Leveraging Readily Available Web Data to Predict Surface Water Bacteria Levels Using Domain Adaptation

Anonymous Author
Institute
City, Country
email@affiliation.org

ABSTRACT

Surface water contaminated by fecal bacteria can cause diarrheal illness, threatening human health (especially among children). In recent years, supervised machine learning (ML) has been used to predict fecal indicator bacteria (FIB) levels. Training ML models to predict FIB levels is challenging and, in some cases, impractical ; Although environmental data are readily available on the web, FIB levels are costly and time-consuming to measure, making it hard to create labeled datasets specially in regions with limited resources. In this paper, we introduce a water quality dataset collected from beaches in Chicago, USA, and present strong ML baselines to predict historical FIB levels on this dataset. Specifically, we use Random Forest (RF), extreme gradient boosting (XGBoost), and attentive interpretable tabular learning (TabNet) models to establish strong baselines in supervised learning and transfer learning settings on our Chicago beaches dataset. We further show that applying supervised and unsupervised domain adaptation methods enhances transfer learning performance in general. Specifically, our results show that the supervised methods, especially RF, are very promising and that domain adaptation could be successfully employed to model FIB levels in settings where FIB levels are rarely measured.

CCS CONCEPTS

• **Do Not Use This Code → Generate the Correct Terms for Your Paper https://dl.acm.org/ccs**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

KEYWORDS

Water Quality Prediction, Fecal Indicator Bacteria, Transfer Learning, Domain Adaptation

ACM Reference Format:
Anonymous Author. 2018. Leveraging Readily Available Web Data to Predict Surface Water Bacteria Levels Using Domain Adaptation. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation emai (Conference acronym 'XX)*. ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

Conference acronym 'XX, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

High levels of bacteria in surface waters bring major challenges in global public health due to causing diarrheal disease, particularly among children as a more vulnerable population. For example, in 2019, unsafe drinking water accounted for an estimated 419,000 deaths of children under the age of 5 years in many countries (especially low- and middle-income countries).

Importantly, levels of the fecal indicator bacteria (FIB) known as enterococci (ENT), measured using the rapid quantitative polymerase chain reaction (qPCR) method, predict the risk of contracting diarrheal illness among swimmers at beaches [33]. Thus, the United Nations Sustainable Development Goals (SDGs) point to the need to identify water quality hazards and protect water sources (SDG 6 Targets 1 and 3), to communicate those hazards to communities (SDG 6 Target 6.b), to protect public health (SDG 3) and to design sustainable urban systems to protect water sources (SDG 11). However, measuring FIB levels in many areas of the world (e.g., in rural areas of high-income countries or low- and middle-income countries) is challenging due to fiscal, technical, and logistical challenges. Even in locations where testing is done, there is generally a 24-hour interval between the time of water sampling and the time bacterial culture results are available. By the time results are available, water quality can change substantially, limiting the value of the FIB measurements for alerting the public about hazardous conditions [9].

A variety of machine learning (ML) approaches have been developed to make use of weather and other environmental variables that are available in near-real time to predict FIB levels in high-income countries [11, 12, 18, 26, 27]. Many of the FIB predictive variables - such as precipitation, temperature, solar irradiation, and wave height - are available from U.S. and international meteorological/atmospheric agencies. However, ML models of surface water quality that have been developed so far using data from one location may not be directly transferable to other locations for which labeled data is scarce. This lack of direct transferability could be due to several reasons, including "domain shift" or "distributional shift", which refers to the fact that the distribution of FIB and/or predictive variables may be different between beaches used to train the models and beaches used for model prediction. To account for distributional differences, domain adaptation (DA) approaches [21] for adapting a model from a data-rich "source" beach (such as a beach with surface waters abundant in historical FIB data) to a data-poor "target" beach (for which FIB data is limited, if at all available) represent an attractive solution, and we aim to explore it in this work.

Towards this goal, our main contributions are as follows: 1) We first introduce a surface water quality dataset consisting of approximately 10,000 FIB samples collected at several Chicago beaches.

To our knowledge, this is the largest dataset for developing ML models for surface water quality prediction based on environmental variables in supervised, transfer learning and domain adaptation settings; 2) Using the newly assembled dataset, we establish strong supervised baselines using ensemble learning and neural network models that are trained and evaluated on data from the same beach (or group of beaches); 3) We also develop transfer learning baselines, i.e., supervised learning models trained on a "source" beach (or group of beaches) and tested on "target" beaches not included in the training set; and 4) Finally, we established domain adaptation baselines, where models trained on data from "source" beaches are adapted to "target" beaches by using a small amount of FIB target data and/or unlabeled target environmental data (i.e., data for which FIB levels are not available).

We have collected and measured FIB levels, retrieved environmental data from the web, and developed transfer learning models enhanced with domain adaptation techniques. The study is aimed to eliminate the need for time-consuming and costly FIB measurements, a particularly significant advantage in regions with limited resources. Our study extends to the development of a user interface that can utilize the available environmental data on the web and employ our trained models to predict water quality for any given location.

Our comprehensive analysis is done for a regression setting (to predict the FIB levels). Our dataset, together with the baselines developed, represents an important first step towards the development of models for predicting surface water bacteria levels, trained and evaluated in domain adaptation settings when labeled data are limited or not available in the target domain. This has a high social impact on improving public health.

2 RELATED WORK

Chicago, USA, is home to some of the most "data rich" beaches globally, with decades of intensive FIB monitoring of 15+ beaches per day, at least five days/week, throughout the 100-day summer "beach season." Linear regression methods were initially used to predict FIB levels [20, 28, 34]. Since then, random forest methods have been used to identify predictors to be used in statistical models of FIB at Chicago beaches [8, 14, 19]. More recently, Lucius et al. [2019] utilized machine learning to predict levels of *E. coli* bacteria (a type of FIB) at Chicago beaches based on qPCR testing of Enterococcus (ENT) together with water and atmospheric variables. Twenty Chicago beaches were first grouped into 5 clusters. For each cluster, one "feature" location was selected based on the maximum number of historical culture-based exceedances and used as a proxy location for other locations in the same cluster. An RF model was trained for the "feature" location in each cluster using data collected between 2006 and 2015, and validated on data from 2016 collected at the other locations in the cluster. In the second phase of the project, models trained on data from years 2015-2016 were tested live on newly collected data in 2017. Experimental results showed that the proposed hybrid model that used ENT data from the "feature" beach to predict *E. coli* at other similar beaches improved sensitivity from 3.4% to 11.2% compared with a prior-day nowcast model. While this work considered a transfer learning scenario between a feature beach and other correlated beaches in its cluster, it did make use of ENT (in addition to environmental variables) to predict *E. coli*. As opposed to that, we explore transfer learning and domain adaptation when *only* environmental variables are used for predicting ENT (without including any other types of FIB as predictors).

Guo and Lee [2021] classified FIB exceedances (0/1) determined by thresholding FIB levels measured at beaches in Hong Kong using the EasyEnsemble (EE) algorithm [17], an ensemble of AdaBoost learners trained on different balanced bootstrap samples.[1] Likewise, FIB levels were modeled using ML approaches for sites in Croatia by Grbčić et al. [2022] and southern California by Searcy and Boehm [2021]. However, none of these works studied domain adaptation approaches, although the work by Grbčić et al. [2022] explored the use of transfer learning between a source beach and a target beach.

Publicly available datasets for predicting FIB concentrations in water samples include Guo and Lee [2021] and Searcy and Boehm [2021]. Guo and Lee [2021] published a 30-year *E. coli* dataset relating to three locations in Hong Kong, China, containing 3939 Enterococcus samples and including up to 8 environmental features such as past rainfall and previous day's solar radiation and used it to study class-imbalance methods for predicting high levels of *E. coli*. Searcy and Boehm [2021] published a 20-year dataset relating to three locations in the U.S. state of California, containing 4805 culture-based Enterococcus and *E. coli* samples from both high-frequency and routine monitoring, while including up to 33 environmental features. They used this dataset to study high-frequency sampling toward assessing water quality at sites with little or no historical routine monitoring data. The City of Chicago has published ENT and EC data online going back to 2006, although the city does not publish accompanying environmental features related to the data [6]. The Water Quality Portal, developed by the U.S. Environmental Protection Agency, U.S. Geological Survey, and National Water Quality Monitoring Council also provide FIB data from numerous sites located throughout the United States, although environmental data are typically not included [24]. Bourel et al. [2021] utilized a simulated dataset which can later be regenerated using their published codebase. In general, however, datasets used in previous research have often gone unpublished or made publicly available [3, 11, 13–16, 18, 22, 23, 25, 30, 31, 36, 37].

In our paper, we construct a dataset based on Chicago beaches for predicting FIB levels (specifically, ENT) based on environmental conditions. To our knowledge, it is the largest dataset for this task and we make it publicly available. We also established strong supervised, transfer learning, and domain adaptation baselines using our dataset.

3 CHICAGO SURFACE WATER QUALITY DATASET

As a first significant contribution of this work, we assemble a large surface water bacteria level dataset to further research in this area. This dataset contains daily measurements of the concentration of Enterococcus (ENT) from 19 beaches in Chicago from 2017 to 2022, and will be made publicly available upon publication of this work.

common attributes

[1] https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.EasyEnsembleClassifier.html

Features	Descriptions	Physical Characteristics	Descriptions
awind	alongshore component of wind speed	shape_Length	length of the beach in GIS
owind	offshore component of wind speed	shape_Area	area of the beach in GIS
WVHT	significant wave height in meters	avg_Beach_slopethe	average slope across the beach
Wtemp_B	sea surface temperature	beach_len	length of the beach in miles
atemp	air temperature	start_latitude	the latitude of the start of the beach
dtemp	dew point temperature	start_longitude	the longitude of the start of the beach
PrecipSum6	Amount of rain in the past 6 hours	end_latitude	the latitude of the end of the beach
Precip24	Amount of rain in the past 24 hours	end_longitude	the longitude of the end of the beach
lograin3T	log10 transform of past 3 days rainfall	dog_park	If there are dog parks near the beach
lograin7T	log10 transform of past 7 days rainfall	nature_park	If there are nature parks near the beach
wet3	if it rained more than 0.1' in the past 3 days	sports_field	If there are sport fields near the beach
wet7	if it rained more than 0.1' in the past 7 days	water_feature	if there is a water feature near the beach
dtide_1	change in tide in the last hour	groin_name	name of the groin
dtide_2	change in tide in the last 2 hour	length	length of the groin in meters
tide_gtm	if the value of tide is greater than mean tide	width	width of the groin in meters
tide	the water level in feet above or below the mean lower low water	height	height of the groin in meters
DPD	dominant wave period, seconds, is the period with the maximum wave energy	geometry	shape of the groin
comment	Visual characteristics such as presence of sand, mud, residue, particles, wood chips, dirt, plants, etc.	beach	the beach at which the groin is located
turbidity	an expression of the optical property that causes light to be scattered and absorbed rather than transmitted in straight lines through the sample		
rad (DNI)	The amount of solar radiation received per unit area by a perpendicular surface		

Table 1: Description of features collected and physical characteristics of beaches.

3.1 Data Collection

To create the dataset, we aimed to extract the same features as in the existing California high-frequency water quality dataset [26]; however, some attributes from the California dataset were not available for Chicago beaches, while other attributes were only available for the Chicago beaches (but not in the California dataset).

The attributes in our Chicago dataset (shown in Table 1) include various environmental characteristics such as wind, wave, precipitation, tide, solar radiation, air and water temperature, and turbidity. Also, a set of physical beach characteristics (such as length, width, adjacent parks, and dog beaches) are gathered for each of the beaches included in the dataset.

Samples were collected at 6 AM every morning to measure the ENT levels during the beach season (for approximately 100 days from late May to early September) and analyzed for ENT levels using the qPCR method [32]. Results were available by 1:00 PM and used for water quality advisories at beaches and on the Chicago Park District's websites and social media outlets.

FIB levels and turbidity were generated by the water microbiology laboratory on behalf of the Chicago Park District. Hourly precipitation, wind, and temperature data were gathered from the Midwest Regional Climate Center for the Midway Airport weather station, approximately 15 km from the shore of Lake Michigan. Wind direction and wind speed were converted to the speed of wind perpendicular to the beach angle. Wave and tide data were obtained from the National Oceanographic and Atmospheric Administration (NOAA) National Data Buoy Center for buoy 45198, Ohio street, and Calumet Harbor buoy. Solar radiation for each beach group (beaches are grouped based on their location) was obtained from the National Solar Radiation Database using the coordinates of a beach near the group's center.

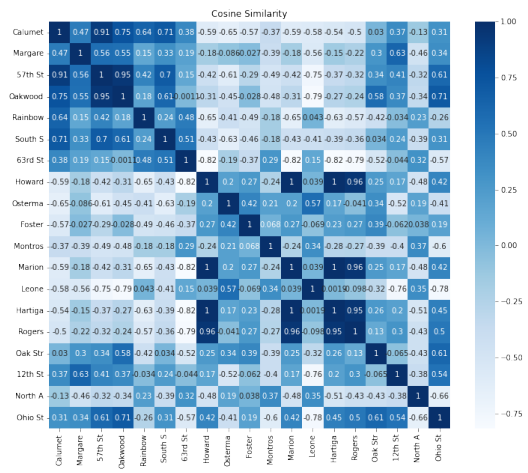
Physical characteristics of individual beaches (such as length, width, adjacent parks, and dog beaches) were obtained from Chicago Park District GIS files where available by analysing these characteristics using ArcGIS. The beach slope was calculated from NOAA Digital Elevation models in ArcGIS. Additionally, data on the groins (shore perpendicular structures for beach maintenance) and jetties of beaches were obtained from [5]. A short description of each of the features collected and physical characteristics of the beaches are available in Table 1.

3.2 Statistical Analysis of FIB Data from Beaches

We divided the 19 Chicago beaches into three groups based on their location: Southern (SB), Central (CB), and Northern beaches (NB). For the transfer learning task (training on a group of beaches and testing on another group), the beaches in one group are expected to be similar to one another and different from the beaches in other groups. To test whether or not the current splits satisfy this expectation, we made use of two metrics to calculate the similarity and divergence between the beaches:

- Cosine Similarity: We calculated pairwise cosine similarity between the physical characteristics of every two beaches. The results are shown in Figure 1 (Left).

$$Cosine\ Similarity = \frac{\vec{a}.\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$$



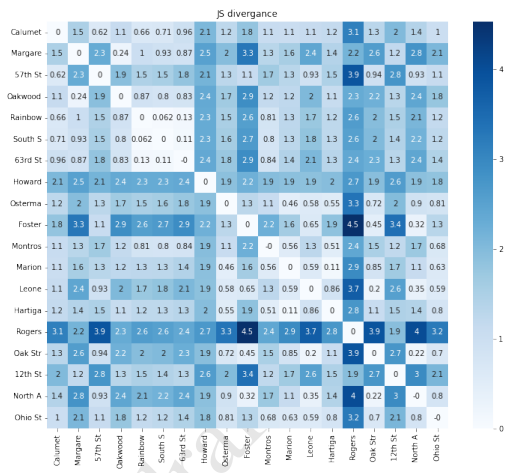


Figure 1: Beach similarity/divergence. Southern beaches are the first seven beaches from Calmut beach to 63rd street beach, Howard beach to Rogers beach are Northern beaches, and the rest are central beaches. (Left) Cosine similarity based on the physical characteristics of the beaches; (Right) JS divergence based on the data collected at each beach.

Features	ENT	atemp	dtemp	DPD	WVHT	Wtemp_B	awind	owind	turbidity	comment	Rad
Mean	439.640	21.193	17.647	3.979	0.260	20.844	-0.129	-0.725	3.483	0.132	351.164
STD	3.350	3.328	1.309	0.220	2.882	3.020	3.151	311.460	4.376	0.297	252.721
Missing%	0.0	74.795	74.795	51.897	46.614	29.049	24.496	24.496	20.800	20.800	37.7471
Features	**tide**	**tide_gtm**	**dtide_1**	**dtide_2**	**PrecipSum6**	**Precip24**	**lograin3T**	**wet3**	**lograin7T**	**wet7**	
Mean	177.061	0.531	0.001	0.001	0.027	0.116	-1.708	0.503	-0.553	0.833	
STD	0.190	0.499	0.078	0.083	0.135	0.291	1.674	0.498	1.143	0.371	
Missing%	5.792	5.792	5.792	5.792	2.237	4.643	10.944	10.944	21.440	21.440	

Table 2: Statistical properties of the features in our Chicago water quality dataset.

where $\overrightarrow{a}$ and $\overrightarrow{b}$ represent samples in the dataset.

- Jensen-Shannon (JS) divergence: We calculated the pairwise JS divergence between the data collected at beaches. The results are shown in Figure 1 (Right).

$$D_{JS}(p\|q) = \frac{1}{2}D_{KL}(p\|\frac{p+q}{2}) + \frac{1}{2}D_{KL}(q\|\frac{p+q}{2})$$

$$\text{where } D_{KL}(p\|q) = \sum_{x\in X} P(x)log(\frac{P(x)}{Q(x)})$$

The heat maps in Figures 1 (Left) and (Right) illustrate the similarity between physical characteristics of beaches and divergence between the data collected at beaches, respectively. As expected, these similarity/divergence metrics suggest that the current groups may benefit from transfer learning and domain adaptation. Also, the heat maps suggest that the cosine similarity based on physical characteristics is aligned with the JS divergence obtained using the actual collected data. They also show high similarity (low distance) between beaches in a group, especially Southern and Northern beaches, and relatively high divergence between beaches that are not in the same group.

3.3 Preprocessing

As part of our data preprocessing, all numerical features were scaled between zero and one using the formula below:

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

After assembling the dataset, some features were unavailable for some dates or beaches, especially during 2020 when FIB monitoring was not performed as beaches were closed due to the COVID-19 pandemic. We removed the features with more than 50 percent missing values to address the missing data issue. The remaining missing values were replaced by the mean of the attributes. Tables 2 and 3 show the statistical properties of the features and the number of available data in each location, respectively.

Following [26], the values for the regression target, ENT, were transformed using a $\log_{10}$ transformation.

3.4 Benchmark Subsets

The train/validation/test split is based on the timeline, i.e., models were trained and tuned on older samples and tested on more recent samples. Specifically, training and validation were performed on

SB		NB		CB	
Beaches	**#**	**Beaches**	**#**	**Beaches**	**#**
Calumet	591	Howard	301	Oak Str	512
Margare	587	Osterman	522	12th St	588
57th St	520	Foster	518	North A	580
Oakwood	501	Montros	587	Ohio St	585
Rainbow	593	Marion	522		
South S	593	Leone	520		
63rd St	584	Hartiga	525		
		Rogers	285		
Train	2518	**Train**	2452	**Train**	1432
Validation	629	**Train**	613	**Train**	358
Test	822	**Test**	715	**Test**	475

Table 3: The available data samples per beach in the dataset. The last two rows show how many samples are available in the train and test sets.

data spanning 2017-2021 (excluding 2020), while year 2022 data were used for testing. Table 3 displays the number of train and test samples for each group of beaches.

The train/validation/test split will be made available to ensure reproducibility and to enable further improvements on the task of predicting surface water quality in transfer learning and domain adaptation settings.

4 METHODOLOGY

4.1 Supervised Machine Learning Models

We consider the following machine learning models for our the regression tasks. Following [26, 27] ensemble trees and gradient boosting methods show promising results in water quality regression tasks, so we chose Random Forest (RF) and XGBoost [4]. RF is an ensemble learning model that combines the output of multiple decision tree models, and XGBoost is a gradient boosting algorithm. Furthermore, we used TabNet to have baseline models in various categories of models; TabNet model[1] is a strong neural natwork model for tabular data that uses an attention mechanism to weigh each feature's importance selectively.

4.2 Supervised and Unsupervised Domain Adaptation

In supervised domain adaptation (SDA), labeled source data (X_S, y_S) is used together with a small amount of labeled target data (X_T, y_T) to train a model h for predicting future target data. We experiment with two supervised DA approaches, balanced weighting [35] and feature augmentation [7].

In unsupervised domain adaptation (UDA), labeled source data (X_S, y_S) and unlabeled target data X_T are used to train a model h for predicting future target data. We experiment with two unsupervised DA approaches, correlation alignment [29] and subspace alignment [10].

Balanced Weighting. In the BWT approach [35], a model h is trained to minimize a modified loss which accounts for both agreement between predicted and ground truth values on target data, as well as agreement on source data. However, the loss on target data $\mathcal{L}(h(X_T), y_T)$ and the loss on source data $\mathcal{L}(h(X_S), y_S)$ have different weights to reflect the importance of the target relative to the source. Formally, the model h is obtained by minimizing the following modified loss:

$$\min_h (1-\gamma)\mathcal{L}(h(X_S), y_S) + \gamma\mathcal{L}(h(X_T), y_T)$$

where γ is a tunable hyper-parameter that defines the extent to which target training data should be prioritized. Both labeled source data (assumed to be large) and labeled target data (assumed to be limited) are thus utilized when training a BWT model.

Feature Augmentation. In the FA approach [7], the source and target training data are augmented by creating three versions of the feature set: a version that is shared between source and target (representing features that are predictive for both source and target data), a version specific to source (which has null values in the target) and a version specific to target (which has null values in the source). Specifically, for source the features $\mathbf{x}$ are transformed into $\tilde{X}_S = (\mathbf{x}, \mathbf{x}, \vec{0})$, while for target the features are transformed into $\tilde{X}_T = (\mathbf{x}, \vec{0}, \mathbf{x})$. A model h is trained on the combined feature-augmented source and target data by minimizing the standard loss:

$$\min_h \mathcal{L}(h(\tilde{X}_S \cup \tilde{X}_T), (y_S \cup y_T))$$

Similar to BWT, in FA both labeled source and target data are utilized to train an FA model.

Correlation Alignment. In CORAL [29], the goal is to minimize the domain variance between the source and target data by aligning the source and target distributions through the means of second order statistics estimated solely from unlabeled data. Specifically, covariance statistics C_S and C_T are estimated for source and target data, respectively. The source covariance matrix C_S is used to perform source "whitening", i.e., to transform the source data such that its covariance matrix becomes the identity. Subsequently, the target covariance matrix C_T is used to "re-color" the source data, so that the source distribution becomes similar to the target distribution. Formally, a linear transformation A of the source data can be obtained as a solution to the following minimization problem:

$$\min_A \left\| A^T C_S A - C_T \right\|_F^2$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm of a matrix and is used as a distance metric. The transformed source data can then be used to train models for the target domain, given that the two domains have now similar distributions.

Subspace Alignment. In SA [10], the goal is to reduce the domain variance by aligning the source and target subspaces represented by their respective eigenvectors induced using principal component analysis (PCA). This is achieved by identifying a transformation matrix M that transforms the source subspace base E_S into the target subspace base E_T (where E_S and E_T are given by the eigenvectors corresponding to the highest k eigenvalues in source and target, respectively). The transformation matrix M used in the subspace alignment is obtained as a solution to the following minimization problem:

$$\min_M \|E_S M - E_T\|_F^2$$

Model	Train R2	Test R2	Train RMSE	Test RMSE	Train rRMSE	Test rRMSE
RF	0.662	0.327	0.408	0.557	0.241	0.292
MLP	0.235	0.184	0.615	0.613	0.332	0.335
XGBoost	0.677	0.363	0.399	0.542	0.221	0.318
TabNet	0.251	0.190	0.608	0.611	0.309	0.309

Table 4: Regression results for supervised models. The performance is recorded using relative root mean squared error (rRMSE), root mean squared error (RMSE) and R2 score. Random Forest (RF), Multi Layer Perceptron (MLP), XGBoost and TabNet are employed to calculate the values for logENT. Our main metric for regression is rRMSE (the lower the values the better). The performance in blue corresponds to models that perform better in this task and are used in the other experiments while the weaker model is shown in red.

RF	XGBoost	TabNet
Most Important		
turbidity	beach	awind
tide	tide	turbidity
awind	turbidity	Wtemp_B
owind	dayofyear	WVHT
dayofyear	beach_area	beach_area
beach	awind	dtide_2
WVHT	dtide_2	PrecipSum6
dtide_2	owind	owind
comment	dtide_1	wet7
Wtemp_B	WVHT	Precip24
dtide_1	lograin7T	beach
lograin7T	comment	lograin7T
beach_area	Wtemp_B	tide
lograin3T	Precip24	wet3
Precip24	lograin3T	tide_gtm
PrecipSum6	wet3	comment
wet7	PrecipSum6	dtide_1
tide_gtm	tide_gtm	lograin3T
wet3	wet7	dayofyear
Least Important		

Table 5: Feature Importance

where $\|\cdot\|_F^2$, as before, denotes the Frobenius norm of a matrix. As in the case of CORAL, the transformed source data is subsequently used to train a model for the target data.

5 EXPERIMENTS AND RESULTS

5.1 Metrics

To evaluate the models, we chose relative root mean squared error (rRMSE) as the main evaluation metric, defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_{pred} - y)^2}$$

$$rRMSE = \frac{RMSE}{\bar{y}}$$

Using rRMSE we are able to compare the results across different datasets in future studies.

5.2 Experiment 1: Supervised Baseline Models

The results of experiment one for our regression task are shown in Tables 4. In this experiment, we trained our supervised learning models on the whole dataset, using attributes with no more than 50 percent missing data. All models were trained using k-fold cross-validation and grid search for hyper-parameter tuning.

In this experiment we explored RF, MLP, XGBoost, and TabNet models for predicting logENT values in a regresion setting. As can be seen in Table 4, in our regression task, the RF, XGBoost and TabNet models showed more promising results.

Each of the RF, XGBoost, and TabNet models produce a list of the most important features; the features are sorted based on their importance for each model in Table 5. Turbidity, awind, and tide seem to have more information than the other features for ENT prediction. The feature importance results will enhance our ability to prioritize the important features in the data collection process for future studies In RF and XGBoost, features importance is computed by averaging the amount that each attribute split point improves the loss function among the trees. In TabNet, the features' importance is computed by analysing the weights and network's attention values.

5.3 Experiment 2: Transfer Learning

In this experiment, we split the dataset based on locations; southern, central, and northern beaches. Each of the models used (RF, XGBoost, and TabNet, based on the supervised baselines in Experiment 1) is trained on each of the three locations and tested on test data of all of the locations. Also, we used the train/test splits mentioned in Table 3.

Our regression results for the second experiment are shown in Table 6, in the TL (transfer learning) rows. As expected, the results (rRMSE) of transfer learning are worse than the results of the supervised experiment (e.g., rRMSE for the RF model in baseline experiments in Table 4 is 0.292). Note that as we discussed, the results in Tables 4 represent lower bounds for the transfer learning and domain adaptation results (note that the lower the rRMSE the better).

In Table 4, as expected, most regressors perform relatively better when we test them on the same location where the model is trained. In the transfer learning task, the regressors trained on southern beaches perform better than regressors trained on the other groups of beaches. Interestingly, the models trained on SB data reached a higher rRMSE score on CB and NB test sets than those trained on

Trained on			SB			NB			CB		
Model	**Approaches**		**NB**	**CB**	**SB**	**NB**	**CB**	**SB**	**NB**	**CB**	**SB**
RF	TL		0.401	0.426	0.318	0.385	0.426	0.328	0.431	0.426	0.393
	DA	FA	0.385	0.415	0.312	0.386	0.415	0.313	0.388	0.415	0.313
		BWT	0.392	0.422	0.312	0.383	0.420	0.320	0.397	0.416	0.330
		CORAL	0.405	0.440	0.316	0.385	0.426	0.318	0.406	0.419	0.420
		SA	0.399	0.432	0.316	0.387	0.429	0.324	0.413	0.415	0.378
XGBoost	TL		0.405	0.431	0.320	0.402	0.442	0.352	0.419	0.423	0.377
	DA	FA	0.386	0.416	0.322	0.404	0.436	0.373	0.385	0.432	0.353
		BWT	0.398	0.426	0.342	0.411	0.444	0.373	0.426	0.429	0.382
		CORAL	0.398	0.431	0.319	0.396	0.437	0.461	0.411	0.416	0.346
		SA	0.421	0.428	0.329	0.398	0.434	0.326	0.432	0.425	0.496
TabNet	TL		0.411	0.439	0.327	0.401	0.441	0.324	0.399	0.415	0.338
	DA	FA	0.405	0.442	0.312	0.390	0.416	0.318	0.399	0.422	0.319
		BWT	0.419	0.498	0.338	0.408	0.530	0.357	0.402	0.426	0.341
		CORAL	0.417	0.440	0.327	0.401	0.435	0.324	0.405	0.416	0.334
		SA	0.438	0.434	0.360	0.385	0.431	0.321	0.396	0.416	0.357

Table 6: Regression results measured by rRMSE for the domain adaptation task. An improvement over the transfer learning model is marked with green, while a decrease in performance is signaled using red. The lower the performance the better.

CB and tested on NB and SB or trained on NB and tested on CB and SB. This can be due to less available data in CB and NB groups (see Table 3). Although there is more similarity between NB and CB data based on the data similarity and divergence metrics in Figure 1, the models trained on southern beaches perform better on CB and NB test data; for our regression task, the rRMSE for the models trained on SB and tested on CB is lower than those trained on NB and tested on CB. The SB models can transfer information better than CB and NB models, and the NB models are in second place.

5.4 Experiment 3: Supervised and Unsupervised Domain Adaptation

Finally, we analyze two supervised and two unsupervised domain adaptation algorithms. FA (Feature Augmentation) and BWT (Balanced Weighting) algorithms are supervised DA models, and CORAL (Correlation Alignment) and SA (Subspace Alignment) are used as unsupervised domain adaptation algorithms. The SDA algorithms take 10 percent of labeled target training data (selected at random), while the UDA models use all the target training data as unlabeled.

The hyper-parameters of the BWT and CORAL models were fine-tuned in the training process. Gamma hyper-parameter in BWT and Lambda hyper-parameter in CORAL correspond to the importance given to the target labeled data and the intensity of adaptation, respectively.

By analyzing the results of the Experiments 2 and 3 (Tables 6 — DA rows and TL), it can be seen that in the majority of the cases considered, the domain adaptation methods helped improve the results. The regression performance is still better for the southern beaches (i.e., when models trained on NB and CB data attempt to predict the water quality in SB locations).

Among the models used as base learners, the RF model benefits more from domain adaptation. Comparing the supervised and unsupervised domain adaptation methods in Table 6, we can see that the FA method has lower (better) rRMSE results as compared with the other DA methods. Overall, FA produces consistently improved results for all three base learners (RF, XGBoost, and TabNet). Less improvements can be observed using BWT, CORAL, and SA with XGBoost and TabNet models. The RF results were improved using all four DA methods, but the SDA methods have better rRMSE performance as compared to UDA.

6 DEPLOYED WEB APPLICATION

As an additional contribution, we developed a web application that utilizes the most accurate predictive model to estimate the FIB levels for a given location. Currently, the environmental variables should be given to the web application and the beach's location. Still, ultimately, the goal of this user interface is to retrieve the environmental data through the related data sources for each variable, apply the preprocessing steps, and feed them to the model to predict the concentration of FIB.

Moreover, our web application incorporates a safety assessment feature that categorizes FIB concentration levels into three distinct categories: red, yellow, and green. When the predicted FIB concentration falls below 300, it is classified as 'green,' signifying a safe condition. Concentrations between 300 and 800 are flagged as 'yellow,' indicating a slightly unsafe range. In cases where FIB levels exceed 800, the application designates them as 'red,' highlighting a hazardous situation.

7 CONCLUSION AND DISCUSSION

Our research is motivated by the global challenge of waterborne diseases and a paucity of FIB monitoring data, which is predictive of waterborne disease occurrence. We first introduced a surface water quality dataset consisting of approximately 10,000 FIB samples collected from Chicago beaches, which to our knowledge, is the largest dataset of its kind. With this dataset, we explored machine learning models to predict FIB levels in surface waters using weather and other types of data that are available from weather station websites to examine transfer learning and the applicability of domain adaptation on the data collected.

Using data from Chicago beaches, we followed the previous studies and trained and validated models for predicting the levels of FIB in a regression setup. Compared to the other published datasets [12, 26], the Chicago water quality dataset is the largest in the number of samples, as we collected data from 19 beaches over 100 days of "beach season" from 2017-2022.

We employed a group of ensemble learning, gradient boosting, and neural network models; our experiments and analysis demonstrate that the results for RF, XGBoost, and TabNet models were promising in both baseline supervised learning and transfer learning settings. We examine transfer learning by grouping the beaches into three groups and training the models in each group separately. Applying Supervised DA (FA, BWT) and Unsupervised DA (CORAL, SA) methods enhanced the transfer learning performance further.

In our regression task, the UDA methods did not improve the XGBoost and TabNet performance, but the results of the SDA methods were promising, especially with RF models.

A limitation of this work is that Chicago beaches are largely protected from wastewater discharge except following very heavy precipitation; that may not be the case in many surface waters elsewhere. Nevertheless, our findings suggest opportunities to extend domain adaptation work to water quality monitoring in settings where FIB levels are rarely measured, and hence, we believe our work has a strong social impact for improving public health in locations worldwide.

REFERENCES

[1] Sercan Ö Arik and Tomas Pfister. 2021. TabNet: Attentive Interpretable Tabular Learning. *Proc. Conf. AAAI Artif. Intell.* 35, 8 (May 2021), 6679–6687.

[2] Mathias Bourel, Angel M Segura, Carolina Crisci, Guzmán López, Lia Sampognaro, Victoria Vidal, Carla Kruk, Claudia Piccini, and Gonzalo Perera. 2021. Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Research* 202 (2021), 117450.

[3] Wesley Brooks, Steven Corsi, Michael Fienen, and Rebecca Carvin. 2016. Predicting recreational water quality advisories: A comparison of statistical methods. *Environmental modelling & software* 76 (2016), 81–94.

[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. (March 2016). arXiv:1603.02754 [cs.LG]

[5] M Chrzastowski. 2004. History of the Uniquely Designed Groins Along the Chicago Lakeshore. *Journal of Coastal Research* (2004), 19–38.

[6] The City of Chicago. 2023. Beach Lab Data, City of Chicago, Data Portal. https://data.cityofchicago.org/Parks-Recreation/Beach-Lab-Data/2ivx-z93u.

[7] Hal Daumé III. 2009. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815* (2009).

[8] Samuel Dorevitch, Stephanie DeFlorio-Barker, Rachael M Jones, and Li Liu. 2015. Water quality as a predictor of gastrointestinal illness following incidental contact water recreation. *Water Res.* 83 (Oct. 2015), 94–103.

[9] Samuel Dorevitch, Abhilasha Shrestha, Stephanie DeFlorio-Barker, Cathy Breitenbach, and Ira Heimler. 2017. Monitoring urban beaches with qPCR vs. culture measures of fecal indicator bacteria: Implications for public notification. *ENVIRONMENTAL HEALTH* 16 (MAY 12 2017). https://doi.org/10.1186/s12940-017-0256-y

[10] Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. 2013. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*. 2960–2967.

[11] Luka Grbčić, Siniša Družeta, Goran Mauša, Tomislav Lipić, Darija Vukić Lušić, Marta Alvir, Ivana Lučin, Ante Sikirica, Davor Davidović, Vanja Travaš, et al. 2022. Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis. *Environmental Modelling & Software* 155 (2022), 105458.

[12] Jiuhao Guo and Joseph H W Lee. 2021. Development of predictive models for "very poor" beach water quality gradings using class-imbalance learning. *Environ. Sci. Technol.* 55, 21 (Nov. 2021), 14990–15000.

[13] Wiley C Jennings, Eunice C Chern, Diane O'Donohue, Michael G Kellogg, and Alexandria B Boehm. 2018. Frequent detection of a human fecal indicator in the urban ocean: environmental drivers and covariation with enterococci. *Environmental Science: Processes & Impacts* 20, 3 (2018), 480–492.

[14] Rachael M Jones, Li Liu, and Samuel Dorevitch. 2013. Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection. *Environ. Monit. Assess.* 185, 3 (March 2013), 2355–2366.

[15] Lingbo Li, Jundong Qiao, Guan Yu, Leizhi Wang, Hong-Yi Li, Chen Liao, and Zhenduo Zhu. 2022. Interpretable tree-based ensemble model for predicting beach water quality. *Water Research* 211 (2022), 118078.

[16] Lubo Liu, Mantha S Phanikumar, Stephanie L Molloy, Richard L Whitman, Dawn A Shively, Meredith B Nevers, David J Schwab, and Joan B Rose. 2006. Modeling the transport and inactivation of E. coli and enterococci in the near-shore region of Lake Michigan. *Environmental science & technology* 40, 16 (2006), 5022–5028.

[17] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. 2008. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 39, 2 (2008), 539–550.

[18] Nick Lucius, Kevin Rose, Callin Osborn, Matt E Sweeney, Renel Chesak, Scott Beslow, and Tom Schenk Jr. 2019. Predicting E. coli concentrations using limited qPCR deployments at Chicago beaches. *Water research X* 2 (2019), 100016.

[19] Chiping Nieh, Samuel Dorevitch, Li C Liu, and Rachael M Jones. 2014. Evaluation of imputation methods for microbial surface water quality studies. *Environ. Sci. Process. Impacts* 16, 5 (May 2014), 1145–1153.

[20] Greg A Olyphant and Richard L Whitman. 2004. Elements of a predictive model for determining beach closures on a real time basis: the case of 63rd Street Beach Chicago. *Environ. Monit. Assess.* 98, 1-3 (Nov. 2004), 175–190.

[21] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering* 22, 10 (2009), 1345–1359.

[22] Anjaneyulu Panidhapu, Ziyu Li, Atefeh Aliashrafi, and Nicolás M Peleato. 2020. Integration of weather conditions for predicting microbial water quality using Bayesian Belief Networks. *Water research* 170 (2020), 115349.

[23] Yongeun Park, Minjeong Kim, Yakov Pachepsky, Seoung-Hwa Choi, Jeong-Goo Cho, Junho Jeon, and Kyung Hwa Cho. 2018. Development of a nowcasting system using machine learning approaches to predict fecal contamination levels at recreational beaches in Korea. *Journal of environmental quality* 47, 5 (2018), 1094–1102.

[24] Emily K Read, Lindsay Carr, Laura De Cicco, Hilary A Dugan, Paul C Hanson, Julia A Hart, James Kreft, Jordan S Read, and Luke A Winslow. 2017. Water quality data for national-scale aquatic research: The Water Quality Portal. *Water Resources Research* 53, 2 (2017), 1735–1745.

[25] Derek Rothenheber and Stephen Jones. 2018. Enterococcal concentrations in a coastal ecosystem are a function of fecal source input, environmental conditions, and environmental sources. *Applied and Environmental Microbiology* 84, 17 (2018), e01038–18.

[26] Ryan T Searcy and Alexandria B Boehm. 2021. A day at the beach: Enabling coastal water quality prediction with high-frequency sampling and data-driven models. *Environmental Science & Technology* 55, 3 (2021), 1908–1918.

[27] Ryan T Searcy and Alexandria B Boehm. 2022. Know before you go: Data-driven beach water quality forecasting. *Environ. Sci. Technol.* (Dec. 2022).

[28] Dawn A Shively, Meredith B Nevers, Cathy Breitenbach, Mantha S Phanikumar, Kasia Przybyla-Kelly, Ashley M Spoljaric, and Richard L Whitman. 2016. Prototypic automated continuous recreational water quality monitoring of nine Chicago beaches. *J. Environ. Manage.* 166 (Jan. 2016), 285–293.

[29] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[30] Justin W Telech, Kristen P Brenner, Rich Haugland, Elizabeth Sams, Alfred P Dufour, Larry Wymer, and Timothy J Wade. 2009. Modeling Enterococcus densities measured by quantitative polymerase chain reaction and membrane filtration using environmental conditions at four Great Lakes beaches. *Water Research* 43, 19 (2009), 4947–4955.

[31] W Thoe, M Gold, A Griesbach, M Grimmer, ML Taggart, and AB Boehm. 2014. Predicting water quality at Santa Monica Beach: evaluation of five different models for public notification of unsafe swimming conditions. *Water research* 67 (2014), 105–117.

[32] U.S. Environmental Protection Agency. 2020. 1: Enterococci in Water by TaqMan® Quantitative Polymerase Chain Reaction (qPCR) with Internal Amplification Control (IAC) Assay. In *Method 1609*.
[33] Timothy J Wade, Rebecca L Calderon, Elizabeth Sams, Michael Beach, Kristen P Brenner, Ann H Williams, and Alfred P Dufour. 2006. Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness. *Environ. Health Perspect.* 114, 1 (Jan. 2006), 24–28.
[34] Richard L Whitman and Meredith B Nevers. 2008. Summer E. coli patterns and responses along 23 Chicago beaches. *Environ. Sci. Technol.* 42, 24 (Dec. 2008), 9217–9224.
[35] Pengcheng Wu and Thomas G Dietterich. 2004. Improving SVM accuracy by training on auxiliary data sources. In *Proceedings of the twenty-first international conference on Machine learning*. 110.
[36] Juan Zhang, Han Qiu, Xiaoyu Li, Jie Niu, Meredith B Nevers, Xiaonong Hu, and Mantha S Phanikumar. 2018. Real-time nowcasting of microbiological water quality at recreational beaches: a wavelet and artificial neural network-based hybrid modeling approach. *Environmental science & technology* 52, 15 (2018), 8446–8455.
[37] Zaihong Zhang, Zhiqiang Deng, and Kelly A Rusch. 2012. Development of predictive models for determining enterococci levels at Gulf Coast beaches. *Water research* 46, 2 (2012), 465–474.

A RESEARCH METHODS

A.1 Part One

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi malesuada, quam in pulvinar varius, metus nunc fermentum urna, id sollicitudin purus odio sit amet enim. Aliquam ullamcorper eu ipsum vel mollis. Curabitur quis dictum nisl. Phasellus vel semper risus, et lacinia dolor. Integer ultricies commodo sem nec semper.

A.2 Part Two

Etiam commodo feugiat nisl pulvinar pellentesque. Etiam auctor sodales ligula, non varius nibh pulvinar semper. Suspendisse nec lectus non ipsum convallis congue hendrerit vitae sapien. Donec at laoreet eros. Vivamus non purus placerat, scelerisque diam eu, cursus ante. Etiam aliquam tortor auctor efficitur mattis.

B ONLINE RESOURCES

Nam id fermentum dui. Suspendisse sagittis tortor a nulla mollis, in pulvinar ex pretium. Sed interdum orci quis metus euismod, et sagittis enim maximus. Vestibulum gravida massa ut felis suscipit congue. Quisque mattis elit a risus ultrices commodo venenatis eget dui. Etiam sagittis eleifend elementum.

Nam interdum magna at lectus dignissim, ac dignissim lorem rhoncus. Maecenas eu arcu ac neque placerat aliquam. Nunc pulvinar massa et mattis lacinia.

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009