AN INTELLIGENT MOBILE APPLICATION FOR PREDICTING PATHOGENIC BACTERIA LEVELS AND WATER QUALITY IN INLAND AND COASTAL BEACHES USING MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Eileen Weiyun Ho[1], Armando Contreras[2]

[1]Lexington High School, 251 Waltham St, Lexington, MA 02421
[2]Computer Science Department, California State Polytechnic University, Pomona, CA 91768

ABSTRACT

Indicator organisms such as Escherichia coli (E. coli) are vital for monitoring microbiological water quality [16]. However, current testing methods are reactive, which may cause delays in reporting E. coli levels after contamination. This can make timely interventions difficult, especially in locations lacking in testing infrastructure. Our proposal involves the creation of a machine learning-based algorithm and application that predicts and displays microbiological water quality and any potential infractions.

Our research examined correlations between E. coli levels, date, and temperature. We found that E. coli levels peaked in July, modeled by an exponential trendline; temperature showed a strong correlation, likely due to its influence in the other variables. We also validated our app's predictions of E. coli levels using data from the Massachusetts Department of Public Health (MDPH) data. Our application had an average prediction difference of 2 units across 50 locations. These findings suggest reliable, real-time water safety information. Through machine learning, our application aims to provide proactive insights into water quality to enhance public health and safety.

KEYWORDS

Pathogenic Bacteria Prediction, Water Quality Prediction, Machine Learning, Environmental Health and Safety

1. INTRODUCTION

Indicator organisms are widely used in water quality monitoring to evaluate the presence of pathogens in a body of water. Escherichia coli (E. coli) is a Cultural fecal indicator bacteria (FIB) that is commonly used to determine the presence of pathogenic bacteria [1]. While many indicator microorganisms exist, E. coli is currently the water sector standard due to its specificity and ease of testing [2]. Contact or accidental consumption of E. coli can result in symptoms such as dysentery, chills, fever, headache, and muscular pain. In addition, an increase in E. coli levels suggest a proportional increase of other pathogenic bacteria.

David C. Wyld et al. (Eds): IoTE, CNDC, DSA, AIAA, NLPTA, DPPR - 2024
pp. 83-94, 2024. - CS & IT - CSCP 2024 DOI: 10.5121/csit.2024.142307

Stormwater and sewer runoff are common pathways in which pathogenic bacteria enter a water body. Correspondingly, the survival and movement of E. coli populations are strongly influenced by local climatic factors, such as precipitation and temperature. For example, warm and wet conditions exacerbate E. coli concentrations in surface water [3]. Given the connection between certain climatic conditions (such as increased precipitation and temperature), climate change has served only to intensify this issue. Between 1958 and 2012, the Northeast United States experienced a 72% increase in the amount of rainfall measured during heavy precipitation events. This region also saw the highest increase in annual precipitation in the United States since 1991 at 8% (Walsh et al., 2014) [4]. Total seasonal precipitation, as well as the frequency and intensity of heavy storms are both projected to increase in the future [5]. Moreover, climate change in the 21st century will increase the frequency of extreme climatic events, such as high-temperature anomalies and "heat waves" [6].

Despite this growing public health risk, there is a lack of centralized resources and prediction technologies that efficiently warn the public of these pathogenic insurgencies. This has placed the burden on our public health system, traditional monitoring systems, and ultimately, our communities.

The first methodology focuses on evaluating microbiological water quality for drinking purposes using models like Exponential Smoothing, ARIMA, and machine learning. This methodology is based on a study conducted at a key point in the Göta älv river in Sweden. However, due to its smaller scope and geographical specificity, our project expands the scope to cover both drinking and recreational water, while also utilizing a larger dataset from the entire state of Massachusetts. The second examined the relationship between coliform bacteria and various in-situ water quality parameters, which used logistic regression models. Their study was focused on coliform bacteria and relied on direct water quality parameters. In contrast, this project uses ex-situ precipitation and temperature data to predict E. coli levels, which are more relevant to modern water quality standards.

The last methodology compared machine learning algorithms for predicting E. coli in agricultural pond waters. Their study was limited to two ponds over a short period of time, and focused on comparing models instead of applying them. The current project applies machine learning models across a much larger geographical region and over a longer period of time, focusing on simple applications rather than just comparing models.

Our proposal involves the creation of a machine learning-based application that predicts and displays pathogenic bacteria levels and water quality. Due to the periodic manual testing process and variability of testing frequency among water bodies, current water quality data collection is reactive [7]. In other words, E. coli levels reported above the threshold are delayed with respect to the moment of contamination, which hinders timely interventions and alerts. Our application takes a proactive approach, analyzing water quality data to provide a continuous stream of information. By doing this, a larger window of time exists in which the public can be notified and interventions can be conducted. This way, residents and authorities receive both real-time and prospective data that can lead to efficient preventative measures, ultimately enhancing public safety.

While beach water quality information can be searched on the internet, it can be challenging to find for individuals who may not be well versed on such endeavors. By implementing the algorithm in a mobile application, a repository of centralized information is widely and easily available for individuals to easily use. Moreover, our algorithm's scalability allows it to be expanded across a large number of locations, adapting to regional climate and water bodies. This ensures that all communities, whether urban or rural, have access to data-driven and up-to-date

public health information. This feature is especially beneficial for communities that lack regular testing resources and services, serving as a vital tool for local residents and authorities.

In summary, our machine-learning based service aims to be a solution that provides a continuous stream of data in a proactive manner, thereby making health data more accessible to the public. By working in conjunction with traditional testing methods, our proposal enables communities to conduct effective management strategies for water quality and public health.

We conducted four experiments to analyze the relationship between microbiological water quality and various factors, and to verify the efficacy of our application. First, we aimed to determine whether a correlation existed between temporal frame and E. coli level. We determined that data could be best modeled using an exponential trendline, indicating that dates in the middle of the summer correlated with higher bacterial levels. We also examined the correlation between daily temperature and E. coli levels, finding only a weak relationship between the two.

For our second experiment, we tested the accuracy of our application's predictions by comparing predicted data to measured data. Across 50 test locations, the average difference between the app's predictions and MDPH data was 2 units, demonstrating high reliability.

Our third experiment aimed to determine the machine learning algorithm that most accurately predicted microbiological water quality levels. To do this, we compared three algorithms: linear regression, polynomial regression, and support vector machine. All algorithms were trained using the same dataset that included five variables: date, beach type, indicator level, organism, and violation. A random train-test split was performed, allocating 20% of the data for testing. The performance of each model was evaluated using the R^2 value from the score() method. We determined the support vector machine as the optimal choice.

Finally, we determined whether the inclusion of climatic data, in our case precipitation, would affect the results of our application's predictions in a positive manner as the prediction score was higher when comparing 0.25 without climatic data to 0.91 when utilizing it.

2. CHALLENGES

In order to build the project, a few challenges have been identified as follows.

2.1. Data Sampling

When collecting and processing water quality information, potential issues concerning data sampling must be addressed. The frequency of water quality testing is variable among water bodies, depending on their size, location, susceptibility to contamination, and other factors. This means there is potential for statistical bias introduced by irregular sampling intervals. To combat this, we could use resampling methods that aggregate data into consistent intervals, allowing us to more uniformly assess it. Issues may also arise when integrating information from different sources into one dataset. We can mitigate this issue by standardizing the data, including both weather and water quality data, to a common temporal framework.

2.2. The Accuracy of Predictions

When implementing our machine learning algorithm, we need to address potential data variations that can reduce the accuracy of our predictions. In particular, we need to address the variability that exists between different water bodies, such as size, salinity, or geographical location [8]. To

address this, we could develop a machine learning model that adjusts its parameters based on characteristics of a given water body. Moreover, because of the seasonal variation that occurs in weather patterns and water quality behavior, it may be useful to restrict our research to the months of May to September. This strategy is also preferable because coast water quality data is predominantly collected throughout this duration.

2.3. Presenting Data

Along with data prediction, presenting our data in a format accessible to the public is also pertinent. To ensure that the data is available to the public in a clear and understandable format, we could implement smart design features that allow intuitive and efficient access to information. For example, we could create a visual representation of data that visualizes complex data simply. We verify that our data is up-to-date and reliable, update schedules, timestamps, and confidence estimates could be used. Additionally, we can create elements such as filtering and user feedback mechanisms to increase usability and reliability of our application.

3. SOLUTION

Our system is composed of three key components. (1) Data prediction, (2) data collection and processing, and (3) data presentation. To predict a body's microbiological water quality, our application uses a machine learning algorithm that relies on training data from the Massachusetts Department of Public Health (MDPH). This dataset is collected throughout the year from Massachusetts water bodies and includes parameters such as beach type, indicator level, and beach location. This data is publicly available and collected primarily by local boards of health across the state. This data was exported as a CSV file and used to train our algorithm, inputted to return whether water quality thresholds were likely to be violated. Our application connects to an AI server build using Python's Flask framework. This model is based off of the Support Vector Machine (SVM), a supervised machine learning algorithm [9].

We aimed to display our data in a presentable manner. We implemented several features to ensure this: (a) map search and (b) list search. The map search allows users to utilize an interactive map interface to search for water quality. The filter feature allows users to sort by location name or geography. Moreover, we implemented a feature where users can find the nearest beaches to them within a 50000 km radius from the user. These options ensure that users can more quickly and effectively access relevant water quality data based on their current needs. Moreover, we ensure that people regardless of technical or scientific background are able to intuitively glean information.

Our application was developed using Thunkable, a platform that uses block coding to streamline the app development process. We selected this platform for its ease of use, simplicity, and array of development tools. We used location services available through the Thunkable website's Google maps feature. This combination of features is designed to provide predicted data in a readable format.

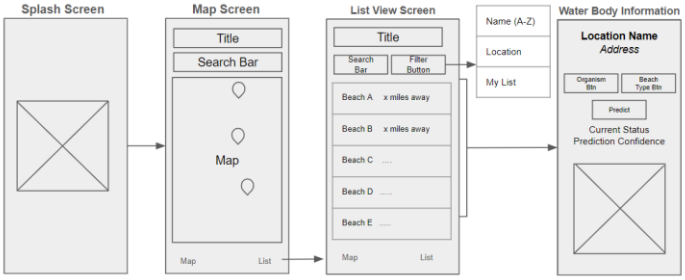

Figure 1. Layout for the application, including map, and list

The first component of our program is the prediction algorithm that allows our application to predict the microbiological quality of water for a user who requests data at a specific time and location. To do this, our application connects to a Python server and uses machine learning to provide users with predicted data.

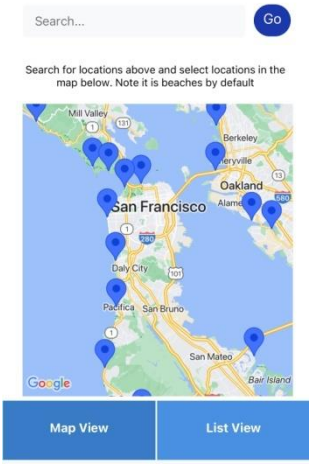

Figure 2. Graphic User Interface of the Map screen

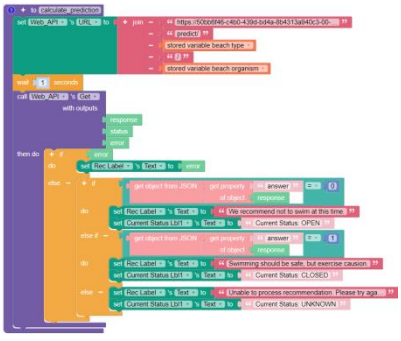

Figure 3. Code excerpt depicting data request and return

Our application connects an AI server build using Python's Flask framework. This machine learning model is based off of the Support Vector Machine (SVM), a supervised machine learning algorithm. As described in section 3.1 B, training data for our machine learning model was derived from the Massachusetts Department of Public Health (MDPH), which contains monitoring data for thousands of water bodies in the state. Figure 3 depicts the process in which

information is (1) requested by and (2) returned to the user. When a request is submitted for prediction, our application creates a GET request which it sends to an AI server through a URL; parameters such as beach type and location are included within this request. Using these parameters, the server responds by approximating the microbiological quality of the water body, marking whether thresholds are likely to have been violated. If successful, the server returns a JSON file which is then parsed and displayed for users. This process relies on key Python libraries like sklearn for the machine learning model, pandas for data handling, and Flask for managing the server.

The second key component of our system is data collection and processing. This helps ensure that the machine learning algorithm has enough information to predict microbiological water quality levels.

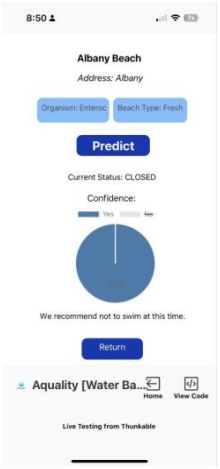

Figure 4. Graphic User Interface of the Prediction screen

Figure 5. Screenshot of code

As depicted in Figure 5, our program fits collected data into the model by transforming parameters into a valid format and assigning them to either an X or Y variable. Then, our application runs a function that returns the predicted value. In addition to processing collected data, our application also collects and processes location data. This feature allows the user to more easily find water bodies around them without needing to search manually. By using Thunkable's Google Maps feature, our application sends the user's position to the Flask server, which then searches for bodies of water within a 50 thousand mile radius.

In order to display information in a professional and readable manner we need to keep in mind how we can make it easy for users to understand. This includes displaying key information such

as precipitation levels, elevation, and bacteria levels, which are crucial for determining water quality.

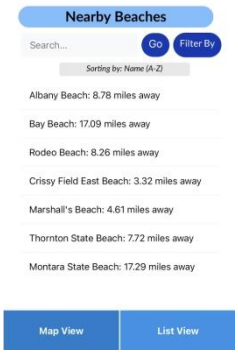

Figure 6. Application list sort feature

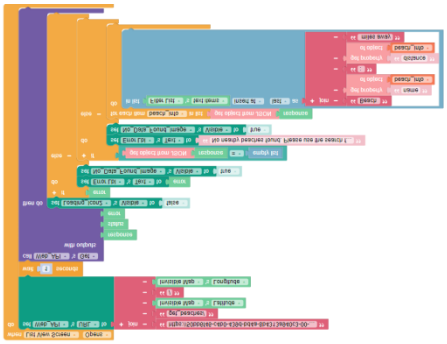

Figure 7. Application location

The app offers a feature that provides users with all the essential information they need about a location. This can be seen in the code above as it displays the latitude and longitude, as well as the name of the beach or body of water the user is viewing. Additionally, the app gives details about specific organisms, like E. coli, which can show if the water is contaminated. To help users understand whether it's safe for activities like swimming, the app indicates whether or not it is safe for human beings to swim given the level of bacteria in the water. This makes it easy for users to decide if they should take any precautions. The interface is designed to be user-friendly, so even people without a science background can easily navigate and understand the information. Overall, this feature allows users to make smart, informed choices about water safety, whether they're planning a beach day or just checking the quality of nearby water sources.

4. EXPERIMENT

4.1. Experiment 1

A potential blind spot in our program is the correlation between date and E. coli levels. Therefore, it is important to verify that there indeed exists a connection between time and microbiological water quality.

To investigate the correlation between date and E. coli levels, we conducted an analysis of data sourced from the Massachusetts historical beach database, supplied by the Massachusetts Department of Public Health. We selected 415 entries that spanned various temporal periods. Our

data was plotted with dates on the x-axis and E. coli levels on the y-axis, and we used an exponential trendline to demonstrate any potential correlations. We identified and omitted outliers to enhance data clarity, focusing on consistent data points. This experiment was conducted with the aim to help us assess how E. coli levels fluctuate over time, enabling us to better understand the correlation between the temporal framework and E. coli levels.

Additionally, we correlated temperature data from the same time periods, specifically looking at maximum, minimum, and average temperatures for the days corresponding to our E. coli measurements. By integrating this temperature data, we aimed to assess how thermal variations may influence bacterial levels.

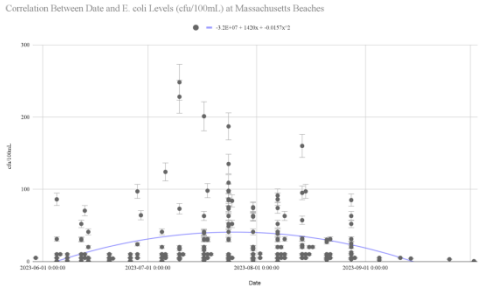

Figure 8. E. coli Levels Over Time at Massachusetts Beaches with Exponential Trendline

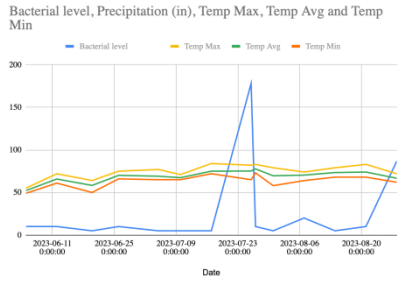

Figure 9. Correlation Between E. coli Levels and Temperature Variations

The first dataset spans from June 5, 2023, to October 2, 2023, covering nearly four months. The mean bacterial level is 30.71 cfu/100ml, while the median is significantly lower at 5 cfu/100ml, indicating a skewed distribution with high values. The range of 1183.5 highlights considerable variation, with 7 outliers being omitted from the graph. This combination of statistics suggests that while most readings are low, there are notable spikes in bacterial levels.

In the second data set, the spike on July 26 is particularly notable, suggesting potential environmental or ecological changes that warrant further investigation, such as local events or pollution sources.

As summer progresses, temperatures generally rise, with the highest average temperature of 75°F recorded on July 26. There appears to be a weak correlation between higher temperatures and increased bacterial levels, especially evident during this spike. This relationship highlights the potential impact of temperature on bacterial growth in the environment.

4.2. Experiment 2

It is also important to verify that our chosen algorithm accurately predicts water quality information. To set up the experiment, data is collected from the MDPH over a period of time, focusing on E. coli levels in various bodies of water. Then, testing would ensure utilizing the app's predictions for those same locations. By comparing the MDPH data with the app's results, this can test its accuracy. MDPH data serves as our control because it's a trusted source of water quality information. This experiment is set up this way to ensure the app's predictions are reliable. Consistently accurate predictions will confirm that the app provides trustworthy information to its users.

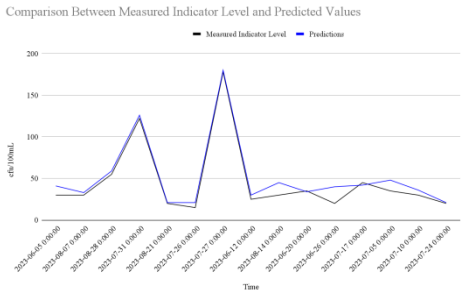

Figure 10. Validation of App Predictions Against MDPH Water Quality Data

To demonstrate the accuracy of the app's predictor, we ran several tests comparing its E. coli predictions to real data from the Massachusetts Department of Public Health (MDPH). For example, at the Marine beach type, the MDPH reported an E. coli level of 20, and the app predicted 23, showing a close match. Similarly, at a Fresh beach type, MDPH data showed a level of 5, while the app predicted 8. At River C, the actual reading was 20, and the app predicted 19. Overall, across 50 different test locations, the average difference between the app's predictions and the MDPH data was an average of only 2 units. This small difference demonstrates that the app's algorithm is reliable, and its consistent accuracy proves that it can use both historical and real-time data effectively. The results suggest that users can trust the app to give accurate water safety information. We did note an outlier with a maximum reading of 24,200 and a minimum of 0.5, but the high value seems to indicate a special case or misreading.

4.3. Experiment 3

Given that a variety of machine learning algorithms can be utilized to predict water quality levels, it is important to select the model that most accurately predicts results.

In order to select a machine learning algorithm that will most successfully return results, we designed this experiment to compare the efficacy of several algorithms. For this experiment, we selected three: linear regression, polynomial regression, and the support vector machine. The machines were trained based on the same dataset, including 5 variables: date, beach type, indicator level, organism, and violation. To evaluate the results of each algorithm, we created a random train-test split using the train_test_split function, which automatically divides the dataset into training and test sets. The test set was set to 0.2 of the entire dataset. The score() method is used to evaluate the performance of a model, which calculates the R^2 value.

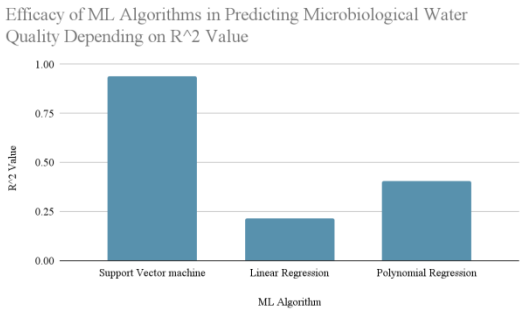

Figure 11. Comparison of Machine Learning Algorithms for Predicting Water Quality Levels

The tested models for predicting the water quality levels returned notable differences in their efficacy. The linear regression model returned a R^2 value of 0.22, demonstrating a weak relationship between predicted and actual data. The polynomial model scored higher than the linear model, returning a score of 0.41. While this indicates an improvement over linear regression, it still remains relatively inaccurate. Finally, the support vector machine returned a 0.94, outperforming both prior models. In summary, the support vector returned a score 0.53 units higher than the polynomial regression model and 0.72 units higher than the linear regression model. On this basis, we can determine that the SVM model is the best alternative as the preferred machine learning algorithm for predicting water quality levels. As such we decided to utilize it moving forward.

4.4. Experiment 4

It is also important to consider the variables that are taken into account when predicting our data. Depending on whether parameters like temperature and weather are accounted for in the dataset, algorithm accuracy could differ.

In order to determine the optimal number and type of variables to include in our dataset, we designed an experiment that utilizes feature set selection to most successfully produce results. Feature selection (FS) is a process of selecting relevant features to obtain the best performing subset of information. To this, we used a python program to merge the water quality sampling data with weather data using the same temporal framework [10]. Given this combined dataset, we calculated the violation status for the X variables (1) Date, organism versus (2) Date, organism, precipitation. The same machine learning model was used for both datasets, the support vector machine.

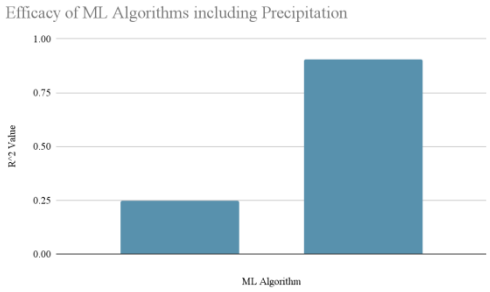

Figure 12. Comparison of Model Performance with and without Precipitation Data

The results of this experiment demonstrated that feature selection does influence the efficacy of the machine learning algorithm employed. More specifically, the inclusion of weather in the dataset (in the case of this experiment, precipitation) in the core dataset of date and organism allowed for an 0.25 in predictive performance. This depicts a relatively low relationship between predicted and actual results. The second model, which included weather data, returned an R-squared value of 0.91. Given the returned R-squared values, we determined that the second model is better by comparison. Based on the results of this experiment, we ultimately chose not to utilize precipitation in our machine learning model.

5. RELATED WORK

Sokolova et al. (2021) evaluated the microbiological water quality of surface water for the purpose of safe human consumption [11]. This study was conducted at the water intake point of a treatment plant located on the Göta älv river in Sweden. Their baseline approach utilized Exponential Smoothing and ARIMA (Autoregressive Integrated Moving Average), as well as several machine learning models. While this study focuses primarily on drinking water quality, our research expands the scope of microbiological water quality to recreational and swimmable purposes. In this sense, our scope covers both drinking water quality and surface water quality. In addition, we are using a wider dataset, the State of Massachusetts, encompassing an area of around 27,363 square kilometers.

Simon Appah Aram et al (2021) analyzes coliform bacteria in relation to 19 water quality parameters, including pH, nitrates, phosphates, etc. [12] This study examined the likelihood of coliform contamination using nested binary logistic regression models and a dataset of water quality data from domestic and commercial water supply systems. While this study uses related in-situ water quality parameters to predict coliform abundance, our research uses ex-situ precipitation and temperature data to predict E. coli abundance. Because E. coli is the primary indicator microorganism in the current water quality testing standard, our research is further tailored to this trend. Moreover, our dataset allows for a more applicable proactive approach due to lack of water monitoring infrastructure in certain areas.

Stocker et al. (2022) aims to predict E. coli abundance in agricultural pond waters by comparing various machine learning algorithms with multiple linear regression [13]. In this study, two irrigation ponds were sampled biweekly from May to August for two years, with analysis using various ML methods in predicting E. coli levels. In addition, this study assesses the effectiveness of specific parameters for water quality. Our research is similar in that it covers surface waters not for human consumption. Moreover, our study covers a larger geographical slope, and focuses on the application rather than the comparison of ML models in microbiological water quality prediction.

6. CONCLUSIONS

There are several ways that our research could be enhanced. One area in which our research is limited is in the geographical scope of our dataset. Our data is localized to Massachusetts water bodies, meaning that the state specific climate and water characteristics may influence the microbiological water quality differently than in other locations. Expanding our dataset to include a greater variety of climate types and water body types by expanding the geographical scope of our data collection may produce more accurate results for other areas [14]. Moreover, while E. coli and other Estrecheria bacteria are currently the standard for water quality monitoring, accounting for other indicators such as fecal and total coliform could provide a larger dataset and scope [15]. In addition, more analysis should be conducted to determine other influencing factors

of microbiological water quality. Using these factors, a deeper statistical analysis of these factors and water quality can be conducted. By doing this, it is possible that more nuanced trends and correlations can be uncovered.

We determined that in comparison to other machine learning algorithms we tested, the SVM (support vector machine) was most effective in predicting E. coli levels if given a suitable dataset. Moreover, we also concluded that temporal framework and certain climate data may influence prediction results.

REFERENCES

[1] Motlagh, Amir M., and Zhengjian Yang. "Detection and occurrence of indicator organisms and pathogens." Water Environment Research 91.10 (2019): 1402-1408.
[2] Odonkor, Stephen T., and Joseph K. Ampofo. "Escherichia coli as an indicator of bacteriological quality of water: an overview." Microbiology research 4.1 (2013): e2.
[3] Li, Rui, Gabriel Filippelli, and Lixin Wang. "Precipitation and discharge changes drive increases in Escherichia coli concentrations in an urban stream." Science of The Total Environment 886 (2023): 163892.
[4] Walsh, John, et al. "Ch. 2: Our changing climate." Climate change impacts in the United States: The third national climate assessment (2014): 19-67.
[5] Assessment, Climate. "Fourth national climate assessment." US Global Change Research Program: Washington, DC, USA (2018).
[6] Davariashtiyani, Ali, et al. "Exponential increases in high-temperature extremes in North America." Scientific Reports 13.1 (2023): 19177.
[7] Torres, Camilo, et al. "Evaluation of sampling frequency impact on the accuracy of water quality status as determined considering different water quality monitoring objectives." Environmental monitoring and assessment 194.7 (2022): 489.
[8] Galbraith, Lisa M., and Carolyn W. Burns. "Linking land-use, water body type and water quality in southern New Zealand." Landscape Ecology 22 (2007): 231-241.
[9] Suthaharan, Shan, and Shan Suthaharan. "Support vector machine." Machine learning models and algorithms for big data classification: thinking with examples for effective learning (2016): 207-235.
[10] Bouchlaghem, Younes, Yassine Akhiat, and Souad Amjad. "Feature selection: a review and comparative study." E3S web of conferences. Vol. 351. EDP Sciences, 2022.
[11] Sokolova, Ekaterina, et al. "Data-driven models for predicting microbial water quality in the drinking water source using E. coli monitoring and hydrometeorological data." Science of the Total Environment 802 (2022): 149798.
[12] Aram, Simon Appah, Benjamin M. Saalidong, and Patrick Osei Lartey. "Comparative assessment of the relationship between coliform bacteria and water geochemistry in surface and ground water systems." Plos one 16.9 (2021): e0257715.
[13] Stocker, Matthew D., Yakov A. Pachepsky, and Robert L. Hill. "Prediction of E. coli concentrations in agricultural pond waters: application and comparison of machine learning algorithms." Frontiers in Artificial Intelligence 4 (2022): 768650.
[14] Nõges, Tiina. "Relationships between morphometry, geographic location and water quality parameters of European lakes." Hydrobiologia 633.1 (2009): 33-43.
[15] Elmund, G. Keith, Martin J. Allen, and Eugene W. Rice. "Comparison of Escherichia coli, total coliform, and fecal coliform populations as indicators of wastewater treatment efficiency." Water Environment Research 71.3 (1999): 332-339.
[16] Vasavada, Purnendu C., Alvin Lee, and Roy Betts. "Conventional and novel rapid methods for detection and enumeration of microorganisms." Food safety engineering (2020): 85-128.