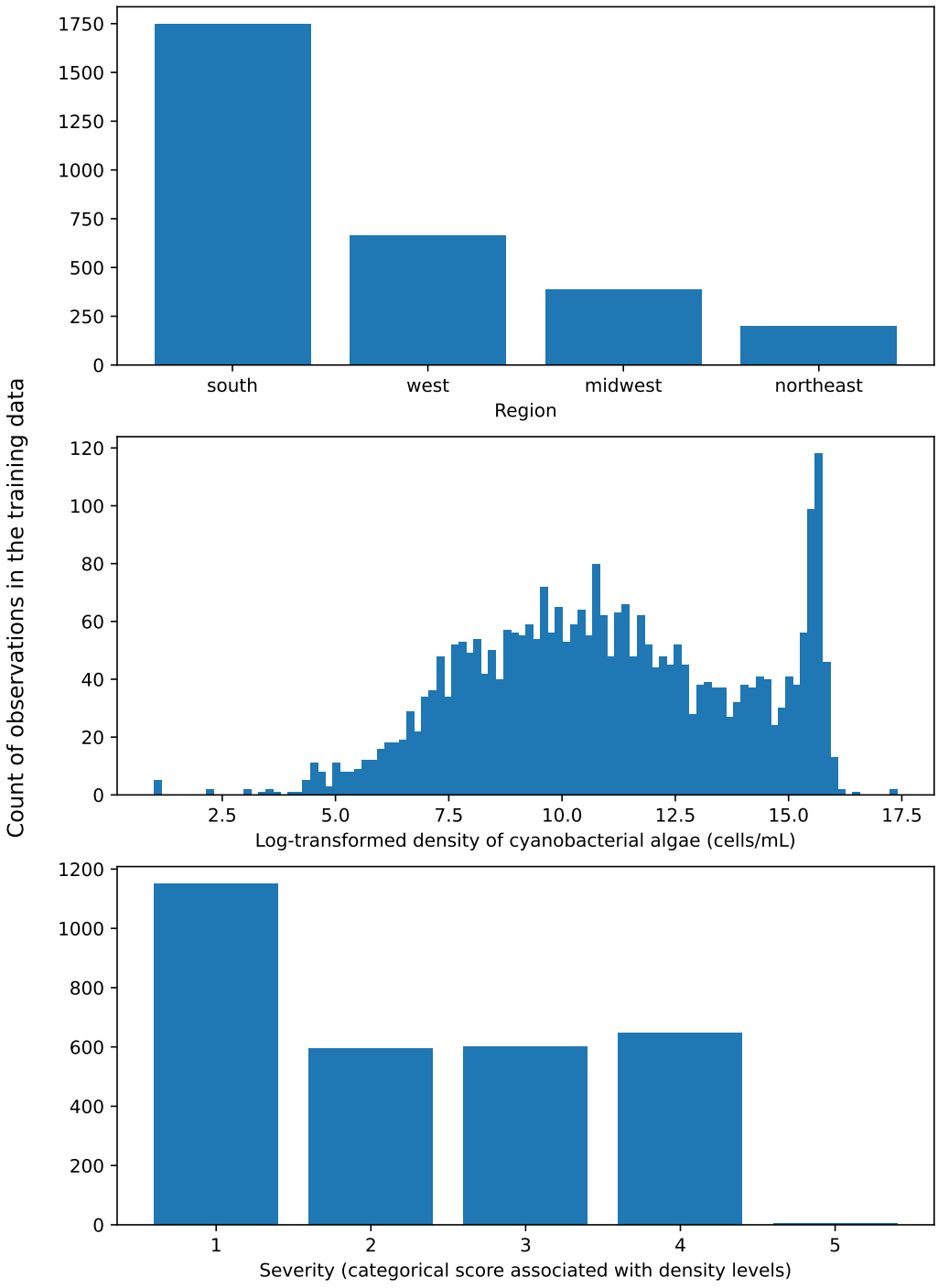

Figure A1: Histograms of input features in the training data

Severity (Binarized)	Severity	Percent of Total
Low	1	44
High	4	21
	2	19
	3	16
	5	< 1

Table A1: Frequency of binarized severity based on estimated risk level in the **original training** dataset

Subgroup	Percent of records in the test data (n=1,167)
Poverty rate	14
Low income	45
Percent non-White	14
Percent Hispanic or Latino	26

Table A2: Descriptive statistics of records in census tracts above the statewide average for each subgroup

	Root mean-squared error
Midwest	0.77
Northeast	1.05
South	0.83
West	0.26
Average overall	0.73

Table A3: Performance using Region-Averaged RMSE

Table A3 table replicates the performance metric used by DrivenData to evaluate the models: the average RMSE across four regions in the U.S. using the categorical outcome.

	Metrics	Poverty rate	Low income	Percent non-White	Percent Hispanic or Latino
Difference	FNR	0.01	0.06	0.05	0.07
	FPR	0.04	0.12	0.04	0.01
	Demographic parity	0.07	0.13	0.05	0.14
Ratio	FNR	0.89	0.38	0.56	0.26
	FPR	0.91	0.79	0.92	0.98
	Demographic parity	0.92	0.85	0.93	0.84
	Equalized odds	0.91	0.79	0.92	0.93

Table A4: Evaluation of fairness metrics using FairLearn

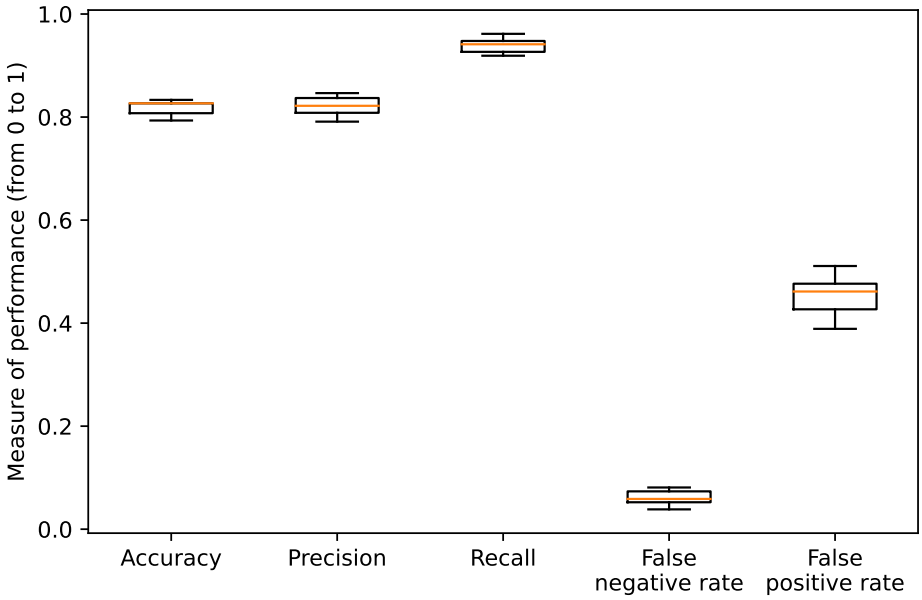

Figure A2: Performance over 10 random splits overall to assess robustness

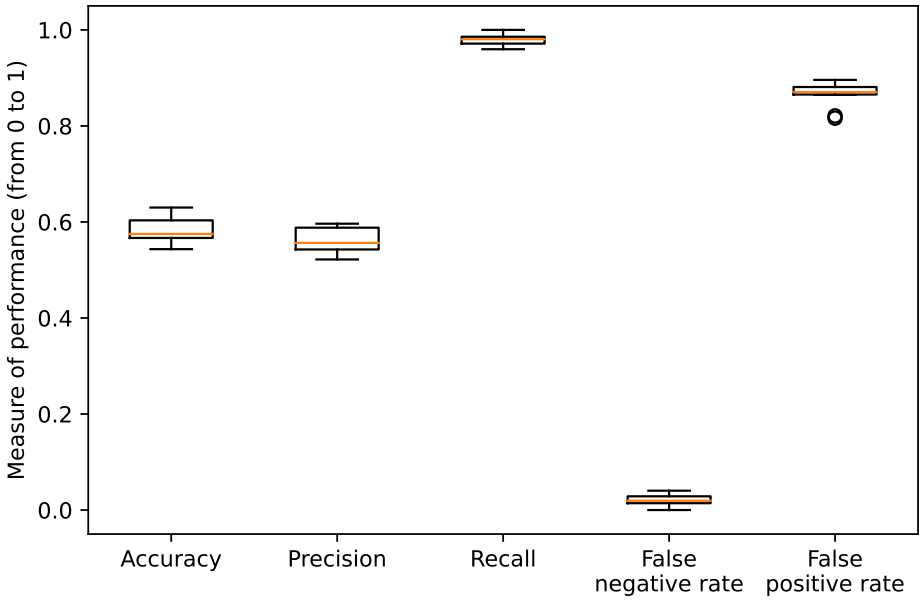

Figure A3: Performance over 10 random splits overall to assess robustness on pre-2016 data

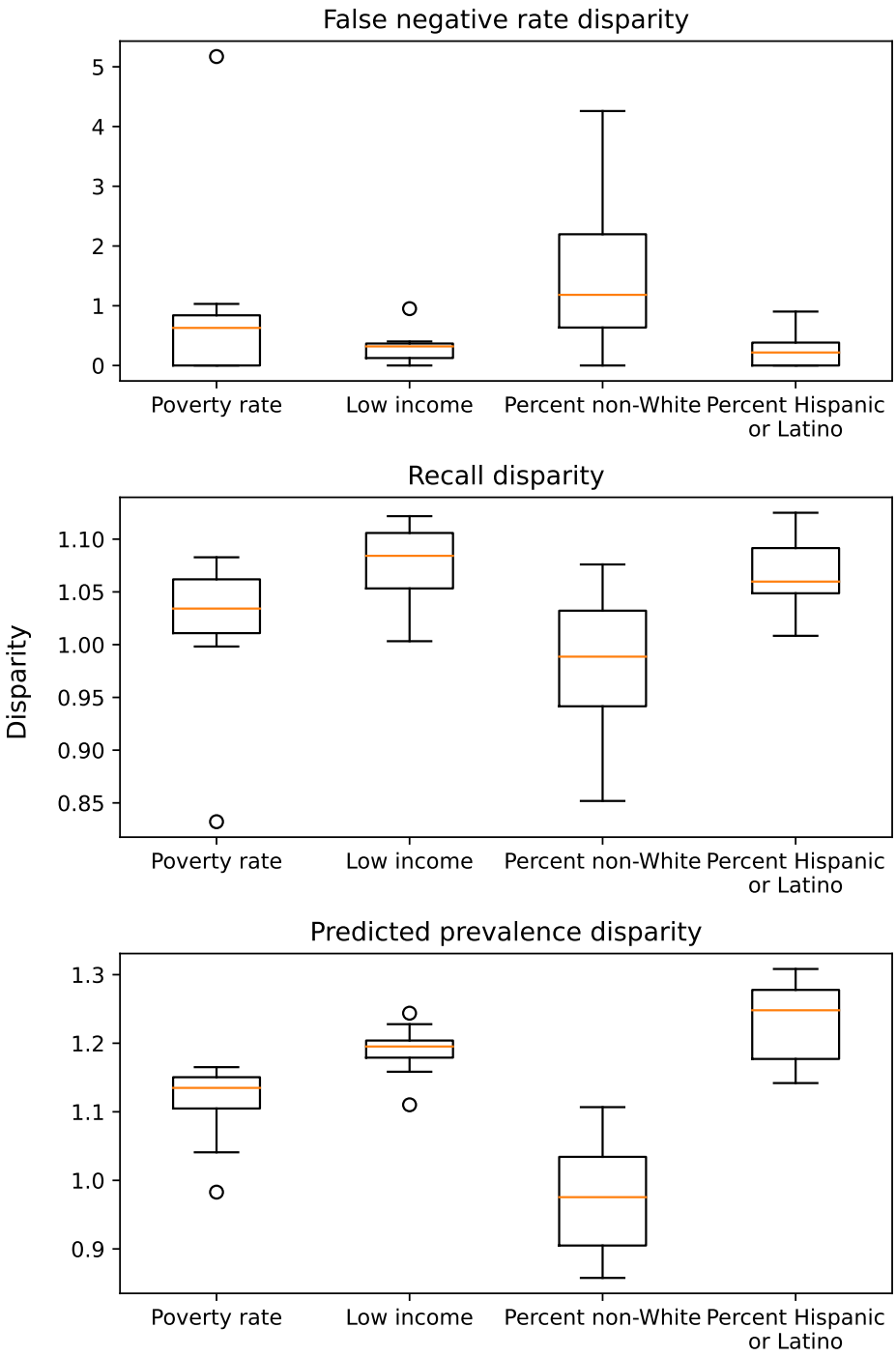

Figure A4: Performance over 10 random splits across subgroups to assess robustness

References

S.-S. Baek, J. Pyo, Y. S. Kwon, S.-J. Chun, S. H. Baek, C.-Y. Ahn, H.-M. Oh, Y. O. Kim, and K. H. Cho. Deep learning for simulating harmful algal blooms using ocean numerical model. *Frontiers in Marine Science*, 8:729954, 2021.

R. Ghani. Aequitas. http://www.datasciencepublicpolicy.org/our-work/tools-guides/aequitas/, accessed April 17, 2023.

T. Gorham, E. D. Root, Y. Jia, C. Shum, and J. Lee. Relationship between cyanobacterial bloom impacted drinking water sources and hepatocellular carcinoma incidence rates. *Harmful algae*, 95:101801, 2020.

S. J. Granger, J. A. Qunicke, P. Harris, A. L. Collins, and M. S. Blackwell. Comparison of high frequency, in-situ water quality analysers and sensors with conventional water sample collection and laboratory analyses: Phosphorus and nitrogen species. *Hydrology and Earth System Sciences Discussions*, 22(1):1–33, 2018. doi: 10.5194/hess-2017-684.

P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018.

L. A. Schaider, L. Swetschinski, C. Campbell, and R. A. Rudel. Environmental justice and drinking water quality: are there socioeconomic disparities in nitrate levels in us drinking water? *Environmental Health*, 18:1–15, 2019.

WHO. Guidelines for safe recreational water environments. *Coastal and Fresh Waters*, 1:1–219, 2003.

M. Wines and M. Cramer. 2020 census undercounted hispanic, black and native american residents. *The New York Times*. URL https://www.nytimes.com/2022/03/10/us/census-undercounted-population.html.

X. Ye, Z. Lai, and D. Li. Prediction of the cyanobacteria coverage in time-series images based on convolutional neural network. In *Proceedings of the 4th International Conference on Control and Computer Vision*, pages 153–158, 2021.