Predicting Surface Water Bacteria Levels Using Transfer Learning and Domain Adaptation

Ali Elahi
Department of Computer Science
University of Illinois Chicago
Chicago, IL, USA
aelahi6@uic.edu

David Shumway
Department of Computer Science
University of Illinois Chicago
Chicago, IL, USA
dshumw2@uic.edu

Megan Kowalcyk, Abhilasha Shrestha
Environmental and Occupational Health Sciences
University of Illinois Chicago
Chicago, IL, USA
{mkowal33, ashres2}@uic.edu

Nikita Gautam, Doina Caragea
Department of Computer Science
Kansas State University
Manhattan, KS, USA
{ngautam,dcaragea}@ksu.edu

Cornelia Caragea
Department of Computer Science
University of Illinois Chicago
Chicago, IL, USA
cornelia@uic.edu

Samuel Dorevitch
Environmental and Occupational Health Sciences
University of Illinois Chicago
Chicago, IL, USA
sdorevit@uic.edu

***Abstract*—Surface water contaminated by fecal bacteria can cause diarrheal illness, threatening human's health (especially among children). In recent years, supervised machine learning (ML) has been used to predict fecal indicator bacteria (FIB) levels. However, training ML models is challenging and, in some cases, even impractical due to sparsity of labeled data in all locations (e.g., in rural areas or low-income countries). In this paper, we introduce the largest water quality dataset available collected from beaches in Chicago and San Diego, USA. We utilized various models to predict historical FIB levels on this dataset establishing strong baseline models for supervised learning and transfer learning. Our models include Random Forest (RF), extreme gradient boosting (XGBoost), and attention-based tabular deep learning (TabNet) models. Additionally, given the widespread use of large language models (LLMs), we have fine-tuned the LLaMA3-8B model for regression in a tabular-to-text setting. Our results show that supervised and unsupervised domain adaptation methods can enhance transfer learning performance. Specifically, the supervised methods, especially RF, represent a promising solution for FIB level prediction, while domain adaptation could be successfully employed to predict FIB levels in locations where they are rarely measured.**

***Index Terms*—Water Quality Prediction, Fecal Indicator Bacteria, Transfer Learning, Domain Adaptation**

I. Introduction

High levels of bacteria in surface waters bring major challenges in global public health due to causing diarrheal disease, particularly among children and other more vulnerable populations. 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 [1]. 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-income countries) is impractical 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 [2].

A variety of machine learning (ML) approaches have been developed to make use of weather and other environmental variables to predict FIB levels in high-income countries [3]–[7]. Many of the FIB predictive variables - such as precipitation, temperature, solar irradiation, and wave height - are available on the Web 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 [8] 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, which we aim to explore.

Towards this goal, our main contributions are as follows:

1) We first introduce a surface water quality dataset consisting of approximately 20,000 FIB samples collected at several Chicago and San Diego beaches, together

with the corresponding meteorological and water predictive variables retrieved. To our knowledge, this is the largest dataset (with other published datasets containing less than 10% of our sample size) 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-based models and deep attention-based tabular networks, and LLM models that are trained and evaluated on data from the same location (or group of beaches at a location);
3) We also develop transfer learning baselines, i.e., supervised learning models trained on "source" beaches and tested on "target" beaches not included in the training;
4) We establish 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);
5) Finally, as a proof of concept, we design a Web app that retrieves environmental data for a location of interest from the Web and invokes our models to make predictions about the risk of infection at the location.

An overarching, important contribution of our work is addressing the time-consuming and costly process of collecting FIB data. We demonstrate that through domain adaptation, it is no longer necessary to collect large amounts of FIB data for every city or beach. By developing models in one location and leveraging large amounts of readily available weather and environmental data from other locations, we can predict water quality across various locations with minimal or no additional FIB data collection.

II. 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 [9]–[11]. Since then, random forest methods have been used to identify predictors to be used in statistical models of FIB at Chicago beaches [12]–[14]. More recently, Lucius et. al. [3] 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 et. al. [5] classified FIB exceedances (0/1) determined by thresholding FIB levels measured at beaches in Hong Kong using the EasyEnsemble (EE) algorithm [15], 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. [7] and southern California by Searcy et. al. [4]. However, none of these works studied domain adaptation approaches, although the work by Grbčić et. al. [7] 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 et. al. [5] and Searcy et. al. [4]. Guo et. al. [5] 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 et. al. [4] 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 [16]. 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 [17]. Bourel [18] utilized a simulated dataset which can be regenerated using their published codebase. In general, however, datasets used in previous work have often gone unpublished, and are not publicly available [3], [7], [12], [19]–[29].

In our paper, we construct a dataset based on beaches located in the United States for predicting FIB levels (specifically, ENT) based on environmental conditions. To our knowledge, this is the largest dataset for this task, and we will make it publicly available upon publication of this work. We also established strong supervised, transfer learning, and domain adaptation baselines using our dataset. In addition, we designed a Web app, with our models in the back-end, to be

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

TABLE I: Description and statistical properties of features.

Features			
Name	**Descriptions**	**Chicago: $\mathrm{Mean}_{\mathrm{STD}}$ (%Missing)**	**San Diego: $\mathrm{Mean}_{\mathrm{STD}}$ (%Missing)**
awind (m/s)	alongshore component of wind speed	$-0.129_{3.151}(24.496)$	$0.001_{1.975}(0.000)$
owind (m/s)	offshore component of wind speed	$-0.724_{3.151}(24.496)$	$0.764_{1.717}(0.000)$
WVHT (m)	significant wave height in meters	$0.260_{2.882}(46.614)$	$1.034_{0.474}$
Wtemp_B ($^{\circ}C$)	sea surface temperature	$20.844_{3.020}(29.049)$	$18.864_{2.917}(0.000)$
atemp ($^{\circ}C$)	air temperature	$21.193_{3.328}(74.795)$	$18.870_{4.070}(0.000)$
dtemp ($^{\circ}C$)	dew point temperature	$17.647_{1.309}(74.795)$	Feature Not Available
lograin3T (Inch)	log10 transform of past 3 days rainfall	$-1.708_{1.674}(10.944)$	$1.209_{4.831}(0.000)$
lograin7T (Inch)	log10 transform of past 7 days rainfall	$-0.553_{1.143}(21.440)$	$3.687_{9.942}(0.000)$
wet3 (–)	if it rained more than 0.1' in the past 3 days	$0.503_{0.498}(10.944)$	$0.184_{0.388}(0.000)$
wet7 (–)	if it rained more than 0.1' in the past 7 days	$0.844_{0.371}(21.440)$	$0.366_{0.481}(0.000)$
dtide_1 (Feet)	change in tide in the last hour	$0.001_{0.078}(5.792)$	Feature Not Available
dtide_2 (Feet)	change in tide in the last 2 hour	$0.001_{0.083}(5.792)$	Feature Not Available
tide_gtm (–)	if the value of tide is greater than mean tide	$0.531_{0.499}(5.792)$	$0.604_{0.488}(0.000)$
tide (Feet)	the water level in feet above or below the mean lower low water	$177.061_{0.190}(5.792)$	$3.594_{1.712}(0.000)$
DPD (Sec)	dominant wave period, seconds, is the period with the maximum wave energy	$3.979_{0.220}(51.897)$	Feature Not Available
comment (–)	Visual characteristics such as presence of sand, mud, residue, particles, wood chips, dirt, plants, etc.	Not Applicable	Feature Not Available
turbidity (NTU)	an expression of the optical property that causes light to be scattered and absorbed rather than transmitted in straight lines through the sample	$3.483_{4.376}(20.800)$	San Diego dataset has an alternative feature called Visibility. turbidity and visibility (measured by meters) are not comparable. $8730_{14819.2}(0.000)$
rad (watts/meter^2)	The amount of solar radiation received per unit area by a perpendicular surface	$351.164_{252.721}(37.747)$	$2942.115_{3328.580}(0.000)$
ENT (CCE/100 mL)	Concentration of ENT bacteria. dimention: calibrator cell equivalent per 100 ml.	$439.640_{3.350}(0.000)$	$546.841_{2119.159}(0.000)$

used to collect environmental data from the Web and to make predictions about water quality in real-time.

III. 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. The dataset contain daily measurements of the concentration of Enterococcus (ENT) from 19 beaches in Chicago from 2017 to 2022, and 14 beaches in San Diego from 2014 to 2021. To create the Chicago and San Diego subsets, we aimed to extract features similar to those in the existing California high-frequency water quality dataset [4]. However, some attributes from the California dataset were not available for beaches in our datasets, while some attributes included in our datasets are not present in the California dataset. The attributes in our datasets (shown in Table I) include various environmental characteristics such as wind, wave, precipitation, tide, solar radiation, air and water temperature, and turbidity.

Note also that not all features from the Chicago dataset were available for San Diego, and vice versa. For instance, turbidity data is collected for Chicago, while a similar feature called visibility is collected for San Diego, but these two are not directly comparable. Additionally, certain features such as comment, previous hours tide, and DPD, present in the Chicago dataset, are not available for San Diego.

A. Chicago Data

Water samples were collected at Chicago beaches 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 [30]. 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.

B. San Diego Data

Beaches included in the San Diego subset were those within 25 km of San Diego Bay. Because FIB levels at those beaches rarely exceeded the 2012 EPA Recreational Water Quality Criteria, we further restricted the dataset to beaches with a 90th percentile ENT value of 100 or greater. Additionally, we excluded from the dataset beaches with fewer than 50 days of beach monitoring, resulting in a dataset of 14 San Diego beaches. FIB levels for San Diego beaches were obtained from the EPA BEACON database [31]. Like the Chicago data, enterococci measurement were the FIB metric analyzed; unlike the Chicago data, enterococci were measured using culture, rather than qPCR method. Many San Diego area beaches contain more than one sampling location which are generally more than 1km apart; FIB values from sampling locations were analyzed individually rather than averaging the values

up to the beach level. Water samples collected at atypical times (before 6AM or after 2PM) were excluded. Because the 2014 BEACON data did not include San Diego area water sample collection times, we substituted median hour value for sample collection in the other years, 10AM, for San Diego beaches in 2014. On days for which more than one water sample was collected per sampling location, we calculated the mean ENT of measurements made during a three-hour interval beginning with the collection of the first water sample that day at that location. The average number of samples at a given monitoring location over the 8 year period (2014-2021) was 624 (230 min, 1031 max). ENT values were linked to weather data temporally and spatially. Wind direction was reported relative to the beach angle in order to calculate the onshore and offshore wind speed. Beach angles were manually calculated by viewing the geocode of the monitoring location on a map alongside shoreline data available from NOAA [32].

Total prior day solar direct normal irradiance (DNI) data from a single location near San Diego were linked to the ENT values. Wind speed (m/s), wind direction (degree), air temperature (C), wave height (m), and water temperature (C) from the prior hour before FIB sampling were also linked to the ENT values. Total precipitation (mm) data in the 72-hours (3 days) and 168-hours (7 days) preceding FIB sampling were further linked to ENT values. Tide level was interpolated between the highest and lowest values (m) and linked to ENT values on the hour of sampling. Solar DNI data were retrieved from the NREL National Solar Radiation Database. Buoy data was retrieved from NDBC. Wave height and water temperature data were available from a number of nearby buoys with the distance to buoy per observation being on average between 12.2 km for wave height (332 meters min, 268.0 km max) and 9.3 km for water temperature (1.4 km min, 46.7 km max). As buoy data were not always available from a given buoy, data from the nearest buoy were used for each individual FIB sample taken at a given monitoring location. Wind speed, wind direction, and air temperature data were collected from the weather station at San Diego International Airport, available from NOAA's Global Hourly Integrated Surface Database. Verified high and low tide level data were collected from two nearby tide stations via the NOAA Tides and Currents CO-OPS API for data retrieval. The average distance from a monitoring location to its nearest tide level station was 8.0 km (284 meters min, 20.3 km max).

C. Statistical Analysis of FIB Data from Beaches

We designed two types of transfer learning experiments. In the first set, we trained models on data from one city and evaluated their performance on data from another city. In the second set, we divided the 19 Chicago beaches into three groups based on their geographical location: Southern (SB), Central (CB), and Northern (NB) beaches. For this transfer learning task—where models are trained on one group of beaches and tested on another—it is expected that the beaches within the same group will share similar characteristics, while being distinct from beaches in other groups.

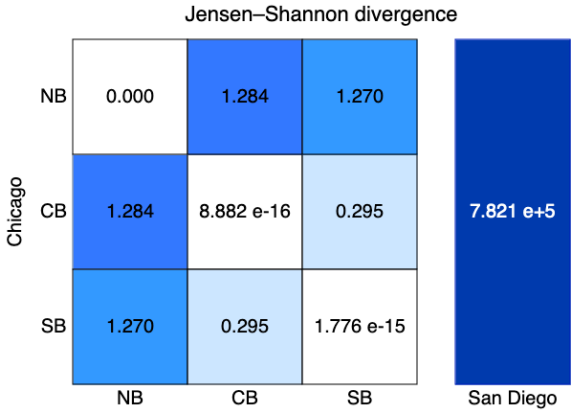

Fig. 1: This heatmap shows the JS divergence between Chicago groups of beaches and between Chicago and San Diego. The darker color indicates a higher divergence and less similarity between the distributions of data.

To ensure compatibility between the San Diego and Chicago datasets for transfer learning, we removed any features that were not common between them. It is important to note that different feature sets were used in the Chicago vs. San Diego experiments compared to the beach group experiments.

To test whether or not the data from two cities or two group of beaches are statistically different, we used the Jensen-Shannon (JS) divergence metric.We calculated the pairwise JS divergence between data collected at groups of beaches as:

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

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

The JS divergence metric results are shown in Fig. 1 and suggest that the current groups may benefit from transfer learning and domain adaptation. The right section of the figure corresponding to the Chicago-San Diego comparison using the JS divergence shows a high divergence between beaches in different cities.

D. 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. For the remaining missing values, we applied mean imputation for each attribute and added a small amount of noise to preserve variability. Tables I show the statistical properties of the features and the number of available data in each location, respectively. Following [4], the values for the regression target, ENT, were $\log_{10}$ transformed.

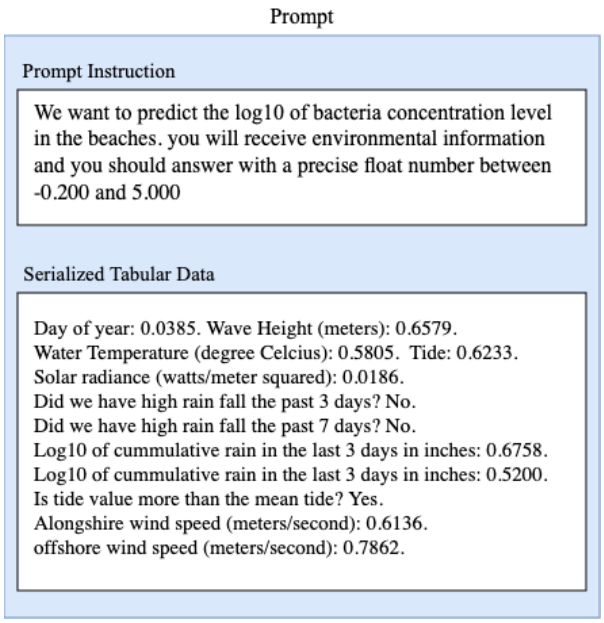

Fig. 2: A sample text prompt used to train and test the LLaMA-3 model.

E. 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 data spanning 2017-2019, while data from 2021 and later were used for testing. 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.

IV. Methodology

A. Supervised Machine Learning Models

a) Traditional Machine Learning Models: We consider the following machine learning models for our regression tasks. Following [4], [6], ensemble trees and gradient boosting methods show promising results in water quality regression tasks, so we chose Random Forest (RF) and XGBoost [33]. RF is an ensemble learning model that combines the output of multiple decision tree models, and XGBoost is a gradient-boosting algorithm.

b) Deep Learning Models: We used TabNet as another baseline models. TabNet [34] is a strong deep neural network model for tabular data that uses an attention mechanism to weigh each feature's importance selectively.

In addition to tabular data models, we also used a large language model in our study, specifically Llama-3, fine-tuned on our tabular training data converted to text, due to the current trends of using language models for regression and classification tasks based on tabular data. We generated text prompts using a format of {Prompt Question}+{Serialized Features}. A sample prompt is shown in Figure 2. We fine-tuned the LLaMA-3 model using the LoRa adaptor [35].

B. 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 [36] and feature augmentation [37]. 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 [38], and subspace alignment [39].

a) Balanced Weighting: In the BWT approach [36], 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 (limited) are thus utilized when training a BWT model.

C. Supervised Baseline Models

a) Feature Augmentation: In the FA approach [37], 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/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 used to train an FA model.

b) Correlation Alignment: In CORAL [38], 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 and target distributions become similar.

TABLE II: Regression results for supervised models. The performance is recorded using relative rRMSE (the lower the values the better). Random Forest (RF), XGBoost, TabNet, and LLaMA-3-8B models are employed to calculate the values for logENT. The performance in blue corresponds to models that perform better in this task.

Methods \ Source Beaches	NB	CB	SB	Chicago	San Diego
RF	0.334	0.358	0.311	0.374	0.597
XGBoost	0.341	0.402	0.353	0.378	0.593
TabNet	0.364	0.358	0.331	0.421	0.624
LLaMA-3	0.358	0.383	0.363	0.393	0.965

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.

c) Subspace Alignment: In SA [39], 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. 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 in the subspace alignment is obtained as a solution to the following minimization problem:

$$\min_{M} \|E_S M - E_T\|_F^2$$

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

d) Domain Adaptation for LLaMA models: To apply domain adaptation for the LLaMA-3 models, we augmented the source training data by adding a sample of 10% instances from the target training dataset (the same instances that were used in SDA). Additionally, we added a new feature called "location" to the feature set of domain adaptation training and testing data.

V. Experiments and Results

We trained our models utilizing attributes with no more than 50 percent missing data. Training involved employing k-fold cross-validation and grid search for hyper-parameter tuning. To ensure the results' statistical significance, the training and testing processes were executed five times, and the average performance across these runs is reported.

A. Metrics

To assess the performance of the models, we employed the relative root mean squared error (rRMSE) as the primary evaluation metric. This metric is defined as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{1}^{N}(y_{pred} - y)^2}, \text{ and } rRMSE = \frac{RMSE}{\bar{y}}$$

The utilization of rRMSE facilitates comparisons across datasets in different experiments. Normalizing RMSE by the average target value ($\bar{y}$) results in rRMSE values in the [0,1] interval, thus being independent of the target variable scale.

B. Supervised Learning

In our regression task, we explored RF, XGBoost, TabNet and Llama-3-8B (8 billion parameters) models for predicting logENT values. The results of this experiment are shown in Table II. We trained our supervised models on the Chicago and San Diego datasets, and we split the Chicago dataset based on groups of beaches and trained and tested the models on each group.

As indicated in Table II, the Random Forest (RF) and XGBoost models demonstrated more promising results for our regression task. Random Forest performed best across most source domains, with the exception of the San Diego dataset, where XGBoost outperformed it. Interestingly, the LLaMA-3 models performed strongly on the Chicago beach groups, outperforming TabNet in 3 out of 5 cases and even surpassing XGBoost in one instance.

Generally, we observe a higher rRMSE for the San Diego dataset, which might be attributed to differences in data collection methods for ENT values that results in difference in data distribution, or the inherent characteristics of the dataset.

C. Transfer Learning

For our transfer learning experiments, we transfer from one group of Chicago beaches to another and one, or from one city to another. Same timeline-based splits as in the supervised case are used for this experiment. Our regression results for this second experiment are shown in Table III, in the TL (transfer learning) rows, and for easier comparison, the rRMSE values for supervised models are also mentioned in the Table III–Supervised rows. As expected, the rRMSE results are worse for TL as compared to the supervised learning. Note that, the results in Tables II represent lower bounds for the transfer learning and domain adaptation results.

In Table III, for the transfer learning task, regressors trained on CB generally outperform those trained on NB and SB data. For example, the RF model transferred from CB to NB outperforms the model transferred from SB to NB (0.326 vs. 0.368). This trend holds true when transferring from CB and

TABLE III: Regression results measured by rRMSE for transfer learning and domain adaptation between Chicago groups of beaches. An improvement over the transfer learning model is marked with green. Best domain adaptation method for each test set is showed using **bold** font. The lower rRMSE scores show a better performance. Timeline split using 2018 and prior for train/development and post 2018 for test.

Trained on			**NB**		**CB**		**SB**		**Chicago**	**San Diego**
Model	**Approaches**		**CB**	**SB**	**NB**	**SB**	**NB**	**CB**	**San Diego**	**Chicago**
RF	Supervised		NB–NB: 0.334		CB–CB: 0.358		SB–SB: 0.311		Chi–Chi: 0.374	San–San: 0.597
	TL		0.426	0.323	0.326	0.317	0.368	0.462	0.747	0.528
	DA	FA	0.360	0.312	0.325	0.311	0.329	**0.353**	0.588	0.374
		BWT	0.391	0.315	**0.313**	**0.301**	0.340	0.419	**0.587**	0.375
		CORAL	0.475	0.330	0.328	0.326	0.368	0.479	0.747	0.511
		SA	0.456	0.333	0.326	0.308	0.353	0.443	0.799	0.448
XGBoost	Supervised		NB–NB: 0.341		CB–CB: 0.402		SB–SB: 0.353		Chi–Chi: 0.378	San–San: 0.593
	TL		0.436	0.361	0.350	0.353	0.419	0.51	0.76	0.449
	DA	FA	0.391	0.341	0.329	0.339	0.334	0.393	0.616	**0.362**
		BWT	0.402	0.339	0.341	0.331	0.359	0.448	0.615	0.365
		CORAL	0.468	0.358	0.359	0.352	0.424	0.513	0.777	0.444
		SA	0.456	0.362	0.366	0.344	0.372	0.473	0.813	0.483
TabNet	Supervised		NB–NB: 0.364		CB–CB: 0.358		SB–SB: 0.331		Chi–Chi: 0.421	San–San: 0.624
	TL		0.443	0.325	0.334	0.319	0.386	0.471	1.018	0.623
	DA	FA	0.422	0.334	0.340	0.319	0.373	0.404	0.594	0.422
		BWT	0.405	0.355	0.347	0.324	0.389	0.432	0.650	0.439
		CORAL	0.451	0.320	0.348	0.323	0.386	0.465	0.784	0.572
		SA	0.421	0.326	0.333	0.332	0.368	0.425	0.809	0.586
LLaMA-3 8B	Supervised		NB–NB: 0.358		CB–CB: 0.383		SB–SB: 0.363		Chi–Chi: 0.393	San–San: 0.965
	TL		0.388	0.349	0.352	0.342	0.367	0.388	0.780	0.864
	DA		0.382	0.317	0.345	0.336	0.344	0.392	0.771	0.742

NB to SB, and applies to XGBoost, TabNet, and LLaMA-3 models as well. Therefore, CB serves as the best source domain for predicting NB and SB test data. Between NB and SB as source datasets, NB models consistently outperform SB models. For instance, the RF model trained on NB achieves an rRMSE of 0.426 on CB, while the RF model trained on SB has an rRMSE of 0.462. This trend is also observed for XGBoost and TabNet, but not for LLaMA-3. Thus, the best source domain for the CB dataset is NB.

To analyze Chicago and San Diego transfer learning, we observe a significant increase in rRMSE when transferring between the two cities compared to supervised learning within the same city. For instance, there is an increase from 0.374 to 0.528 when transferring from Chicago (source) to Chicago (target) versus from San Diego (source) to Chicago (target) using the RF model or when transferring from Chicago (source) to San Diego (target), with an increase from 0.624 to 1.017 for the TabNet model. This trend generally occurs for all of the models. This poor performance in transfer learning between Chicago and San Diego can be explained by the substantial differences between the two datasets, as highlighted in Figure 1, since the cities differ significantly—one is next to a saltwater body, and the other is beside a freshwater lake.

We observe that the RF models consistently outperform all other machine learning methods, including LLaMA-3, in both the Chicago beach group experiments and the Chicago-San Diego transfer learning experiments. TabNet and LLaMA-3 show comparable performances, with each outperforming the other in different cases.

D. Supervised and Unsupervised Domain Adaptation

Finally, we analyze two supervised and two unsupervised DA algorithms for out traditional machine learning algorithms and TabNet model, and an augmentation based domain adaptation method for LLaMA-3 models. FA (Feature Augmentation) and BWT (Balanced Weighting) are supervised DA (SDA) algorithms, and CORAL (Correlation Alignment) and SA (Subspace Alignment) are unsupervised DA (UDA) algorithms. The SDA algorithms use 10 percent of labeled target training data (selected at random), while the UDA models use all the target training data as unlabeled. LLaMA-3 domain adaptation models receive 10 percent of the target's training data, which are then added to the source training data to enhance model performance during the adaptation process.

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 in Tables III — specifically, DA rows and TL, it can be seen that in the majority of the cases considered, the domain adaptation methods helped improve the

results. Based on Table III—DA rows, CB models remain the best source for NB and SB as the target after domain adaptation, similar to the results seen in transfer learning. However, in some cases (with TabNet and LLaMA-3), transferring from NB to SB results in a lower rRMSE compared to transferring from CB to SB. For predicting CB as the target domain, NB continues to be the best source domain for most of the cases, but the best predictor for CB after domain adaptation is trained on SB data.

Chicago-San Diego domain adaptation demonstrates a significant improvement in performance. A particularly interesting observation is that, after the drastic increase in rRMSE observed during transfer learning, domain adaptation models actually achieved better performance than supervised models. For example, transferring the RF model from San Diego to Chicago with domain adaptation reduced the rRMSE from 0.747 to 0.587, even surpassing the supervised learning models trained directly on the San Diego dataset (0.597).

The majority of models showed improvements with FA and BWT, while fewer improvements were observed with CORAL and SA. This is understandable, as CORAL and SA are UDA methods, which are generally expected to perform worse than SDA approaches. Additionally, the augmentation based domain adaptation method applied to the LLaMA-3 model demonstrated a significant improvement (except one out of eight cases), further proving that this approach can be effectively utilized with LLMs for DA in a regression setting.

VI. Web Application

To showcase the potential use of our models, we developed a proof-of-concept Web application that utilizes the most accurate predictive model in the back-end to estimate the FIB levels for a given location. While the environmental variables could be retrieved directly from the Web for a location of interest, in the current prototype, they are provided by the user or filled in using default values (representing the most common values in our dataset). Once the values of the environmental variables are filled in, the user can push the "Predict" button, and the best model is invoked to make a prediction on the current data point. The results are presented to the user in an easy-to-understand format, captured by three values, *Green* (safe), *Orange* (warning), and *Red* (danger), which represent three different categories of FIB levels. Specifically, *Green* represents FIB concentrations smaller than 300 cce (calibrator cell equivalent) and do not pose any danger, *Orange* represents concentrations between 300 cce and 800 cce, indicating slightly unsafe water, while *Red* represents concentrations that exceed 800 cce and highlight hazardous conditions. As a future extension of the app, we envision forecasting the water quality at a location for several days based on forecasted environmental variables so that the users can plan recreational water activities ahead of time. Alternatively, the app could show the water quality at several beaches around a location of interest, in case the user wants to target different beaches in that area.

Ultimately, the application should be capable of retrieving weather and environmental data based on a given location and utilizing pre-trained models to make predictions. Fig. 3 provides a preview of our web application. In the top image, the results for a query regarding a beach in Chicago are shown. The second and third images display additional information about water quality, such as the quality of nearby beaches and predicted water quality for the following days, based on retrieved environmental data.

VII. Conclusions 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 20,000 FIB samples collected from Chicago and San Diego beaches, which to our knowledge, is the largest dataset of its kind. With this dataset, we explored a diverse range of machine learning models and the regression capabilities of large language models to predict FIB levels in surface waters. We used weather and other types of data to investigate transfer learning and assess the effectiveness of domain adaptation on the collected data.

We employed a group of ensemble learning, gradient boosting, and neural network models; our experiments and analysis demonstrate that the results for RF, XGBoost models were promising in both baseline supervised learning and transfer learning settings. LLaMA-3 model cannot outperform the RF and XGBoost models but surprisingly the results are close. Note that we used the smaller version of the model (8B). It is expected that a larger model might outperform traditional ML models. We examined transfer learning within Chicago dataset by grouping the beaches into three groups and training the models in each group separately, and between the cities. Applying SDA and UDA 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. For our LLaMA-3 model we introduced an augmentation based domain adaptation method that significantly improved the results.

A limitation of this work is that Chicago and San Diego 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 DA work to water quality monitoring in settings where FIB levels are rarely measured; the expensive process of gathering vast FIB data can be transited to minimal or none using proposed domain adaptation models, more over, the existance of these models will allow us to use the massive weather and environmental data available on the web to predict the bacteria levels. We believe our work has a strong social impact on improving public health in locations worldwide. The outcomes of our study could be especially useful for vulnerable populations, e.g., children and the elderly.

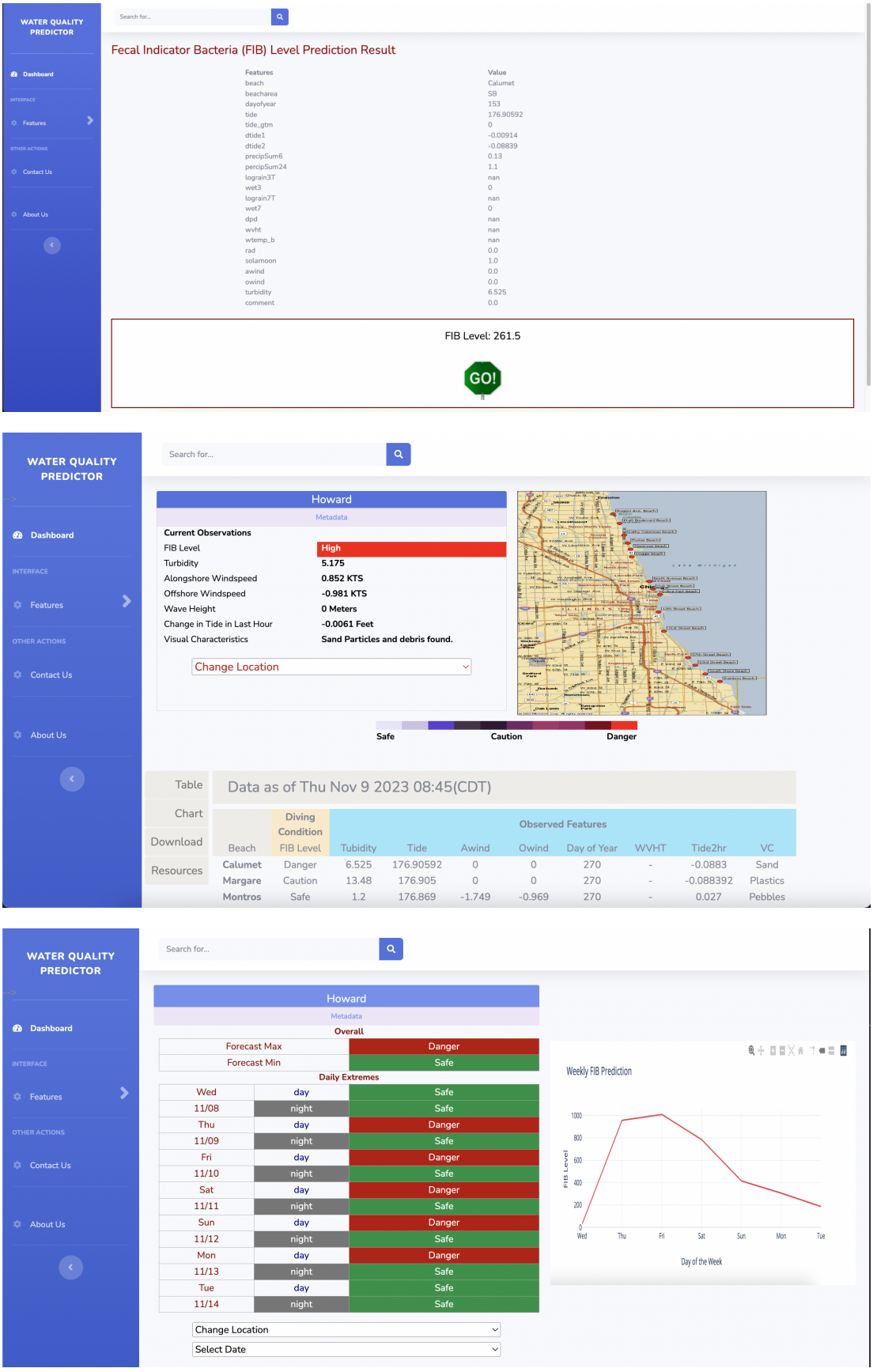

Fig. 3: Web Application Interface: The web application will receive the location and environmental information from the user and displays the prediction results. The results are summarized using three values: safe/green, warning/yellow, danger/red.

ACKNOWLEDGMENTS

This research is sponsored by the Department of the Navy, Office of Naval Research under ONR award number N00014-21-1-2286.

REFERENCES

[1] T. J. Wade, R. L. Calderon, E. Sams, M. Beach, K. P. Brenner, A. H. Williams, and A. P. Dufour, "Rapidly measured indicators of recreational water quality are predictive of swimming-associated gastrointestinal illness," *Environ. Health Perspect.*, vol. 114, no. 1, pp. 24–28, Jan. 2006.

[2] S. Dorevitch, A. Shrestha, S. DeFlorio-Barker, C. Breitenbach, and I. Heimler, "Monitoring urban beaches with qpcr vs. culture measures of fecal indicator bacteria: Implications for public notification," *ENVIRONMENTAL HEALTH*, vol. 16, MAY 12 2017.

[3] N. Lucius, K. Rose, C. Osborn, M. E. Sweeney, R. Chesak, S. Beslow, and T. Schenk Jr, "Predicting e. coli concentrations using limited qpcr deployments at chicago beaches," *Water research X*, vol. 2, p. 100016, 2019.

[4] R. T. Searcy and A. B. Boehm, "A day at the beach: Enabling coastal water quality prediction with high-frequency sampling and data-driven models," *Environmental Science & Technology*, vol. 55, no. 3, pp. 1908–1918, 2021.

[5] J. Guo and J. H. W. Lee, "Development of predictive models for "very poor" beach water quality gradings using class-imbalance learning," *Environ. Sci. Technol.*, vol. 55, no. 21, pp. 14 990–15 000, Nov. 2021.

[6] R. T. Searcy and A. B. Boehm, "Know before you go: Data-driven beach water quality forecasting," *Environ. Sci. Technol.*, Dec. 2022.

[7] L. Grbčić, S. Družeta, G. Mauša, T. Lipić, D. V. Lušić, M. Alvir, I. Lučin, A. Sikirica, D. Davidović, V. Travaš *et al.*, "Coastal water quality prediction based on machine learning with feature interpretation and spatio-temporal analysis," *Environmental Modelling & Software*, vol. 155, p. 105458, 2022.

[8] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[9] G. A. Olyphant and R. L. Whitman, "Elements of a predictive model for determining beach closures on a real time basis: the case of 63rd street beach chicago," *Environ. Monit. Assess.*, vol. 98, no. 1-3, pp. 175–190, Nov. 2004.

[10] D. A. Shively, M. B. Nevers, C. Breitenbach, M. S. Phanikumar, K. Przybyla-Kelly, A. M. Spoljaric, and R. L. Whitman, "Prototypic automated continuous recreational water quality monitoring of nine chicago beaches," *J. Environ. Manage.*, vol. 166, pp. 285–293, Jan. 2016.

[11] R. L. Whitman and M. B. Nevers, "Summer e. coli patterns and responses along 23 chicago beaches," *Environ. Sci. Technol.*, vol. 42, no. 24, pp. 9217–9224, Dec. 2008.

[12] R. M. Jones, L. Liu, and S. Dorevitch, "Hydrometeorological variables predict fecal indicator bacteria densities in freshwater: data-driven methods for variable selection," *Environ. Monit. Assess.*, vol. 185, no. 3, pp. 2355–2366, Mar. 2013.

[13] C. Nieh, S. Dorevitch, L. C. Liu, and R. M. Jones, "Evaluation of imputation methods for microbial surface water quality studies," *Environ. Sci. Process. Impacts*, vol. 16, no. 5, pp. 1145–1153, May 2014.

[14] S. Dorevitch, S. DeFlorio-Barker, R. M. Jones, and L. Liu, "Water quality as a predictor of gastrointestinal illness following incidental contact water recreation," *Water Res.*, vol. 83, pp. 94–103, Oct. 2015.

[15] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2008.

[16] T. City of Chicago, "Beach lab data, city of chicago, data portal," https://data.cityofchicago.org/Parks-Recreation/Beach-Lab-Data/2ivx-z93u, 2023.

[17] E. K. Read, L. Carr, L. De Cicco, H. A. Dugan, P. C. Hanson, J. A. Hart, J. Kreft, J. S. Read, and L. A. Winslow, "Water quality data for national-scale aquatic research: The water quality portal," *Water Resources Research*, vol. 53, no. 2, pp. 1735–1745, 2017.

[18] M. Bourel, A. M. Segura, C. Crisci, G. López, L. Sampognaro, V. Vidal, C. Kruk, C. Piccini, and G. Perera, "Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters," *Water Research*, vol. 202, p. 117450, 2021.

[19] L. Liu, M. S. Phanikumar, S. L. Molloy, R. L. Whitman, D. A. Shively, M. B. Nevers, D. J. Schwab, and J. B. Rose, "Modeling the transport and inactivation of e. coli and enterococci in the near-shore region of lake michigan," *Environmental science & technology*, vol. 40, no. 16, pp. 5022–5028, 2006.

[20] J. W. Telech, K. P. Brenner, R. Haugland, E. Sams, A. P. Dufour, L. Wymer, and T. J. Wade, "Modeling enterococcus densities measured by quantitative polymerase chain reaction and membrane filtration using environmental conditions at four great lakes beaches," *Water Research*, vol. 43, no. 19, pp. 4947–4955, 2009.

[21] Z. Zhang, Z. Deng, and K. A. Rusch, "Development of predictive models for determining enterococci levels at gulf coast beaches," *Water research*, vol. 46, no. 2, pp. 465–474, 2012.

[22] W. Thoe, M. Gold, A. Griesbach, M. Grimmer, M. Taggart, and A. Boehm, "Predicting water quality at santa monica beach: evaluation of five different models for public notification of unsafe swimming conditions," *Water research*, vol. 67, pp. 105–117, 2014.

[23] W. Brooks, S. Corsi, M. Fienen, and R. Carvin, "Predicting recreational water quality advisories: A comparison of statistical methods," *Env. modelling & software*, vol. 76, pp. 81–94, 2016.

[24] Y. Park, M. Kim, Y. Pachepsky, S.-H. Choi, J.-G. Cho, J. Jeon, and K. H. Cho, "Development of a nowcasting system using machine learning approaches to predict fecal contamination levels at recreational beaches in korea," *Journal of environmental quality*, vol. 47, no. 5, pp. 1094–1102, 2018.

[25] J. Zhang, H. Qiu, X. Li, J. Niu, M. B. Nevers, X. Hu, and M. S. Phanikumar, "Real-time nowcasting of microbiological water quality at recreational beaches: a wavelet and artificial neural network-based hybrid modeling approach," *Environmental science & technology*, vol. 52, no. 15, pp. 8446–8455, 2018.

[26] D. Rothenheber and S. Jones, "Enterococcal concentrations in a coastal ecosystem are a function of fecal source input, environmental conditions, and environmental sources," *Applied and Environmental Microbiology*, vol. 84, no. 17, pp. e01 038–18, 2018.

[27] W. C. Jennings, E. C. Chern, D. O'Donohue, M. G. Kellogg, and A. B. Boehm, "Frequent detection of a human fecal indicator in the urban ocean: environmental drivers and covariation with enterococci," *Environmental Science: Processes & Impacts*, vol. 20, no. 3, pp. 480–492, 2018.

[28] A. Panidhapu, Z. Li, A. Aliashrafi, and N. M. Peleato, "Integration of weather conditions for predicting microbial water quality using bayesian belief networks," *Water research*, vol. 170, p. 115349, 2020.

[29] L. Li, J. Qiao, G. Yu, L. Wang, H.-Y. Li, C. Liao, and Z. Zhu, "Interpretable tree-based ensemble model for predicting beach water quality," *Water Research*, vol. 211, p. 118078, 2022.

[30] U.S. Environmental Protection Agency, "1: Enterococci in water by TaqMan® quantitative polymerase chain reaction (qPCR) with internal amplification control (IAC) assay," in *Method 1609*, 2020.

[31] ——, "Ambient water quality tools," https://www.epa.gov/waterdata/ambient-water-quality-tools, 2024, accessed: 2024-02-12.

[32] National Oceanic and Atmospheric Administration (NOAA), "Western u.s. shoreline data," https://geodesy.noaa.gov/dist_shoreline/Western.zip, 2024, accessed: 2024-09-04.

[33] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Mar. 2016.

[34] S. Ö. Arik and T. Pfister, "TabNet: Attentive interpretable tabular learning," *Proc. Conf. AAAI Artif. Intell.*, vol. 35, no. 8, pp. 6679–6687, May 2021.

[35] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=nZeVKeeFYf9

[36] P. Wu and T. G. Dietterich, "Improving svm accuracy by training on auxiliary data sources," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 110.

[37] H. Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.

[38] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[39] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 2960–2967.