GMR-216

AI/ML-Based Water Quality Monitoring MobileApp for Predicting *E.coli* in Surface Waters

Introduction/Abstract

E.coli contamination in surface waters has proven to be a significant public health concern, requiring innovative monitoring solutions. Our research presents the design of an AI-driven mobile application to predict whether *E.coli* bacteria are present at levels exceeding acceptable thresholds in surface waters. The methodology employs sensor devices to collect water quality data parameters, such as water temperature, pH, dissolved oxygen and turbidity. A dataset is generated based on these parameters and machine learning (ML) algorithms are applied to evaluate accuracy, precision, recall, and processing time. Additionally, our ML algorithms establish a correlation matrix among water quality parameters to identify the key factors influencing *E.coli* levels. We applied various machine learning techniques to the dataset, including Support Vector Regression (SVR), Random Forest Classification (RFC), XGBoost, and ensemble methods that combine these algorithms. Our findings indicate that the ensemble of Random Forest Classification and XGBoost achieved the highest accuracy. Users can view E. coli predictions based on current sensor values through our Mobile App. Index Terms—Internet of Things (IoT), LoRaWAN, AI/ML algorithms, Ensemble Learning Model, Water Quality Monitoring, Mobile App, Predicting *E.coli*

Objective

To develop an affordable, real-time water quality monitoring system that uses machine learning models to predict *E. coli* contamination based on key water parameters, delivering timely contamination alerts to users via a mobile app.

Experimental Setup

The experimental setup for this paper involves using a LoRaWAN network to transmit water quality data collected from sensors measuring water temperature, pH, dissolved oxygen, and turbidity. Specifically, the setup includes an outdoor LoRaWAN gateway (RAK7289) with a range of up to 10 miles, which forwards sensor data to The Things Network (TTN) server. The TTN server then persists the data in a PostgreSQL database via MQTT, where it is analyzed using machine learning algorithms to predict unacceptable *E. coli* levels. A mobile application displays real-time predictions and sends notifications when *E. coli* levels exceed safe thresholds.

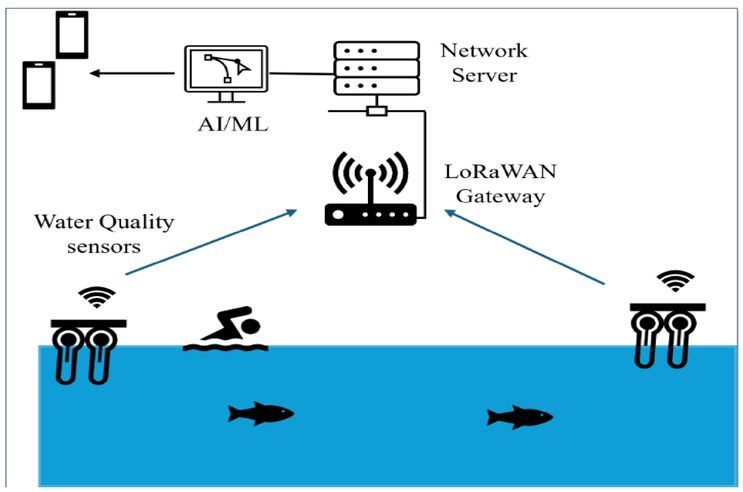

Fig 1: Water Quality Management Architecture

Algorithms Used

The solution leverages AI/ML algorithms to monitor surface waters continuously. It focuses on minimizing false negatives (failing to detect unsafe *E. coli* levels) while tolerating some false positives. The prediction model is built from sensor data, standardized using scikit- learn's StandardScaler, and evaluated through training and testing sets.

Intelligent Models for *E. Coli* Prediction

1.Support Vector Regression (SVR): A method designed for regression tasks that aims to find a function that fits the data within a defined margin of error while minimizing complexity.

2.Random Forest Classification (RFC): An ensemble method using multiple decision trees to improve prediction robustness and reduce overfitting through majority voting.

3.XGBoost: An efficient implementation of gradient boosting that builds trees sequentially, correcting errors from previous trees, and is well-suited for complex datasets.

Dataset Generation:

An initial synthetic dataset was generated using Python, defining acceptable ranges and thresholds for random value generation.

Water Temperature: Maintained between 60°F and 90°F
pH: The acceptable pH level is between 6.0 and 8.5
Dissolved Oxygen: Levels should be kept between 4.0 and 10.0 mg/L,
Turbidity: Keeping turbidity within 0.0 to 5.0 NTU helps ensure clear water.

Real-world data was gathered from a USGS station, which provided detailed records of the same water quality parameters (temperature, turbidity, pH, and dissolved oxygen). This data was used to calculate a water quality index, helping to classify the overall quality of the water. The *E. coli* prediction model worked well with this real-world data, confirming that our dataset is effective for predicting *E. coli* levels.

Hyperparameters are critical for tuning the performance of machine learning models. By adjusting these parameters, we can optimize the model's performance for specific datasets and tasks

Algorithm	Hyperparameters	Values
SVR	kernel	linear
	regularization parameter (C)	1
	epsilon	0.1
RFC	random_state	42
	class_weight	{0: 1, 1: 5}
	n_estimators	100
	max_depth	10
XGBClassifier	random_state	42

Table 1: Hyper parameters and their values

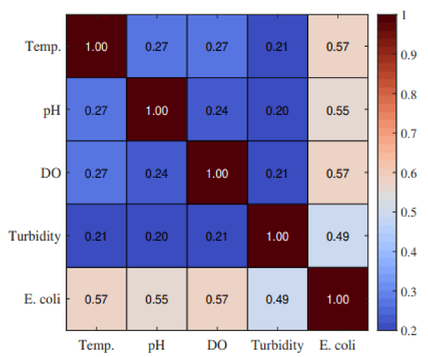

Fig 2: Correlation Matrix of Dataset

From Correlation Matrix, *E.coli* level presence strongly correlates with water temperature and dissolved oxygen. The pH also strongly correlates with *E.coli* level presence, while turbidity has the l

Mobile App for Real-Time Monitoring

The app uses Firebase Authentication for secure user access, Firebase Realtime Database for displaying live sensor data on water quality parameters, and real-time updates without user refresh. It sends alerts if parameters exceed safety thresholds, ensuring timely notifications. Historical data access and graphs enable trend analysis, while an intuitive dashboard and parameter-specific buttons enhance usability for quick and easy monitoring.

Results

The results demonstrate that the ensemble model combining Random Forest Classification (RFC) and XGBoost achieved the highest accuracy at **99.9%**, with a perfect precision of **1.000**, recall of **0.998**, and F1 score of **0.998**. This ensemble outperformed individual models like Support Vector Regression (SVR), RFC, and XGBoost alone. Although the ensemble model showed the highest accuracy, it had a slightly longer execution time than individual algorithms.

The study also generated a correlation matrix for the dataset, showing that E. coli levels strongly correlate with water temperature and dissolved oxygen, with lower correlation to turbidity.

Model	Accuracy	Precision	Recall	F1	Time
SVR	97.65	0.992	0.960	0.976	8.123
RFC	99.7	1.000	0.994	0.997	5.406
XGB	99.8	1.000	0.996	0.998	2.286
SVR + RFC	99.2	1.000	0.984	0.992	14.83
SVR + XGB	99.65	1.000	0.993	0.996	9.443
RFC + XGB	99.9	1.000	0.998	0.998	7.106

Table 2 shows Algorithm performance evaluation results, summarizing accuracy, precision, recall, F1 score, and execution time for each model

Conclusion & Future Scope

The proposed AI/ML-based system provides an effective, low-cost solution for real-time monitoring of *E. coli* contamination in surface waters. By using a combination of LoRaWAN and machine learning algorithms, including an ensemble of RFC and XGBoost, the system achieves high accuracy and minimizes false negatives, ensuring reliable detection of unsafe *E. coli* levels.

The mobile application allows users to access real-time water quality information and receive alerts, contributing to safer recreational water use and improved public health monitoring.

Future work can include collecting a larger volume of real-world data to further refine and validate the model's accuracy and reliability over longer periods. In addition to this,, comparing the system's predictions with laboratory-based results over extended durations to enhance its credibility.

Contact Information

Email ID of Author: kkaruman@students.kennesaw.edu
Email ID of Supervisor: alee146@kennesaw.edu

References

[1] O'Flaherty, E., Solimini, A., Pantanella, F., and Cummins, E. 2019. "The Potential Human Exposure to Antibiotic Resistant-Escherichia coli Through Recreational Water."
[2] Persson, S., Olsen, K., Scheutz, F., Krogfelt, K., and Gerner-Smidt, P. 2007. "A Method for Fast and Simple Detection of Major Diarrhoeagenic Escherichia coli in the Routine Diagnostic Laboratory."
[3] Jabbar, W. A., Mei Ting, T., Hamidun, M. F. I., Kamarudin, A. H. C., Wu, W., Sultan, J., Alsewari, A. A., and Ali, M. A. 2024. "Development of LoRaWAN-Based IoT System for Water Quality Monitoring in Rural Areas."
[4] Weller, D. L., Love, T. M., and Wiedmann, M. 2021. "Interpretability Versus Accuracy: A Comparison of Machine Learning Models Built Using Different Algorithms, Performance Measures, and Features to Predict E. coli Levels in Agricultural Water."
[5] Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., and García-Nieto, J. 2019. "Efficient Water Quality Prediction Using Supervised Machine Learning."

Our research work got approved for **IMCOM 2025**

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Author: Keerthi Priya Karumanchi kkaruman@students.kennesaw.edu
Advisor: Dr. Ahyoung Lee alee146@kennesaw.edu