URBAN WATER QUALITY MONITORING AND ANALYSIS USING EVENT DETECTION SYSTEM AND MACHINE LEARNING METHODS

by

Nabila Nafsin

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

May 2022

ABSTRACT

URBAN WATER QUALITY MONITORIING AND ANALYSIS USING EVENT DETECTION SYSTEM AND MACHINE LEARNING METHODS

by

Nabila Nafsin

The University of Wisconsin-Milwaukee, 2022
Under the Supervision of Professor Jin Li

Water quality is defined as the measure of physical, chemical, and biological characteristics of water. Monitoring water quality is a growing challenge because of accidental or intentional spills of industrial, domestic, and agricultural wastes into surface water. Conventional methods used for measuring water quality parameters are time-consuming and expensive, making real-time contamination detection difficult. Advanced monitoring technology can be employed for real-time monitoring, providing a reliable and cost-effective solution to water management. CANARY event detection system (EDS) has been used in water distribution networks and wastewater treatment plants for detecting anomalous water quality events and proved to be an effective alternative to manual laboratory analysis. This dissertation is directed towards analyzing different methods for real-time water quality monitoring, identifying quality trends, and predicting water quality using CANARY and machine learning (ML) techniques. The research provides an insight into the effectiveness of CANARY and ML algorithms for surface water quality monitoring. Considering the effectiveness of CANARY in real-time contamination event detection, this study evaluated the application of the EDS to river water quality analysis and beach bacterial contamination monitoring. For more efficient water quality data management and pollution control, ML models have been developed for water quality monitoring and

prediction of different water quality variables, including biochemical oxygen demand (BOD_5), total organic carbon (TOC), and *Escherichia coli* (*E. coli)* bacteria. The significance of this dissertation is the first successful application of CANARY to natural source water and the development of novel ensemble-hybrid ML models in predicting surface water quality.

To

my parents

and

especially my niece Adrita

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ANN	Artificial Neural Network
BED	Binomial Event Discriminator
BOD	Biochemical Oxygen Demand
DI	Deionized Water
EDS	Event Detection System
EWS	Early Warning System
FIB	Fecal Indicator Bacteria
GBM	Gradient Boosting Machine
LPCF	Linear Prediction Correction Filter
ML	Machine Learning
MLP	Multilayer Perceptron
MPN	Most Probable Number
MSE	Mean Squared Error
MVNN	Multivariate Nearest Neighbor
PBS	Phosphate Buffered Saline
RF	Random Forest
SCADA	Supervisory Control and Data Acquisition
SVM	Support Vector Machine
TOC	Total Organic Carbon

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude to my advisor Prof. Dr. Jin Li, for the continuous support of my Ph.D. study and research and for her patience, motivation, and immense knowledge. Her guidance helped me in all the time of research and writing of this thesis. I am grateful to her for allowing me to be a part of her research group and for having faith in me over the years during my Ph.D. Besides my advisor, I would like to thank my Ph.D. thesis committee members: Prof. Dr. Qian Liao, Prof. Dr. Yin Wang, Prof. Dr. Tian Zhao, and Prof. Dr. Shangping Xu, for their insightful comments, suggestions, and encouragement which incented me to widen my research from various perspectives.

I thank my fellow labmates for the stimulating discussions on the research projects we worked on together. Also, I thank my friends at the University of Wisconsin-Milwaukee. Thanks to the University of Wisconsin-Milwaukee for providing me with the financial support to pursue my Ph.D. Also, I would like to thank Mandy Bauer, administrative assistant, and Alicia Lopez, support staff of the university support services in the College of Engineering and Applied Science, for providing the necessary technical and administrative assistance throughout my doctorate study.

Last but not the least, I would like to thank my family: my husband, my parents, and my sister for constantly supporting me throughout my Ph.D. and my life in general.

CHAPTER 1: INTRODUCTION

Rapid urbanization and industrial development result in the degradation of water quality of natural source water at an alarming rate. Conventional methods for water quality assessment, including manual sampling and laboratory analysis are time-consuming, expensive, and inefficient. For sustainable water management, it is imperative to monitor and detect pollutants causing contamination in water. Water pollution is classified into several categories, including toxic pollution, organic pollution, nutrient pollution, microbial contamination, sediment pollution, and radiological pollution. The quality of water is generally affected by several physical, chemical, and biological parameters. However, there is no single parameter that can define water quality completely. To ensure the supply of good quality water, treatment of wastewater, and maintaining aquatic life ecosystem, it is crucial to monitor different water quality parameters, such as biochemical oxygen demand (BOD_5), chemical oxygen demand (COD), dissolved oxygen (DO), total organic carbon (TOC), turbidity, bacteria level, etc. Analyses of water quality parameters require expensive lab testing and the results obtained by lab analysis are often inaccurate due to the inconsistencies between the lab environment and the exact field condition during sampling. As of today, advanced monitoring technology, such as early warning event detection systems (EDS) and artificial intelligence (AI)-based machine learning (ML) methods have been employed as an alternative to conventional water quality assessment methods. This research is directed towards analyzing different methods for real-time water quality monitoring, identifying quality trends, and predicting water quality using CANARY EDS and ML techniques.

1.1 Water quality event detection system (EDS)

Safety of drinking water has recently generated considerable interest because of the credible concern that water can be contaminated by chemical, biological, and radiological pollutants. Rapid and early detection of these pollution events is necessary to determine appropriate changes in water treatment and to ensure the safety of water supplied to the consumers. Early warning monitoring system provides timely information on changes in source water quality so that knowledgeable response decisions can be made. Early Warning Systems (EWS) are used mostly on riverine systems where water quality can change more rapidly, but less frequently used for impoundments and rarely used for groundwater. EWSs have been developed to detect a wide range of natural and man-made contamination incidents and generate information that allows decision-makers to take action against the contamination event to mitigate human health risks and other hazardous outcomes.

Water infrastructure of the United States gained increased awareness after the events of September 11, 2001. The U.S. Environmental Protection Agency (EPA), as the lead federal agency for water security, helped develop tools, procedures, and documentation to support water utilities and other agencies in protecting water supply. A primary focus of this effort has been to develop and demonstrate components of drinking water contamination warning systems monitoring and surveillance systems that can detect contamination in time to allow for mitigation of human health and economic consequence. In conjunction with the U.S. EPA, Sandia National Laboratories has developed the CANARY EDS software. CANARY is proposed as a freely available event detection tool to water utilities and researchers to better understand the normal background variability and to identify anomalies that are potentially indicative of contamination incidents. CANARY supports several open-source components, including online water quality

monitoring and contamination event detection. The advantages of CANARY over other EDS include Representational State Transfer (REST) web service friendly, transparency of algorithms, direct integration of operational data into event detection components, centralized processing capability on a single computing system, and supporting multiple sensors' operation from different manufacturers.

1.2 Artificial Intelligence (AI)-based water quality monitoring

Most early warning event detection systems are based on statistical approaches that mainly focus on the fitting of a specific probability model, which does not rely on predictive accuracy. On the contrary, AI-based ML methods focus on finding a generalized predictive pattern with learning algorithms. ML algorithms deal with data with a large number of observations and access and learn from each observation over time to predict the exact combination of actions as well as achieve high accuracy of the model result. For more efficient water quality data management and contamination detection system, AI-based machine learning (ML) methods have been developed. Based on the type of model to be processed and the dataset available, several ML algorithms are used, such as Artificial Neural Network (ANN), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), Random Forest (RF) for water quality monitoring and prediction of significant changes in water quality variables. Besides prediction tasks, ML models are also developed for classification problems, where the algorithm predicts a categorical label (or a discrete value). In the field of hydro-environment research, AI/ML techniques have been successfully employed to detect leaks in water distribution networks, forecasting municipal and agricultural water demand, managing energy consumption in water systems, detecting changes or anomalous water quality events, and predicting water quality

parameters in natural source water. Implementation of AI technology in water monitoring programs supports decisions that lead to more efficient, cost-effective, and sustainable urban water management.

1.3 Objective of the dissertation

The main goal of the dissertation is to analyze physical, chemical, and biological characteristics of urban water (e.g., lake, river) using advanced water quality monitoring technologies, including CANARY EDS and AI-based machine learning (ML) approaches. The objectives can be documented precisely as follows.

Firstly, a review of the increasing implementation of smart water sensor networks and AI technology in urban water systems and recommending integration of sensor networks with advanced data analysis methods such as AI/ML for efficient and reliable urban water management.

Secondly, monitoring and evaluating the concentration and interaction of fecal indicator bacteria (FIB) in beach sand and water at Bradford Beach along Lake Michigan, analyzing the effectiveness of different types of eluents, i.e., DI water and phosphate buffered saline (PBS) to enumerate bacteria from beach sand, and evaluating the effect of the presence of algae on bacteria concentration. This study also suggested that CANARY EDS can potentially be a useful statistical tool and an early warning system for monitoring FIB concentration that can provide timely information on beach contamination and public health outcomes associated with pathogens.

Thirdly, analysis of anomalous water quality events for the Milwaukee River using CANARY event detection system based on the available water quality monitoring data and evaluation of the effectiveness of CANARY in monitoring water quality of natural source water.

This study is the first to apply CANARY software successfully for river water quality monitoring.

Fourthly, implementation and development of novel ensemble-hybrid AI/ML models for predicting biochemical oxygen demand (BOD_5) in a highly polluted river system of a specific region of Bangladesh. This study provides a comprehensive assessment of the employed standalone and ensemble-hybrid ML models for evaluating their ability to predict BOD_5. Identification of the most influential water quality parameters in predicting BOD_5 of the river system is one of the objectives of the study. The purpose of this study is to provide an efficient water data management system using AI/ML techniques and develop improved pollution control strategies to protect natural source water from domestic and industrial pollutants in developing countries such as Bangladesh.

Finally, analysis of organic matter pollution and microbial contamination in the Milwaukee River, Menomonee River, and Kinnickinnic River of the Milwaukee River basin in Wisconsin using ML methods and evaluation of different regression-based ML models' performances to find the most efficient prediction model for total organic carbon (TOC) and *E. coli* bacteria. The developed *E. coli* prediction models explained the variability in living microorganisms' behavior based on the specific physicochemical parameters. The significance of the study is the application of the developed ensemble-hybrid methods for TOC and *E. coli* prediction that can provide a reliable and direct approach to complement existing monitoring techniques in the Milwaukee River system with satisfactory prediction accuracies.

CHAPTER 2: REVIEW OF SMART SENSOR NETWORK AND MACHINE LEARNING TECHNIQUES IN URBAN WATER SYSTEMS

2.1 Introduction

Several contamination events, including accidental and intentional spills of industrial pollutants, municipal and agricultural wastes discharging into natural source water affect the quality of urban water sources. For sustainable urban water management, continuous monitoring of different physical, chemical, and microbial parameters is essential. Generally, water samples are collected by manual field/grab sampling, and laboratory analyses are performed using conventional methods, which are time-consuming, costly, and labor-intensive. Also, analysis of water samples in a laboratory-based environment may not represent the exact field condition, especially water temperature and flow velocity. Therefore, advanced water monitoring strategies are required for prompt responses to accidental events and improved water resource management. Advanced monitoring technologies such as smart sensor networks, information and communication technology (ICT), and artificial intelligence (AI) methods are increasingly used for real-time monitoring of water quality and quantity parameters in the field of drinking water supply, wastewater treatment plants, and natural source water. These technologies can significantly address critical urban water challenges like drought, water quality, reliance on imported water, and population growth. The potentiality of online sensor networks in real-time monitoring of water quality and quantity parameters from multiple locations and remote operation with low power consumption benefits the decision-makers in improving water management and pollution control strategies for urban water systems. The large amount of data collected by online sensor networks provides valuable information and speeds up the deployment of data-driven approaches in urban water management. This chapter overviews the recent

advances and application of smart water monitoring technologies in urban water systems, providing reliable measurements and cost-effective tools for water data management which can support decision-makers. Based on the papers reviewed, we observed the increasing implementation of advanced monitoring technology in areas such as water distribution networks, natural source water, water treatment plants, pipe infrastructure, filtration efficacy in treatment processes, and early flood warnings.

2.2 Urbanization and stream ecosystem

Urbanization is one of the significant causes for environmental change, particularly the urban stream ecosystem. Besides affecting the physical, chemical, and biological characteristics of local streams, urban development processes, such as the conversion of rural land, replacement of natural vegetation and topsoil by an impervious cover, protective structure along streambanks change the water-drainage pattern resulting in storm runoff and flooding. The modified drainage pattern allows a high volume of water with pollution sources, including sediments, fertilizers, and nutrients flowing to the stream and affects the stream hydrology, stream habitat, and stream chemistry. (Coles et al., 2012).

2.3 Smart water technology

For sustainable water management, smart water is potentially a solution to the urban water issues that ensure water efficiency, energy efficiency, and water quality improvements by providing a more effective technology use in decision-making. Smart water technologies are beneficial to water distribution systems for inspection and identification of leakage and corrosion in pipes, measuring water quality parameters to improve treatment processes, and detection of

abnormalities or contamination events. Online sensor monitoring networks integrated with AI technology are increasingly used for real-time monitoring and forecasting of water quality and quantity parameters in water distribution systems, water treatment plants, and natural source water. The recent advancement in online sensor networks and AI technology ensures reliable measurements and cost-effective management of large environmental data and provides data-driven information for water quality management. Figure 2.1 and Figure 2.2 show smart water monitoring system setups that combine real-time monitoring using online sensors, data collection, transmission, and data analysis using advanced technology.

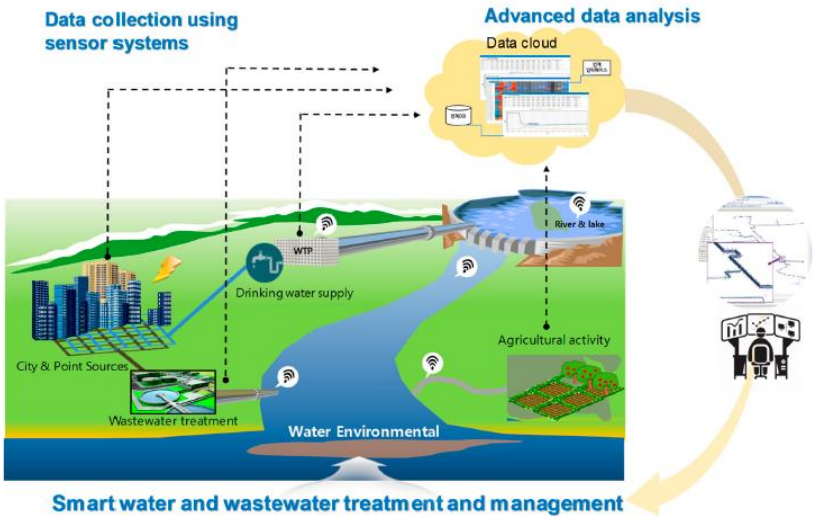


Figure 2.1. Smart water monitoring system combining real-time monitoring, data collection, transmission, and data analysis using advanced technology [Park et al., 2020]

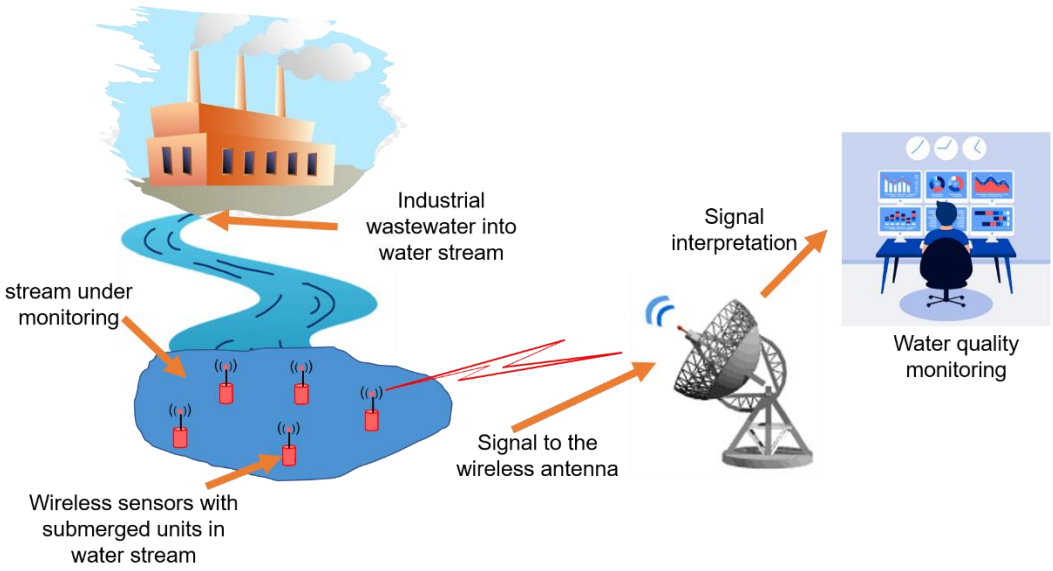


Figure 2.2. Water quality monitoring system setup with wireless sensor network

2.4 Sensor-network based monitoring

A smart sensor network consists of a network of distributed sensors and sensor nodes to acquire data and relay it to a central monitoring station. The sensors can be wired, wireless, or a combination of both, depending on the application and the environment where it is deployed. Generally, a smart sensor network is composed of two primary networks: data acquisition network and data distribution network. In data acquisition networks, sensor nodes collect data, process data, and deliver it to a base station for data processing, analysis, and data management. Sensor information is then transmitted to the data distribution network, allowing remote-controlled operation.

Online sensors are used for monitoring physical, chemical, and biological characteristics of water, including pH, electrical conductivity, turbidity, temperature, total organic carbon (TOC), free chlorine, suspended solids (SS), nutrients, such as nitrogen and phosphorus, representing the degree of contamination of water. Wireless sensors collect and store field data and transmit the

data to a controller or SCADA system. The collected data are then transmitted from the controller system to the data storage cloud, which is finally used for data analysis using advanced technology such as AI. Figure 2.3 shows a smart sensor network integrated with advanced monitoring technology (AI/ML). Several physical, chemical, remote optical, and commercially available real-time monitoring sensors are used for water quality management in the field of drinking water supply systems, lake, and river management, and water resource distributions (Park et al., 2020). Table 2.1 provides an overview of sensing technology for monitoring different water quality/quantity parameters in urban water systems.

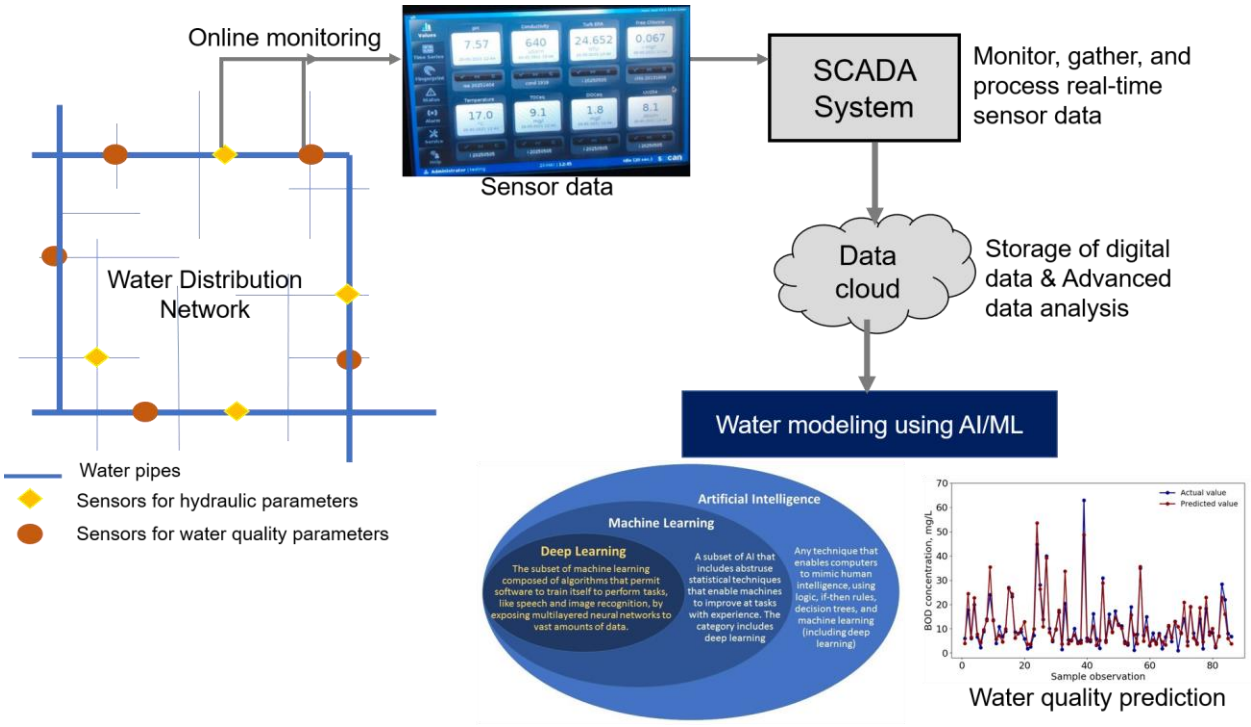


Figure 2.3. Sensor network integrated with Artificial Intelligence/ML

Table 2.1. Wireless sensing technology for urban water monitoring

Parameter	Sensor type	References
Electrical conductivity, Turbidity, Temperature, pH, DO, Oxidation-reduction potential (ORP)	Two electrode conductivity cells, nephelometric (light scattering and absorbing), temperature measuring channels (thermistor), membrane electrode, potentiometric, optical sensor	Postolache et al., 2014; Ritter et al., 2014
COD	Electrochemical sensor	Hassan et al., 2018
Nitrate, phosphate, ammonia	Optical sensor, colorimetric	Pellerin et al., 2014; 2016
Harmful algal bloom (HAB), Chl-a	In situ optical sensor	Adamo et al., 2014
Water flow	Hall effect, doppler effect	Alves et al., 2019
Water pressure, vibrations	Piezoelectric, electromagnetic, capacitive, strain-gauge, accelerometers	Alves et al., 2019
Velocity	Acoustic doppler velocimetry (velocity sensor)	Chanson, 2008
Water level	In situ acoustic sensor	Boon & Brubaker, 2008

2.5 Emerging technology in online sensing

2.5.1 UV-vis spectrometer

Spectrophotometer probes can measure various water quality parameters continuously and directly in the water. The dual light beams, one that passes through the sample solution and the other used as a reference beam ensures long-term stability of the signal produced. In the measuring beam section, which is positioned between emitting and receiving units, the emitted light passes through the medium to be analyzed. Substances present in the medium located in between the measuring windows of the probe adsorb visible and UV light. Internally a second light beam is guided across a comparison pathway with distilled water. This two-beam setup

makes it possible to compensate, with each single measurement, any instrumental effects that could influence the quality of the measurement (e.g., ageing of the light source). The spectrophotometer probe records the complete absorbance spectrum between 190-720 nm (UV-vis) or 190-390 nm (UV) resolving it into 256 wavelengths- resulted in as “fingerprints” (absorbance spectrum) (Figure 2.4). The information in the fingerprints is used to monitor multiple water quality parameters simultaneously. The operating electronics contained in the probe control the entire measuring process and check the measuring signal and calculate the fingerprints and parameters values.

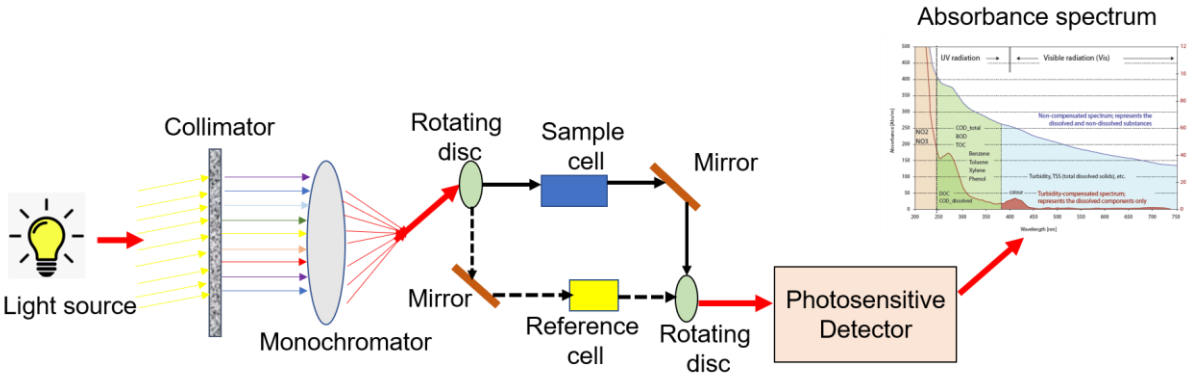


Figure 2.4. Measuring principle of UV-Vis Spectrophotometer

2.5.2 Fluorescence Spectroscopy

Fluorescence spectroscopy analyzes fluorescence from a molecule based on its fluorescent properties. Fluorescence is a type of luminescence caused by photons exciting a molecule, raising it to an electronic excited state. A beam of light passing through an extraction filter excites the electrons in molecules of certain compounds and causes them to emit light. The emitted light beam is directed towards a filter and onto a photosensitive detector for measurement and identification of the molecule or changes in the molecule (Figure 2.5).

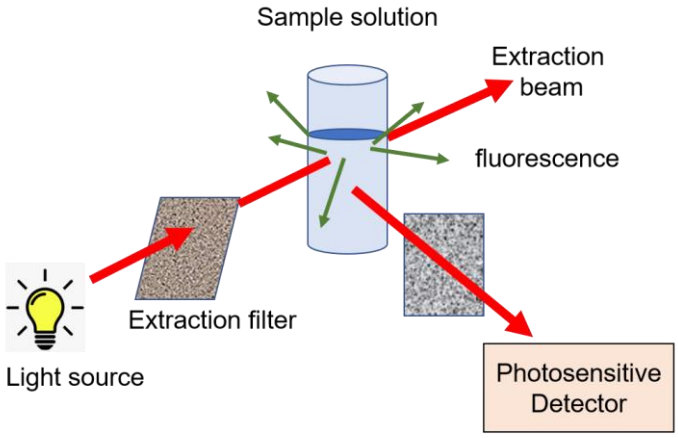


Figure 2.5. Measuring principle of Fluorescence spectroscopy

2.5.3 Remote sensing

Remote sensing technology includes the detection and monitoring of physical characteristics of an area by measuring the reflected and emitted radiation from a satellite or aircraft at a distance (Figure 2.6). In the field of water resource management, the use of remote sensing is advantageous over the conventional approaches used for water quality assessment. Water quality data with high temporal and spatial resolution from multiple surface water sources at a time can be achieved. Also, the advanced technology supports the evaluation of environmental issues and potential health risks by analyzing the changes in water quality and detecting harmful algal blooms. The satellite retrieved data are used for the study of water quality trend analysis and potential impacts of land use and land cover change on water quality. The remote sensing activities including the integration of real-time satellite data into early warning systems can be particularly beneficial to urban water management.

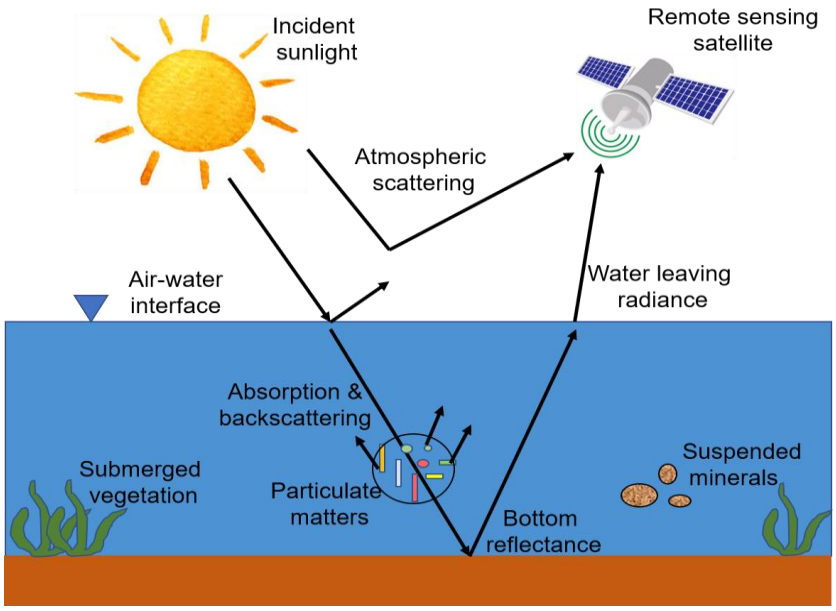


Figure 2.6. Remote sensing technology

2.5.4 Biomonitoring

Biomonitoring includes the measurement of the effects of pollutants and toxic compounds on aquatic organisms from bacteria to fish. Aquatic biomonitor continuously monitors water by detecting changes in fish ventilatory or breathing patterns and movement patterns exposed to the water of interest. In biomonitoring technology, electrical signals generated by muscle movements of individual fish are monitored by carbon block electrodes that suspend above and below each fish. The electrical signals are amplified, filtered, and transferred to a computer for analysis. Besides, fish ventilatory data, other water quality parameters such as pH, conductivity, dissolved oxygen, temperature are monitored using water quality sensors to determine if the fish responses are due to the toxicant or non-toxic water quality variations. Biomonitoring must lead to an integrated strategy for surveillance, early warning, and control of the freshwater ecosystem, which will be able to respond to the different impacts in time and space. Figure 2.7 shows a commercially available biomonitoring technology that allows early warning for chemical contaminants in water.

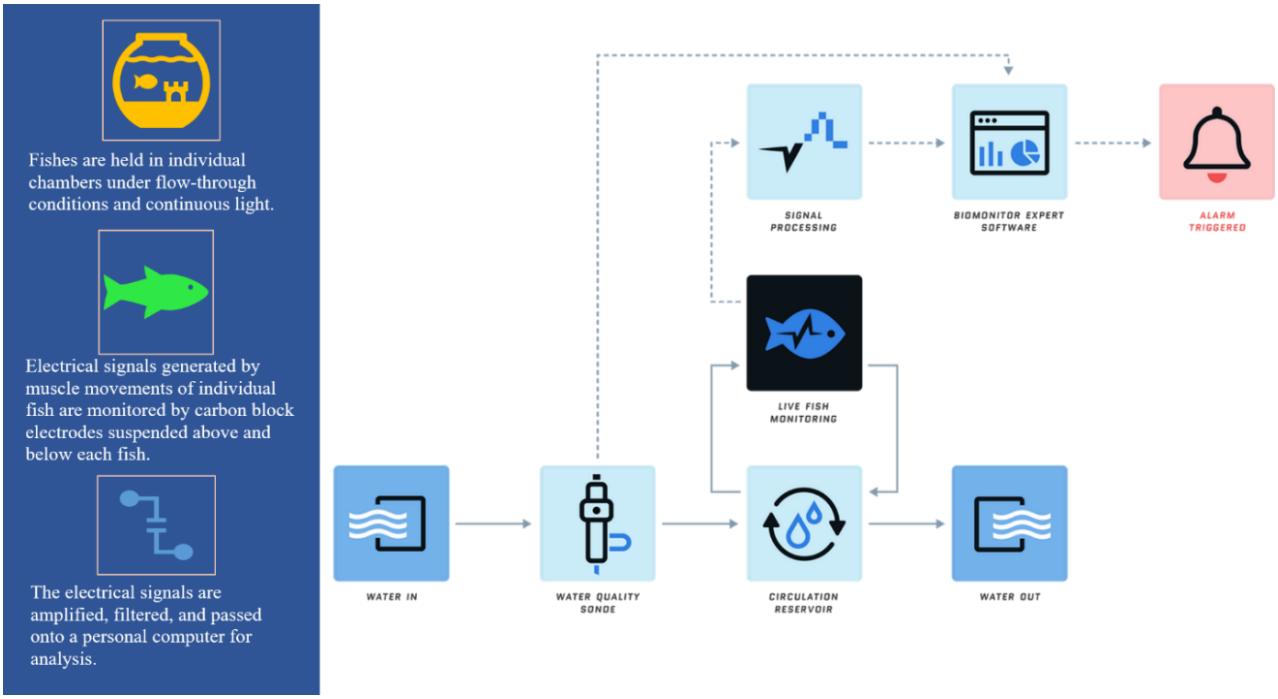


Figure 2.7. Biomonitoring technology (Source: Blue Sources: Early warning for chemical contamination of water. https://www.bluesources.com/#service)

In addition, there are commercially available real-time monitoring sensors that can measure water quality parameters at relatively low costs (Park et al., 2020). An example of a commercially available sensor is pipe::scan manufactured by S::can (Figure 2.8). The pipe::scan is a modular sensor system for monitoring drinking water quality in pipes under pressure. The flow-cell monitoring system can measure up to 10 parameters with a measuring interval of 1 minute (s::can). Key parameters such as Chlorine, Turbidity, and pH amongst many others such as Conductivity, Color, TOC (Total organic Carbon), DOC (Dissolved Organic carbon), UV254, Temperature, Pressure can be tracked as water flows through the network to the end-users allowing the utilities to make proactive decisions with regards to the treatment.

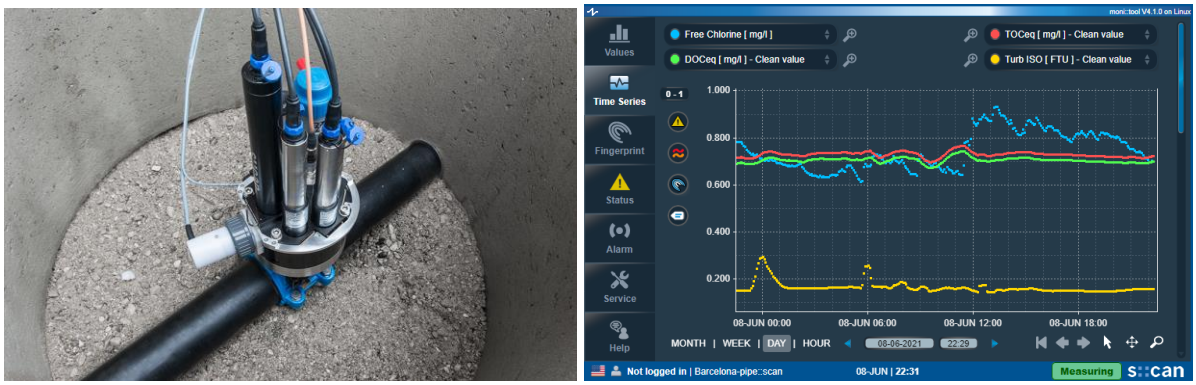


Figure 2.8. Pipe::scan system (manufactured by S::can)

2.6 AI technology in urban water systems

AI-based machine learning (ML) techniques have been widely used in hydro-environment research, providing a potential tool for urban water management. AI/ML models are successfully employed in detecting leaks in water distribution networks, modeling rainfall-runoff transformation, flood prediction, forecasting urban water demand and managing energy consumption in water systems, predicting different water quality parameters in natural source water (rivers and lakes), and identifying anomalies or events in distribution systems. Water planning and management applications such as early flood warnings require the ability to extract data-driven information from a large dataset in real-time. Such environmental monitoring programs can benefit from the potentiality of ML techniques in process optimization and data-driven decision-making (Sun and Scanlon, 2019). Table 2.2 provides an overview of the application of AI/ML technology to urban water systems.

Table 2.2. Application of ML techniques to urban water systems

ML model	Model prediction/ Output	Input variables	Study field	Performance evaluation metrices	References
ANN, SVM, logistic regression, DNN, RNN, LSTM, LDA	Event detection	Chlorine dioxide, pH, conductivity, ORP, turbidity	Water distribution system	Best performing model SVM: F1-score= 0.9891	Muharemi et al., 2019
ANN	Event detection	Total chlorine, electrical conductivity (EC), pH, temperature, total organic carbon (TOC), and turbidity	Water distribution system	Correlation coefficient (R^2: 0.225-0.914); mean squared error (MSE: 0.016-269.119); receiver operating characteristic (ROC:0.549-0.791); and true and false positive rates (TPR:0.085-0.587 and FPR:0.001-0.093)	Perelman et al., 2012
ANN, SVM	Chemical and bio-contamination detection (anomaly detection)	Free chlorine, pH, alkalinity, TOC	Water distribution system	Confusion matrix (classification loss about 4%)	Tinelli & Juran, 2019
Feed-forward back propagation ANN	DO	Flow, temperature, pH, conductivity, BOD	River water	R^2: 0.928, RMSE: 1.52	Sarker & Pandey, 2015
ANN (MLP and RBF)	DO, BOD, COD	EC, pH, Ca, Mg, Na, Turbidity, PO_4, NO_3 and NO_2	River water	**DO:** R^2 = 0.85, RMSE=3.15, MAE=2.7 **BOD:** R^2 = 0.96, RMSE=2.57, MAE=2.05 **COD:** R^2 = 0.94, RMSE=1.99, MAE=1.77	Emamgholizade et al, 2013
CEEMDANCSA, MARS, M5Tree, CEEMDAN-CSA-MARS	TOC	pH, EC, DO, temperature, COD, SS	River	Best performing CEEMDAN-MARS-CSA: R^2 = 0.900, RMSE=0.675, NSE=0.680	Kim et al, 2021

Multiple Linear Regression (MLR)	*E. coli*, Enterococci, Somatic coliphages	Temperature, pH, EC, turbidity, DO, chlorophyll a, filterable solids (FS), nitrate-nitrogen, ammonia-nitrogen, phosphate phosphorous, discharge, rainfall, solar irradiance	River	*E. coli*: R^2 = 0.62- 0.73, Enterococci: R^2 = 0.67- 0.64, Coliphages: R^2 = 0.71- 0.76	Herrig et al., 2015
RF, M5P, RT, REPT, hybrid: combination of standalone with bagging, CVPS, RFC	WQI (Water quality index)	BOD, COD, DO, TS, pH, conductivity, turbidity, Fecal coliform, PO_4, NO_3, NH_3-N	Rivers	Best performing hybrid BA-RT: R^2 = 0.941, RMSE=2.71, MAE=1.87, NSE=0.941, PBIAS=0.500	Bui et al., 2020
DT, NB, LR, LDA, CRT, KNN, SVM, RF, CRF, DCF	WQC (Water quality class)	pH, DO, COD, NH_3-N	Rivers and lakes	DCF: Precision = 1-0.91, Recall=0.95-0.99, F1-score=0.94-0.99	Chen et al., 2020
ANN-MLP	Turbidity	Raw water turbidity, water flow, water retention level, daily rainfall & reservoir temperature	Water treatment plant	R = 0.84, RMSE = 0.49	Rak, 2013
Genetic algorithm (GA), ANN-MLP	Pipe pressure	Storage tank levels, water demand	Pipe infrastructure	Best performing model ANN: RMSE = 0.11	Nazif et al., 2010
ANN	Leakage detection	Flow and pressure	Pipe infrastructure	True positive rate (% of correct leak detection = 75%)	Mounce et al., 2006
MLR, Time series analysis, ANN	Peak daily water demand	Water demand, maximum daily temperature, rainfall	Municipal water demand	Best performing ANN: R^2 = 0.69, AARE=12, MARE=41,	Adamowski, 2008
Neural networks: GRNN,	Monthly water use	Average monthly water bill, population, number of	Municipal water demand	Best performing GRNN: R^2 = 0.933, NRMSE=0.068,	Firat et al., 2009

FFNN, RBNN		households, gross national product, monthly average temperature, monthly total rainfall, monthly average humidity, inflation rate		Efficiency E=0.889	
Radial basis function neural network (RBFNN), BPNN, MR	Permeate flux decline	Particle radius, solution pH, transmembrane pressure, filtration time, ionic strength	Membrane filtration	Best performing RBFNN: R^2 = 0.988, RMSE=0.082	Chen & Kim, 2006
ANN-MLP	Flux decline	Transmembrane pressure, cross-flow velocity, operating time, dynamic fouling	Membrane filtration	R^2 = 0.99	Corbaton et al., 2016
Deep CNN, ANN, SVM	Water flow, water level	Max-temperature, Min-temperature, runoff, water flow	River (water quantity)	**Water flow:** R^2 = 0.99, RMSE=3.46, MAE=1.43 **Water level:** R^2 = 0.99, RMSE=0.06, MAE=0.03 (Deep CNN)	Assem et al., 2017
KNN, SVM, Naïve-Bayes, DNN	Flood prediction	Temperature, rainfall intensity	Early flood warning	Best performing model DNN: Precision =0.95, Recall =0.93, Accuracy = 91.18%, F1 score=0.95, MCC = 0.64	Sankaranarayanan et al., 2020
RBFNN-FA, hybrid SVM-FA	Flood discharge	Monthly river flow	Early flood warning	Best performing SVM-FA: R^2 = 0.9818, RMSE=0.0306, MSE=0.0079	Sahoo et al., 2021

2.7 Conclusion

AI/ML techniques are very promising to continuously monitor water in urban systems since they can operate based on optical readings rather than laboratory analysis. A combination of smart sensors network and AI can prevent disruption of water supply in a city by taking care of a problem before it occurs. However, AI has raised concern among people due to fear of unemployment. Such fear can negatively impact funding for automation as people can influence the various levels of stakeholders. It is a matter of time to adapt and adjust to the benefits of AI. Instead of replacing the workforce, AI can offer faster decision-making capabilities that can be managed by skilled manpower.

2.8 References

1. Adamo, F., Attivissimo, F., Carducci, C. G. C., & Lanzolla, A. M. L. (2014). A smart sensor network for sea water quality monitoring. *IEEE Sensors Journal*, *15*(5), 2514-2522.
2. Adamowski, J. F. (2008). Peak daily water demand forecast modeling using artificial neural networks. Journal of Water Resources Planning and Management, 134(2), 119-128.
3. Alves, A. J. R., Manera, L. T., & Campos, M. V. (2019). Low-cost wireless sensor network applied to real-time monitoring and control of water consumption in residences. Revista Ambiente & Água, 14.
4. Assem, H., Ghariba, S., Makrai, G., Johnston, P., Gill, L., & Pilla, F. (2017, September). Urban water flow and water level prediction based on deep learning. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 317-329). Springer, Cham.

5. Boon, J.D., Brubaker, J.M. (2008). Acoustic-microwave water level sensor comparisons in an estuarine environment. In Proceedings of the OCEANS, Quebec City, QC, Canada, 15–18 September 2008; pp. 1.
6. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. Science of The Total Environment, 721, 137612.
7. Chanson, H. (2008). Acoustic Doppler Velocimetry (ADV) in the Field and Laboratory: Practical Experiences. Proceedings of the International Meeting on Measurements and Hydraulics of Sewer, Brisbane, Austria, 19–25 August 2008, pp. 49–66.
8. Chen, H., & Kim, A. S. (2006). Prediction of permeate flux decline in crossflow membrane filtration of colloidal suspension: a radial basis function neural network approach. Desalination, 192(1-3), 415-428.
9. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., ... & Zhang, Y. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. Water Research, 171, 115454.
10. Coles, J. F., McMahon, G., Bell, A. H., Brown, L. R., Fitzpatrick, F. A., Eikenberry, B. S., ... & Stack, W. P. (2012). Effects of urban development on stream ecosystems in nine metropolitan study areas across the United States. US Geological Survey Circular, 1373, 152.
11. Corbatón-Báguena, M. J., Vincent-Vela, M. C., Gozálvez-Zafrilla, J. M., Álvarez-Blanco, S., Lora-García, J., & Catalán-Martínez, D. (2016). Comparison between artificial neural networks and Hermia's models to assess ultrafiltration performance. Separation and Purification Technology, 170, 434-444.

12. Emamgholizadeh, S., Kashi, H., Maroufpoor, E., & Zalaghi, E. (2013). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. International Journal of Environmental Science and Technology, 11. 645-656.
13. Firat, M., Yurdusev, M. A., & Turan, M. E. (2009). Evaluation of artificial neural network techniques for municipal water consumption modeling. Water resources management, 23(4), 617-632.
14. Hassan, H.H., Badr, I.H., Abdel-Fatah, H.T., Elfeky, E.M., Abdel-Aziz, A.M. (2018). Low-cost chemical oxygen demand sensor based on electrodeposited nano-copper film. Arab. J. Chem. 2018, 11, 171–180.
15. Herrig, I. M., Böer, S. I., Brennholt, N., & Manz, W. (2015). Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. Water research, 85, 148-157.
16. Kim, S., Maleki, N., Rezaie-Balf, M., Singh, V. P., Alizamir, M., Kim, N. W., ... & Kisi, O. (2021). Assessment of the total organic carbon employing the different nature-inspired approaches in the Nakdong River, South Korea. Environmental Monitoring and Assessment, 193(7), 1-22.
17. Mounce, S. R., & Machell, J. (2006). Burst detection using hydraulic data from water distribution systems with artificial neural networks. Urban Water Journal, 3(1), 21-31.
18. Muharemi, F., Logofătua, D., Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. Journal of information and Telecommunication, 3(10): 1-14.
19. Nazif, S., Karamouz, M., Tabesh, M., & Moridi, A. (2010). Pressure management model for urban water distribution networks. Water resources management, 24(3), 437-458.

20. Park, J., Kim, K. T., & Lee, W. H. (2020). Recent Advances in Information and Communications Technology (ICT) and Sensor Technology for Monitoring Water Quality. Water, 12(2), 510.
21. Pellerin, B.A.; Bergamaschi, B.A.; Gilliom, R.J.; Crawford, C.G.; Saraceno, J.; Frederick, C.P.; Downing, B.D.; Murphy, J.C. (2014). Mississippi River nitrate loads from high frequency sensor measurements and regression-based load estimation. Environmental Science & Technology, 48, 12612–12619.
22. Pellerin, B.A.; Stau_er, B.A.; Young, D.A.; Sullivan, D.J.; Bricker, S.B.;Walbridge, M.R.; Clyde, G.A.; Shaw, D.M. (2016). Emerging tools for continuous nutrient monitoring networks: Sensors advancing science and water resources protection. J. Am. Water Resource Association, 52, 993–1008.
23. Perelman, L., Arad, J., Housh, M., Ostfeld, A. (2012). Event detection in water distribution systems from multivariate water quality time series. Environmental Science & Technology, 46(15), 8212-8219.
24. Postolache, O., Pereira, J. D., & Girão, P. S. (2014). Wireless sensor network-based solution for environmental monitoring: Water quality assessment case study. IET Science, Measurement & Technology, 8(6), 610-616.
25. Rak, A. (2013). Water turbidity modelling during water treatment processes using artificial neural networks. International Journal of Water Sciences, 2.
26. Ritter, C., Cottingham, M., Leventhal, J., & Mickelson, A. (2014, October). Remote delay tolerant water quality montoring. In IEEE Global Humanitarian Technology Conference (GHTC 2014) (pp. 462-468). IEEE.

27. s::can Messtechnik Gmbh. (n.d.). Low cost water Quality Monitoring for Multiple Parameters. Retrieved from s::can: http://i-scan.at/index.php/versions/low-cost-online-water-quality-monitoring

28. Sahoo, A., Samantaray, S., & Ghose, D. K. (2021). Prediction of Flood in Barak River using Hybrid Machine Learning Approaches: A Case Study. Journal of the Geological Society of India, 97(2), 186-198.

29. Sankaranarayanan, S., Prabhakar, M., Satish, S., Jain, P., Ramprasad, A., & Krishnan, A. (2020). Flood prediction based on weather parameters using deep learning. Journal of Water and Climate Change, 11(4), 1766-1783.

30. Sarker, A., & Pandey, P. (2015). River Water Quality Modelling Using Artificial Neural Network Technique. Aquatic Procedia, 4, 1070-1077.

31. Sun, A. Y., & Scanlon, B. R. (2019). How can Big Data and machine learning benefit environment and water management: a survey of methods, applications, and future directions. Environmental Research Letters, 14(7), 073001.

32. Tinelli, S., & Juran, I. (2019). Artificial intelligence-based monitoring system of water quality parameters for early detection of non-specific bio-contamination in water distribution systems. Water Supply, 19(6), 1785-1792.

CHAPTER 3: MONITORING *E. COLI* AND ENTEROCOCCI BACTERIA IN LAKE MICHIGAN BEACH SAND

3.1 Introduction

Indicator bacteria are used as surrogates for disease-causing microorganisms because pathogenic microbes are present at very low concentrations in water and direct tests are expensive. There is a direct positive correlation between the concentration of fecal indicator bacteria such as *Escherichia coli* (*E. coli)* and Enterococci and the occurrence of gastrointestinal illness in humans (Wade et al., 2003). Epidemiological studies indicate that illness in swimmers can be quantitatively linked to indicator bacteria in coastal waters contaminated with wastewater and urban runoff (Yamahara et al., 2009). The coliform bacteria *E. coli* and Enterococci from the bacterial group streptococci are found in the feces of humans and other animals. These bacteria can synthesize ATP under aerobic or anaerobic conditions, allowing them to survive in various environments. Other commonly tested fecal indicator bacteria are fecal coliform, total coliform, and fecal streptococci. Previously, by using the ratio of fecal coliform to streptococci, the origin of the contamination (human or non-human) could be determined. However, this method is no longer recommended by the U.S. EPA due to reliability issues.

Guidelines developed by the U.S. EPA indicate that *E. coli* and Enterococci are appropriate indicator bacteria for monitoring recreational water (U.S. EPA, 2018). In marine systems, Enterococci play a crucial role in public health outcomes, and *E. coli* is considered a significant contributor to contamination in freshwater systems (Halliday and Gust, 2011). The presence of *E. coli* is carefully monitored in freshwater bodies in Wisconsin, where different advisory limits are declared for different threshold values of *E. coli* concentration. According to

the Wisconsin DNR, when the *E. coli* concentration at a beach is 235 CFU/100 mL, the health department posts a yellow caution sign. A red "Closed" sign is posted to indicate an elevated public health risk when the concentration level reaches 1000 CFU/100 mL. In 2012, the EPA recommended new statistical standards that incorporate threshold values and the geometric mean for indicator bacteria to determine the water quality for recreational access at the beach (U.S. EPA, 2018).

Conventionally, nearshore waters are tested for indicator bacteria to monitor recreational water quality. However, people with exposure to beach sand without contacting water were found to be affected by *E. coli*, which suggested that exposure to beach sand is a possible reason for the sickness (Alm et al., 2006; Bonilla et al., 2007; Byappanahalli et al., 2003; Ishii et al., 2006). Besides gastrointestinal diseases, other adverse health outcomes associated with beach sand and water include skin, respiratory, and ear infections. Yau et al. (2009) investigated skin-related health conditions due to recreational exposure to fecal indicator bacteria and reported a statistically significant correlation between bacteria levels in marine water and skin-related effects. Beach water can also be affected by cyanobacteria, which can produce cyanotoxins resulting in adverse human health conditions (U.S. EPA, 2012).

Previous studies have indicated that bacteria may be more viable in sand than in water due to easier attachment to sediment particles (Whitman and Nevers, 2003), the presence of sufficient nutrients and fewer predators in sediment (Thupaki et al., 2013), and a lower inactivation rate of bacteria due to less exposure to UV radiation. Additionally, fecal indicator bacteria could survive in sand without a recent source of human fecal contamination (Nevers et al., 2014). Re-suspension of sediments under the water column can significantly increase fecal contamination by releasing bacteria in submerged sediments near sewage outfalls. In response to

nutrient deprivation in submerged sediments, bacteria form endospores, producing highly resistant cells for the preservation of their genetic material. A comprehensive review by Vestby et al. (2020) indicates the potential of bacterial biofilms to cause diseases of the digestive, cardiovascular, auditory, reproductive, urinary, and respiratory systems. Another important factor contributing to the occurrence and survival of indicator bacteria is the presence of *Cladophora* algae. Doucette (1995) investigated the association between bacteria and harmful algal blooms and suggested that the production of pathogenic bacteria and phycotoxins is generally attributed to algal species. The results indicated that *Cladophora* algal mats acted as reservoirs for the indicator bacteria *E. coli* and Enterococci.

The bacterial contamination level of nearshore waters is affected by the concentration of bacteria in beach sand. Researchers have reported that bacteria are mobilized from beach sand to the water column by high tides in California beaches (Yamahara et al., 2007). Several recent studies discussed the dependence of bacterial counts in water on their concentration in beach sand. However, the exact relation between them and the effect of bacterial mobilization on the health of swimmers has not been well established. Understanding the significance of the relationship between the presence of bacteria in sand and public health risk is complicated due to the lack of a widely accepted method for bacterial enumeration in beach sand. There are several published methods for extracting bacteria from sand, including simple handshaking of the sample and complex methods using light sonication, mechanical shakers, and modified buffer solutions. Although shaking methods are most frequently used for bacterial enumeration, they differ in the shaking duration, type of shaking, type of eluent, eluent-to-sand ratio, and eluent composition. Selecting the optimal extraction method with effective eluents is essential for ensuring accurate recovery of indicator bacteria from sand. Boehm et al. (2009) compared

different extraction methods for bacterial enumeration in sand to determine the approach that produced the highest recovery. Five different eluents were tested based on the physical characteristics of the extraction method, and the results indicated that 2-minute handshaking in PBS or DI water with a 30 sec settling time, one rinse step, and a 10:1 eluent-to-sand ratio produced the highest recovery of indicator bacteria. However, selecting the appropriate eluent type based on its physical characteristics depends on the sand composition. Other studies also used DI water or PBS as eluents with a shaking duration of 1-2 min to enumerate *E. coli* and Enterococci in sand (Alm et al., 2003; Bonilla et al., 2007; Hartz et al., 2008; Sampson et al., 2006). In this study, we used DI water and PBS to enumerate bacteria in beach sand using mechanical shaking with a shaking duration of 5 min, settling time of 5 min, eluent-to-sand ratio of 8:1 (200 mL to 25 gm), and PBS composition according to EPA standards: 0.32 grams of sodium dihydrogen phosphate, 1.1 grams of sodium monohydrogen phosphate, 8.5 grams of sodium chloride, and 1 liter of distilled water.

This study was undertaken to enhance the knowledge of fecal indicator bacteria levels in beach sand and develop methods for more accurate predictions of public health outcomes from the measured bacterial concentration. The objectives of this study were to examine the concentration and interaction of fecal indicator bacteria in beach sand and water at Bradford Beach along Lake Michigan; to analyze the effectiveness of different types of eluents, i.e., DI water and PBS, in enumerating bacteria in beach sand; and to evaluate the impact of algae on bacterial concentration. Bradford Beach was used as the sampling location because it is one of the top urban beaches and the most visited place for public recreational activities in Milwaukee. Several contamination sources such as stormwater discharge and sewage overflow, along with a large population of shorebirds, cause bacterial pollution at Bradford Beach. Analysis of bacterial

concentrations at this location could help to develop methods for accurate prediction of public health outcomes as a result of increased contamination with fecal indicator bacteria. The samples were collected in the summer and early fall months because of a consistent trend of elevated bacteria levels in beach sand due to public activities, and 2013 was studied as a representative year. CANARY, a statistical event detection system, was used to identify abnormal conditions, i.e, anomalies in bacterial concentration. Research studies have been performed with CANARY for water quality monitoring in drinking water, surface water, and wastewater systems (Leow et al., 2017; Nafsin and Li, 2021; Perelman et al., 2012). However, studies on the application of CANARY for monitoring bacterial contamination have not been reported. To the best of the authors' knowledge, this study is the first to use CANARY software to monitor bacterial concentrations at a recreational beach. The application of CANARY was evaluated in the analysis of beach water to determine the effectiveness of the software in detecting changes in fecal indicator bacteria concentration and providing timely information on beach contamination and public health risks associated with pathogens.

3.2 Materials and Methods

3.2.1 Study area and sample collection

Bradford Beach (Figure 3.1) is an urban beach along the shoreline of Lake Michigan, Milwaukee, WI, with location coordinates of 43.0313°N, 87.8737°W. Three different transects were used as sampling locations and remained constant throughout the sampling period by using local landmarks to ensure consistency. These three transects were sampled three days per week at 9:15 am. Three samples were obtained from each transect: one sand sample from the swash zone, one sand sample from 6 m inland, and a water sample. The swash zone samples were taken

from the beach surface, whereas the 6 m inland samples were taken from the water table so that all sand samples were saturated. During the sampling period, the water table varied from 0.15 to 0.53 m, depending on the precipitation and local topography. Water samples were taken from a depth of approximately 0.5 m. The water temperature ranged from 11 °C to 23 °C with a mean of 15.4 °C during the sampling period. Water samples and sand samples were stored in 500 mL plastic bottles and plastic Whirl-Pak bags, respectively, and placed in a cooler with ice until laboratory analysis was performed.

Figure 3.1. Study area (Bradford beach, Milwaukee, WI) with the three transect locations (Transect 1, Transect 2 and Transect 3) that were used for sampling.

3.2.2 Sample preparation

The samples homogenized at 200 rpm using an Excella E24 incubator shaker platform for 5 minutes to detach indicator bacteria from sand. Before performing the IDEXX Most Probable

Number (MPN) analysis, the larger particles in the homogenized samples were allowed to settle for 5 minutes. To keep the bacterial count within the detection limit of the sampling equipment, 25 grams of the sand samples were weighed. After mixing thoroughly and weighing, the sand samples were placed in sterile bottles. Each IDEXX plastic bottle was filled with 200 mL of either of the two eluents (i.e., DI water and PBS). After shaking, the samples were allowed to settle. The 200 mL sample was then placed into two sterile 120 mL vessels, each containing 100 mL. One of the vessels was used for testing and enumerating *E. coli*, and the other was used for Enterococci. Using the IDEXX Quanti-Tray/2000 method (APHA, 2017), the bacterial concentration in the sample was determined in units of MPN/100 mL.

Ultra-pure DI water was produced by the Milli-Q Grad 2 system manufactured by Millipore. Using advanced technologies such as ion-exchange resins, activated carbon, an ultra-violet photo-oxidation technique, and 0.22 μm membrane filter, the water was particulate and bacteria-free and had low organics at parts-per-billion levels and high resistivity (8.2 MΩ × cm at 25 °C). Negative control tests for *E. coli* and Enterococci with the two eluents were performed following the IDEXX MPN method. After an incubation period of 24 hours at the specific temperature, the Quanti-Tray wells were found to be colorless as compared to the IDEXX Quanti-Tray 2000 comparator, which was used as the liquid color or fluorescence reference. For eluents, no fluorescence wells were present for Colilert and Enterolert tests, indicating the absence of bacteria in both DI water and PBS eluents.

3.2.3 Visual classification system for algae level

A rating scale was developed to determine the levels of algae in beach water, with ‘0’ for no algae, ‘1’ for low levels, ‘2’ for moderate levels, and ‘3’ for the high levels of algae. Algal

levels were recorded throughout the sampling period with different wave actions near the shore. A rating of '1' indicated a small number of algae visible in the beach water, while a rating of '2' indicated a significant number of algae with wave action on the beach. Because of the thick covering in nearshore water, there was no wave action on the beach that led to stagnation, and for that level of algae, the rating was classified as '3', the highest level on the scale. According to a report by Kasich et al., (2014), in the absence of additional data, visual evidence can be used to characterize algal blooms as severe (significant cell concentration or surface accumulation is present/visible throughout the water column), moderate (bloom is visible), or minor (little visual evidence of bloom). Visual evidence is used as an initial assessment, which is then refined upon the collection of additional information. In this study, the rating scale based on visual classification was mainly used to identify the presence of algae and maintain consistency in the specific sampling location throughout the sampling period.

3.2.4 Data analysis

Analyses of bacterial data from beach sand and water were performed using the U.S. EPA's CANARY (Hart et al., 2007; Haxton et al., 2013) EDS. The software includes statistical tools to analyze water quality data in real time and in offline mode and thereby identify contamination events. The input data for CANARY can be either a comma-separated value (CSV) file in offline mode or a link to a database Supervisory Control and Data Acquisition (SCADA) system in online mode (U.S. EPA, 2014). Linear Prediction Correction Filter (LPCF) algorithm in the offline mode of CANARY was used in this study to analyze bacterial concentrations. Based on a historical dataset, the algorithm predicts the current signal value at each time step. The predicted data are compared to the observed data as soon as the actual signal

value is available at that time step. The difference between the observed and predicted data, defined as the residual, is calculated and normalized within the software. The algorithm detects a data point as an outlier when the residual value is above the user-defined threshold value (Leow et al., 2017). Once the data point is detected as an outlier, it is no longer used for further predictions. Residuals are normalized with a mean of 0.0 and standard deviation of 1.0 because the water quality signals have different units and magnitude, and normalization allows for combining different signal measurements. The data were normalized using equation (3.1).

$$X_s = \frac{X_h - \bar{X}}{\sigma_x} \qquad (3.1)$$

where X_S is the normalized data, X_h is the historical data, σ_x is the standard deviation, and $\bar{X}$ is the mean value.

The software calculates the probability of any contamination event within the Binomial Event Discriminator (BED) window. Depending on the number of outliers detected over a specific time period and whether the probability value exceeds the threshold, the software triggers an "alarm". For the detection of events, the CANARY algorithm uses binomial distribution theory and produces a binary result (either success or failure) that is further extracted by the BED. The extraction of data by the BED reduces the influence of time steps with unexpected data (Murray and Haxton, 2010).

The configuration parameters of the EDS require optimization for a specific set of data to be analyzed. The optimized parameters in the configuration file of CANARY include the history window, the outlier threshold, the event threshold, and the BED window. During the configuration process of the software, the algorithm selects the optimal values of these parameters and adjusts the BED for an increasing or decreasing probability of anomalous water

quality events. Among the configuration parameters, the history window and the threshold are the most significant factors influencing software performance. The optimal configuration parameters selected were history window = 6, outlier threshold = 0.80, BED window = 6, and event threshold = 0.89.

In this study, the measured bacteria levels in beach sand and water were analyzed in the offline mode of CANARY using historical data. However, in online mode, CANARY can analyze datasets in real time, differentiate real events from background variability, and trigger an alarm to alert the operator to take immediate action against possible contamination. Although real-time data in the online mode of CANARY were not used, the effectiveness of the software was observed by using historical data for the detection of anomalies or 'events', which can be useful for the real-time monitoring of bacterial concentrations in beach sand in the future.

3.2.4.1 Model sensitivity analysis

A sensitivity analysis was conducted using LPCF algorithm by investigating the key parameters and determining the optimal parameters for configuration. The optimal window size is determined by running a set of historical data by the LPCF algorithm and predicting the future signal value. The software provides the output with the average absolute value of the residual and the standard deviation of the residual as a function of the window size. The threshold value which is another significant parameter to model performance is defined as a multiplier of signal's standard deviation. Based on the threshold value, the algorithm classifies any water quality signal as background or anomalous water quality. An outlier is detected for a threshold value of 1.0 if the signal value is greater than one standard deviation from the mean value. The decision of CANARY regarding a signal value being acceptable or anomaly is defined by the outlier

threshold parameter. The software triggers an alarm if any water quality signal contributes as an abnormal condition or outlier.

E. coli count in sand samples from the swash zone and from 6m inland at three transects analyzed with eluent 200 mL DI water were used for sensitivity analysis. Different window sizes ranging from 3-time steps to 12-time steps were used for performing the sensitivity analysis with bacteria data. Figure 3.2(a) and Figure 3.2(b) indicate the results from using the six different window sizes. In all test runs, the parameters that control the integration of results were kept constant. An event could be detected when enough outliers were present within a specified BED window. In this analysis, 4 outliers within a BED window of 6 time-steps were necessary before identifying an event. Bacteria data analysis was performed using the six window sizes (time-step: 3, 4.5, 6, 7.5, 9 and 12).

The absolute values of the residual obtained from the CANARY output were used to determine the standard deviation of residual for each window size. The calculated standard deviation was plotted against the corresponding window size. Figure 3.2(a) and Figure 3.2(b) indicate that the standard deviation of residual tends to decrease with the increased window size. Lower values of the performance measure: standard deviation of the residual indicate increased precision and accuracy in future prediction of the signal value. The window size that produces the lowest value of standard deviation of the residuals would be a better choice for selection of 'history window'. However, increased window size would result in longer computational time for updating the parameters. In this consideration and based on the number of sample observations, we avoided a larger window size, and selected a size of 6 that resulted in comparatively lower values of standard deviation for all the three curves on the plot in Figure

3.2(a). It also appears that the value reduces close to their final minimum value for a window size of at least 6 time-steps.

Figure 3.2(c) shows the variation of events with different window sizes. It can be observed that as the window size increased from 3 to 12-time steps, the total number of events decreased. No event was detected for the window size of 12-time steps. Based on the information presented in Figure 3.2(c), the history window was selected as 6-time steps for the bacteria data analysis, which resulted in 8 detected events.

Analysis was carried out using six different outlier threshold values as shown in Figure 3.2(d). The number of detected events decreased with the increase of outlier threshold value, because higher values of the outlier threshold result in less outlier. However, for a higher threshold value, real events might be undetected considering all data as acceptable or good. It is worth including a lower than 1 value in any testing, because in most cases the data respond well to a lower outlier threshold value. For this study, outlier threshold value was selected as 0.80. Event threshold was determined from the BED window and number of outliers within a typical timeframe ranging from 30 min to 120 min, using the binomial probability distribution function. The value of BED window requires to be either less than or at least equal to the history window size, not exceeding the size of history window. Usually, the value of BED window is set within a range of 4 to 18-time steps. In this study, the BED window was set as 6-time steps from which the event threshold was found as 0.89. From this sensitivity analysis with changing parameters, the optimal algorithm configuration parameters for the bacteria data analysis were chosen as: History window 6, Outlier threshold 0.80, BED window 6 and Event Threshold 0.89.

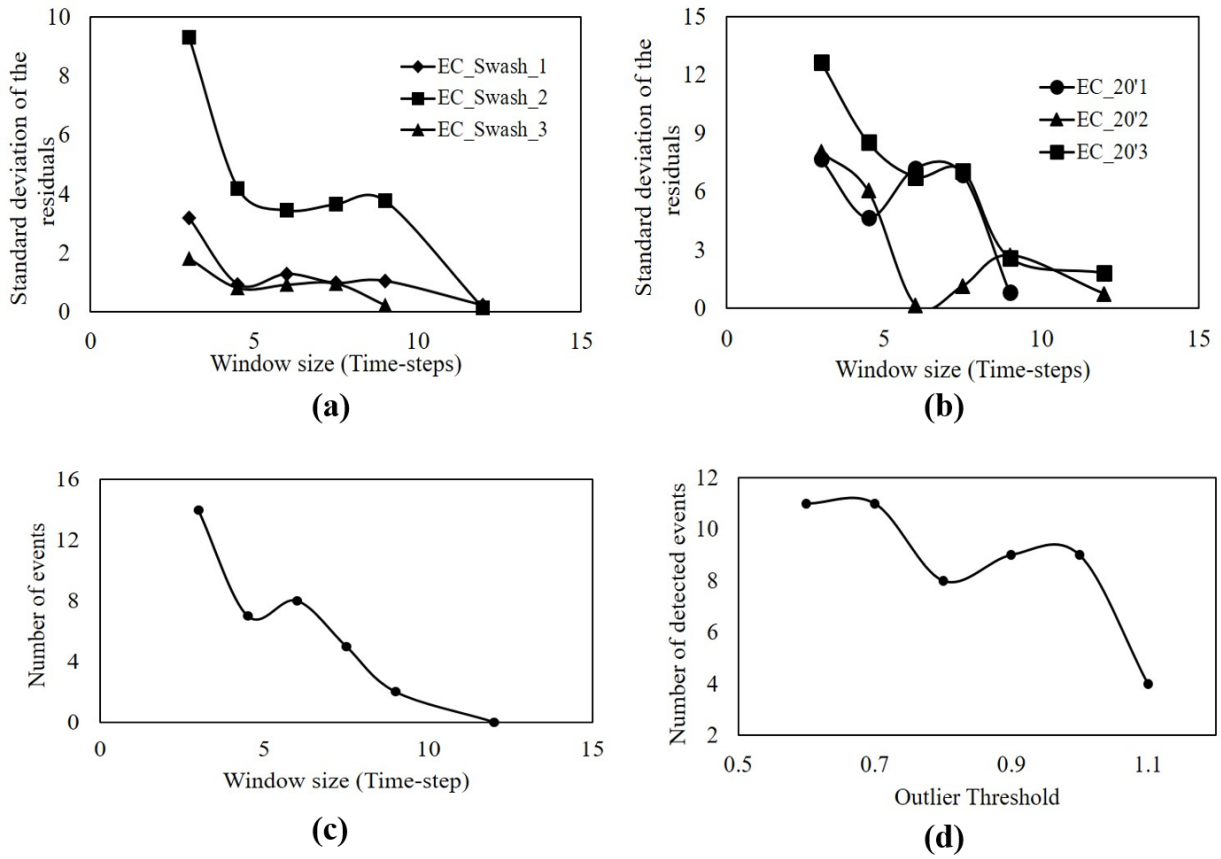


Figure 3.2. (a) Sensitivity analysis with different window sizes for *E. coli* in swash zone sample at three transects. (b) Sensitivity analysis with different window size for *E. coli* in 6m inland samples at three transects. (c) Variation of detected events with window size. (d) Number of detected events with variation of outlier threshold.

3.3 Data analysis and Result

3.3.1 Interaction of bacteria in beach sand and water

CANARY was used to analyze data on *E. coli* and Enterococci concentrations in samples collected from 1st July 2013 to 21st August 2013. Basic statistical parameters, such as the minimum, maximum, mean, standard deviation, and coefficient of variation (CV) of the input *E.*

coli and Enterococci concentrations used in CANARY, are presented in Tables 3.1 and 3.2. Bacterial concentrations at different sampling locations were highly variable, with a coefficient of variation of more than 100%. A high CV indicates that the data are highly dispersed around the mean value. For *E. coli* in 6 m inland sand samples, the highest CVs were found at transect 2 for both eluents with 352.43% for DI water and 392.41% for PBS. For sand samples in the swash zone and 6 m inland, the mean *E. coli* concentrations were greater at transect 2 than transects 1 and 3. Similarly, the mean *E. coli* level in water samples was the greatest at transect 2. For Enterococci in the swash zone, the mean and CV of the bacteria level were higher at transect 3 with both eluents. However, in the 6 m inland samples, transect 2 had the greatest CV and mean concentrations. For Enterococci in water samples, the mean bacterial concentration was 479.73 MPN/100 mL at transect 1, higher than the other two transects. Statistical analysis also indicated significant differences (p-value < 0.05) between the mean values of log MPN of *E. coli* and Enterococci in sand and water samples.

Table 3.1. Basic statistical analysis of ***E. coli*** levels at different sampling locations.

Sample	Sample location	Unit	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation %
Sand sample with DI water	Swash zone 1	MPN/100 mL	1.0	99.0	20.9	21.7	103.9
	Swash zone 2		1.0	1203.3	85.6	260.9	312.2
	Swash zone 3		1.0	121.1	27.6	28.0	101.5
	6 m inland 1		0.0	217.8	31.3	59.1	188.6
	6 m inland 2		0.0	727.0	45.4	159.9	352.4
	6 m inland 3		0.0	62.4	6.1	14.2	233.9
Sand sample with PBS	Swash zone 1	MPN/100 mL	0.0	57.1	16.9	14.8	87.3
	Swash zone 2		0.0	1416.3	87.6	302.2	344.8
	Swash zone 3		2.0	122.3	21.5	27.2	126.9
	6 m inland 1		0.0	209.8	24.6	51.9	211.8
	6 m inland 2		0.0	727.0	39.5	155.1	392.4
	6 m inland 3		0.0	62.4	5.1	13.7	272.1

Water sample	Transect 1	MPN/100 mL	0.0	2419.6	450.2	666.3	147.9
	Transect 2		0.0	1553.1	273.5	363.5	132.9
	Transect 3		2.0	1203.3	141.9	262.8	185.2

Table 3.2. Basic statistical analysis of **Enterococci** levels at different sampling locations.

Sample	Sample location	Unit	Minimum	Maximum	Mean	Standard deviation	Coefficient of variation %
Sand sample with DI water	Swash zone 1	MPN/100 mL	0.0	1986.3	100.3	432.2	431.0
	Swash zone 2		1.0	15.8	5.1	4.3	87.7
	Swash zone 3		0.0	2420.0	169.2	556.1	328.8
	6 m inland 1		0.0	24.3	4.6	6.0	132.2
	6 m inland 2		0.0	78.9	5.4	17.1	319.5
	6 m inland 3		0.0	7.3	1.4	1.9	140.1
Sand sample with PBS	Swash zone 1	MPN/100 mL	0.0	35.9	10.2	9.8	95.9
	Swash zone 2		0.0	37.9	9.9	11.2	111.2
	Swash zone 3		0.0	2425.0	168.4	547.2	324.9
	6 m inland 1		0.0	24.3	4.5	5.7	125.7
	6 m inland 2		0.0	78.9	5.2	16.7	321.1
	6 m inland 3		0.0	53.8	4.9	11.5	230.7
Water sample	Transect 1	MPN/100 mL	0.0	2425.0	479.7	870.9	181.5
	Transect 2		1.0	2419.6	268.6	683.9	254.7
	Transect 3		1.0	261.3	30.9	62.2	201.1

Samples were collected on days with no rainfall and days after rainfall events. The rainfall events and associated rainfall-runoff might have affected the measured bacterial concentrations and caused high variations in the bacteria data over the sampling period. Kleinheinz et al. (2009) showed a significant association between rainfall and elevated bacterial counts. A higher standard deviation than the mean value indicates that the data are not normally distributed, and some of the measured values affected by the rainfall event may be significantly higher than all other data. Whitman and Nevers (2003) also reported higher standard deviations than mean *E. coli* counts in water samples, which resulted in CV greater than 100%. In this

study, quality control was assured by following the U.S. EPA statistics for microbiological analysis in the USEPA microbiology methods manual, Part IV, A–C (Bordner et al., 1978). Before performing statistical analysis, the original data were converted into logarithms to obtain a symmetric distribution resembling the normal distribution. The best measure of central tendency for microbiological data is the log transform.

Table 3.3 shows the results from CANARY for both bacteria in sand samples with different eluents and in water samples. The results indicate that DI water produced a higher number of events than PBS for *E. coli*. The mean values of the measured *E. coli* concentration in sand samples at different locations were also greater for DI water than PBS (Table 3.1). The detailed statistical analysis supporting this finding is provided in Table 3.4. For *E. coli*, CANARY identified a total of eight events with DI water and only four events with PBS. However, for Enterococci bacteria, a total of 10 events were identified with DI water, and a total of 11 events were detected with PBS. Both eluents resulted in the same number of events (four events) for *E. coli* in the swash zone, whereas no events were detected with the PBS eluent in samples from 6 m inland. In addition, when using DI water, a higher number of events were detected for Enterococci than *E. coli* in both the swash zone and 6 m inland samples, while with PBS, similar numbers of events were detected in the swash zone for both *E. coli* and Enterococci. With PBS, a significantly higher number of events were found for Enterococci than *E. coli* in 6 m inland samples.

Table 3.3. Results of CANARY output (number of detected 'Events') for sand samples in swash zone and in 6 m inland and water samples.

Fecal indicator bacteria	**Swash zone**		**6 m inland**		**Sand sample**		**Water sample**
	DI	**PBS**	**DI**	**PBS**	**DI**	**PBS**	
E. coli	4	4	4	0	8	4	5
Enterococci	5	4	5	7	10	11	5

Analysis was performed using CANARY to evaluate the potential of freshwater sand to act as a reservoir of indicator bacteria. This was accomplished by comparing the number of detected events by CANARY that resulted in anomalous water quality in water and sand samples in both the swash zone and 6 m inland for *E. coli* and Enterococci bacteria. The results indicate that a higher number of events were found for the indicator bacteria in sand samples than in water samples, as shown in Table 3.3 and Figure 3.3. Figure 3.3 shows the signal plots of the bacterial data, as well as the event probability plot. The event plot shows the probability that an event will occur at each time step as computed by CANARY. The dots on the event probability plot indicate events. In addition, the dots on the signal plots represent the bacterial data that contributed to the identification of an event at that time. Five events were detected in water samples for both *E. coli* and Enterococci bacteria, while in sand samples, more than five events were detected. In addition, when using PBS, Enterococci appeared to accumulate in sand to a greater extent than did *E. coli*. These results suggest that freshwater beach sand can be evaluated further for its potential to serve as a reservoir for the survival of indicator bacteria. Additional plots for CANARY outputs are provided in Figure 2.4-Figure 2.7.

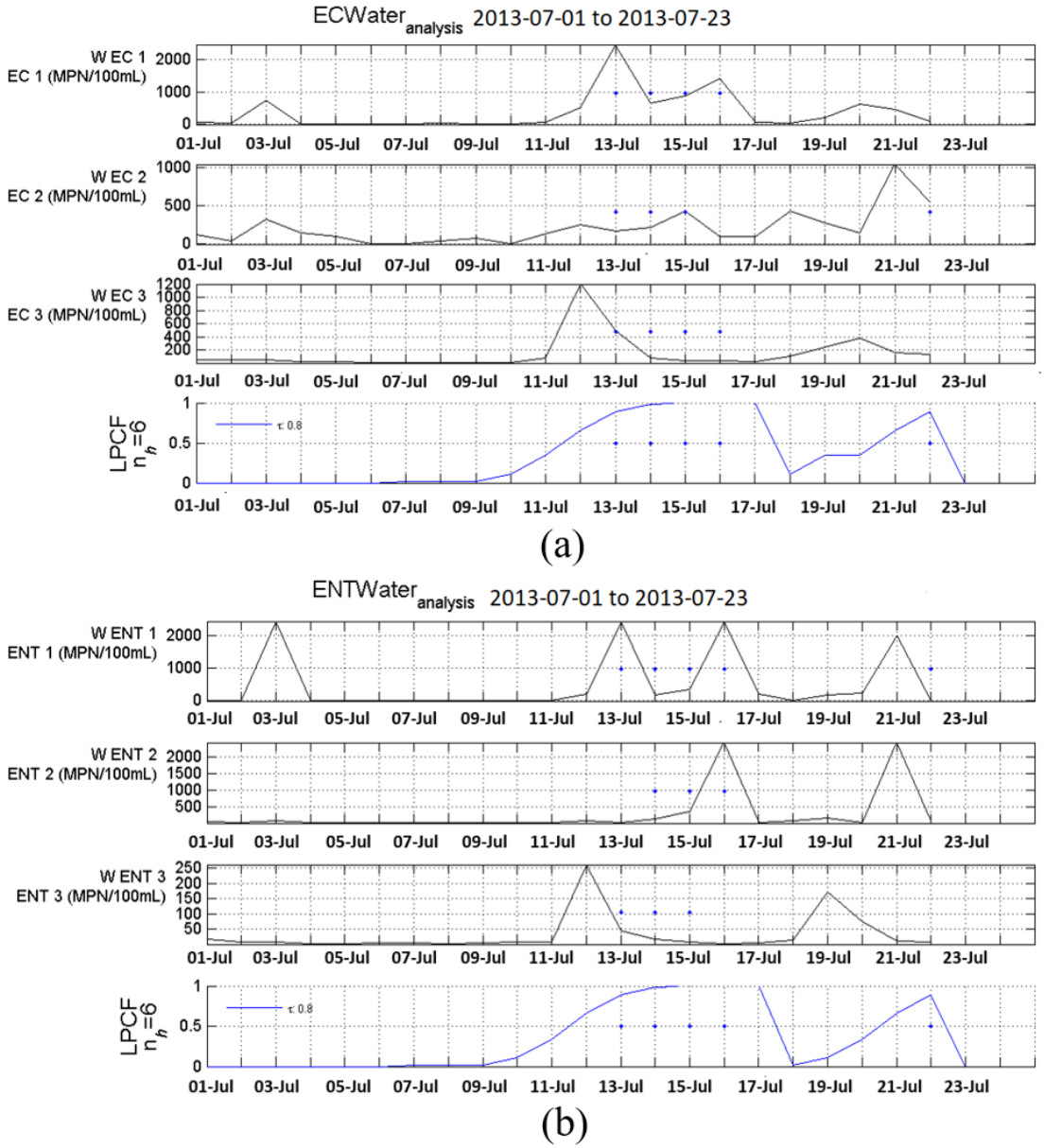


Figure 3.3: CANARY output for (a) *E. coli* and (b) Enterococci count in water sample at three transect locations with probability of event plot indicating total number of detected events 5 for both bacteria during sampling period. The dots in the probability plot indicate 'events' and the dots in the signal plots indicate the 'outliers'

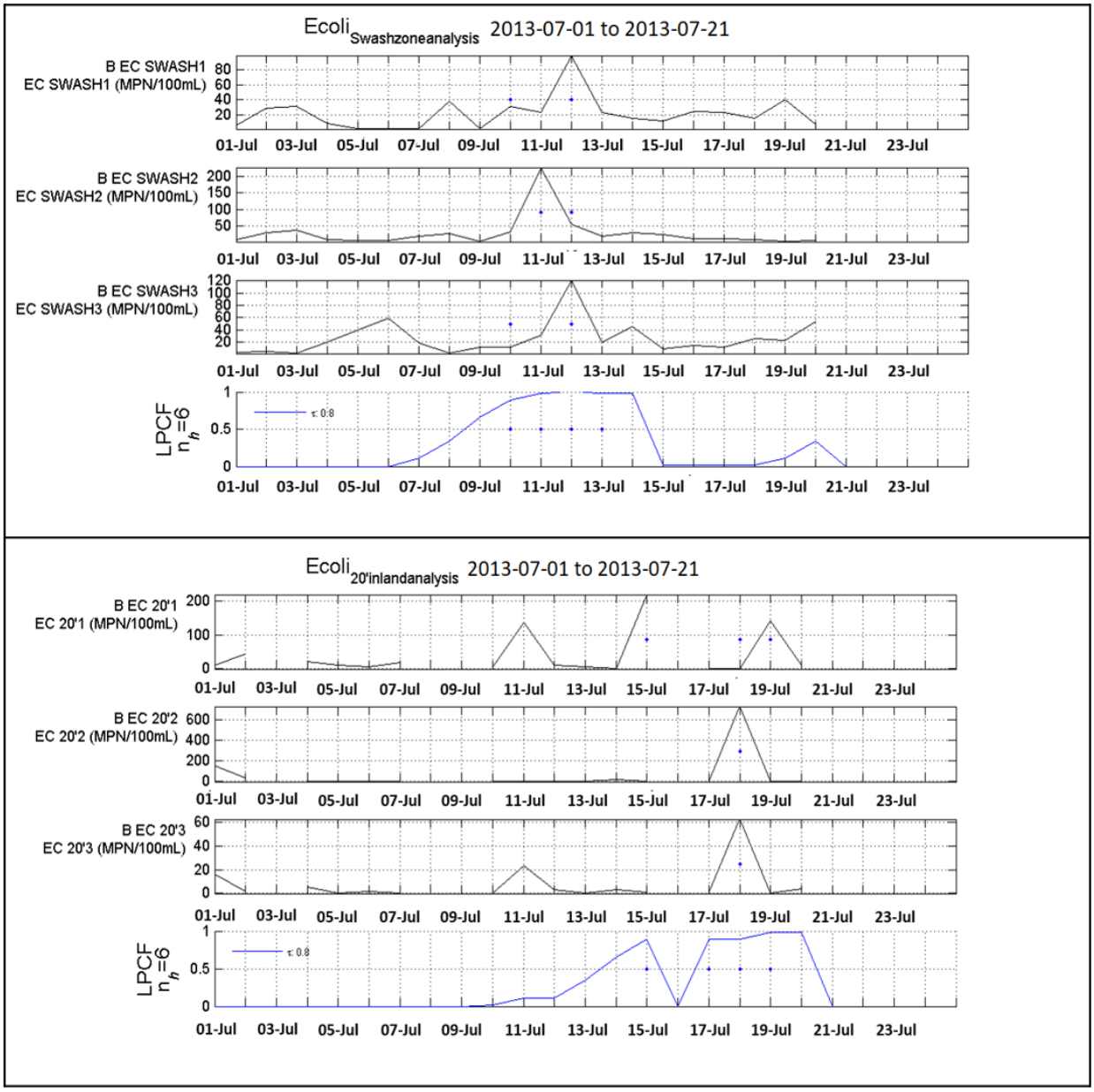


Figure 3.4. CANARY output for *E. coli* count in sand samples in swash zone (Top) and 6m inland (bottom) at each of the three transect with eluent DI water during sampling period; Probability of event plot showing total number of detected events 8. (4 detected events for both of the swash zone and 6m inland). The dots in the probability plot indicate 'events' and the dots in the signal plots indicate the 'outliers'

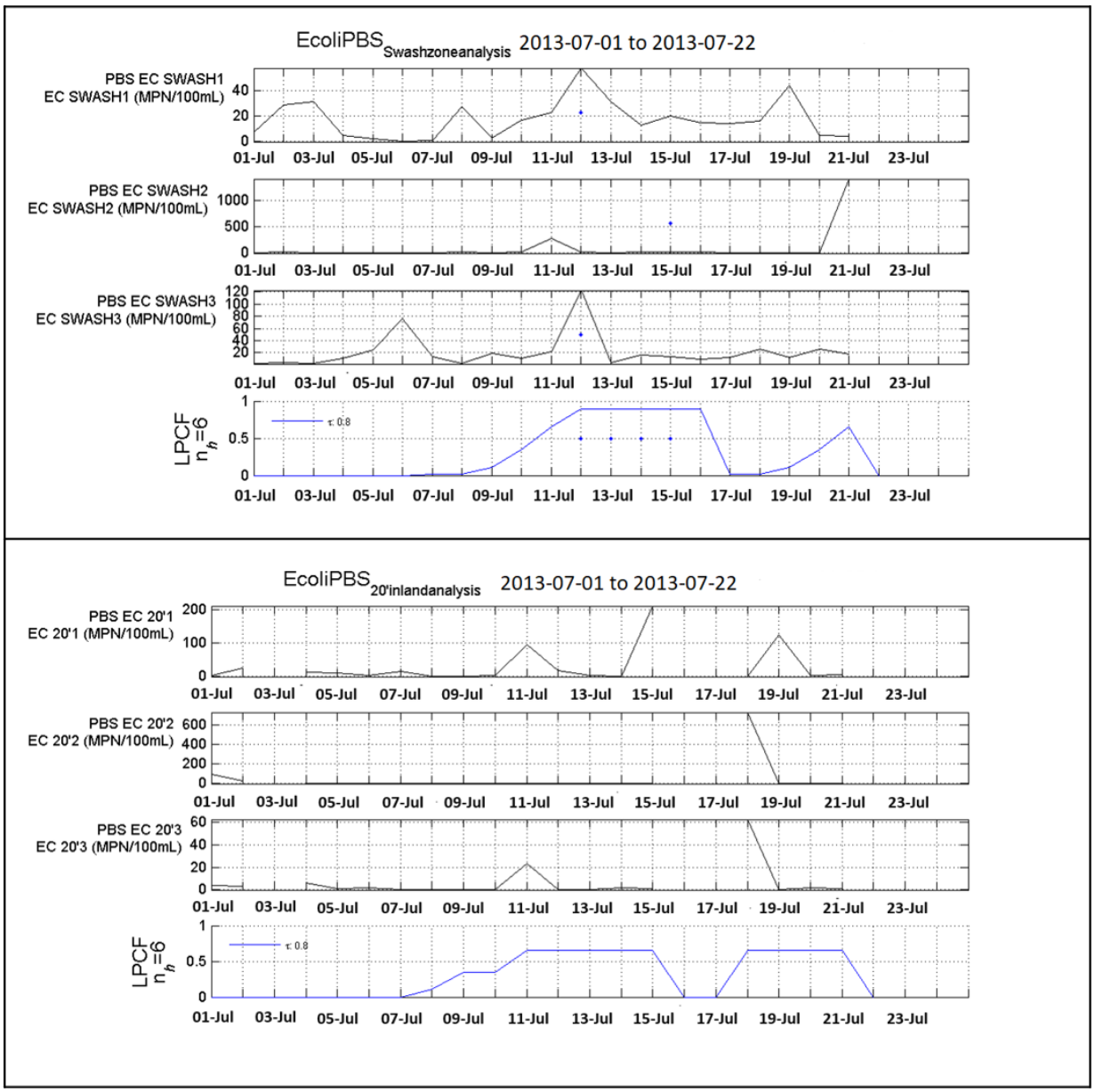


Figure 3.5. CANARY output for *E. coli* count in sand samples in swash zone (Top) and 6m inland (bottom) at each of the three transect with eluent PBS during sampling period; Probability of event plot showing total number of detected events 4 (4 detected events in Swash zone and 0 event for 6m inland). The dots in the probability plot indicate 'events' and the dots in the signal plots indicate the 'outliers'

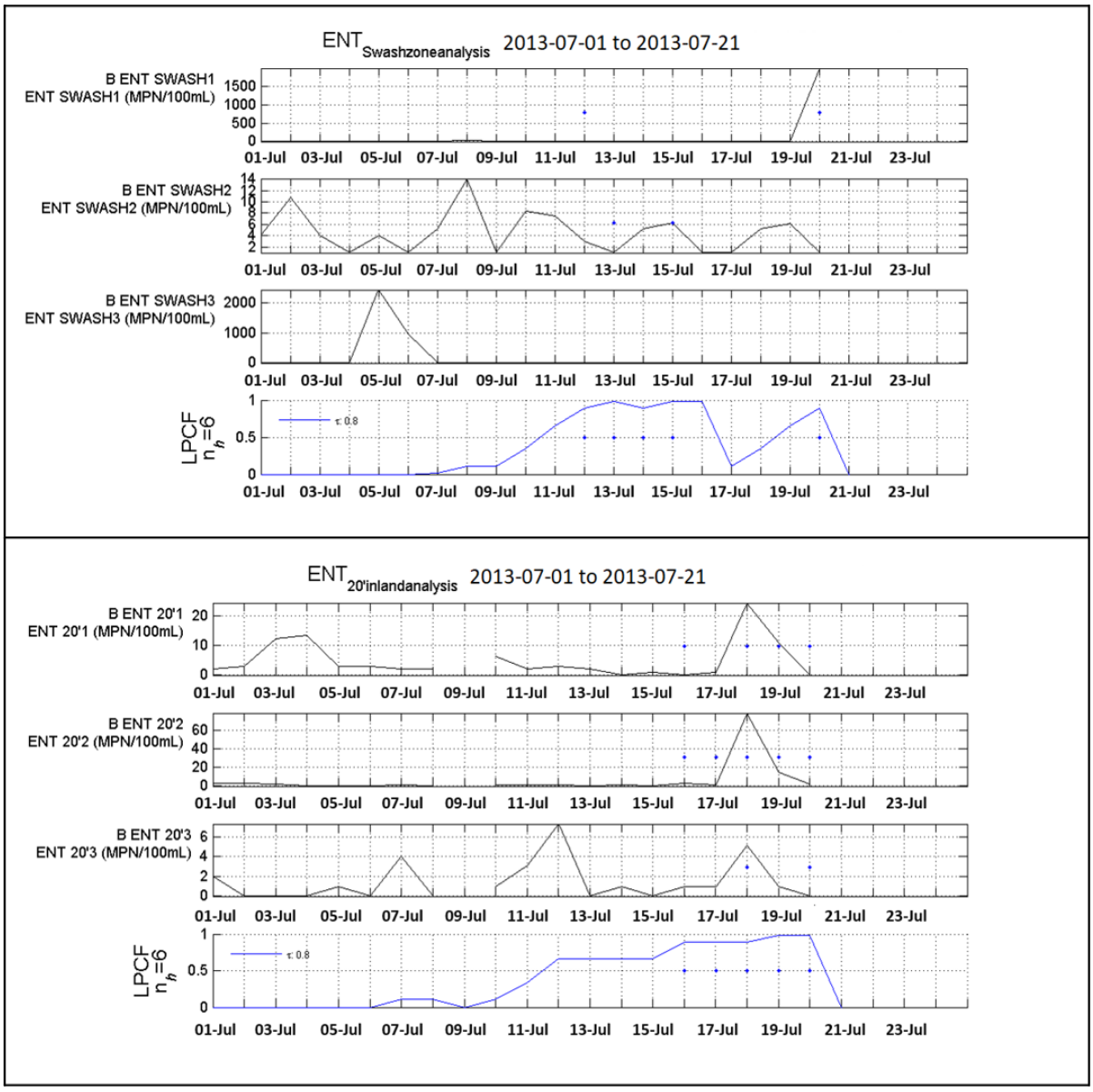


Figure 3.6. CANARY output for Enterococci count in sand samples in swash zone (Top) and 6m inland (bottom) at each of the three transect with eluent DI water during sampling period; Probability of event plot showing total number of detected events 10 (5 detected events for swash zone and 5 events for 6m inland). The dots in the probability plot indicate 'events' and the dots in the signal plots indicate the 'outliers'

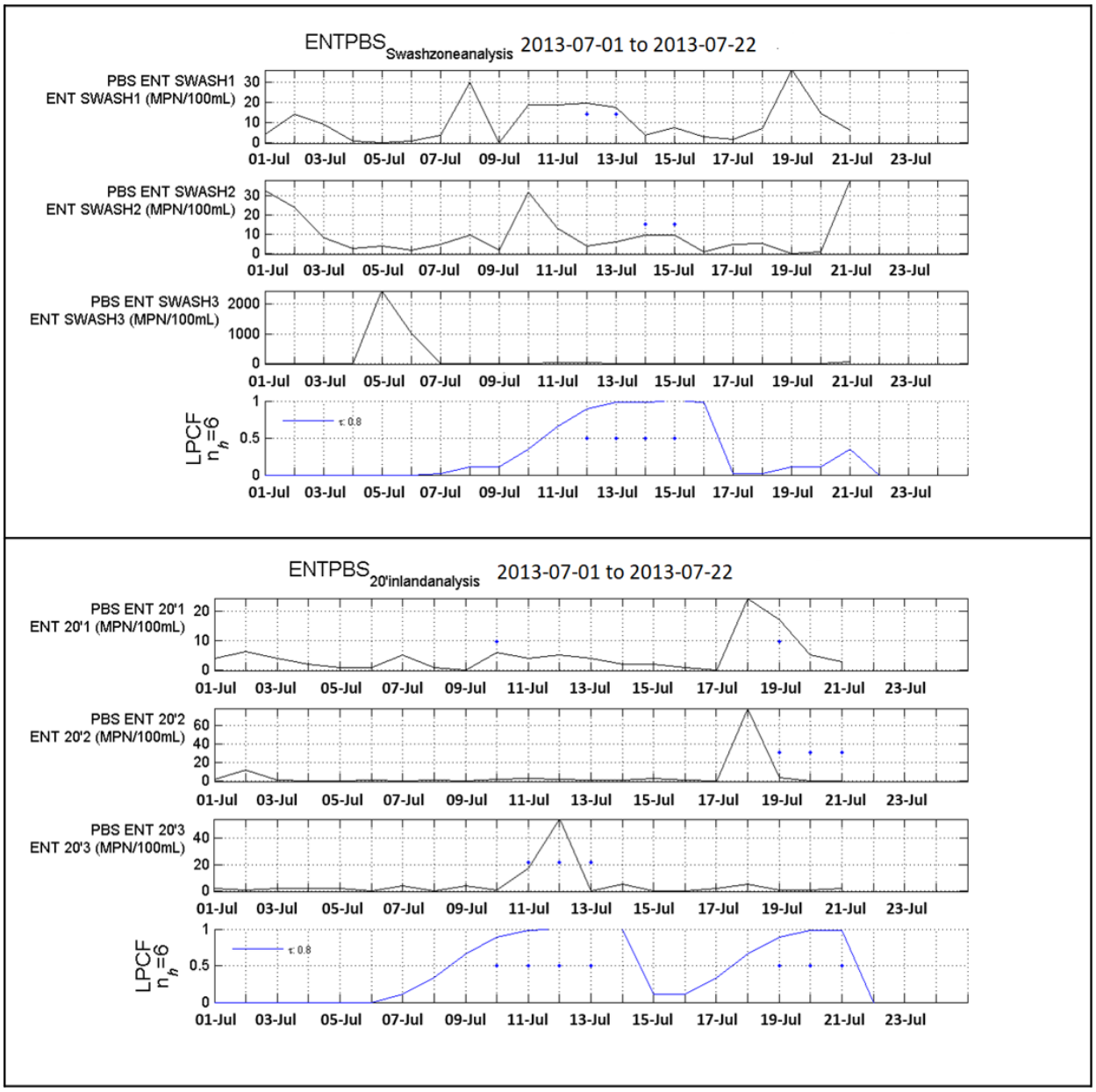


Figure 3.7. CANARY output for Enterococci count in sand samples in swash zone (Top) and 6m inland (bottom) at each of the three transect with eluent PBS during sampling period; Probability of event plot showing total number of detected events 11 (4 detected events in swash zone and 7 events in 6m). The dots in the probability plot indicate 'events' and the dots in the signal plots indicate the 'outliers'

3.3.2 Effect of eluents

A paired t-test analysis was performed using Minitab statistical software for both bacteria comparing DI water and PBS eluents as shown in Table 3.4. The result showed differences in mean value (log MPN/100 g sand) for both bacteria. When using DI water, the mean value of log MPN was found as 1.84 for *E. coli* and 1.24 for Enterococci, while with PBS the mean value of log MPN was 1.69 and 1.46 respectively. The differences in mean value of log MPN between the two eluents were statistically significant with p-value < 0.05 for 95% confidence level. A hypothesis testing was also performed to analyze if the mean log MPN value of DI greater than PBS. The results concluded that for *E. coli*, the mean of log MPN of DI was greater than PBS at 0.05 level of significance and the difference in mean value between the eluents was found as 0.1457 log unit. For Enterococci, there was not enough statistical evidence to conclude that the mean log MPN of DI was significantly greater than the mean of PBS. Instead, for Enterococci, the mean log MPN of PBS was greater than DI water at 0.05 level of significance ($p< 0.001$). From the statistical analysis, it appears that DI water produced higher *E. coli* counts while PBS produced higher Enterococci counts. For eluent comparison, the total number of samples used for *E. coli* was 113 and for Enterococci 127 during the overall sampling duration. Figure 3.8 shows the log MPN values for all samples analyzed with both eluents. It indicated that the ratio for each sample between the two eluents was highly variable and the variability was found as statistically significant with p-value=0.00603 < 0.05 with 95% confidence level.

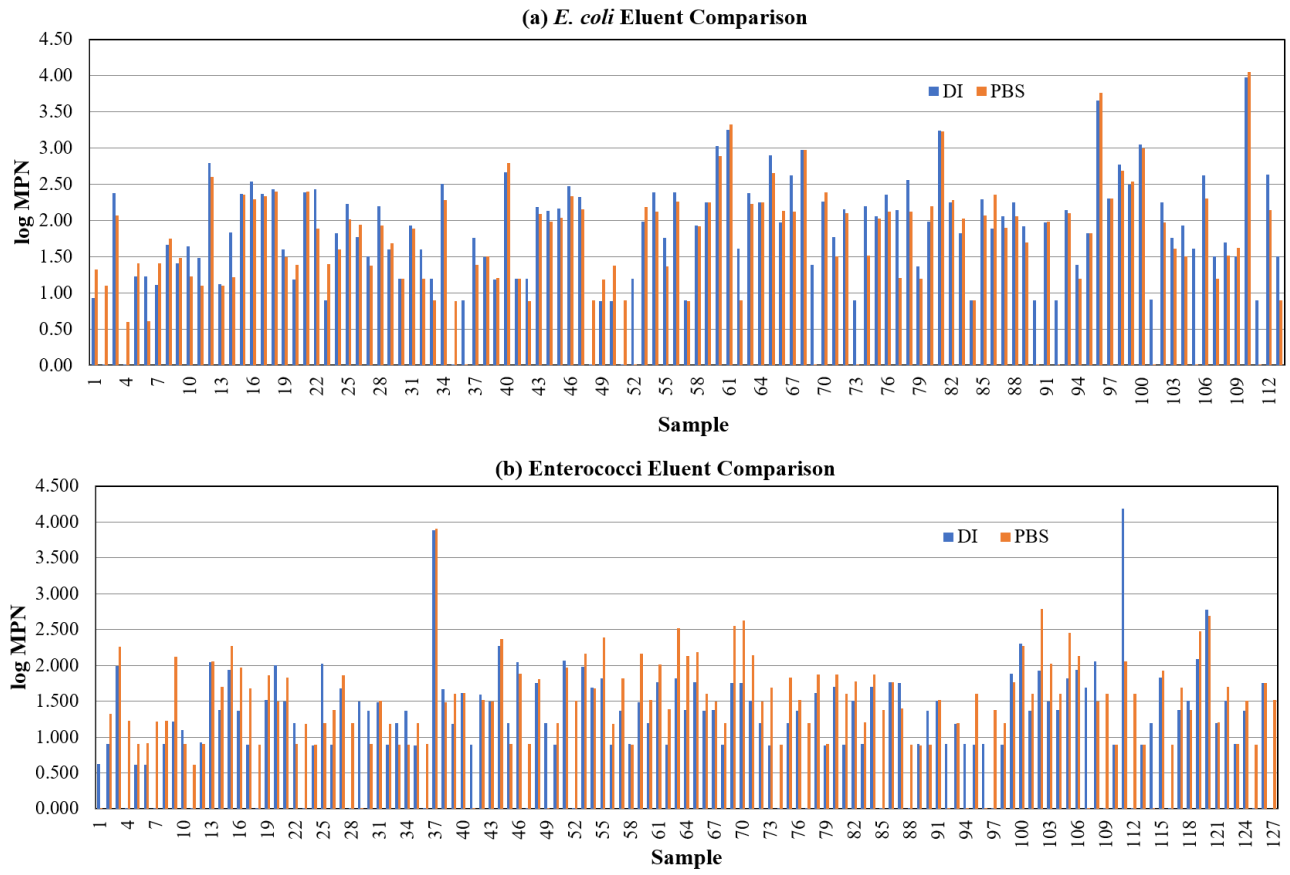
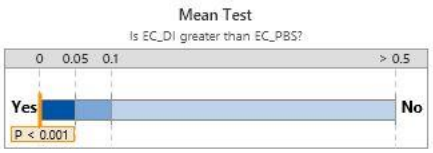
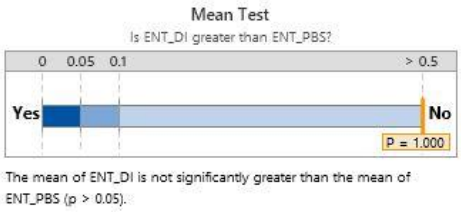


Figure 3.8. Eluent comparison for (a) *E. coli* and (b) Enterococci; log MPN values for total number of samples analyzed.

Table 3.4. Paired t-test analysis for *E. coli* and Enterococci with DI water and PBS eluent

	E. coli		**Enterococci**	
	DI water	**PBS**	**DI water**	**PBS**
Mean	1.8385	1.6928	1.2415	1.4562
Standard deviation	0.7548	0.8074	0.7237	0.6612
SE mean	0.071	0.076	0.0642	0.0587
Number of samples	113	113	127	127
T-value	3.64		3.85	
P-value	0.000411		0.00019	
Hypothesis testing: (Mean test: Is DI greater than PBS?)	*E. coli*: The mean for DI is significantly greater than PBS ($p<0.05$)		Enterococci: The mean for DI is not significantly greater than PBS ($p > 0.05$)	





The log-transformed MPN of bacterial colonies were plotted to show the effects of eluent. Results displayed in Figure 3.9 show that on average, deionized water provided higher MPN than PBS for *E. coli*. An attempt was made to establish a direct ratio between the two eluents for each group of bacteria. The slope and R-squared value for *E. coli* are 0.92 and 0.73, indicating a fairly strong linear relationship between the MPN generated from DI water and PBS. The relationship between the results generated from different eluents for Enterococci is less strong. In addition. Pearson Correlation analysis was performed in Minitab software between the two eluents for *E. coli* and Enterococci bacteria as shown in Table 3.5. For *E. coli*, a strong positive correlation (R^2=0.854) was observed between the two eluents while for Enterococci, the correlation between the eluents was comparatively weak (R^2=0.59). The correlation coefficient results were statistically significant with p-value<0.05 with 95% confidence interval. The results indicate that for *E. coli*, it is possible to demonstrate the difference between the eluents accurately due to the well-correlated dataset. However, for Enterococci, it is difficult to develop the relationship between the two eluents with the highly varied dataset.

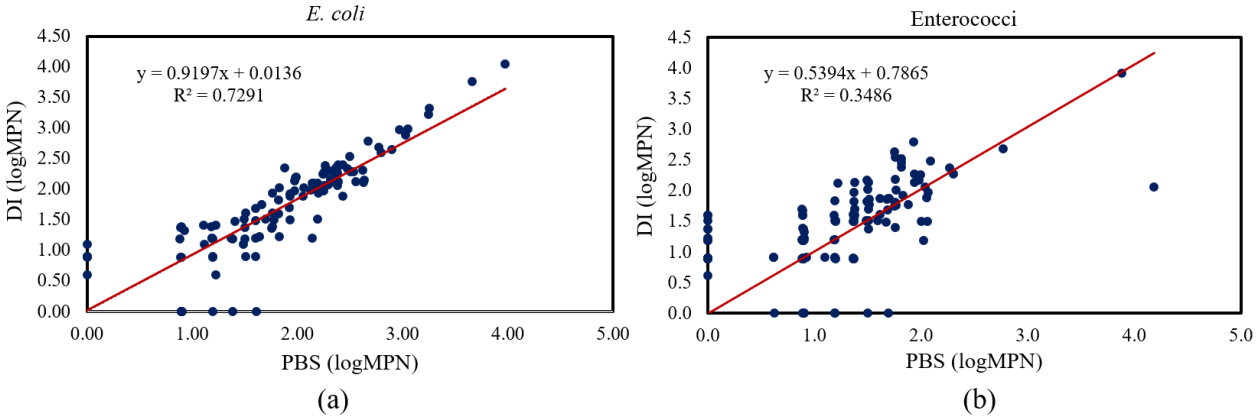


Figure 3.9. (a) Liner regression of *E. coli* concentration in sand using DI water and PBS as eluents, (b) Linear regression of Enterococci concentration in sand using DI water and PBS as eluents.

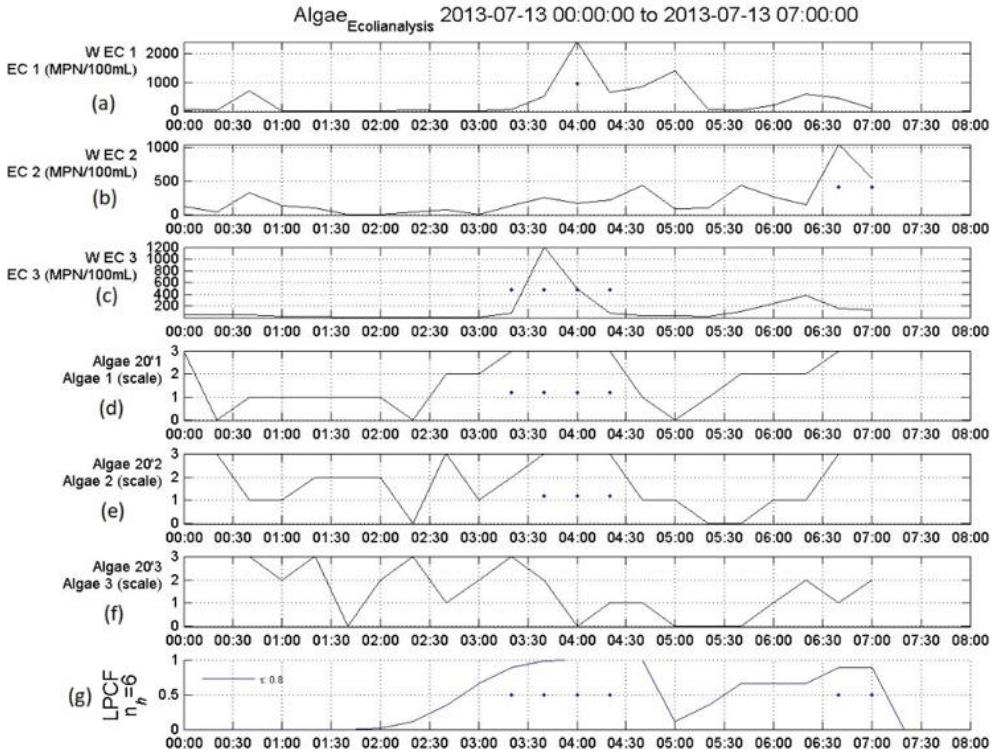
Table 3.5. Pearson Correlation analysis between the two eluents for *E. coli* and Enterococci

Sample 1	Sample 2	No. of samples	Correlation	95% CI	P-Value
E. coli (DI)	*E. coli* (PBS)	113	0.854	(0.794, 0.897)	0.000
Enterococci (PBS)	Enterococci (DI)	127	0.590	(0.464,0.693)	0.000

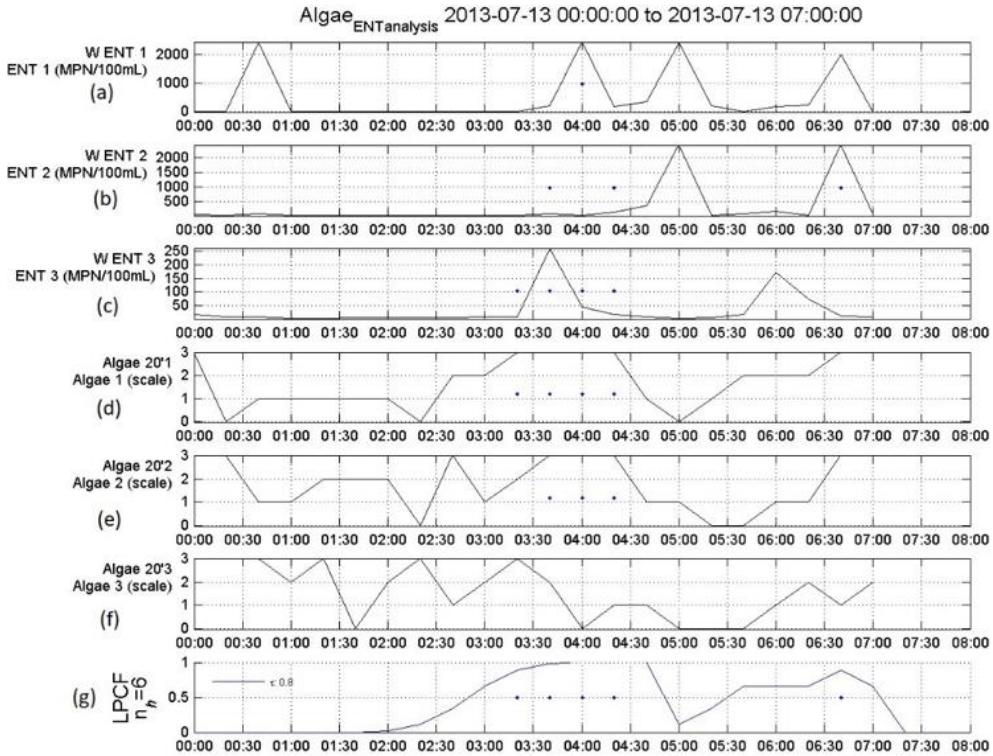
3.3.3 Impact of presence of algae level on bacteria concentration

CANARY was used to analysis the impact of algae on bacteria concentration in water sample. Figure 3.10 shows CANARY output for *E. coli* and Enterococci count with algae level during the sampling duration. The results indicated a correlation between algae levels and bacterial counts as the events with the highest bacterial counts often occurred on time periods with elevated level of algal presence. It appears that bacteria concentration deviated from its baseline and detected as anomaly during the time step where the presence of algae was reported as high (scale 3). Consequently, events were detected as abnormal water quality by CANARY

during time step between 03:20 to 04:20 for *E. coli* as there were deviations from normal behavior at transect 1 and 3. The algae level was also found to be higher over that period. Moreover, at transect 2, the presence of algae was higher that resulted in detected events during time step 06:40 to 07:00 with significant variations in *E. coli* concentration. The presence of high levels of algae showed similar effect on Enterococci at the three transects. Figure 3.10 indicates that during time step between 3:20 to 4:20, events were detected and there were baseline variations of Enterococci count at transect 1,2 and 3.


Algae Ecolianalysis 2013-07-13 00:00:00 to 2013-07-13 07:00:00
W EC 1
EC 1 (MPN/100mL)
(a)
W EC 2
EC 2 (MPN/100mL)
(b)
W EC 3
EC 3 (MPN/100mL)
(c)
Algae 20'1
Algae 1 (scale)
(d)
Algae 20'2
Algae 2 (scale)
(e)
Algae 20'3
Algae 3 (scale)
(f)
(g) LPCF n_h=6
00:00 00:30 01:00 01:30 02:00 02:30 03:00 03:30 04:00 04:30 05:00 05:30 06:00 06:30 07:00 07:30 08:00

(A)


Algae ENTanalysis 2013-07-13 00:00:00 to 2013-07-13 07:00:00
W ENT 1
ENT 1 (MPN/100mL)
(a)
W ENT 2
ENT 2 (MPN/100mL)
(b)
W ENT 3
ENT 3 (MPN/100mL)
(c)
Algae 20'1
Algae 1 (scale)
(d)
Algae 20'2
Algae 2 (scale)
(e)
Algae 20'3
Algae 3 (scale)
(f)
(g) LPCF n_h=6
00:00 00:30 01:00 01:30 02:00 02:30 03:00 03:30 04:00 04:30 05:00 05:30 06:00 06:30 07:00 07:30 08:00

(B)

Figure 3.10. **(A)** CANARY output for *E. coli* count in water sample as well as level of Algae at three transects during the sampling period; (a), (b), (c) Plot of *E. coli* count at transect 1,2,3 respectively, (d), (e), (f) Plot of algae level at transect 1,2,3 respectively (g) Probability of event plot indicating 6 events detected by CANARY. **(B)** CANARY output for Enterococci count in water sample as well as level of algae during the sampling period; (a), (b), (c) Plot of Enterococci count at transect 1,2,3 respectively, (d), (e), (f) Plot of algae level at transect 1,2,3 respectively (g) Probability of event plot indicating 5 events detected by CANARY.

In addition, average bacteria count during summer from water sample and sand sample in swash zone indicated a positive correlation between the algal presence and bacteria count. From Table 3.6, it appears that the gradually increasing rating scale of algae at the three transect locations was positively correlated as the bacteria average count increased with the level of algae. However, there was an exception to that relation for Enterococci count in water sample. In addition, results from Pearson correlation analysis indicated strong positive correlation between algae level and bacteria concentration with R- value within a range of 80%-95% as shown in Table 3.7. The correlation between algae level and *E. coli* in water and swash zone sample were statistically significant at 0.05 level of confidence. However, for Enterococci, the correlation was not statistically significant (p-value > 0.05). The reason for this might be the small sample size. With the increased sample sizes, it would be possible to interpret the results for Enterococci more significantly.

Table 3.6. Average bacteria count (MPN/100 mL) with different algae level (scale 0-3) in water sample and sand sample in swash zone.

Bacteria	Sampling location	Algae level (scale)			
		0	**1**	**2**	**3**
E. coli	Water	25	102	262	612
	Swash Zone	126	232	257	618
Enterococci	Water	7	115	57	507
	Swash zone	15	50	81	1045

Table 3.7. Pearson Correlation analysis between the algae level and bacteria count

Sample 1	**Sample 2**	**Correlation**	**P-Value**
E. coli in water	Algae level	0.952	0.048
E. coli in swash zone	Algae level	0.905	0.035
Enterococci in water	Algae level	0.817	0.183
Enterococci in swash zone	Algae level	0.808	0.192

3.4 Discussion

In this study, slightly higher *E. coli* counts were measured when DI water was used as the eluent compared to PBS. An important factor in bacterial adhesion to sand in a beach environment is the electrostatic interactions between the water chemistry and bacteria, which can be predicted by traditional Derjaguin–Landau–Verwey–Overbeek (DLVO) theory. Under most conditions, *E. coli*, Enterococci, and quartz sand are characterized by electronegativity, with Enterococci being the most electronegative (Chen and Walker, 2012, Yamahara et al., 2007).

The surface charge of mineral particles depends on the pH of water. Given the fact that the point of zero charge (PZC) of quartz is 3 and the pH of DI and PBS solution is circumneutral, it is difficult to explain the difference in bacterial elutriation between DI and PBS. It appears that the bacteria–mineral particle interaction is determined by the cell membrane charge based on ionic strength and ionic species. Yee et al. (1999) demonstrated the adsorption of bacteria onto mineral surfaces as a function of time, pH, ionic strength, and the bacteria/mineral mass ratio. The data indicated that ionic strength had a strong impact on the interaction between the bacteria and mineral surface. In addition, the adsorption behavior was controlled by the chemical speciation on the surfaces. In this study, the results indicate that the attachment of *E. coli* to sand was greater than that of Enterococci, which agrees with the energy profiles calculated by DLVO theory, with *E. coli* predicted to have a lower energy barrier to attachment than Enterococci. Another factor that affects bacterial adhesion to sand is water ionic strength. The repulsive force between sand and indicator bacteria, both of which are electronegative, decreases as the water ionic strength increases. This explains the slightly higher detachment rate for *E. coli* when DI water was used as the eluent compared to PBS.

In addition to eluent ionic strength, phosphate in PBS, appears to affect bacterial adhesion as the zeta potentials of quartz sand and indicator bacteria decrease because of phosphate. Previous research showed that the addition of phosphate caused the release of attached *E. coli* from quartz sand (Wang et al., 2011). It is also possible that phosphate competes with bacteria for binding sites on the surface of sand particles. As a result, bacterial adhesion to the sand surface may be reduced. Although the attachment of bacteria to sand is an important factor, the survival of bacteria in a particular eluent is also important, i.e., the osmotic pressure on *E. coli*, which is a Gram-negative bacteria, can increase because of deionized water (Bayer,

1967). Although there is a possibility of increased *E. coli* cell death in DI water, DI water was found to be a more effective eluent than PBS in this study. This indicates that, for *E. coli* in DI water, bacterial adhesion efficiency has a greater effect than cell death when compared to PBS. Enterococci, as Gram-negative bacteria, can endure osmotic stress and are therefore mostly unaffected by the osmotic stress in DI water (Lleo et al., 2005).

Interactions between indicator bacteria and environmental factors are very complicated, which emphasizes the need to understand the role of the local environment as a source or sink for bacteria. People generally prefer sandy environments when visiting a water body for swimming and recreational purposes. This highlights the importance of understanding the interactions of environmental factors with beach sand, water, and bather health. Bacteria in water can be affected by many environmental factors. Sand can provide a protected environment that enables the survival of indicator bacteria. In a favorable environment with sufficient nutrients available, less sunlight inactivation, and protozoan grazing, the microbial community may grow, resulting in higher bacterial counts.

It has been observed that bacteria can survive longer in sand than in water because they can more easily adhere to sediment particles than free particles in water (Whitman et al., 2003a). Alm et al. (2003) reported that the cultivatable concentrations of *E. coli* and Enterococci in wet sands were 3-17 times and 4-38 times higher, respectively, than those in water. The presence of bacteria in sand can have significant effects on bacterial counts in water, which requires future research to provide a better understanding.

The results of this study indicate that the presence of algae is an important factor for higher microbial concentrations in beach water. A previous study reported the presence of green algae (*Cladophora glomerate*) along the shore of Lake Michigan beaches during the summer

months. Indicator bacteria such as *E. coli* and Enterococci can grow on the cell wall of *Cladophora* algae, which provides favorable conditions that provide sufficient nutrients, a grazing surface, and a suitable environment for attachment (Whitman et al., 2003a). Wave and wind actions may release microorganisms into the nearshore of beach sand and water. However, some other potential contamination sources, such as combined and sanitary sewer overflow, stormwater discharge from outfalls near beaches, wildlife residing or visiting in or near beaches, and runoff from parking lots and other impervious areas near beaches, may contribute to the higher bacteria (*E. coli*) concentrations detected at the sampling location. Moreover, rainfall events that occurred during the summer months and the stormwater runoff associated with the rainfall may lead to the transportation of contaminants, including fecal coliform, to beach sand and water. Beach water can be contaminated with fecal matter from humans, livestock, and other wildlife and result in microbial contamination of beach sand.

3.5 Conclusion

This study was undertaken to analyze bacterial concentrations and develop methods to accurately predict public health outcomes as a result of increased contamination with fecal indicator bacteria. The statistical analysis indicated the possibility of establishing a relationship between the *E. coli* results obtained using the two eluents, i.e., DI water and PBS; however, for Enterococci, the results were less promising due to high variation in the ratios of enumerated bacteria between the two methods. The ability to establish a ratio of bacterial counts among eluents would be a convenient tool with which to compare data collected using different eluents. CANARY may be useful as an early warning system for monitoring beach contamination and may help to identify abnormal conditions.

3.6 References

1. Alm, E. W., Burke, J., & Hagan, E. (2006). Persistence and potential growth of the fecal indicator bacteria, Escherichia coli, in shoreline sand at Lake Huron. *Journal of Great Lakes Research*, *32*(2), 401-405.
2. Alm, E. W., Burke, J., & Spain, A. (2003). Fecal indicator bacteria are abundant in wet sand at freshwater beaches. *Water research*, *37*(16), 3978-3982.
3. APHA & Federation, W. E. (2005). Standard methods for the examination of water and wastewater. *American Public Health Association (APHA): Washington, DC, USA*, *21*.
4. Bayer, M. E. (1967). Response of cell walls of Escherichia coli to a sudden reduction of the environmental osmotic pressure. *Journal of bacteriology*, *93*(3), 1104-1112.
5. Boehm, A. B., Griffith, J., McGee, C., Edge, T. A., Solo-Gabriele, H. M., Whitman, R., ... & Weisberg, S. B. (2009). Faecal indicator bacteria enumeration in beach sand: a comparison study of extraction methods in medium to coarse sands. *Journal of applied microbiology*, *107*(5), 1740-1750.
6. Bonilla, T. D., Nowosielski, K., Cuvelier, M., Hartz, A., Green, M., Esiobu, N., ... & Rogerson, A. (2007). Prevalence and distribution of fecal indicator organisms in South Florida beach sand and preliminary assessment of health effects associated with beach sand exposure. *Marine pollution bulletin*, *54*(9), 1472-1482.
7. Bordner, R., Winter, J. A., & Scarpino, P. (Eds.). (1978). *Microbiological methods for monitoring the environment: water and wastes* (Vol. 600). Environmental Protection Agency, Office of Research and Development, Environmental Monitoring and Support Laboratory.

8. Byappanahalli, M., Fowler, M., Shively, D., & Whitman, R. (2003). Ubiquity and persistence of Escherichia coli in a Midwestern coastal stream. *Applied and Environmental Microbiology*, *69*(8), 4549-4555.
9. Chen, G., & Walker, S. L. (2012). Fecal indicator bacteria transport and deposition in saturated and unsaturated porous media. *Environmental Science & Technology*, *46*(16), 8782-8790.
10. Doucette, G. J. (1995). Interactions between bacteria and harmful algae: a review. *Natural toxins*, *3*(2), 65-74.
11. Halliday, E., & Gast, R. J. (2011). Bacteria in beach sands: an emerging challenge in protecting coastal water quality and bather health. *Environmental science & technology*, *45*(2), 370-379.
12. Hart, D., McKenna, S. A., Klise, K., Cruz, V., & Wilson, M. (2007). CANARY: a water quality event detection algorithm development tool. In *World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat* (pp. 1-9).
13. Hartz, A., Cuvelier, M., Nowosielski, K., Bonilla, T. D., Green, M., Esiobu, N., ... & Rogerson, A. (2008). Survival potential of Escherichia coli and enterococci in subtropical beach sand: implications for water quality managers. *Journal of environmental quality*, *37*(3), 898-905.
14. Haxton T, Murray R, Hager J. (2013). CANARY Training Tutorials. Report EPA/600/R-13/201. U.S. Environmental Protection Agency Washington, DC: Office of Research and Development Publication.

15. Ishii, S., Ksoll, W. B., Hicks, R. E., & Sadowsky, M. J. (2006). Presence and growth of naturalized Escherichia coli in temperate soils from Lake Superior watersheds. *Applied and environmental microbiology*, *72*(1), 612-621.
16. Kasich, J. R., Taylor, M., & Butler, C. W. (2014). Public water system harmful algal bloom response strategy. *Ohio Environmental Protection Agency*.
17. Kleinheinz, G. T., McDermott, C. M., Hughes, S., & Brown, A. (2009). Effects of rainfall on E. coli concentrations at Door County, Wisconsin beaches. *International journal of microbiology*, *2009*.
18. Leow, A., Burkhardt, J., Platten III, W. E., Zimmerman, B., Brinkman, N. E., Turner, A., ... & Garland, J. (2017). Application of the CANARY event detection software for real-time performance monitoring of decentralized water reuse systems. *Environmental science: water research & technology*, *3*(2), 224-234.
19. Lleò, M. D. M., Bonato, B., Benedetti, D., & Canepari, P. (2005). Survival of enterococcal species in aquatic environments. *FEMS Microbiology Ecology*, *54*(2), 189-196.
20. Murray R, Haxton T. (2010). Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems: Development, Testing, and Application of CANARY. Report EPA/600/R-10/036. U.S. Environmental Protection Agency Washington, DC: Office of Research and Development Publication.
21. Nevers, M. B., Byappanahalli, M. N., Edge, T. A., & Whitman, R. L. (2014). Beach science in the Great Lakes. *Journal of Great Lakes Research*, *40*(1), 1-14.
22. Perelman, L., Arad, J., Housh, M., & Ostfeld, A. (2012). Event detection in water distribution systems from multivariate water quality time series. *Environmental science & technology*, *46*(15), 8212-8219.

23. Sampson, R. W., Swiatnicki, S. A., Osinga, V. L., Supita, J. L., McDermott, C. M., & Kleinheinz, G. (2006). Effects of temperature and sand on E. coli survival in a northern lake water microcosm. *Journal of Water and Health*, *4*(3), 389-393.
24. Thupaki, P., Phanikumar, M. S., Schwab, D. J., Nevers, M. B., & Whitman, R. L. (2013). Evaluating the role of sediment-bacteria interactions on Escherichia coli concentrations at beaches in southern Lake Michigan. *Journal of geophysical research: oceans*, *118*(12), 7049-7065.
25. U.S. EPA. (2014). Configuring Online Monitoring Event Detection Systems. Report EPA/600/R-14/254. U.S. Environmental Protection Agency Washington, DC: Office of Research and Development Publication.
26. U.S. EPA. (2012). Cyanobacteria and Cyanotoxins: Information for Drinking Water Systems. Report EPA/810/F-11/001. U.S. Environmental Protection Agency Washington, DC: Office of Water.
27. U.S. EPA. (2018). Five-Year Review of the 2012 Recreational Water Quality Criteria. Report EPA/823/R-18/001. U.S. Environmental Protection Agency Washington, DC: Office of Research and Development Publication.
28. Vestby, L. K., Grønseth, T., Simm, R., & Nesse, L. L. (2020). Bacterial biofilm and its role in the pathogenesis of disease. *Antibiotics*, *9*(2), 59.
29. Wade, T. J., Pai, N., Eisenberg, J. N., & Colford Jr, J. M. (2003). Do US Environmental Protection Agency water quality guidelines for recreational waters prevent gastrointestinal illness? A systematic review and meta-analysis. *Environmental health perspectives*, *111*(8), 1102-1109.

30. Wang, L., Xu, S., & Li, J. (2011). Effects of phosphate on the transport of Escherichia coli O157: H7 in saturated quartz sand. *Environmental science & technology*, *45*(22), 9566-9573.

31. Whitman, R. L., & Nevers, M. B. (2003). Foreshore sand as a source of Escherichia coli in nearshore water of a Lake Michigan beach. *Applied and environmental microbiology*, *69*(9), 5555-5562.

32. Whitman, R. L., Shively, D. A., Pawlik, H., Nevers, M. B., & Byappanahalli, M. N. (2003a). Occurrence of Escherichia coli and enterococci in Cladophora (Chlorophyta) in nearshore water and beach sand of Lake Michigan. *Applied and Environmental Microbiology*, *69*(8), 4714-4719.

33. Yamahara, K. M., Layton, B. A., Santoro, A. E., & Boehm, A. B. (2007). Beach sands along the California coast are diffuse sources of fecal bacteria to coastal waters. *Environmental science & technology*, *41*(13), 4515-4521.

34. Yamahara, K. M., Walters, S. P., & Boehm, A. B. (2009). Growth of enterococci in unaltered, unseeded beach sands subjected to tidal wetting. *Applied and environmental microbiology*, *75*(6), 1517-1524.

35. Yau, V., Wade, T. J., de Wilde, C. K., & Colford, J. M. (2009). Skin-related symptoms following exposure to recreational water: a systematic review and meta-analysis. *Water Quality, Exposure and Health*, *1*(2), 79-103.

36. Yee, N., Fein, J. B., & Daughney, C. J. (2000). Experimental study of the pH, ionic strength, and reversibility behavior of bacteria–mineral adsorption. *Geochimica et Cosmochimica Acta*, *64*(4), 609-617.

CHAPTER 4: USING CANARY EVENT DETECTION SOFTWARE FOR WATER QUALITY ANALYSIS IN THE MILWAUKEE RIVER

4.1 Introduction

Urban water sources are susceptible to various contamination events as a result of natural occurrences such as algal bloom, urban runoff, accidental industrial and transportation-related spills, and intentional spills such as illegal disposal of organic chemicals and hazardous materials. Early Warning Systems (EWS) are used mostly on riverine systems where water quality can change more rapidly (Gullick et al., 2004), thus provide timely information on any changes in source water quality that allows decision-makers to take action against the contamination event to mitigate human health risks and other hazardous outcomes (Gullick et al., 2003). In 1977, the first modern EWS to detect source water contamination was developed after a leakage from a chemical storage facility into the Ohio River, which led to the development of the Ohio River Valley Water Sanitation Commission (ORSANCO) Organics Detection System (ODS). Enhanced with advanced monitoring and modeling technologies, it served as a model in the development of subsequent regional source water EWS (U.S. EPA, 2017). There are several applications of EWSs that enclose multiple surface water sources including the Delaware River Basin, Lake Erie, the Lake Huron to Lake Erie corridor, the Lower Mississippi River Basin, the Ohio River Basin, the Susquehanna River Basin, the Upper Mississippi River Basin and others (Michigan DEQ, 2007; Regional Response, 2008; Stumpf et al., 2012; Louisiana DEQ, 2014; Exchange Network, 2015). EWS has become an integral component of the water operation system, allowing the detection of incidental contamination events as well as naturally occurring accidental events. According to Water Research Foundation, there are several components in an effective EWS, including a mechanism for detection of the presence of contaminants, a means

for confirming and determining the nature and concentration of the contamination in the water distribution system, a communication protocol for transferring information about the event, mechanisms that allow response to the presence of the contamination in the effluent water to reduce impacts on end-users and an institutional framework with a centralized unit for managing contamination events (Grayman et al., 2001; Hasan et al., 2004).

4.1.1 Event detection system (EDS)

One of the important components of EWS is 'detection' which is a mechanism for identifying the presence of contaminant events in source water. Detection includes continuous, periodic, or sporadic monitoring and reporting of contamination incidents. The development of automated EDS enables the use of time-series information from different monitoring systems to identify anomalous behavior indicating and reporting as contamination events to the EWS operators (U.S. EPA, 2017). EDS uses different algorithms to detect any event or abnormality in data. Signals from multiple water quality sensors are transmitted to a central database, commonly Supervisory Control and Data Acquisition (SCADA), and stored in the database. EDS can access data and analyze real-time data to identify and filter out variations from background water quality. An alarm for an event is triggered if any significant change in water quality is identified with a duration of a one-time step or more. Several EDSs are available for water quality monitoring and event detection, including:

- CANARY- Sandia National Laboratory and US EPA
- GuardianBlue Early Warning System- HACH company
- BlueBoxTM – Whitewater Security (Barbanti et al., 1995)
- OptiEDS- OptiWater

Among these systems, CANARY, developed by Sandia National Laboratory and U.S.EPA (Hart et al., 2007; U.S. EPA, 2008), is a freely available event detection tool to water utilities and researchers to better understand the normal background variability and to identify anomalies that are potentially indicative of contamination incidents. Others are commercial EDSs often supplied with hardware systems sold by companies and with limited public documentation (Zaefferer, 2012).

4.1.2 CANARY EDS

CANARY consists of a two-stage approach to event detection. The first stage is the prediction of future water quality value referred to as the state estimation. The predicted value is then compared with the observed value in the second stage. The difference between the predicted and observed values is known as the residual and classification is made to determine whether the water quality at that specified time-step is expected or anomalous. CANARY algorithms process the data at each time step to identify the periods of anomalous water quality and the result indicates the probability of water quality event existing at that time-step (U.S. EPA, 2010). CANARY can be run in online mode or offline mode. In offline mode, previously collected datasets are examined, whereas, in online mode, the distribution network SCADA system is used to provide data from real-time monitoring. The offline mode is used with the best combination of event detection parameters. There are no limits within CANARY to the maximum number of water quality signals at any monitoring location or the number of monitoring locations analyzed simultaneously (McKenna et al., 2008).

CANARY EDS was developed to enhance drinking water contamination warning systems that can detect anomalous events in treated drinking water and mitigate the risk of

human health. In water industries, the anomalous events are referred to as significant changes or degradation of water quality within a water distribution network due to accidental or intentional contamination events including pipe breaks that introduce material into water system, sewer cross-connection, leaching, and injection of chemicals and biological contaminants that require response decisions and rapid actions from water utility operators. Surface water sources are also susceptible to various anomalous or contamination events such as chemical, fuel, and sewage spills that cause significant deviations in measured values of water quality signals than the expected values. CANARY is designed to detect these contamination events as a period of anomalous activity with multiple outliers detected over a minimum specified number of time-steps. Using time-series data from multiple locations and sensors, CANARY can differentiate real water quality events from background variability based on statistical models. A real water quality event is referred to as a real-time anomaly that occurred in water quality monitoring data due to the occurrence of contamination events in the source water. CANARY identifies an 'event' and triggers an alarm when there is a sudden and significant change in average measured values over multiple time-steps that deviates from the expected background water quality.

Several research studies have been made with CANARY for drinking water systems. A previous study (Leow et al., 2017) used CANARY for wastewater treatment monitoring. However, studies on the application of CANARY in surface water monitoring are yet to be developed. To the best of the authors' knowledge, this study is the first to use CANARY software for river water quality monitoring. We performed analyses in the offline mode of CANARY where previous historical data were used to observe any sudden significant change in the averaged measured values over multiple time-steps. However, in online mode, CANARY can analyze datasets in real-time, differentiate real water quality events from background variability,

and alert the operator of possible contamination by triggering alarm very quickly. This study suggested that CANARY can be useful for detecting significant deviations in water quality signals due to the occurrence of anomalous activities in surface water. The application of the software can be more effective in online mode for automatic real-time event detection. Although in this study, CANARY was not used in the online mode using real-time data with the occurrence of known contamination events such as industrial, municipal, and sewage waste spill, the effectiveness of the software was observed using historical data for detection of anomaly or 'events' in water quality signals that can be useful for real-time monitoring of surface water quality in future. Several commercially available EDS systems focus on connecting the capabilities of the EDS to their proprietary sensor hardware. These systems operate with sensors from a single manufacturer that is incompatible with water utilities where sensors from different manufacturers need to be operated simultaneously for water quality monitoring. CANARY can work with any sensor hardware from any manufacturer that reduces additional installation costs of utilities for specific sensor hardware. CANARY performs centralized analysis, which allows transmission of data to a single location, while some other EDS require installation at the actual monitoring location for performing local analyses. CANARY is a free publicly available software that is accessible to water utilities and researchers, while other EDS software packages are commercially available with limited public documentation.

The objective of the study is to analyze anomalous water quality events for Milwaukee river using CANARY based on the available monitoring data of pH, conductivity, and turbidity. Also, this study was conducted to get an insight into the effectiveness of the application of CANARY to natural source water for detecting real-time water quality events.

4.2 Materials and methods

4.2.1 Study area and data collection

The study area is located along Milwaukee River in Milwaukee, Wisconsin (Lat 43°01'28", long 87°53'54", in SW 1/4 NE 1/4 sec.33, T.7 N., R.22 E., Milwaukee County, Hydrologic Unit 04040003), with a drainage area of 872 square miles (Figure 4.1). Sources of contamination impacting the stream site include a large population of shorebirds and stormwater outfalls that discharge surface runoff from nearby streets. Water quality data, i.e., pH, turbidity, and conductivity from Milwaukee River were collected from monitoring station 04087170 of USGS National Water Information System. River water quality data during the summer months of 2018, 2019, and 2020 were used for data analysis with CANARY. Data from two sampling periods: August 1-7, 2019 and August 25-27, with a data interval of 10 minutes, were also used in this study to evaluate the effects of rainfall on water quality parameters.

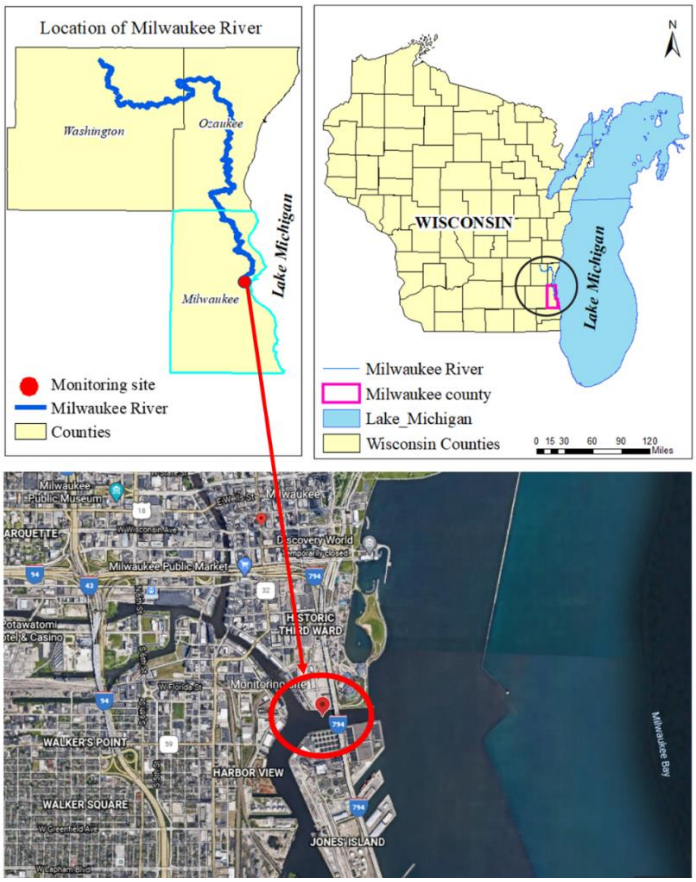


Figure 4.1. Location of the Milwaukee River and monitoring site in Wisconsin

4.2.2 Data analysis model

CANARY (U.S. EPA 2010; Hagar et.al 2013) was used to analyze river water quality data in the offline mode using both Linear Prediction Correction Filter (LPCF), the most commonly used algorithm to identify the onset of anomalous water quality period, and Multivariate Nearest Neighbor (MVNN) algorithm. LPCF produces an alarm when the water quality signals have a prolonged deviation from the baseline. Since the water signals had different magnitude and units (e.g., mg/L, NTU, etc.), the data were normalized to have a mean of 0.0 and standard deviation of 1.0, so that different signal measurements can be combined. A data point is considered an outlier when the normalized residual, calculated as the difference

between the actual and predicted value exceeds the threshold value (Leow et al., 2017). Once identified as an outlier, the data point would not be used for predicting future signal values. A probability was calculated based on the number of outliers in the user-defined binomial event discriminator (BED) window. An alarm was created when the probability exceeded the user-defined threshold. For event detection, CANARY uses binomial distribution theory that produces a binary result for each time-step (U.S. EPA, 2010). The results of multiple successive time steps were combined into a time-integrated probability of event P (event) by employing the BED. P (event) was defined as a function of the number of outliers within the BED window, the length of the BED integration window, and the probability of an outlier at any given time step under the background water quality conditions. Due to the integration of results over multiple time steps, an additional lag time occurs between the true onset of the events and the time at which an event was detected (U.S. EPA, 2010).

Using the MVNN algorithm, the newest measured water quality data was compared against all the other water quality data in the history window and the newest measured value was plotted against the other data. The normalized water quality data were mapped in an m-dimensional multivariate space, where m was the number of water quality signals. Water quality values in the history window were classified into different classes or clusters. The distance between a new data point and the centroid of each existing cluster was calculated as a Euclidian measure. The residual was calculated as the Euclidean distance from the new point to the closest neighborhood historical data point. The measured distance was not a function of any individual water quality signal but a combined measure of distance using all water quality signals (U.S. EPA, 2010; U.S. EPA, 2012).

Analysis of a specific set of data using CANARY requires the optimization of configuration parameters, which involves selecting the ideal set of parameters for the event detection algorithm and adjusting the BED parameters to increase or decrease the probability of generating an alarm. The configuration file consists of the optimized parameters, i.e., history window, BED window, outlier threshold, and event threshold, with the window size and the threshold most significant to the model performance (U.S. EPA, 2013; U.S. EPA, 2014).

4.2.3 Statistical analysis

Statistical analysis was carried out using Minitab to evaluate whether rainfall events had any impact on river water quality. One sample-Z test was performed to measure the significance (p-value) with a 95% confidence level and the effect of rainfall was evaluated. A simple linear regression method was used to analyze the correlation between water quality parameters and temperature during the sampling period.

4.3 Results and discussion

4.3.1 Model sensitivity analysis

Figure 4.2 shows the sensitivity analysis regarding window size and threshold using pH, conductivity, and turbidity data from Milwaukee River during the sampling period of August 25-August 27, 2019. Figure 4.2(a) shows the relationship between the standard deviation of the residuals and window sizes ranging from 36-time steps to 216-time steps with the LPCF algorithm. The parameters that control the integration of results across consecutive time steps

using the BED algorithm were held constant across all test runs. CANARY outputs give the absolute value of the residual between the observed and predicted water quality data. From these residual values, the standard deviation of the residual for each of the window sizes was calculated and plotted against the corresponding window size. Lower value of the performance measure: standard deviation of the residual indicates increased precision and accuracy in future prediction of the signal value. The window size that produces the lowest value of the standard deviation of the residuals would be a better choice for the selection of 'history window'. However, increased window size would result in a longer computational time for updating the parameters and predicting the future water quality at each time step. Also, if the history window becomes too large the computational load could increase to the point of making real-time estimation impractical (U.S. EPA, 2010). In this consideration and based on the number of sample observations, we avoided a larger window size. Results shown in Figure 4.2(a) indicate that a window size of at least 72-time steps (720 minutes) is required to reduce the standard deviation of the residuals to near their final minimum value for each of the water quality parameters, i.e., previous water quality data from the past 12 hours could be used for prediction of the future water quality value at the next time-step.

From Figure 4.2(b), it can be observed that the window size significantly influences the total number of events reported. With the larger window length, there would be a possibility of false-negative alarms meaning that no events will be detected during a true event condition. The threshold value, a multiplier of a signal's standard deviation, is required to classify residuals as being indicative of either background water quality or outlier. An outlier threshold value of 1.0 indicates that a signal value greater than one standard deviation from the mean signal value is labeled as an outlier that can contribute to an alarm (U.S. EPA, 2014). Figure 4.2(c) indicates

that the number of detected events decreases with the increase of outlier threshold value as a higher value of the outlier threshold results in fewer outliers. The accuracy of the prediction will degrade if too many outliers are excluded from the estimation of future value (U.S. EPA, 2010). In most cases, the data respond well to a lower outlier threshold value, so it is worth including a threshold value lower than one. In the river water quality analysis, the outlier threshold value was set to be 0.80. For the typical timeframe range: 30 minutes to 120 minutes, the event threshold was calculated from the BED window and the number of outliers, using the binomial probability distribution function.

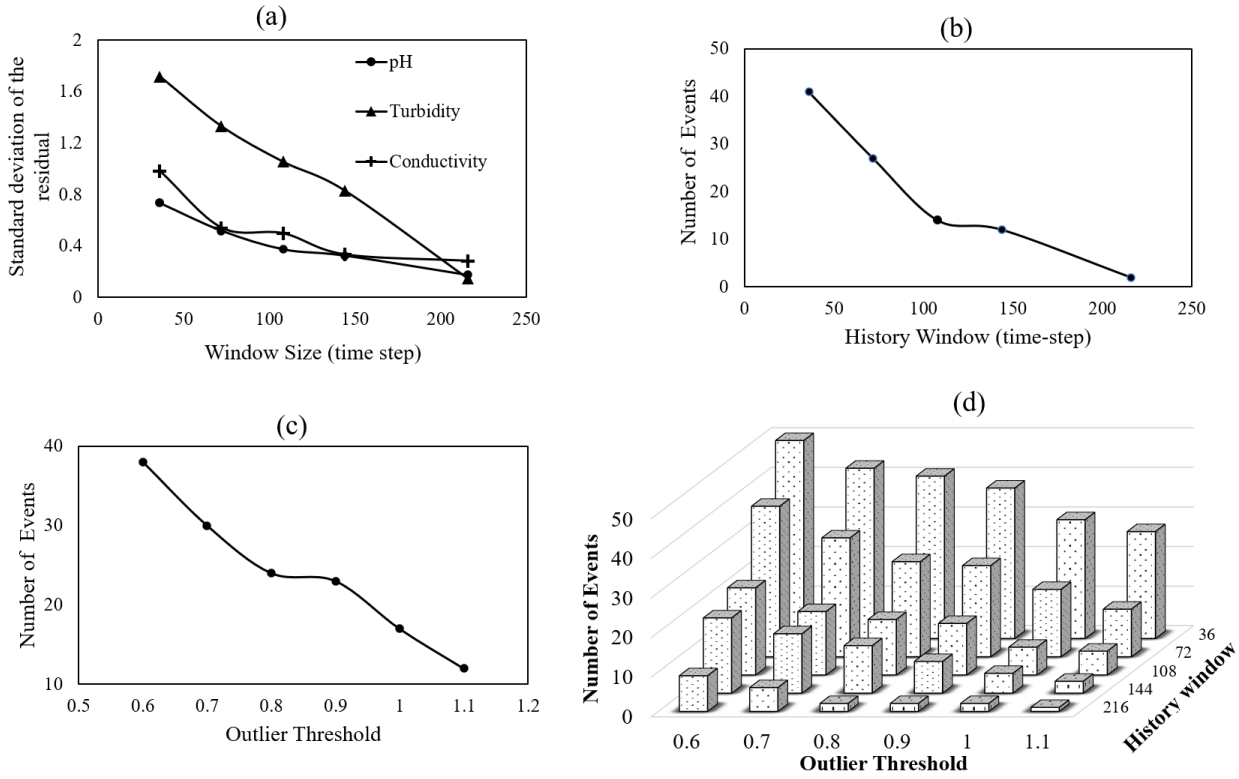


Figure 4.2. (a) Standard deviation of residuals with different window sizes (b) Number of detected events with different window sizes (c) Number of detected events with variations of outlier threshold (d) Total detected events for different combinations of configuration parameters

Figure 4.2(d) examines the trend of detected events as a result of 30 different combinations of history window and outlier threshold parameter values. It indicates that the decrease in the number of events is close to a linear function of the outlier threshold parameter, while the number of events decreases more rapidly with an increasing history window. In the case of data with no available real events, it might be more effective to allow for a slightly larger number of detected events to ensure detection is sensitive enough to report true hazardous contamination events (U.S. EPA, 2013a). Table 4.1 shows the optimal/selected algorithm configuration parameters for the river data analysis with a data interval of 10 minutes selected based on the sensitivity analysis.

Table 4.1. CANARY algorithm configuration parameters

Configuration parameters	**Optimal/Selected value**
History Window	72
Outlier Threshold	0.80
BED window	3
Event Threshold	0.875

4.3.2 Analysis of Milwaukee River water quality

Analysis was made on Milwaukee River water quality data during two time periods, early and late August 2019. The overall duration for the late August sampling period was 3 days with an interval of 10 min. The date-time start for the data collection was 08/25/2019 00:00:00 (mm/dd/yyyy HH: MM: SS) and the date time-stop was 08/27/2019 23:50:00. The analysis was performed in the offline mode using both LPCF and MVNN algorithm. The graphical interface shows where events were detected and the raw data values at that time. An “EVENT” is the cumulative set of outliers. Each signal outlier was marked with a blue dot in the signal plot, and

the “EVENT” was each of the sets of blue dots in the event probability plots. The number of blue dots in the event determined the duration of that event.

Figure 4.3 to Figure 4.10 show the detailed water quality analysis during the time-period from 08/25/2019 08:00:00 to 08/27 22:50:00 with both LPCF and MVNN algorisms. In Figure 4.3, two events were detected with LPCF and one with MVNN from 08/25/2019 08:00:00 to 15:50:00. More outliers were detected during 08/25/2019 16:00:00 to 23:55:00, three with LPCF and one with MVNN (Figure 4.4). It appears that during these two time periods, the number of detected events with the LPCF algorithm was higher than the ones reported with the MVNN algorithm. However, both algorithms reported the same number of events (three) during the time period 08/26/2019 00:00:00 to 07:50:00 (Figure 4.5). From 08/26/2019 08:00:00 to 08/26/2019 15:50:00, no event was detected with LPCF while two events were found with the MVNN algorithm (Figure 4.6). It can be observed that from 08/26/2019 after time-step 16:00:00 to 08/27/2019 23:50:00, a significantly higher number of events was detected (Figure 4.7 to Figure 4.10) than the previous time periods. This indicates that there was a sudden change in water quality during that time period. As a result, CANARY reported a total number of 24 events with pH contributing to 17 of the 24 events, turbidity 12 of the 24 events, and conductivity 14 of the 24 events with the LPCF algorithm. All 24 events might have contributions from one or all three signals. With the MVNN algorithm, the same number of total events were detected with pH contributing to 18 of the 24 events, turbidity 14 of the 24 events, and conductivity 16 of the total 24 events. It appears that both LPCF and MVNN algorithms detected the same number of total events although there was a small variation in the number of contributing events from each signal.

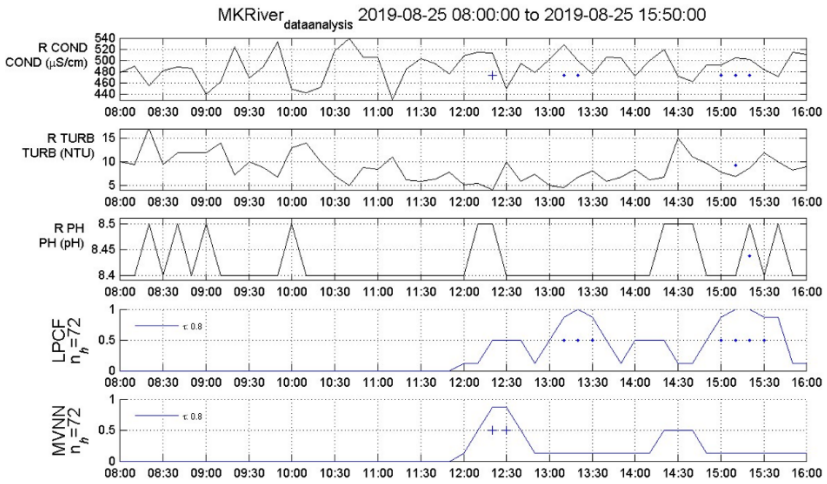


Figure 4.3. CANARY output (hourly variation) from time period 2019-08-25 08:00:00 to 2019-08-25 15:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 2 (LPCF) and 1 (MVNN).

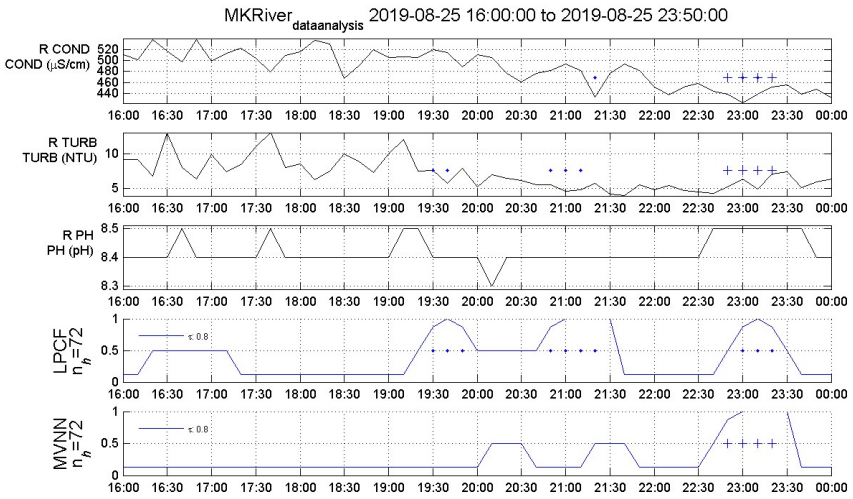


Figure 4.4. CANARY output (hourly variation) from time period 2019-08-25 16:00:00 to 2019-08-25 23:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 3 (LPCF) and 1 (MVNN).

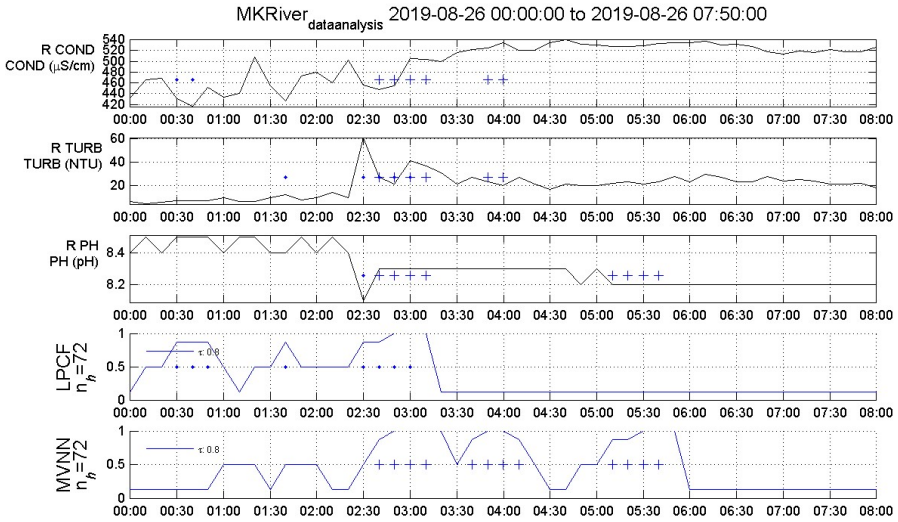


Figure 4.5. CANARY output (hourly variation) from time period 2019-08-26 00:00:00 to 2019-08-26 07:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 3 (LPCF) and 3 (MVNN).

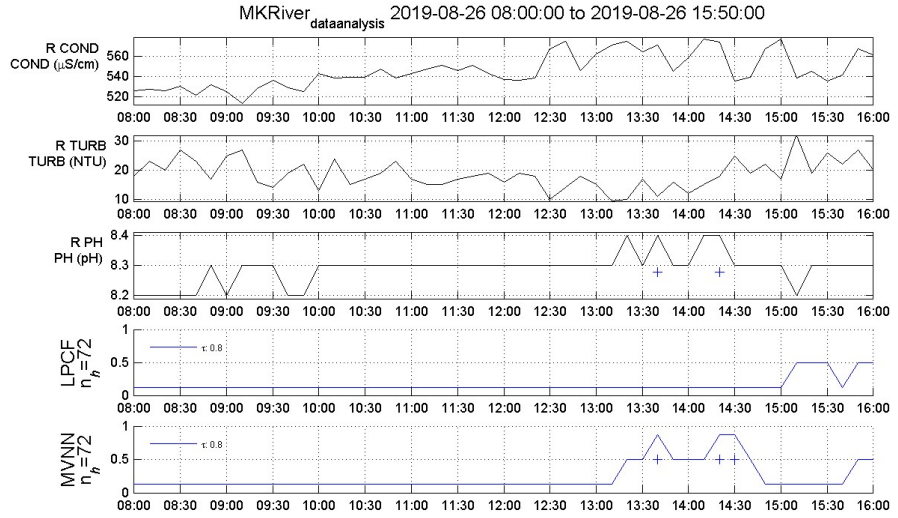


Figure 4.6. CANARY output (hourly variation) from time period 2019-08-26 08:00:00 to 2019-08-26 15:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 0 (LPCF) and 2 (MVNN).

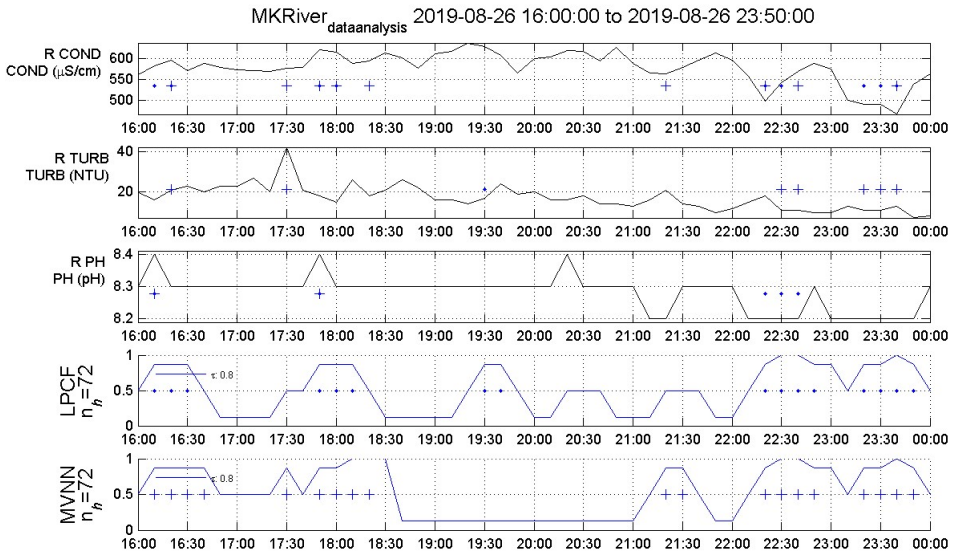


Figure 4.7. CANARY output (hourly variation) from time period 2019-08-26 16:00:00 to 2019-08-26 23:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 5 (LPCF) and 6 (MVNN).

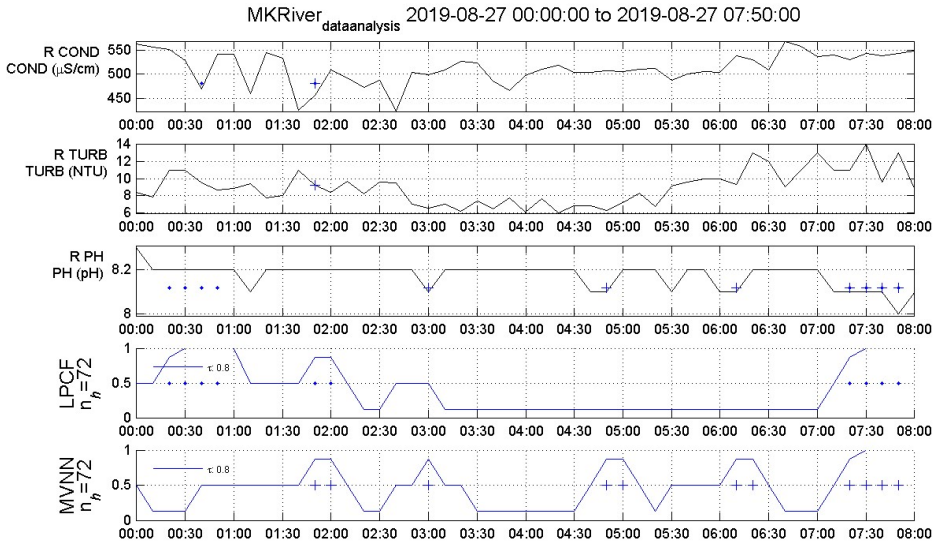


Figure 4.8. CANARY output (hourly variation) from time period 2019-08-27 00:00:00 to 2019-08-27 07:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 3 (LPCF) and 5 (MVNN).

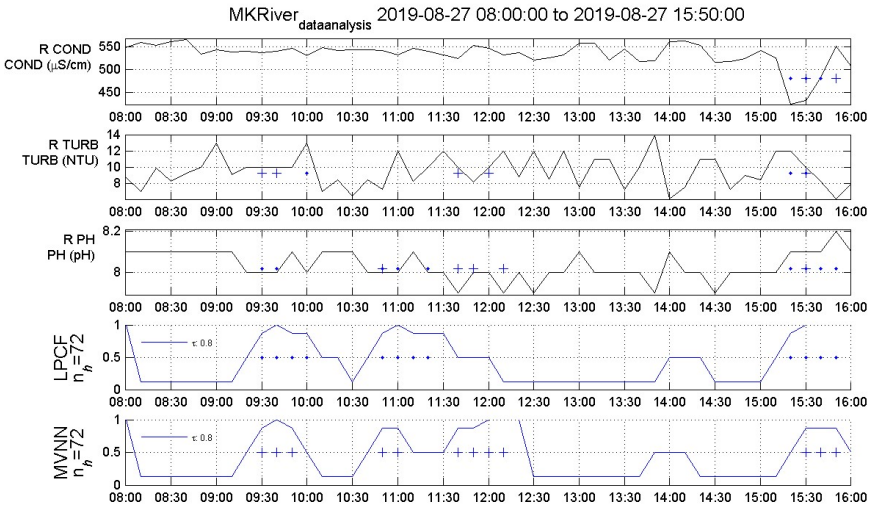


Figure 4.9. CANARY output (hourly variation) from time period 2019-08-27 08:00:00 to 2019-08-27 15:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 3 (LPCF) and 4 (MVNN).

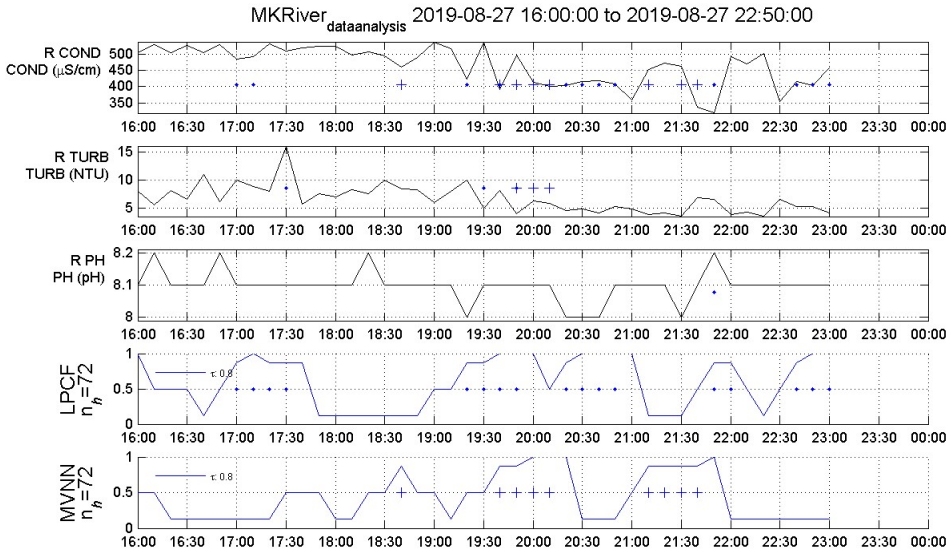


Figure 4.10. CANARY output (hourly variation) from time period 2019-08-27 16:00:00 to 2019-08-27 22:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Number of detected events: 5 (LPCF) and 3 (MVNN)

Additional analysis was made on CANARY during the week of August 1- 7, 2019 with a data interval of 10 min (Figure 4.11). For data from 08/01/2019 00:00:00 to 08/07/2019 22:00:00 the water quality parameters showed deviation or abnormality resulting in 21 total events with pH contributing to 13 of the total 21 events, turbidity 13 of the 21 events and conductivity 9 of the 21 events with LPCF. Using the MVNN algorithm, a total of 12 events were detected with each of the water quality signal contributing to 7 of the 12 events. It appears that during the first week of August, the LPCF algorithm detected higher number of events than the MVNN. Even though the two algorithms identified different number of events, each of them detected more events during the late August sampling period than in the early August period.

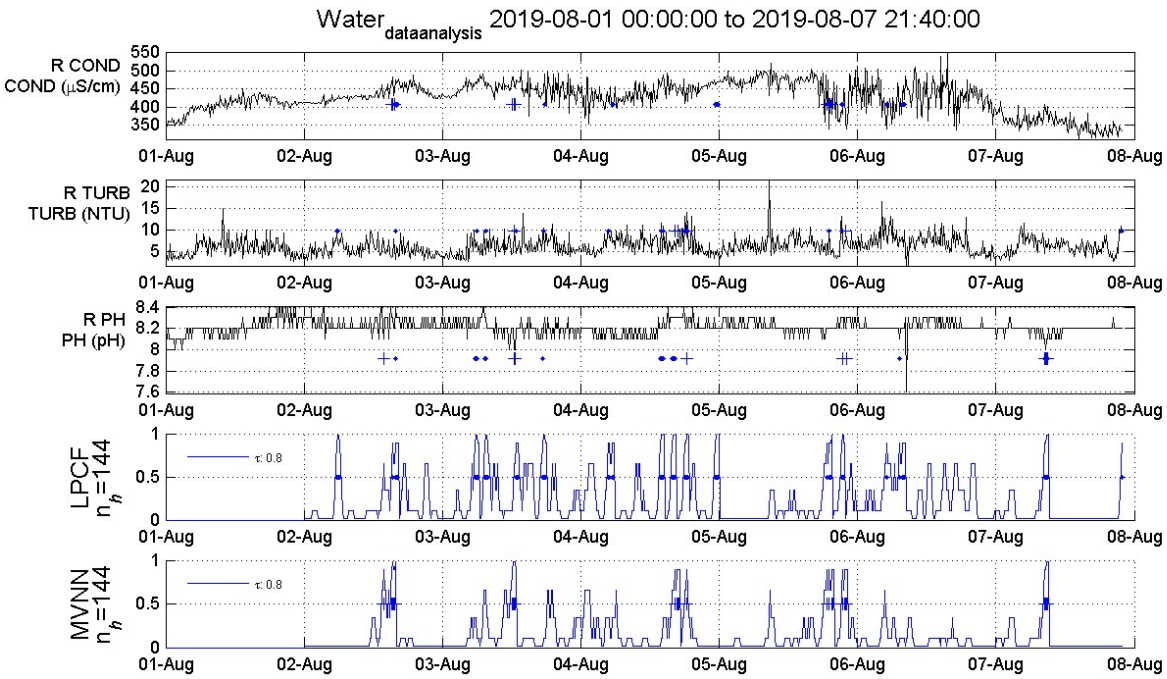


Figure 4.11. CANARY output (weekly variation) from time period 2019-08-01 00:00:00 to 2019-08-07 21:40:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Total Number of detected events: 21 (LPCF) and 12 (MVNN)

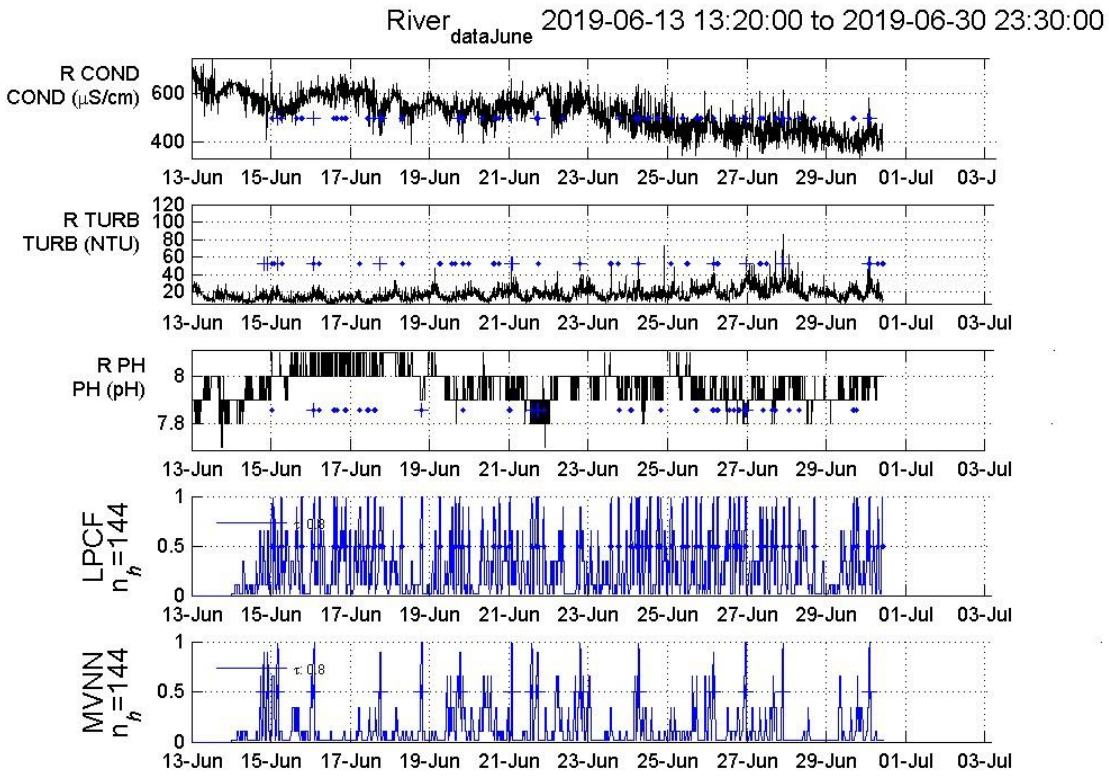


Figure 4.12. CANARY output from time period 2019-06-13 00:00:00 to 2019-06-30 23:30:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Total number of detected events: 67 (LPCF) and 17 (MVNN).

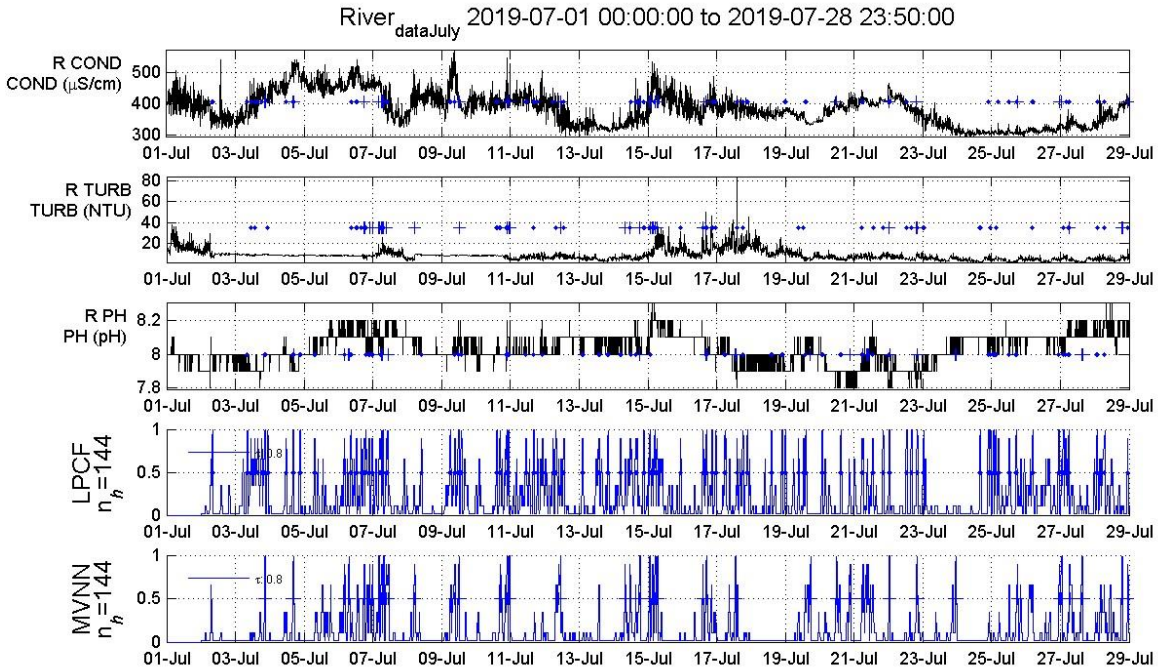


Figure 4.13. CANARY output from time period 2019-07-01 00:00:00 to 2019-07-28 23:50:00 with three signals (conductivity, Turbidity, pH) and two probability of event plots using LPCF and MVNN algorithms for Milwaukee River data. Total Number of detected events:120 (LPCF) and 54 (MVNN).

The results from the EDS indicated significant variation in water quality with a considerable number of detected events, especially during late August. Table 4.2 show the maximum, minimum and mean value of each of the water quality parameters during August 1-7, 2019 and August 25-27, 2019 respectively. It can be observed that the ranges of maximum and minimum values for the signals in late August were much higher than those of early August, especially for turbidity and conductivity. The mean value of turbidity increased from 6.18 NTU to 11.64 NTU. There was also a rapid change in mean conductivity value from early to late

August (426.3 µS/cm to 509.7 µS/cm). The analysis suggests that the river water quality changed mostly due to variation in turbidity and conductivity from the beginning to the end of August.

Table 4.2. Maximum, minimum, and mean value of water quality parameters

Parameter	**August 1 - August 7**			**August 25 - August 27**		
	Max	Min	Mean	Max	Min	Mean
pH	8.4	7.6	8.21	8.6	7.9	8.28
Conductivity (µS/cm)	551	313	426.3	637	319	509.7
Turbidity (NTU)	21.6	1.6	6.18	61	3.5	11.64
Temperature ^{o}C	20	10.7	15.5	22.5	14.2	21.2

Summer seasonal data analyses were performed using CANARY during 2018, 2019, and 2020. The CANARY outputs for the sampling period from mid-June to July 2019 are shown in Figure 4.12 and Figure 4.13. It appears that LPCF detected a higher number of events than MVNN algorithm. During the time period of 6/13/2019 to 6/30/2019, a total of 67 events were detected with pH contributing to 44 events, turbidity 44 events, and conductivity 58 events when using LPCF algorithm; while with MVNN algorithm, only a total of 17 events were detected with pH contributing to 7 events, turbidity 13 events and conductivity 14 events. During July, a total of 120 events were detected with LPCF algorithm with pH contributing to 68 events, turbidity 65 events, and conductivity 62 events; while with MVNN, a total of 54 events were detected with pH contributing to 32 events, turbidity 34 events, and conductivity 35 events. Similarly, we performed analysis using pH, conductivity, and turbidity monitoring data during the summer months of 2018 and 2020 and the results indicated that LPCF detected a higher number of events than MVNN as shown in Table 4.3.

Table 4.3. Total number of detected events by LPCF and MVNN algorithm during different sampling periods and the number of detected events contributed by each water quality signal

Sampling period	Algorithm	Total detected Events	Contributing signals		
			Conductivity	Turbidity	pH
June,2018	LPCF	53	37	41	38
	MVNN	15	12	11	13
July, 2018	LPCF	28	18	16	17
	MVNN	17	12	10	17
June,2019	LPCF	67	58	44	44
	MVNN	17	14	13	7
July, 2019	LPCF	120	62	65	68
	MVNN	54	35	34	32
June,2020	LPCF	47	35	33	25
	MVNN	15	11	11	8
July, 2020	LPCF	35	18	20	18
	MVNN	23	12	12	20

Analysis was performed using historical data and the information on the causes of the occurrence of the water quality events detected by CANARY during the sampling period was not available. CANARY identifies the presence of anomaly or events in water quality data, not the causes of anomaly. However, the significant deviations or the presence of anomalies in water quality signals could be due to wastewater discharge from several non-point and point contamination sources, including combined and sanitary sewer overflows, stormwater discharge, chemical and fuel spill, surface runoff from nearby streets, and seasonal effects to the region. Besides, the ‘events’ can be detected due to some invalid alerts such as normal variability, sensor or communication problems, and other insignificant causes which are plausible in most event

detection systems. A previous study (U.S. EPA, 2013a) also indicated that although CANARY resulted in some invalid alerts, the performance of the software in detecting valid alerts was fairly consistent.

The variation between the number of events reported from the LPCF and MVNN algorithms may be attributed to their different residual calculation mechanisms. With any two water quality signals, the LPCF algorithm will select the maximum value at each time-step for comparison to the threshold if the normalized residual between the observed and predicted value for each signal is one; whereas the MVNN algorithm calculates the Euclidean distance between the current and nearest previous observation. If the predicted water quality value is one standard deviation for both the signals, then the Euclidean distance for MVNN will be $\sqrt{2}\ or\ 1.41$. For a threshold value between 1 and 1.4, only the MVNN algorithm will identify the time-step as an outlier. In theory, a larger threshold is required for MVNN to get the same results as LPCF based on their threshold calculation mechanisms. (U.S. EPA, 2010)

It is also observed that more "blue dots" in the event plot can be found than the number of outliers at a few time steps. The 'dots' in the CANARY probability event plots are referred to as the detected 'events' by the CANARY algorithm. For example, on 08/26/2019 between time steps 22:30:00 and 23:00:00 (Figure 4.7), the probability plot for LPCF showed more blue dots than the outliers detected over that time-step, particularly at time step 22:50:00 where an additional blue dot can be found with no detected outliers in the signal plots. Depending on the value of the event threshold (tau in the probability plot), there can be more blue dots in the event than outliers in that event. For LPCF, at 22:30, one dot was found for conductivity and three for pH, but four dots were found in the event plot. This is because the event threshold is a moving window, and the probability leveled off but still exceeded the predefined threshold. The

probability of an event is a function of the number of outliers within the BED, length of the BED integration window, and probability of an outlier at any given time step. The probability was indicating some number of outliers/BED window exists, as a result, it sometimes gave the appearance of a prolonged event.

4.3.3 Statistical analysis of rainfall effect on water quality

Several climate factors are considered as possible triggers for the variation in river water quality. Precipitation data from NOAA online weather were collected over the sampling periods. During the first week of August 2019, there was only a trace amount of daily precipitation resulting in total accumulated precipitation of 1.20 inches, while over the sampling period of August 25-27, 2019 the total accumulated precipitation was 1.70 inches. On August 26, there was a sudden increase in accumulated precipitation value due to a rainfall event that occurred at the end of that day. As a result, there was a sudden change in the water quality signals during and after the rainfall event. CANARY also reported a higher number of events during the time period 08/26/2019 16:00:00 to 08/27/2019 23:50:00.

Seasonal water quality data of Milwaukee River during summer 2019 (June-August) were collected from the USGS data source. The mean seasonal values of pH, turbidity, and conductivity were determined by averaging all individual parameter values from which the mean values were found as pH 8.08, turbidity 9.58 NTU, conductivity 439.98 µS/cm. During our sampling duration of August 25-27, the mean and standard deviation (µ, σ) of each of these water parameters were found as pH 8.28 (0.15), turbidity 11.64 (7.06), and conductivity 509.71 (48.33). Statistical analysis (hypotheses testing: One sample Z-test) was performed with Minitab

to evaluate if there was a statistical difference between the seasonal mean and sample mean and significance (p-value) was measured. For the hypothesis testing, the null (H_o) and alternate hypotheses (H_1) were set as:

H_0: Rainfall does not affect water quality

H_1: Rainfall does affect water quality

The statistical analysis indicated that there was a statistically significant difference between the seasonal mean and sampling period mean of the water quality parameters, which concludes rainfall did affect water quality (p-value=0.000<0.05). Several potential contamination sources may be associated with the rainfall event, as the monitoring site is located in an urban area. These sources include overflows from combined and sanitary sewers, discharges of stormwater from outfalls near the stream, runoff from parking lots, and other impervious areas adjacent to the stream, which may transport many pollutants into the river.

4.3.4 Regression analysis of water quality with temperature

From CANARY results (Table 4.3), it can be observed that the total number of detected events increased from mid-June to July 2019, which might be correlated to some extent with temperature and precipitation over that period. The NOAA Online Weather Data showed that there was an increase in temperature from mid-June to July. During mid-June, the temperature range was observed as 11-29 ^oC, while during July, the range was 14-34 ^oC. Also, the accumulated precipitation gradually increased from 0.19 inches to 5.51 inches through mid-June to July 2019. The accumulated precipitation also increased through the summer months of the years 2018 and 2020 (0.27 inches to 8.45 inches through June-July,2018 and 0.21 inches to 7.28

inches through June-July,2020). However, during the summer months of 2018 and 2020, a higher number of detected events were found in June than July. It appears that temperature and rainfall were not the only contributing factors for significant deviations or anomalies in water quality. Other natural, accidental, or human-induced contamination incidents including chemical, fuel, and sewage spill, stormwater discharge from adjacent streets and parking areas might contribute to the occurrence of water quality events during that period.

Simple linear regression was used to analyze the correlation between the water quality parameters and water temperature. For this model, the explanatory variable was water temperature (^{o}C) while the response variables were pH, turbidity (NTU), and conductivity (µS/cm). The best regression model was used to predict the response variable for a given explanatory variable. From the regression analysis, it was observed that for the explanatory variable (water temperature) the p-value was 0.000 which is less than alpha = 0.05 for each of the water quality signals. This indicates that the model was significant to explain the relationship between temperature and water quality parameters. The R-squared value for conductivity was 43.02%, for pH 18.69 %, and for turbidity 9.43%, which indicates that conductivity has a stronger correlation with temperature than pH and turbidity.

4.4 Conclusion

This study provides an analysis of anomalous water quality events for Milwaukee River using CANARY, an EDS software. The ability of the software in detecting changes relative to water quality values that were expected to occur at specific time was analyzed. This study also provides an insight into the effectiveness of the application of CANARY for automatic real-time

event detection in natural source water. Using time-series data from multiple locations and sensors, CANARY is capable of differentiating real water quality events from background variability based on statistical models. Using CANARY software, anomalous water quality events could be quickly detected at any time-step, and from that, the signals that contributed to the water quality event at the specific time and location could be identified. In most cases of this study, the LPCF algorithm detected a higher number of events than the MVNN algorithm, which may be attributed to their different residual calculation mechanisms. For future work, advanced analytics techniques such as machine learning (ML) methods can be applied to anomaly detection in water quality data for more accurate, faster, and better decision support mechanisms. CANARY is based on a statistical approach that mainly focused on the fitting of a specific probability model, which does not rely on predictive accuracy. ML is focused on finding a generalized predictive pattern with learning algorithms. ML algorithms deal with data with a large number of observations and access and learn from each observation over time to predict the exact combination of actions as well as achieve high accuracy of the model result. In the future study, the river water quality data can be used as independent variables and machine learning algorithms can be applied to develop and train a model for the prediction of the target value of event.

4.5 References

1. Barbanti, A., Bondi, A., Gasparoni, F., & Morandi, D. (1995, October). BLUE BOX: a system for long term unattended environmental monitoring of marine basins-prototype architecture and test results. In *'Challenges of Our Changing Global Environment'. Conference Proceedings. OCEANS'95 MTS/IEEE* (Vol. 1, pp. 642-649). IEEE.
2. Exchange Network Philadelphia Water. (2015). The Delaware Valley early warning system, Exchange Network National Meeting, Philadelphia PA.
3. Grayman, W. M., Deininger, R. A., Males, R. M. (2001). Design of early warning and predictive source water monitoring systems, first ed. American Water Works Association Research Foundation, Denver.
4. Gullick, R. W., Gaffney, L. J., Crockett, C. S., Schulte, J., & Gavin, A. J. (2004). Developing regional early warning systems for US source waters. *Journal-American Water Works Association*, *96*(6), 68-82.
5. Gullick, R. W., Grayman, W. M., Deininger, R. A., & Males, R. M. (2003). Design of early warning monitoring systems for source waters. *Journal-American Water Works Association*, *95*(11), 58-72.
6. Hagar, J., Murray, R., Haxton, T., Hall, J., & McKenna, S. (2013). Using the CANARY event detection software to enhance security and improve water quality. In *World Environmental and Water Resources Congress 2013: Showcasing the Future* (pp. 989-1003).
7. Hart, D., McKenna, S. A., Klise, K., Cruz, V., & Wilson, M. (2007). CANARY: a water quality event detection algorithm development tool. In *World Environmental and Water Resources Congress 2007: Restoring Our Natural Habitat* (pp. 1-9).

8. Hasan, J., States, S., Deininger, R. (2004). Safeguarding the security of public water supplies using early warning systems: A brief review. *Journal of Contemporary Water Research and Education*, 129, 27-33.
9. Leow, A., Burkhardt, J., Platten III, W. E., Zimmerman, B., Brinkman, N. E., Turner, A., ... & Garland, J. (2017). Application of the CANARY event detection software for real-time performance monitoring of decentralized water reuse systems. *Environmental science: water research & technology*, *3*(2), 224-234.
10. Louisiana DEQ. (2014). Early Warning System helps protect the Mississippi River as a drinking source. https://deq.louisiana.gov/assets/docs/Newsletters/DiscoverDEQNewsletter-Issue34-November2014.pdf (accessed 20 September 2019)
11. McKenna, S. A., Hart, D. (2008). On-Line Identification of Adverse Water Quality Events from Monitoring of Surrogate Data: CANARY Software. In Singapore International Water Week Conference, June 23rd-27th, Singapore.
12. Michigan DEQ. (2007). Real Time Monitoring Program - Protecting the drinking water source in the Huron to Erie Corridor. https://www.michigan.gov/documents/deq/deq-wb-wws-BrockHowardCIPRTMtalk5-30-07_237078_7.pdf (accessed 26 February 2021).
13. Regional Response Team III PA. (2008). Delaware Valley early warning system (EWS). https://www.slideserve.com/rayya/delaware-valley-early-warning-system-ews (accessed 15 March 2020).
14. Stumpf, R.P., Wynne, T.T., Baker, D.B., Fahnenstiel, G.L. (2012). Interannual Variability of Cyanobacterial Blooms in Lake Erie. *PLOS ONE*, 8, e42444.
15. U.S. EPA. (2008). Interim Guidance on Developing an Operational Strategy for Contamination Warning Systems.

https://nepis.epa.gov/Exe/ZyNET.exe/P1008EAE.TXT?ZyActionD=ZyDocument&Client=EPA&Index=2006+Thru+2010&Docs=&Query=&Time=&EndTime=&SearchMethod=1&TocRestrict=n&Toc=&TocEntry=&QField=&QFieldYear=&QFieldMonth=&QFieldDay=&IntQFieldOp=0&ExtQFieldOp=0&XmlQuery=&File=D%3A%5Czyfiles%5CIndex%20Data%5C06thru10%5CTxt%5C00000019%5CP1008EAE.txt&User=ANONYMOUS&Password=anonymous&SortMethod=h%7C-&MaximumDocuments=1&FuzzyDegree=0&ImageQuality=r75g8/r75g8/x150y150g16/i425&Display=hpfr&DefSeekPage=x&SearchBack=ZyActionL&Back=ZyActionS&BackDesc=Results%20page&MaximumPages=1&ZyEntry=1&SeekPage=x&ZyPURL (accessed 20 March 2019).

16. U.S. EPA. (2010). Water Quality Event Detection Systems for Drinking Water Contamination Warning Systems: Development, Testing, and Application of CANARY. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NHSRC&dirEntryId=221394 (accessed 15 May 2019).
17. U.S. EPA. (2012). CANARY User's Manual Version 4.3.2. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NHSRC&address=nhsrc/si/&dirEntryId=253555 (accessed 15 March 2019).
18. U.S. EPA. (2013). CANARY Training Tutorials. https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=261777&Lab=NHSRC&fed_org_id=1253&subject=Homeland%20Security%20Research&view=desc&sortBy=pubDateYear&showCriteria=1&count=25&searchall=%27CANARY%27 (accessed 15 May 2019).
19. U.S. EPA. (2013a). Water quality event detection system challenge: methodology and findings. https://19january2017snapshot.epa.gov/sites/production/files/2015-

07/documents/water_quality_event_detection_system_challenge_methodology_and_findings.pdf (accessed 20 June 2019).

20. U.S. EPA. (2014). Configuring Online Monitoring Event Detection Systems. https://cfpub.epa.gov/si/si_public_record_report.cfm?dirEntryId=287299&Lab=NHSRC&fed_org_id=1253&subject=Homeland%20Security%20Research&view=desc&sortBy=pubDateYear&showCriteria=1&count=25&searchall=%27CANARY%27 (accessed 20 June 2019).
21. U.S. EPA. (2017). Drinking Water Treatment Source Water Early Warning System State of the Science Review. https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NHSRC&dirEntryId=339641 (accessed 10 March 2019).
22. Zaefferer, M. (2012). Optimization and Empirical Analysis of an Event Detection Software for Water Quality Monitoring. Master Thesis. Germany: Cologne University of Applied Science.

CHAPTER 5: PREDICTION OF FIVE-DAY BIOCHEMICAL OXYGEN DEMAND IN THE BURIGANGA RIVER OF BANGLADESH USING NOVEL HYBRID MACHINE LEARNING ALGORITHMS

5.1 Introduction

Rapid urbanization, climate change, and population growth in developing countries create an immense pressure on water infrastructures and natural resources, including water supply and distribution, urban drainage, solid waste collection, and disposal. The performance of water management and pollution control services in developing countries is poor and as a result, urban populations in these countries lack access to safe drinking water and face shortages of sources of water. In Bangladesh, the river water system is tremendously dynamic and considered to be the most important natural resource for domestic, irrigation, and recreational purposes. Several industries and urban centers are developed along the banks of the rivers. The discharges from the rivers and their tributaries are mostly influenced by domestic, industrial, and agricultural runoff and seasonal fluctuations particularly during monsoon periods (Department of Environment, 2016). The inland surface water quality is gradually degrading because of the rapid expansion of population and polluting industries in cities and the increased use of fertilizers that affect the surrounding land, local communities, and aquatic ecosystems. The extreme examples of these effects are near Dhaka City of Bangladesh where industrial, agricultural, and municipal wastes are discharged into the nearby rivers without proper treatment. The population growth rate in Dhaka combined with limited financial resources constrains the development of sustainable and adequate urban water infrastructure. This research was initiated to enhance the capacities of water management agencies and increase their knowledge on the application of advanced

technology for water pollution control, therefore providing a technical solution for water quality monitoring and modeling.

Among all the rivers near Dhaka, the Buriganga River in the southwestern region of Dhaka is the most polluted with significant contamination sources including municipal, industrial and agricultural effluents, and discharge of untreated wastewater from tanneries and other outfalls. The Buriganga River is mainly fed by an upper tributary of Turag River which is also influenced by industrial effluents, agrochemicals, fecal pollution, ship breaking and lube oil discharge, and seasonal effects. Continuous changes in water quality of the river system make the environment unsuitable for aquatic lives. For efficient water management and restoration of river water quality, it is essential to monitor key parameters: pH, turbidity, conductivity, dissolved oxygen (DO), 5-day BOD, and chemical oxygen demand (COD). One of the most important parameters BOD_5, defined as the amount of dissolved oxygen required by the microorganisms to decompose organic matter at a specific temperature (20^o C) requires much time and effort to detect. Also, BOD_5 is a major criterion for controlling stream pollution with severe conditions of organic loading in water. A higher BOD_5 level indicates organic pollution with less DO in water to maintain aquatic life ecosystems. Therefore, it is essential to continuously monitor BOD_5 in river water systems.

In recent years, the Buriganga River system was found to be highly polluted with a significantly higher BOD_5 level and a lower DO level than the standard limits (Papry, 2019). Traditional methods of BOD_5 have some potential limitations for water management and research due to the time constraint associated with the 48-hour sample holding time from the collection and a five-day incubation period (Ma *et al*., 2020). When the water quality changes rapidly and unexpectedly, the results obtained from the 5-day BOD test would no longer be

relevant to the current situation for water management and treatment. In addition, maintaining a constant temperature over the 5-day incubation period may be difficult due to frequent power interruptions in developing countries such as Bangladesh. Therefore, there is a need to improve the water quality monitoring program with a more accurate, reliable, and cost-effective method for measuring water quality parameters.

For more efficient water management, pollution control, and eliminating the shortcomings of the traditional approaches, machine learning (ML) modeling tools have been developed to predict water quality parameters more efficiently. ML methods were used for developing event detection models in water distribution systems by Muharemi *et al*. (2018), Perelman *et al*., (2012), Silvia and Ilan (2019), and Zou *et al*. (2019). Other studies (Nafsin and Li, 2021; Nafsin *et al.*, 2021) used statistical methods and event detection system for surface water quality monitoring and analysis. In river water systems, Haghiabi *et al*. (2018), Hayder *et al*. (2020), Sarker and Pandey (2015) developed and investigated the performances of different ML models for predicting water quality parameters. Venkateswarlu *et al*. (2020), Anmala and Turuganti (2021) developed water quality prediction models to determine the impacts of climate and land use on stream water quality using different ML methods such as PCA, CCA, ANN, decision tree (DT), and extreme machine learning (EML) algorithm. Other studies (Ahmed *et al*., 2019; Babbar, 2017; Bui *et al*., 2020; Chen *et al*., 2020; Hameed *et al*., 2016; Sakizadeh, 2016; Wang *et al*., 2012) used ML techniques to predict water quality based on WQI (Water Quality Index). Among several ML techniques, some specific ML methods have been explored for the prediction of BOD (Baki *et al*., 2019; Dogan *et al*., 2009; Fathima *et al*., 2014; Khaled *et al*., 2017; Kim *et al*., 2020; Noori *et al*., 2015; Solgi *et al*., 2017). Emamgholizadeh *et al*. (2013) applied ANN and ANFIS models for predicting BOD_5, COD, and DO in a river system. Raheli *et*

al. (2017) investigated the predictive ability of a hybrid forecasting model integrated with MLP in predicting BOD_5 and DO in a river system. Jafari *et al.* (2020) and Ma *et al.* (2020) implemented a hybrid wavelet-genetic programming (WGP) method and deep learning methods for BOD_5 prediction of surface water, respectively. In industrial wastewater treatment plants, ML techniques have been applied to predict effluent water quality such as BOD_5, COD, and TDS, and to forecast removal efficiency of different water quality parameters (El-Rawy *et al.*, 2021; Sharafati *et al.*, 2020).

Uddin and Jeong (2021) reviewed the Dhaka city urban river water pollution based on research data during the past 40 years. The study confirmed the growth of industrialization as one of the main reasons for environmental pollution affecting aquatic life and human health in the area. Sarker *et al.* (2015) analyzed the pollution in Buriganga River from tannery effluent and found its connection with DO, COD, BOD, and electrical conductivity (EC). Moreover, several other studies (Ahammed *et al*., 2016; Ahmed *et al*., 2010; Bhuiyan *et al*., 2015; Kamal *et al*., 1999; Rahman *et al*., 2010; Saifullah *et al*., 2012) investigated the impact of contamination sources on the Buriganga river system in Dhaka and provided the state of water quality of the river. Moniruzzaman *et al*. (2009) evaluated the spatial distribution of physicochemical parameters of the Buriganga river through GIS technology. Whitehead *et al*. (2019) developed a dynamic model to simulate heavy metals along the Buriganga river system in Central Dhaka. To the best of the authors' knowledge, studies are yet to be conducted to examine the application of artificial intelligence (AI) based models for forecasting BOD_5 in the highly polluted rivers of Bangladesh. The data-driven ML models can provide more accurate and timely information on significant changes in BOD_5 levels of a river that allow decision-makers better water management and water quality restoration. ML algorithms learn from historical observations and

predict the future value, thus reducing the longer detection time of BOD_5 measurement. The models can improve their prediction performance by re-training the model with newly available sample observations. The data-driven models analyze a large volume of real-time water quality data, find the complex relationship between water quality variables, and identify the parameters that are most influential for predicting the output. ML models can be effective tools for continuous monitoring of water quality parameters in rivers near Dhaka city. The properly tested and optimized predictive models are accurate, reliable, and robust to outliers and work well with a limited and large dataset.

A number of ML algorithms have been applied in predicting water quality parameters, each standalone ML algorithm has its strengths and shortcomings. For example, Artificial Neural Network (ANN) exhibits excellent performance in identifying complex patterns and nonlinear relationships in datasets but performs poorly for small datasets and when the testing data is outside the range of the training dataset (Bui *et al*., 2020). In Support Vector Machine (SVM), the kernel functions can arrange the datapoints in multidimensional feature space which enables it to find complex relations between the features and output. However, this algorithm is not robust against noisy data and often underperforms for large datasets with variations. On the contrary, Random Forest (RF) is robust to outliers and works efficiently on large datasets with a reduced chance of overfitting. Gradient Boosting Machine (GBM) often underperforms with noisy data resulting in overfitting problems. With proper tuning of hyperparameters, GBM performs satisfactorily due to the boosting mechanism that builds one decision tree at a time and improves performance significantly by learning from the mistakes of the previous tree.

The purpose of the research is to provide an efficient water data management system using AI/ML techniques and develop improved pollution control strategies to protect natural

source water from domestic and industrial pollutants in developing countries such as Bangladesh. The performance efficiencies of several ML models were evaluated to obtain the most efficient prediction model of BOD_5. The BOD_5 of the highly polluted Buriganaga River system were predicted using four standalone traditional ML algorithms: ANN, SVM, RF, and GBM and six novel hybrid models: RF-SVM, ANN-SVM, GBM-SVM, RF-ANN, GBM-ANN, and RF-GBM. A comprehensive assessment of the ML techniques for evaluating BOD_5 was conducted. The hybridization of the traditional ML algorithms optimized with specific hyperparameters, provides a more accurate, robust, and reliable prediction of the output. Additionally, the most influential water quality parameters in predicting BOD_5 of the river system were identified.

5.2. Materials and Methods

5.2.1 Study area

The Buriganga River flows through the southwest region of Dhaka, Bangladesh (Figure 5.1). The catchment area of the river is 253 km^2 with a length, width, and depth of 27 km, 400 m, and 10 m, respectively. The flow and quality of the river water are influenced by the upper tributary of Turag River at the northwestern boundary of Dhaka City. Thousands of factories located around the Buriganga River System release their wastes into the rivers and make the river system highly polluted. The Department of Environment (DoE) of Bangladesh reported a daily discharge of above 60,000 m^3 of industrial wastes into the water bodies of Dhaka City. The annual discharge of toxic wastes and sludge from textile industries are about 56 million tons and 0.5 million tons, respectively. The most significant source of contamination appears to be from

tanneries with a daily discharge of 22,000 m^3 of wastes. Disposal of the ever-increasing industrial, municipal wastes and agricultural runoff into the rivers deteriorated the Buriganga River System ecologically and made it one of the most polluted rivers in Bangladesh. (Department of Environment, 2016)

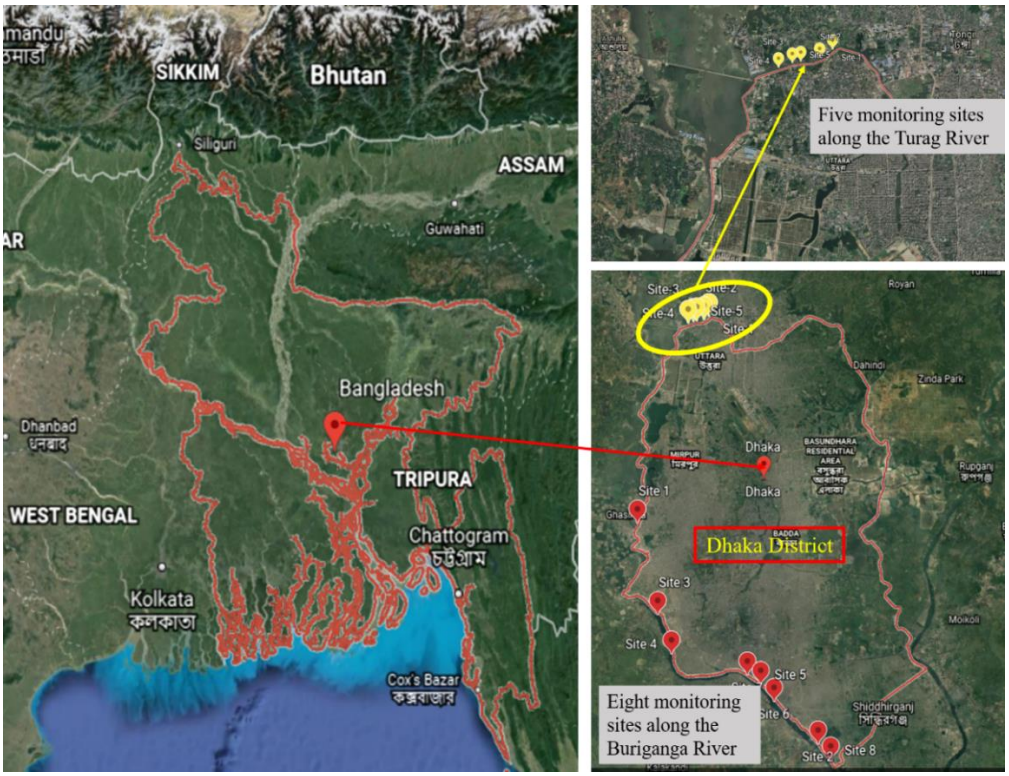


Figure 5.1. Monitoring locations along Buriganga River and Turag River in Dhaka, Bangladesh (Buriganga River: eight monitoring sites indicated as 'red' location points (bottom right) and Turag River: five monitoring sites indicated as 'yellow' location points (top right))

5.2.2 Input variables and data preparation

Water quality data of the Buriganga River System were collected from the DoE under the Ministry of Environment, Forest, and Climate Change, Bangladesh. Monthly monitoring data over 3 years (2015-2017) from 8 monitoring stations of the Buriganaga River and 5 monitoring

stations of the Turag River (Figure 5.1) were used as the inputs to the ML models. The complete dataset composed of 341 sample observations with 12 water quality parameters including pH, turbidity, EC, temperature, DO, chloride, total alkalinity, total solids (TS), total dissolved solids (TDS), suspended solids (SS), COD, and BOD_5. Initially, all 11 water quality parameters were considered as the input to develop the prediction models for BOD_5. To reduce the number of input features and select the most effective variables, the feature importance computed by a tree-based algorithm that indicated the most important features was used to predict BOD_5. The water quality parameters were measured using standard methods: Modified Winkler's method for DO, dilution method for BOD, closed reflux colorimetric method for COD, the argentometric method for chloride, Nephelometric method for turbidity, and gravimetric method for TDS and SS determination. For pH, alkalinity, and electrical conductivity, standard methods for the examination of water and wastewater were followed (Department of Environment, 2016; Federation and APHA, 2005).

Data preprocessing including data splitting and data normalization (feature scaling) is an important step for developing ML models. To ensure the models' prediction accuracy, data was splitted into two sets: a training set for model building and validation and a test set for model performance evaluation. In this study, 75% of the dataset (255 samples) with the output variable was considered as training set and 25% of the dataset (86 samples) was used as the test set. 75% of the training set (191 of 255 samples) was considered as training and 25% as validation set (64 of 255 samples). The dataset was rescaled to make data representation more suitable for the ML algorithms. In the *scikit_learn* tool of the python ML library, the 'robust scaler' was used to standardize the features using the median, 25^{th} and 75^{th} quartiles, bringing all the features to the

same magnitude. This standardization method is robust to the presence of outliers, which is common in the water quality dataset.

The monitoring stations used in this study are scattered throughout the area near the Buriganga river and Turag river, where several industries are located. We combined all the data from the 13 monitoring stations to develop more generalized prediction models that contained local or seasonal variations in water quality parameters. The data from different locations help to improve the models' ability to adapt properly to a new or previously unseen dataset. Using the combined data from the 13 monitoring stations, we performed basic statistical analysis to visualize the water quality data of the river system, and the complete dataset provided an insight into the pollution status of the rivers near Dhaka city.

5.2.3 Machine learning models

Supervised machine learning algorithms: ANN, SVM, RF, and GBM were trained to develop regression models for the prediction of BOD_5. The *scikit-learn* (V0.21) machine learning toolkit in the python program was used for the analysis.

Artificial Neural Network (ANN) model consists of simple units that develop mathematical decisions, analyze complex problems, and provide accurate results. A simple neural network includes three layers of neurons: an input layer, a hidden layer, and an output layer. Increasing the number of hidden layers increases model complexity and prediction accuracy. In the model development process, the algorithm takes several independent variables as inputs that are run through an activation function in the hidden layers and converts into an output. In this analysis,

the Multilayer Perceptron (MLP) feed-forward network was used. The model architecture of the MLP network is presented in Figure 5.2.

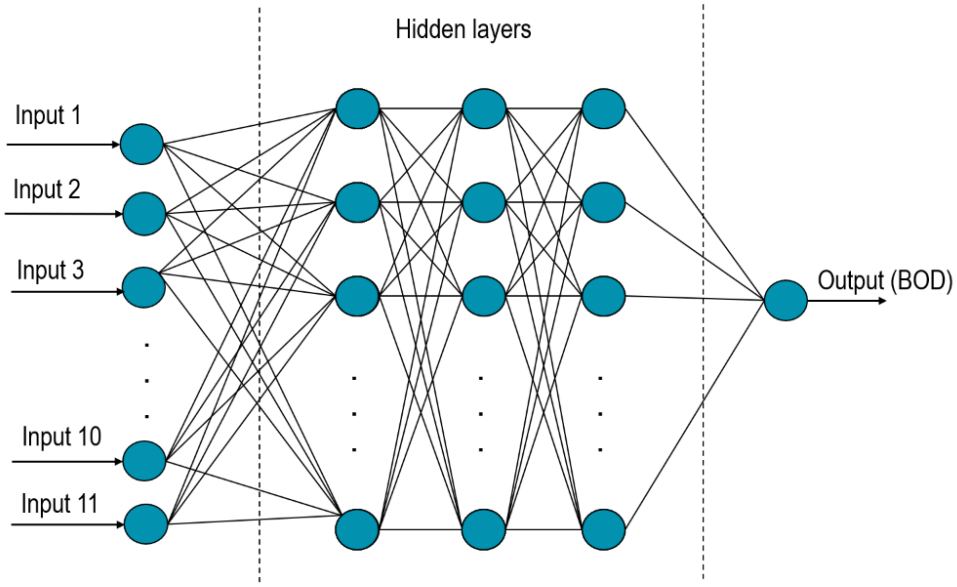


Figure 5.2. Multilayer Perceptron (MLP) model architecture

Support vector machine (SVM) is widely used as a support vector regressor (SVR) in regression problems. The algorithm finds a decision boundary or hyperplane that includes the maximum number of data points. Based on the training dataset, the model identifies a function that approximates the mapping of the input domain to real numbers. In SVM, different kernel functions such as RBF, linear, sigmoid, and polynomial are used to convert 2-dimensional input data into higher dimensional feature space.

Random forest (RF) is an ensemble decision tree algorithm that uses several individual trees with random choices for each tree to build decision models. Each tree in the model makes a prediction, and the final prediction result is obtained by averaging results from all the decision trees. The two important parameters for model building are the number of decision trees and the number of features for best splits in trees that indicate the degree of randomness of each tree.

Gradient boosting machine (GBM) combines multiple decision trees to build a more efficient prediction model and learns from the mistakes of previous trees. Each tree in the model only works on a portion of the data and provides a good prediction. The model performance is improved by adding several decision trees iteratively. Although GBM works well without feature scaling, it is more sensitive to parameter tuning. The key parameters such as the number of decision trees and the learning rate are highly interconnected and control the model's complexity. (Muller and Guido, 2016).

Six novel hybrid models were developed by integrating the standalone ML algorithms with optimization techniques for improving the efficiency of the prediction models. The hybrid models, including RF-SVM, ANN-SVM, GBM-SVM, RF-ANN, GBM-ANN, and RF-GBM, were optimized to find the best model parameters resulting in a high prediction success. The hybridization process took two base algorithms to develop a method with greater flexibility as compared to a standalone ML method. The results of SVM were improved by hybridizing the SVM algorithm with RF. When developing the RF-SVM model, the specific parameters for the RF algorithm such as n-estimator and max-features were used. For SVM, the kernel-specific parameters (C, gamma, and kernel) were defined. A voting regressor, which is an ensemble meta-estimator was employed to fit the RF and SVM regressor each on the dataset. The individual predictions of each model were averaged to achieve the final prediction. The hyperparameters for both models were optimized to achieve a high prediction accuracy of the developed hybrid model. ANN-SVM hybrid model was developed by integrating the SVM model with the MLP-ANN regressor and optimizing the hyperparameters of ANN (i.e., hidden layer size with the number of nodes in each layer, activation function, solver, and alpha) and SVM. Using the voting regressor ensemble method, the GBM-SVM hybrid model was

developed by combining the predictions of GBM and SVM models, and the average prediction of the contributing models was calculated. The hyperparameters of GBM (i.e., learning rate and max depth) were optimized along with the specific parameters of SVM. To develop the RF-ANN hybrid model, the RF model was integrated with ANN (MLP) using a voting regressor estimator, and the ensembled model was fit to the training dataset to predict the output with a high prediction success. In addition, for the GBM-ANN and RF-GBM hybrid ensemble learning model, the sub-models were defined with their corresponding specific parameters, and the voting regressor estimator determined the average prediction of the developed hybrid models.

5.2.4 Performance evaluation criteria

The performances of the regression models were evaluated by determining the error rate of the prediction and goodness-of-fit of the model that indicated how well the regression models fit the observed dataset. Statistical metrics coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) were used to evaluate the performances of the regression models.

The coefficient of determination measures the degree to which the model explains the variability in the dependent variable through the independent variables. The higher the value of R^2, the better the model fits the observed data. The mathematical representation of the statistical measure is shown in equation (5.1).

$$R^2 = \frac{\sum(y_{predicted} - y_{average})^2}{\sum(y_{actual} - y_{average})^2} \qquad (5.1)$$

Mean absolute error (MAE) measures the average magnitude of the errors between the actual and predicted value. It is calculated by measuring absolute error for each value considering equal weights to the error and dividing the sum of the error by the total number of sample observations, as shown in equation 5.2.

$$MAE = \frac{1}{n}\sum_{1}^{n}\left(\left\|y_{actual} - y_{predicted}\right\|\right) \qquad (5.2)$$

The quadratic scoring metric (RMSE) measures the average magnitude of the squared error. It indicates the standard deviation of residuals that measures the distance between the actual data and the regression line. The lower the value of the error, the more accurately the model can predict the variable. The RMSE can be calculated using equation (5.3).

$$RMSE = \sqrt{\frac{1}{n}\sum_{1}^{n}(y_{actual} - y_{predicted})^2} \qquad (5.3)$$

5.3. Results and Discussion

5.3.1 Analysis of water quality of the Buriganga river system

The water quality parameters of the Buriganga River System were analyzed using basic statistical measures including minimum, maximum, mean, standard deviation, and the coefficient of variation (CV). The basic statistical analysis of the normally distributed water quality data is shown in Table 5.1. The CVs were high for most of the parameters, particularly, BOD_5, DO, turbidity, chloride, and COD, with %CV of 110.38, 100.12, 106.07, 93.29, and 90.39 respectively. The BOD_5 concentrations varied within a range of 0.8-86 mg/L with a mean value of 13.42 mg/L. The BOD_5 levels in the river system were significantly higher than the maximum

standard limit of Bangladesh Environmental Quality Standard (EQS) for inland surface water quality (standard: 6 mg/L or less). The discharge of untreated wastewater from tanneries was the likely reason for the high BOD_5 level. Another important water quality parameter, turbidity varied significantly with a minimum value of 3.6 NTU and a maximum value of 250 NTU. The level of turbidity in the river system was higher than the standard limit of 10 NTU. The maximum and minimum DO level was found as 0 mg/L and 6.3 mg/L, respectively. During the sampling period, the level of DO was significantly lower than the EQS standard limit (standard: 6 mg/L or more). The depletion of DO level was caused by the reduced water flow in the dry season, discharge of untreated industrial effluent, municipal wastes, and tannery wastes into the river. The DO levels were found to be slightly higher during the wet season. The concentration of COD in the river was mostly within the Bangladesh EQS standard limit (200 mg/L) for industrial wastewater after treatment. The COD levels varied within a range of 4.83-258 mg/L with a mean value of 38.99 mg/L. The maximum and minimum chloride concentration was found to be 135.6 mg/L and 5 mg/L with a mean chloride concentration of 35.76, which is below the EQS maximum standard limit (600 mg/L) for industrial wastewater after treatment. TDS of the Buriganga River System varied from 52.6 to 930 mg/L with a mean value of 253 mg/L. The TDS level was found within the standard limit (2100 mg/L) for wastes from industrial units. pH in the river system varied from 5.76 to 8.79 while the standard limit for inland surface water quality for fisheries is 6.5-8.5. Considerable variation was observed with conductivity level from 107.5-1767 µS/cm with a mean of 497.3 µS/cm and a standard deviation of 406.63 µS/cm. The conductivity was mostly within the EQS standard limit of 1200 µS/cm. The results indicated that among all the measured parameters, turbidity and BOD_5 contributed more significantly to the

contamination of the Buriganga River System during the sampling period with concentration levels exceeding the EQS standards.

Table 5.1. Statistical analysis of measured water quality parameters of the Buriganga River System

Parameter	Unit	ECR* standard	Minimum	Maximum	Mean	Standard deviation	CV%
Temperature	^{o}C	20.0-30.0	18.40	31.90	26.40	3.14	11.89
pH		6.5-8.5	5.76	8.79	7.29	0.38	5.17
Conductivity	µS/cm	1200.0	107.50	1767.00	497.30	406.63	81.77
Chloride	mg/L	600.0	5.00	135.60	35.76	33.36	93.29
T. Alkalinity	mg/L	N/A	38.00	450.00	115.74	83.24	71.92
Turbidity	NTU	10.0	3.60	250.00	39.41	41.80	106.07
TS	mg/L	N/A	60.00	1069.00	290.40	228.82	78.79
TDS	mg/L	2100.0	52.60	930.00	253.06	209.74	82.88
SS	mg/L	150.0	4.00	297.00	73.77	34.79	79.49
DO	mg/L	6 or more	0.00	6.30	2.14	2.146	100.12
COD	mg/L	200.0	4.83	258.01	43.14	38.99	90.39
BOD_5	mg/L	6 or less	0.80	86.00	13.42	14.80	110.38

*The Environment Conservation Rules, Bangladesh, 1997: Environmental Quality Standards (EQS) for inland surface water

The variation of BOD_5 levels in the highly polluted Buriganga River System (Buriganga and Turag) during different months of the year 2015 and 2016 are shown in Figure 5.3. The BOD_5 level in each month of the year represents the average of the BOD_5 concentrations from different monitoring stations of the river. A 2-sample t-test analysis indicated that the mean of the measured BOD_5 in the Turag River was significantly greater than the Buriganga River with an observed difference of 8.18 mg/L (T value = 1.85, p-value =0.038 <0.05). The reason for the comparatively higher BOD_5 level in the Turag River is due to the river flowing through a dense industrial and urbanized area. In both years, BOD_5 exceeded the tolerable limits from January to May. A similar trend is observed between the sampling periods of 2015 and 2016 with a reduced BOD_5 level in the wet season, particularly in June and July, and then again increased in the dry

season from November to April. The seasonal variation of river water flow and the operation of tanneries might affect the variability of the water quality. During the dry season with low flow conditions, a small amount of organic waste addition might change the water quality significantly, resulting in a higher BOD_5 concentration, whereas in the wet season, the dilution by high flows would result in lower concentration of pollutants.

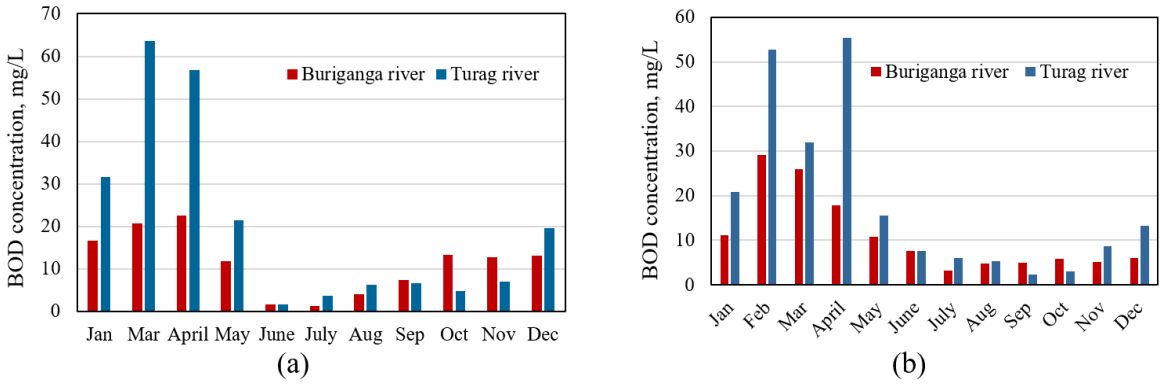


Figure 5.3. Concentration of 5-day BOD in highly polluted rivers during different months of the year (a) 2015; (b) 2016

5.3.2 Feature selection

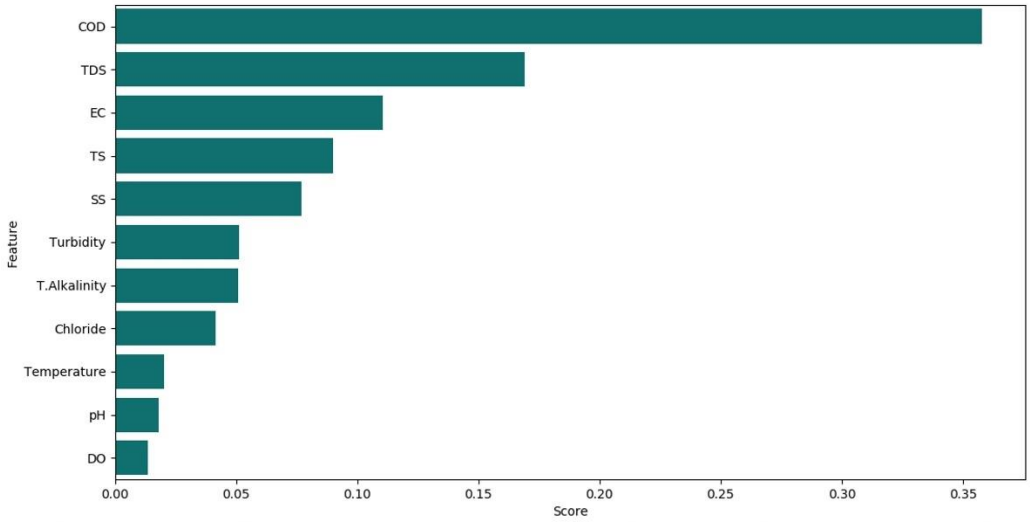
Pearson's correlation analysis was performed to identify relationships between the input features and the predicted BOD_5. Table 5.2 indicates are significant correlations (p-value<0.05) between BOD_5 and each input feature with 95% confidence levels. COD, TS, TDS, and EC had strong positive correlations to BOD_5 with correlation coefficients (R-value) of 0.865, 0.743, 0.738, and 0.726 respectively, while turbidity and suspended solids (SS) had moderate positive correlations with R-values of 0.55. DO and temperature were moderately negatively correlated with BOD_5, while chloride and total alkalinity were moderately positively correlated with BOD_5. Among all the parameters, pH exhibited the lowest value of correlation coefficient (R= 0.16) indicating a weak association between pH and BOD_5.

Table 5.2. Pearson's correlation coefficient (R-value) between BOD_5 and other parameters at 0.05 level of significance

Parameter	Correlation coefficient	P-value	Parameter	Correlation coefficient	P-value
Temperature	-0.323	0.000	TS	0.743	0.000
pH	0.160	0.003	TDS	0.738	0.000
EC	0.726	0.000	SS	0.550	0.000
Chloride	0.696	0.000	DO	-0.574	0.000
T. Alkalinity	0.642	0.000	COD	0.865	0.000
Turbidity	0.553	0.000			

Although the correlation analysis gave an insight into the linear association between the variables, Pearson's correlation is not adequate to explain the variability in a complex dataset where non-linearity may exist. The correlation coefficient is also not resistant to the presence of outliers in the dataset which is common to water quality data. For more efficient feature selection in ML, RF algorithm that uses individual decision trees was used and appropriate split points were selected to measure the importance of each feature. The relative importance of the feature is estimated as the expected fraction of the samples the feature contributes to (Muharemi *et al.*, 2019). Feature importance computed by the RF algorithm indicated the importance of each feature in predicting the output. Feature importance assigned a score between 0 and 1 to each feature and was normalized to sum to 1. A feature importance score of 1 indicated the feature contributed to the perfect prediction of the output, while 0 indicated the feature as unused in the prediction. Figure 5.4 presents the score of each feature in predicting BOD_5. The results show that COD was the most influential feature in predicting BOD_5, with a feature importance score of 0.35. TDS and EC had feature scores of 0.17 and 0.12, while TS and SS had lower scores of about 0.10. The lowest scores were less than 0.05 for chloride, temperature, pH, and DO.

However, low scores did not necessarily mean that the features were uninformative, they might be associated with other important features, and the algorithm encoded the same information from other features.



COD: Chemical Oxygen Demand; TDS: Total Dissolved Solids; EC: Electrical Conductivity; TS: Total Solids; SS: Suspended Solids; T. Alkalinity: Total Alkalinity; DO: Dissolved Oxygen

Figure 5.4. Feature importance using Random Forest

From the results, it can be observed that COD was closely associated with BOD_5. BOD is a measure of the biochemical oxidation of organic matter while COD measures the chemical oxidation of organic and inorganic matter which typically results in a higher value of COD than BOD_5. From statistical analysis, the mean value of the measured COD was greater than the mean of BOD_5 with an observed difference of 29.72 mg/L (p-value<0.05). BOD_5 and solids are related in a manner that a portion of the total solids, i.e., volatile suspended solids are biodegradable and contributes to BOD_5. The dissolved organic matter in TDS may contribute to BOD_5 resulting in a significant correlation between the variables, as indicated in Table 5.2. To improve the prediction accuracy of the models, six water quality parameters (COD, EC, TDS, turbidity, TS, and SS) with higher feature scores were selected for predicting BOD_5. The variables with lower feature importance scores were not considered as these parameters might be correlated with other

important input features. For example, DO was associated negatively with COD (R= -0.535, p-value<0.05) as with the increase in COD, the DO in water decreased. Also, alkalinity and conductivity were positively correlated (R= 0.764, p-value<0.05) as strongly acidic or alkaline substances were more capable of conducting electricity. Chloride was associated with TDS (R=0.844) as dissolved solids might contain mineral salts of chloride ions such as sodium chloride and potassium chloride. The correlation analysis indicated a weak correlation (R=0.16, p-value<0.05) between BOD_5 and pH.

The selection of the best combination of input variables influences the performance of the prediction models (Bui *et al*., 2020). In this study, a variety of input combinations were used, and the models were trained with the specific features. The models were evaluated on the test set and the best combination of variables was selected based on the lowest RMSE score. The results of the feature importance computed by the RF algorithm provided an insight into the importance of each input variable for predicting the output. The ML models were also evaluated based on RMSE with different combinations of input variables extracted from the feature importance, as shown in Table 5.3. Category 1 in Table 5.3 shows all the input variables used for predicting the output, while categories 2 to 7 indicate the combinations of variables that resulted in higher feature importance scores. Our goal was to select the best combination of input features from the feature importance. The models were optimized for each category of input combinations to achieve the lowest RMSE. For RF and SVM, the lowest RMSE value was found for category 3 COD, TDS, EC, TS, SS, and turbidity. For the GBM model, the input combination that resulted in the lowest RMSE was category 2, with an addition of the input feature alkalinity to category 3. For ANN, category 5 with input variables COD, TDS, EC, and TS resulted in the lowest RMSE. In this analysis, category 3 was considered the optimal input combination that could predict

BOD_5 with the lowest RMSE for most of the ML models. It appears that the best combination was not consistent for all the models, which could be attributed to the diverse structure of each ML algorithm and differences in their learning ability from the specific dataset of the features.

Table 5.3. Selection of the optimal input combination based on RMSE

Category	**Input variable combinations**	**RMSE score**			
		ANN	**SVM**	**RF**	**GBM**
1	Temperature, pH, EC, Chloride, Alkalinity, Turbidity, TS, TDS, SS, DO, COD	5.145	5.438	4.459	5.734
2	COD, TDS, EC, TS, SS, Turbidity, Alkalinity	5.294	3.795	4.041	**3.849**
3	COD, TDS, EC, TS, SS, Turbidity	3.805	**3.564**	**3.946**	4.507
4	COD, TDS, EC, TS, SS	3.965	3.627	4.159	4.211
5	COD, TDS, EC, TS	**3.588**	4.144	4.349	4.528
6	COD, TDS, EC	3.955	4.455	4.814	5.267
7	COD, TDS	3.726	4.843	4.517	4.560

5.3.3 Optimization of ML models

The employed ML models were optimized using the grid search method to select the combination of the hyperparameters that resulted in the best model performance. The optimization process with five-fold cross-validation improved the generalization performance of the models when using an unseen dataset and reduced the chance of overfitting. All possible combinations of the hyperparameters were examined and the optimal values are shown in Table 5.4.

Table 5.4. ML models' key parameters selection

Model's key parameters	Optimal values									
	ANN	**SVM**	**RF**	**GBM**	**RF-SVM**	**GBM-SVM**	**RF-GBM**	**RF-ANN**	**GBM-ANN**	**ANN-SVM**
Hidden layer	(15,15,15)							(10,10)	(10,10)	(10,10)
activation	'relu'							'relu'	'relu'	'relu'
alpha	0.1							0.061	0.06	0.92
solver	'adam'							'lbfgs'	'lbfgs'	'lbfgs'
kernel		sigmoid			'rbf'	'rbf'				'rbf'
C		100			500	500				500
gamma		0.001			0.05	0.05				0.05
n-estimators			50		5		5	5		
max-features			4		3		3	3		
max-depth				2		5	5		5	
learning rate				0.01		0.92	0.089		0.862	

5.3.4 Model performance evaluation

The prediction performances of the ML models were evaluated using the statistical performance metrics: R^2, RMSE, and MAE. For regression problems, R^2-value is considered as the most important for model performance evaluation as it indicates the percentage of variation in the data point around the fitted regression line of the model. The goodness-of-fit or R^2 ranges from 0 to 1, where 1 indicates that the model can perfectly explain all the variability in the data without any error, and 0 indicates that the model is not able to explain the variability. An R^2 -value < 0.5 indicates unsatisfactory model execution in predicting the output, while an R^2 -value ranging from 0.6 to 0.9 indicates a satisfactory model execution (Hasan *et al.,* 2021). A higher R^2 value > 0.9 indicates that the model is satisfactory and a good fit for the data. The smaller the value of MAE and RMSE, the more efficient the model is in predicting the output. Initially, all the water quality parameters were used to develop prediction models for BOD_5 and the results

for the performance metrics for the models are presented in Table 5.5. The results indicate that the R^2-values of the models were satisfactory within a range of 0.70-0.82. RF resulted in an R^2 value of 0.82 that indicates that the model execution is satisfactory for predicting the output. RF exhibited the lowest MAE and RMSE. ANN, SVM, and GBM also performed well with R^2 values of 0.756, 0.727, and 0.701 respectively. The performance metrics indicate that the employed regression algorithms can efficiently predict BOD_5 from all the input variables.

Table 5.5. Model performances for prediction of BOD_5 using all input features

ML Algorithm	**MAE**	**RMSE**	**R^2-value**
RF	2.7599	4.4592	0.8165
ANN	3.4845	5.1455	0.7557
SVM	3.4027	5.4388	0.7270
GBM	3.7385	5.7337	0.7012

The models' efficiency in the prediction of BOD_5 was further improved by using only the feature importance with the optimal input combination: COD, EC, TDS, TS, SS, and turbidity. Table 5.6 shows the performance metrics of the 4 traditional standalone and 6 hybrid ML models with feature importance. The performance evaluation matrices for the standalone ML algorithms were higher comparing to the models in Table 5.5. Feature importance significantly improved the performances of the prediction models, especially for SVM and ANN with a reduced prediction error (RMSE and MAE). When using only the important features, the percentage decreases in RMSE of the SVM, ANN, RF, and GBM were 34.4%, 26.06%, 11.5%, and 21.4%, respectively. Similarly, the standalone ML models with feature importance achieved 4.0%-30% lower MAE compared with the models with all input variables. It appears that the performances of the regression standalone algorithms were significantly improved with R^2-values ranging from

0.81 to 0.88. The prediction success of the SVM, ANN, RF, and GBM increased from 72% to 88%, 75% to 87%, 82% to 86%, and 70% to 81%, respectively.

Table 5.6. Model performances for 4 standalone and 6 hybrid algorithms using feature importance

Algorithms	**MAE**	**RMSE**	**R^2**	**Rank order**
SVM	2.446	3.564	0.883	4
ANN	2.552	3.805	0.866	6
RF	2.490	3.946	0.856	9
GBM	2.609	4.507	0.813	10
RF-SVM	**2.230**	**3.158**	**0.908**	**1**
ANN-SVM	2.199	3.356	0.896	2
GBM-SVM	2.499	3.486	0.888	3
RF-ANN	2.445	3.738	0.871	5
GBM-ANN	2.664	3.841	0.864	7
RF-GBM	2.572	3.888	0.861	8

In addition, the performances of the hybrid models developed from the traditional standalone ML algorithms were evaluated for their prediction of BOD_5. Based on the prediction success, the employed ten ML models were ranked for their performance in predicting BOD_5 as shown in Table 5.6. The results indicate that the hybrid models successfully increased the prediction performance as compared to the standalone models with a range of R^2-values 86%-91%. RF-SVM outperformed others with the highest prediction accuracy ($R^2 = 0.908$). For some nonlinear models, the coefficient of determination might not always be appropriate. Besides R^2 value, other goodness-of-fit measure such as standard errors should be considered for measuring prediction performance of nonlinear models. A smaller value of error represents a better model. RMSE and MAE represent different aspects of model performance. RMSE indicates how spread out the predicted values are from the regression fit line, while MAE indicates the average value

of the residuals. Using the RMSE score, the standard deviation of the unexplained variance can be interpreted. In this study, the RMSE score was used to evaluate the performance efficiency of the models. From the results, it appears that the models with high R^2 values also resulted in low RMSE scores. The three best performing hybrid models, RF-SVM, ANN-SVM, and GBM-SVM resulted in the lowest RMSE scores of 3.158, 3.356, and 3.486 respectively.

The hybrid model RF-SVM benefitted the bagging mechanism of the RF algorithm with reduced model complexity that overfits the training data. The strength of the kernel function used in SVM made the hybrid model more powerful by converting the original two-dimensional input data into high dimensional feature space. The combination of the bagging mechanism and rearrangement of the data into a higher dimensional feature space by kernel function enables the RF-SVM hybrid model to perform more efficiently in predicting the output using the collected water quality data, compared to other employed standalone and hybrid models. The output of the RF model depends mainly on the strength and correlation between individual trees developed randomly in the forest without being optimized, while SVM uses an optimized function through minimal sequential optimization. The integration of SVM with RF allowed identifying feasible solutions of the output through iterative computation of the input data using the optimized function while eliminating the chance of overfitting. As a result, the integration of the RF and SVM models made a more powerful hybrid model that outperformed other employed models.

Among the 10 employed models, the best three prediction models were the hybrid models RF-SVM, ANN-SVM, and GBM-SVM. It appears that the best performing standalone model SVM achieved a greater prediction success of above 89% because of hybridization. In addition, the prediction success of the relatively poorer standalone RF and GBM algorithms improved significantly when integrated with other standalone algorithms such as ANN and SVM. The

prediction success for the hybrid RF-GBM model increased from 85.6 % (for standalone RF) and 81.3 % (for standalone GBM) to 86.1%. The hybridization process enhanced the performances of the models by overcoming substantial weaknesses of each standalone model and ensured a more robust, accurate, and reliable prediction of the output variable as compared to standalone models. However, the hybrid models may not be successful in all cases as the level of prediction success of the hybrid models depends on the prediction accuracy of the base algorithms (Bui *et al*. 2020). For example, the standalone model ANN (R^2 = 0.866) had higher prediction success on its own than the hybrid GBM-ANN (R^2 = 0.864) models. For the specific input variables combination, the best-performing prediction model for BOD_5 was considered as the hybrid RF-SVM model. However, the resulted best performing algorithm might not be the "best" in all circumstances. According to the model structure, data structure, and input variables, each algorithm should be tested to find the most efficient prediction model in different situations.

The scatter plots in Figure 5.5 and Figure 5.6 indicate the linear correlation between the actual and predicted BOD_5 for the employed standalone and hybrid ML models when using only the feature importance. The plots clearly explained the variability of the dependent variable through the independent variables. It can be observed that for all the models, most of the data points were the best fit with the regression line, except for a few data points that deviated from the line. The degree of error between the actual and predicted BOD_5 values was exhibited by the deviation of the data points from the regression line. The results indicate that the employed ML models can explain the percent of variance satisfactorily. The time variation graphs in Figure 5.5 and Figure 5.6 represent the predicted value of BOD_5 against the actual value during the testing phase, verifying the prediction performances of the employed ML algorithms. The predicted values of BOD_5 for each sample observation in the test set were close to the actual BOD_5 which resulted in

satisfactory R^2 values for the models. The small deviation between the predicted and actual data indicated that the ML models generalized well on the test dataset.

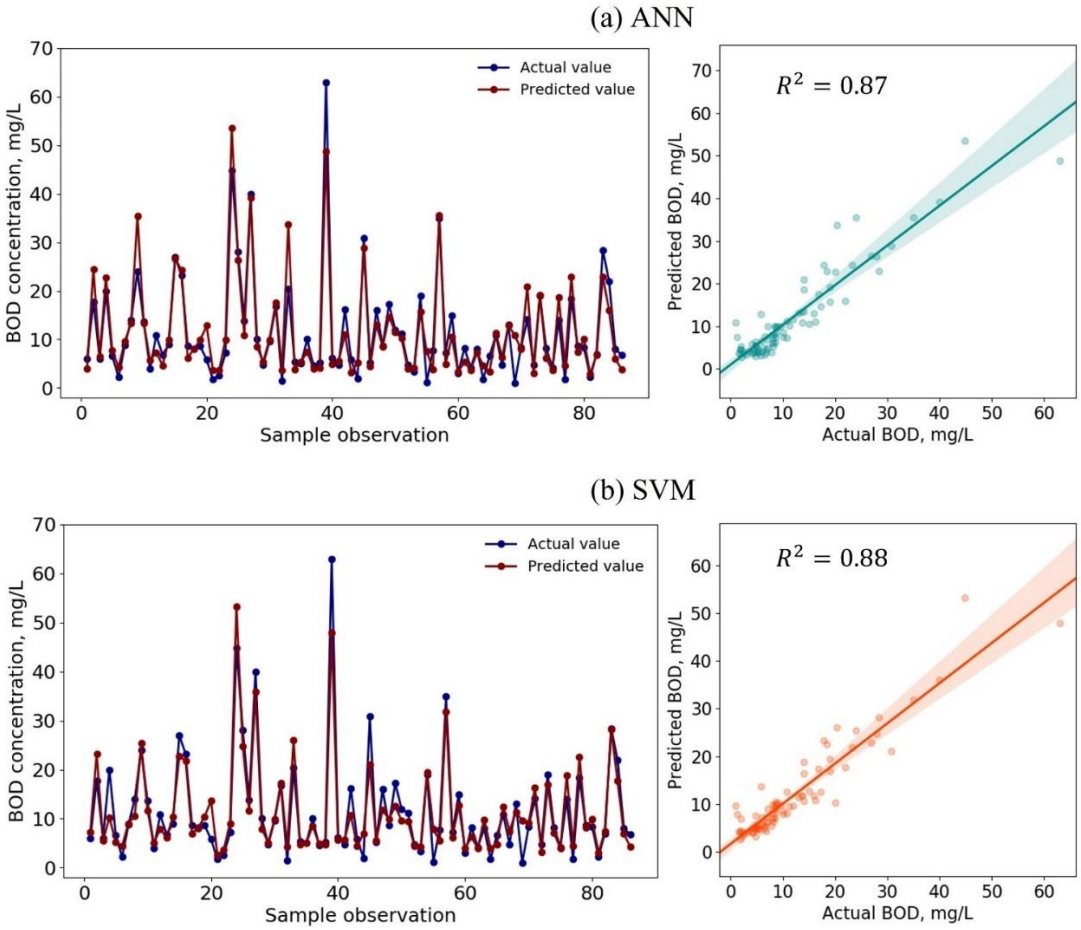
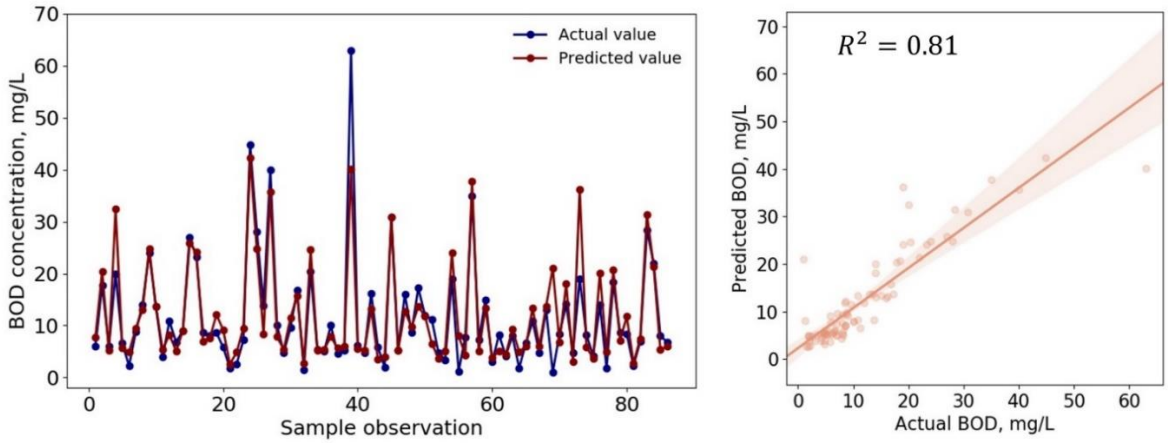


Figure 5.5. Comparison between actual and predicted 5-day BOD concentrations during the testing phase for the traditional standalone ML algorithms (left: Time variation graphs; right: scatter plots)

(c) GBM



(d) RF

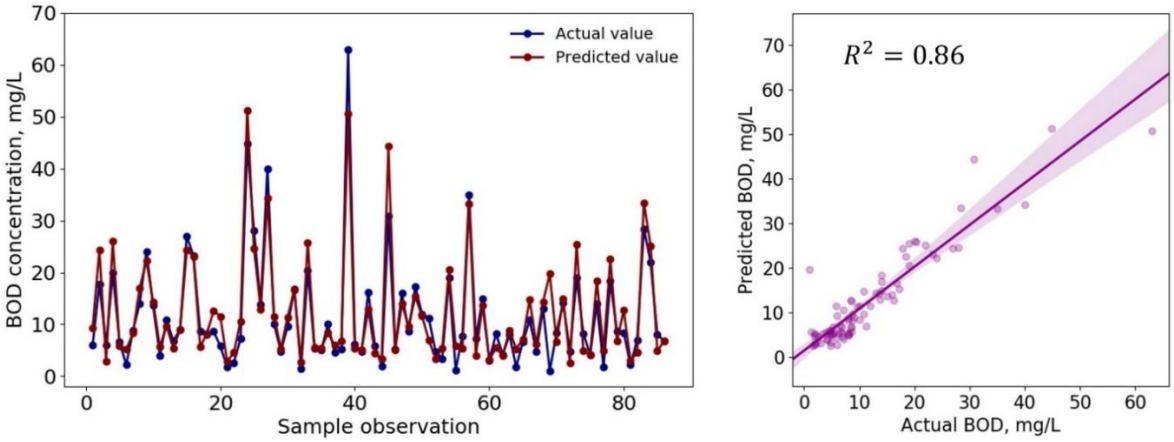


Figure 5.5. (continued)

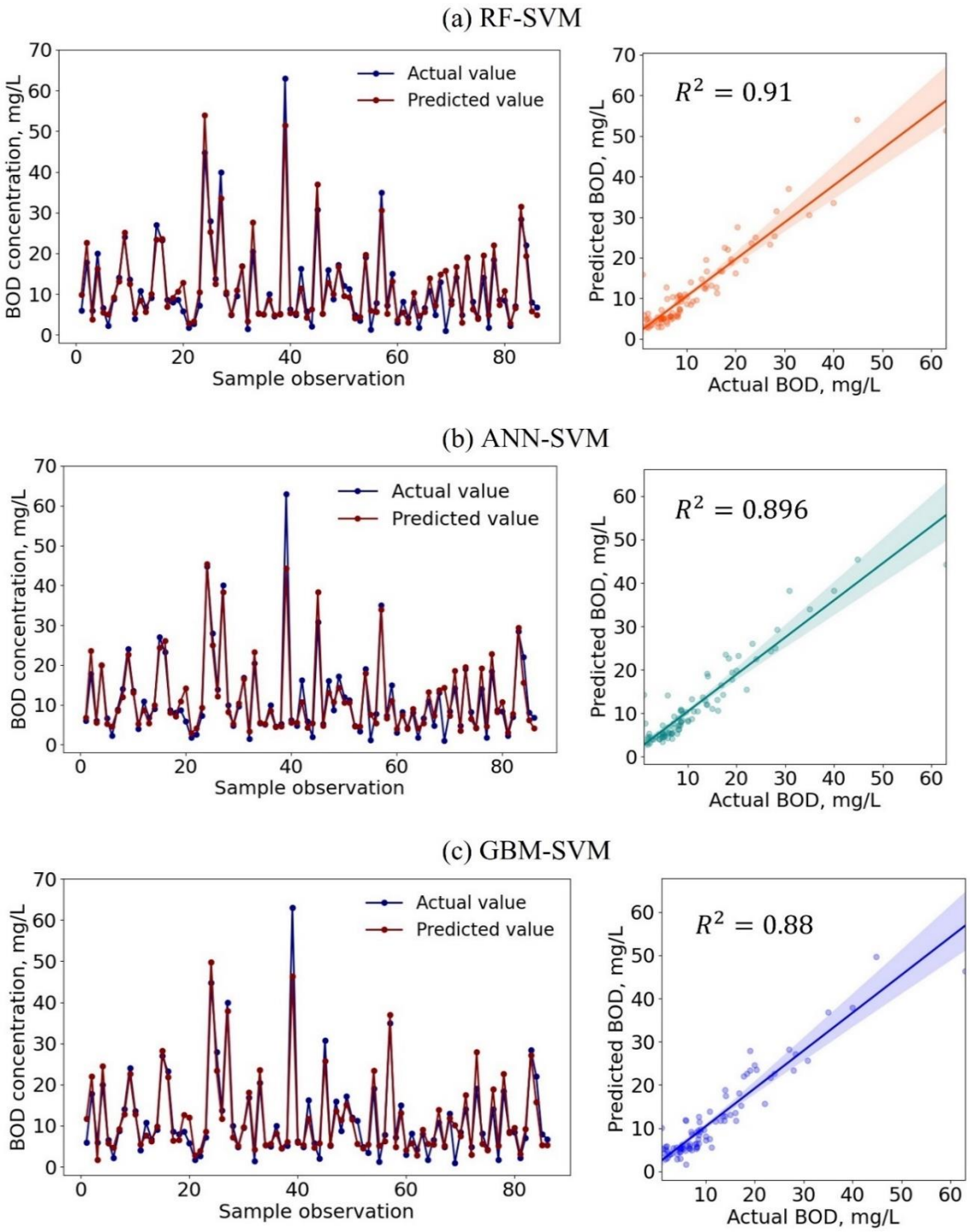


Figure 5.6. Comparison between actual and predicted 5-day BOD concentrations during the testing phase for the novel hybrid ML algorithms (left: Time variation graphs; right: scatter plots)

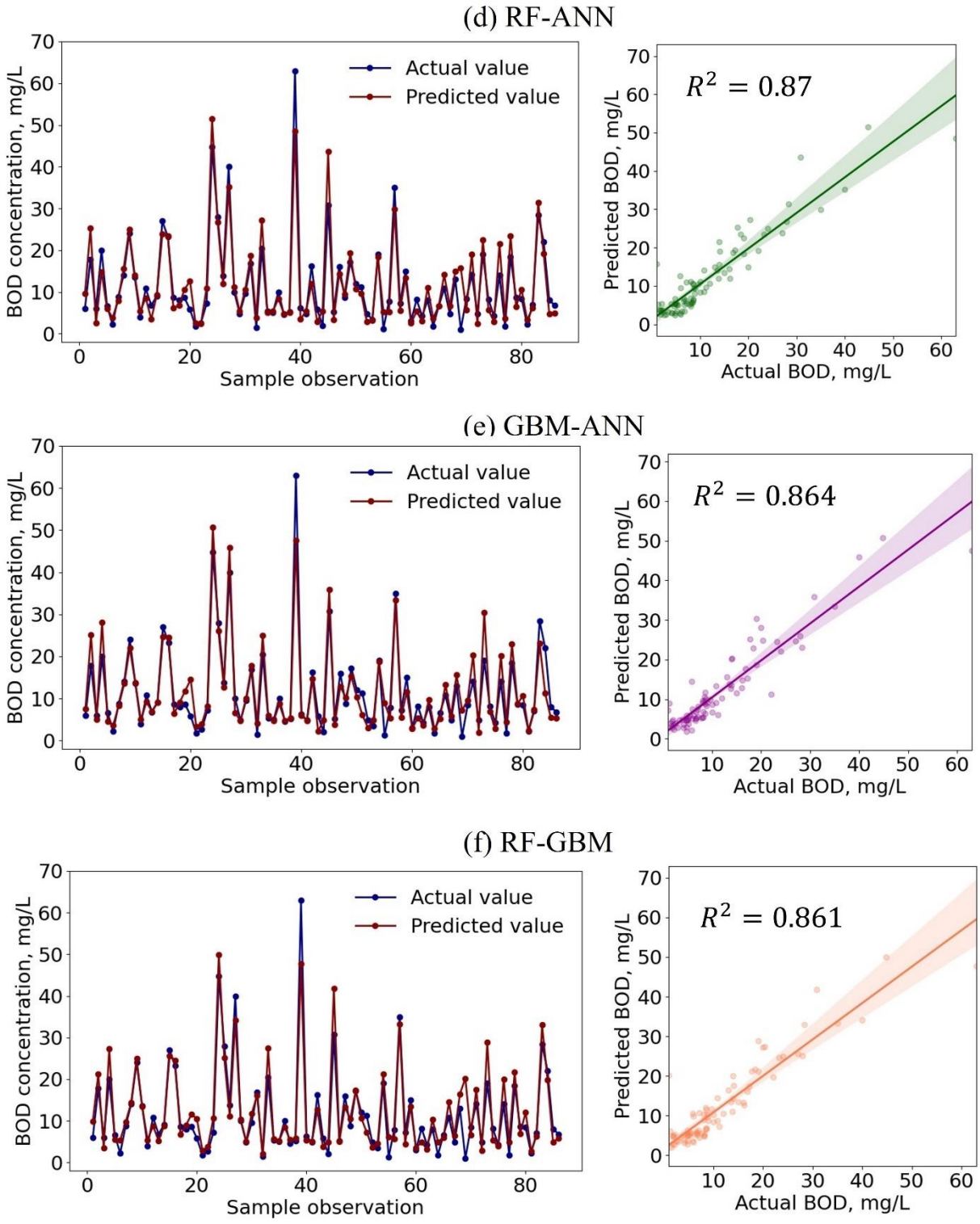


Figure 5.6. (continued).

Several studies (Baki et al., 2019; Dogan et al., 2009; Oliveira-Esquerre et al., 2002) used ML techniques such as classical regression analysis (CRA), multivariate adaptive regression splines (MARS), artificial bee colony (ABC), teaching-learning based optimization (TLBO), ANN, and PCA for predicting BOD in water systems that resulted in an R^2 value within a range of 0.36-0.87. The algorithms used in this study are some of the strongest ML algorithms with diverse working mechanisms. For example, RF and GBM are decision tree-based algorithms, and SVM works with a decision boundary or hyperplane for finding complex relationships between the input and output. A total of 10 ML models, including the hybrid models, were developed to compare the performances of the models and gain insights into the models' performances on the same input dataset. The results show that the hybridization process improved the prediction accuracy by 90% over the standalone models (ANN, SVM, GBM, and RF) with the prediction accuracy within a range of 70%-82%.

Initially the Pearson's correlation analysis was performed between the predicted variable BOD_5 and the input variables. However, Pearson analysis is unable to explain the variability in a complex dataset where non-linearity may exist. Using the decision tree-based RF algorithm, feature importance analysis was performed that captures the non-linear relationship in the dataset and identifies the important features for predicting the output. Simple linear regression was applied to the dataset using scatter plots, fit regression models to calculate the R-squared value, and found the R-squared value less than or about 0.5 for most of the parameters such as TDS, SS, TS, conductivity, turbidity etc. Usually if the R-squared value was close to 1.0, the dataset could be seen as linear data set. As there was non-linearity in the dataset, some non-linear ML algorithms were considered such as neural networks, SVM, RF, or decision tree-based algorithms. In SVM algorithm, hyperplanes are used for non-linearly separable datasets by

projecting the dataset in higher dimension in which it is linearly separable. Similarly, RF and GBM algorithms develops decision trees with non-linear data structure.

The same dataset was used to build several models to compare how different models perform with the specific dataset. For example, ANN exhibits excellent performance in identifying complex patterns and nonlinear relationships in datasets but performs poorly for relatively small datasets. On the other hand, SVM underperforms with a larger dataset. With the specific size of the dataset, the goal was to achieve the best performance of the models by proper optimization and tuning of hyperparameters and running each model with the optimum input variable combination. RF model applies the technique of bagging mechanism (bootstrap aggregating) to decision tree learners that enable the model to perform satisfactorily on a relatively small dataset with proper tuning of hyperparameters. To overcome the issues related to the small dataset, the ANN model was integrated with other models that work well with the relatively small dataset such as SVM and RF algorithm, and developed hybrid models RF-ANN and SVM-ANN. The standalone ANN model had prediction success of 86.6%, but when integrated with SVM and RF, the prediction accuracy increased to 89.6% and 87.1%, respectively. This study provides an insight into the successful application of the models on the available dataset that would potentially be used for similar research with a limited or larger dataset.

5.4 Conclusion

Four standalone and six novel hybrid ML models were used to predict BOD_5 in a highly polluted river of Bangladesh using different water quality parameters. The employed ML models

accurately and directly measured BOD_5 and provided data-driven decisions by extracting predictive information from the historical dataset. The results showed that the hybrid ML models achieved better prediction performance by integrating two different ML algorithms as compared to the standalone traditional ML algorithms. The overall evaluation indicated that the three hybrid models RF-SVM, ANN-SVM, and GBM-SVM provided the most reliable measurement of BOD_5 and successfully increased the prediction accuracy among all the employed models, with a prediction success of 91%, 89.6%, and 88.8% respectively. Several water quality parameters were used for developing the models with different input combinations from category 1 to category 7. Each category of input variable combination was evaluated based on the RMSE, and the category that resulted in the lowest RMSE was selected as the optimum input variable combination. The feature importance computed by the RF algorithm indicated COD, EC, TDS, TS, SS, and turbidity to be the most influential variables for predicting BOD_5. The prediction performance of the models was also verified by comparing the predicted BOD_5 with the measured BOD_5, and a small deviation found between the two datasets indicated good model executions. This study suggested that properly tested and optimized ML hybrid models can potentially be used in forecasting BOD_5 in future time-steps based on the historical water quality dataset. This will alert the river water operators about the BOD levels associated with possible future pollution events. However, the considered best algorithm may not be the most accurate in all circumstances. Each model needs to be evaluated to find the best-performing model based on the specific data structure, input features, and target variable. The application of the novel hybrid models provides a more accurate and direct measure of BOD_5 without extensive laboratory analysis including a five-day incubation period and allows timely information to the operators about any river contamination for control purposes.

5.5 References

1. Ahammed, S. S., Tasfina, S., Rabbani, K. A., and Khaleque, M. A. (2016). An investigation into the water quality of Buriganga-A river running through Dhaka. *International Journal of Scientific & Technology Research, 5*(3), 36-41.
2. Ahmad, M. K., Islam, S., Rahman, S., Haque, M., and Islam, M. M. (2010). Heavy metals in water, sediment and some fishes of Buriganga River, Bangladesh. *International Journal of Environmental Research*, 4(2), 321-332.
3. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A., Irfan, R., and García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, 11(11), 2210.
4. Anmala, J., and Turuganti, V. (2021). Comparison of the performance of decision tree (DT) algorithms and extreme learning machine (ELM) model in the prediction of water quality of the Upper Green River watershed. *Water Environment Research*, 1-14.
5. Babbar, R., and Babbar, S. (2017). Predicting river water quality index using data mining techniques. *Environmental Earth Sciences*, 76(14), 1-15.
6. Bakia, O. T., Arasb, E., Akdemirc, U. O., and Yilmaza, B. (2019). Biochemical oxygen demand prediction in wastewater treatment plants by using different regression analysis models. *Desalination & Water Treatment*, 157, 79-89.
7. Bhuiyan, M. A. H., Dampare, S. B., Islam, M. A., and Suzuki, S. (2015). Source apportionment and pollution evaluation of heavy metals in water and sediments of Buriganga River, Bangladesh, using multivariate analysis and pollution evaluation indices. *Environmental monitoring and assessment*, 187(1), 1-21.

8. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., and Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of The Total Environment*, 721, 137612.
9. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., Liu, F., Zuo, M., Zou, X., Wang, J., and Zhang, Y. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water Research*, 171,115454.
10. Department of Environment (DOE). (2016). River water quality report 2015, ISSN: 2226-1575, Natural Resource Management Section, Department of Environment, Bangladesh. viewed 2 February 2021, <https://doe.portal.gov.bd/sites/default/files/files/doe.portal.gov.bd/publications/b1ed783d_9b05_4f36_a83e_2173d8698023/River%20Water%20Quality%20Report%202015.pdf>
11. Dogan, E., Sengorur, B., and Koklu, R. (2009). Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management*, 90(2), 1229-1235.
12. El-Rawy, M., Abd-Ellah, M. K., Fathi, H., & Ahmed, A. K. A. (2021). Forecasting effluent and performance of wastewater treatment plant using different machine learning techniques. *Journal of Water Process Engineering*, 44, 102380.
13. Emamgholizadeh, S., Kashi, H., Maroufpoor, E., and Zalaghi, E. (2013). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology*, 11, 645-656.

14. Fathima, A., Mangai, J. A., and Gulyani, B. B. (2014). An ensemble method for predicting biochemical oxygen demand in river water using data mining techniques. *International Journal of River Basin Management,* 12(4), 357-366.
15. Federation, W.E. and APHA. (2005). *Standard methods for the examination of water and wastewater*. Washington DC: American Public Health Association (APHA).
16. Haghiabi, A. H., Nasrolahi, A. H., Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, 53(1), 3-13.
17. Hameed, M., Sharqi, S.S., Yaseen, Z.M., Afan, H.A., Hussain, A., and Elshafie, A. (2017). Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Computing and Applications*, 28(1), 893-905.
18. Hasan, M.S., Kordijazi, A., Rohatgi, P.K. and Nosonovsky, M. (2021). Triboinformatic modeling of dry friction and wear of aluminum base alloys using machine learning algorithms. *Tribology International*, 161, 107065.
19. Hayder, G., Kurniawan, I., Mustafa, H.M. (2020). Implementation of Machine Learning Methods for Monitoring and Predicting Water Quality Parameters. *Biointerface Research in Applied Chemistry*, 11(2), 9285-9295.
20. Jafari, H., Rajaee, T., & Kisi, O. (2020). Improved water quality prediction with hybrid wavelet-genetic programming model and shannon entropy. *Natural Resources Research*, 29, 3819-3840.
21. Kamal, M. M., Malmgren-Hansen, A., and Badruzzaman, A. B. M. (1999). Assessment of pollution of the River Buriganga, Bangladesh, using a water quality model. *Water science and technology*, 40(2), 129-136.

22. Khaled, B., Abdellah, A., Noureddine, D., Salim, H., and Sabeha, A. (2018). Modelling of biochemical oxygen demand from limited water quality variable by ANFIS using two partition methods. *Water Quality Research Journal*, 53(1), 24-40.
23. Kim, S., Alizamir, M., Zounemat-Kermani, M., Kisi, O., and Singh, V. P. (2020). Assessing the biochemical oxygen demand using neural networks and ensemble tree approaches in South Korea. *Journal of Environmental Management*, 270, 110834.
24. Ma, J., Ding, Y., Cheng, J. C., Jiang, F., and Xu, Z. (2020). Soft detection of 5-day BOD with sparse matrix in city harbor water using deep learning techniques. *Water Research*, 170, 115350.
25. Moniruzzaman, M., Elahi, S. F., and Jahangir, M. A. A. (2009). Study on temporal variation of physicochemical parameters of Buriganga river water through GIS (Geographical Information System) Technology. *Bangladesh Journal of Scientific and Industrial Research*, 44(3), 327-334.
26. Muharemi, F., Logofătu, D., Andersson, C., and Leon, F. (2018). Approaches to building a detection model for water quality: A Case Study. *In*: A. Sieminski, A. K. M. Nunez, Q. T. Ha, eds. *Modern Approaches for Intelligent Information and Database Systems*. Cham, Switzerland: Springer, 173-183.
27. Muharemi, F., Logofătua, D., Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of information and Telecommunication*, 3(10), 1-14.
28. Muller, A.C., and Guido, S. (2016). Introduction to machine learning with Python: A guide for data scientists. O'Reilly Media, Inc.

29. Nafsin, N., and Li, J. (2021). Using CANARY event detection software for water quality analysis in the Milwaukee River. *Journal of Hydro-environment Research*, 38, 117-128.

30. Nafsin, N., Bevers, B., Schruender, R., Liao, Q., and Li, J. (2021). Escherichia coli and Enterococci Bacteria in Lake Michigan Beach Sand. *Environmental Engineering Science*, 39.

31. Noori, R., Yeh, H. D., Abbasi, M., Kachoosangi, F. T., & Moazami, S. (2015). Uncertainty analysis of support vector machine for online prediction of five-day biochemical oxygen demand. *Journal of Hydrology*, 527, 833-843.

32. Oliveira-Esquerre, K. P., Mori, M., & Bruns, R. E. (2002). Simulation of an industrial wastewater treatment plant using artificial neural networks and principal components analysis. *Brazilian Journal of Chemical Engineering*, 19, 365-370.

33. Papry, S. A. (2019). Assessment of water quality of selected rivers in Bangladesh using FEWS software. Thesis (Masters). Bangladesh University of Engineering & Technology.

34. Perelman, L., Arad, J., Housh, M., Ostfeld, A. (2012). Event detection in water distribution systems from multivariate water quality time series. *Environmental Science & Technology*, 46(15), 8212-8219.

35. Raheli, B., Aalami, M. T., El-Shafie, A., Ghorbani, M. A., and Deo, R. C. (2017). Uncertainty assessment of the multilayer perceptron (MLP) neural network model with implementation of the novel hybrid MLP-FFA method for prediction of biochemical oxygen demand and dissolved oxygen: a case study of Langat River. *Environmental Earth Sciences*, 76(14), 1-16.

36. Rahman, M. A., and Bakri, D. A. (2010). A study on selected water quality parameters along the River Buriganga, Bangladesh. *Iranica Journal of Energy & Environment*, 1(2), 81-92.

37. Saifullah, A. S. M., Kabir, M. H., Khatun, A., Roy, S., and Sheikh, M. S. (2012). Investigation of some water quality parameters of the Buriganga River. *Journal of Environmental Science and Natural Resources*, 5(2), 47-52.

38. Sakizadeh, M. (2016). Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*, 2(1),8.

39. Sarker, A., and Pandey, P. (2015). River water quality modelling using artificial neural network technique. *Aquatic Procedia*, 4, 1070-1077.

40. Sarkar, M., Rahman, A. L., Islam, J. B., Ahmed, K. S., Uddin, M. N., & Bhoumik, N. C. (2015). Study of hydrochemistry and pollution status of the Buriganga river, Bangladesh. *Bangladesh Journal of Scientific and Industrial Research*, 50(2), 123-134.

41. Sharafati, A., Asadollah, S. B. H. S., & Hosseinzadeh, M. (2020). The potential of new ensemble machine learning models for effluent quality parameters prediction and related uncertainty. *Process Safety and Environmental Protection*, 140, 68-78.

42. Silvia, T., and Ilan, J. (2019). Artificial intelligence-based monitoring system of water quality parameters for early detection of Non-specific Bio-contamination in water distribution systems. *Water Supply*, 19(6), 1785-1792.

43. Solgi, A., Pourhaghi, A., Bahmani, R., and Zarei, H. (2017). Improving SVR and ANFIS performance using wavelet transform and PCA algorithm for modeling and predicting biochemical oxygen demand (BOD). *Ecohydrology & Hydrobiology*, 17(2), 164-175.

44. Uddin, M. J., & Jeong, Y. K. (2021). Urban river pollution in Bangladesh during last 40 years: potential public health and ecological risk, present policy, and future prospects toward smart water management. Heliyon, 7(2), e06107.

45. Venkateswarlu, T., Anmala, J., and Dharwa. M. (2020). PCA, CCA, and ANN Modeling of Climate and Land-Use Effects on Stream Water Quality of Karst Watershed in Upper Green River, Kentucky, USA. *ASCE, Journal of Hydrologic Engineering*, 25(6), 05020008-1 to 05020008-11.

46. Wang, X., Zhang, F., and Ding, J. (2017). Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific Reports*, 7,12858.

47. Whitehead, P. G., Bussi, G., Peters, R., Hossain, M. A., Softley, L., Shawal, S., Jin, L., Rampley, C.P.N., Holdship, P., Hope, R., and Alabaster, G. (2019). Modelling heavy metals in the Buriganga River system, Dhaka, Bangladesh: impacts of tannery pollution control. *Science of the Total Environment*, 697, 134090.

48. Zou, X. Y., Lin, Y. L., Xu, B., Guo, Z. B., Xia, S. J., Zhang, T. Y., Wang, A.Q., and Gao, N. Y. (2019). A novel event detection model for water distribution systems based on data-driven estimation and support vector machine classification. *Water Resources Management*, 33(13),4569-4581.

CHAPTER 6: PREDICTION OF TOTAL ORGANIC CARBON AND *E. COLI* IN RIVERS WITHIN THE MILWAUKEE RIVER BASIN USING MACHINE LEARNING METHODS

6.1 Introduction

Surface water is one of the most important non-renewable natural resources used for numerous purposes including drinking water, public use, irrigation, and the aquatic environment. However, surface water quality deteriorates because of the discharge of municipal and industrial wastes into rivers, lakes, and reservoirs. Various contaminants such as microbial pollutants, inorganic matter, synthetic and volatile organic compounds, radioactive material enter the source water through different pathways, including point source pollution (i.e., municipal and industrial wastewater outfalls) and nonpoint source pollution (i.e., stormwater runoff and atmospheric deposition). Therefore, timely monitoring of surface water quality is essential to ensure the safety of the source water. Any significant change in the physical, chemical, and biological characteristics of water indicates the presence of contaminants in water. To ensure the supply of good quality water, treatment of wastewater, and maintaining aquatic life ecosystem, it is crucial to monitor water quality parameters, such as biochemical oxygen demand (BOD), chemical oxygen demand (COD), dissolved oxygen (DO), total organic carbon (TOC), bacteria level, etc. Water pollution is classified into several categories, including toxic pollution, organic pollution, nutrient pollution, microbial contamination, sediment pollution, and radiological pollution (Southeastern Wisconsin, 2007). Microbial pollution caused by the presence of bacteria and viruses (i.e., *Giardia*, *E. coli*, *Cryptosporidium*) in sewage treatment plants, septic systems, wildlife, and agricultural livestock operations transmits water-borne infectious diseases and

causes acute illnesses of human health. Organic pollution caused by oxygen demanding organic substances, including carbonaceous and nitrogenous organic compounds can result in DO level depletion in surface water which severely affects fish and aquatic life. Organic compounds generally include chemicals containing carbon that can be naturally produced by organisms or synthetic material, such as detergents, disinfectants, dye agents, flavoring agents, flame retardants, fragrances, insect repellants, plasticizers, and solvents. Synthetic organic contaminants (i.e., atrazine, polychlorinated bi-phenyl (PCBs)) and volatile organic compounds (i.e., benzene, vinyl chloride, and styrene) from agricultural and industrial activities, water reclamation facilities, gas stations, petroleum production, landfills, and stormwater runoff can negatively affect human health and fish and aquatic life. Total organic carbon (TOC) is one of the convenient ways of direct measure of organic contamination in surface water and TOC measurements indicate the number of carbon-containing compounds. A high level of carbon or organic content in surface water stimulates bacterial growth resulting in a rapid rate of organic matter decomposition. Decomposition of organic matter contributes to the depletion of oxygen supply in surface water to levels below the concentration required for maintaining aquatic life and ecosystems. In addition, the concentration of fecal indicator bacteria (FIB) is measured to assess surface water quality for drinking and recreational purposes. Two groups of FIB (fecal coliform and *Escherichia coli* (*E. coli*)) are commonly used as surrogate to indicate the presence of fecal matter and examined in surface water. The presence of high concentrations of FIB indicates a high probability of pathogenic microbial contamination in water. A previous study (Nafsin et al., 2022) analyzed the concentration of *E. coli* and Enterococci bacteria in Lake Michigan beach sand and developed methods for efficient prediction of public health outcomes associated with microbial contamination.

This research focuses on analysis of water quality in terms of organic pollution (TOC concentration) and fecal matter contamination (*E. coli* level) in natural streams within the Milwaukee River basin. The Milwaukee River basin is located around seven counties, including several cities, towns, and villages, serving about 1.3 million people. About 90% of the basin's population live in the southern portion of the basin in the Milwaukee County of Wisconsin. The study area is located within the Milwaukee River basin that includes three major rivers (Milwaukee River, Menomonee River, and Kinnickinnc River) flowing into the harbor of Milwaukee, Wisconsin. The Milwaukee River starts from Fond du Lac County, north of Wisconsin and flows towards the south of Wisconsin in downtown Milwaukee and discharges into Lake Michigan. Menomonee River and Kinnickinnic River are the two main tributaries of the Milwaukee River. The majority of the monitoring stations in this study are located at the southern portion of the Milwaukee River Basin. Urban and agricultural runoff, municipal and industrial point sources, construction site erosion, stream bank erosion, stream and wetland modification, and contaminated sediments are the major contributors of the degradation of stream water quality of the Milwaukee River system. There are four municipal wastewater treatment facilities along the Milwaukee River south watershed. Also, land use has the greatest impact on the degradation of surface water quality of the Milwaukee River system. Due to urbanization and growth of population, rural lands are converted for business and homes, resulting in increased point source and non-point source pollution to surface water. The most urbanized basin, Milwaukee River basin is greatly affected by urban runoff. Because of high percentage of impermeable surfaces (i.e., paved surface) in urban area, rainfall or melting snow washes the pollutants off parking lots, construction sites, and streets. The storm sewers and roadside ditches transport the untreated pollutants directly into the surface water. In urban areas,

stormwater is considered as one of the significant sources of pathogenic microorganisms and organic matter pollution (Burzynski, 2001). Previous study (Paule-Mercado et al., 2016) also indicated that urban runoff had the highest level of fecal contamination with urban land use land cover changes.

To detect surface water contamination rapidly and more accurately, Early Warming Systems (EWS) have been developed. Sensor technology is integrated with an early warning system (EWS) that allows identification of contamination events and creates an alarm on that event so that knowledgeable response decisions can be made to ensure the safety of water supplied to consumers (Dogo et al., 2019; Gullick et al., 2003). Nafsin and Li (2021) and Nafsin et al. (2022) applied statistical event detection software CANARY for analysis of surface water quality of the Milwaukee River and Lake Michigan, respectively. The application of machine learning (ML) techniques is useful in predicting water quality as the models can provide data-driven decisions by extracting predictive information from a large dataset. The models can be re-trained as new data are available and improved their prediction performance accuracy with the increasing training size. The performances of the models can be evaluated using different performance metrics, which makes it possible to validate the models for a future dataset (Babbar and Babbar, 2017). Many researchers used ML models such as Artificial Neural Network (ANN), Logistic Regression (LR), Support Vector Machine (SVM) for water quality monitoring and prediction of water quality variables. Previous literatures indicated that AI can achieve good results in anomaly detection in the data mining research area (Ahmed et al., 2015; Muharemi et al., 2018). Several other studies (Muharemi et al., 2019; Perelman et al., 2012; Silvia and Ilan, 2019; Zou et al., 2019) also developed ML models for contamination or event detection in water distribution systems. Emamgholizadeh et al. (2013) and Sarker and Pandey (2015) investigated

the application of different types of ANN (MLP and RBF model) to river water quality simulation. The authors developed models to predict water quality parameters such as DO, BOD, and COD and evaluated the prediction performances of ANN models using the R^2 value and root mean squared error (RMSE). Other studies (Ahmed et al., 2019; Haghiabi et al., 2018; Hayder et al., 2020, Najah et al., 2013) also investigated the performances of different ML models for predicting water quality parameters of natural source water. Several studies (Ahmed et al., 2019a; Babbar and Babbar, 2017; Bui, et al., 2020; Chen et al., 2020; Hameed, et al., 2016; Sakizadeh, 2016; Wang et al., 2017) used ML techniques to predict water quality based on WQI (Water Quality Index) and WQC (Water quality class). Chen et al. (2020a) explored the increasing popularity of the ANN model for water quality predictions in rivers, lakes, and wastewater treatment plants (WWTP). Many research studies from environmental engineering during the year 2008 to 2019 used ML models to predict parameters such as DO, BOD, COD, pH, and temperature. Based on the review of 151 papers, the authors emphasized 23 water quality parameters that have been used in different modeling problems in lakes, rivers, and WWTP.

Several studies explored different ML techniques in predicting TOC to characterize hydrocarbon potential of source rocks, soil, organic shale, and mudstone (Lawal et al., 2019; Mandal et al., 2021; Ouadfeul & Aliouane, 2015; Rong et al., 2021; Wang et al., 2021). However, to the best of the authors' knowledge limited studies have been made for developing TOC prediction models in natural streams. Yeon et al. (2009), Goz et al. (2019), and Kim et al. (2021) explored the application of ANN, kernel extreme machine learning (KEML), extreme machine learning (EML) models to estimate total organic carbon of rivers. In addition, several studies (Bourel et al., 2021; David et al., 2011; He et al., 2008; Khan et al., 2021; Mohammed &

Seidu, 2018; Paule-Mercado et al., 2016) investigated regression-based techniques and mechanistic models for microbial analysis (fecal contamination) of surface water and groundwater using physicochemical and hydrometeorological parameters. However, development of such predictive models for fecal indicator bacteria analysis based on physicochemical and hydrometeorological parameters is site and bacteria group specific. The survival of FIB can be affected by complex interactions among physicochemical, hydrometeorological parameters, and land use pattern of the study area.

In this study, several standalone and ensemble-hybrid ML algorithms were developed that can potentially be very effective tools in predicting TOC and *E. coli* in natural streams of the Milwaukee River system. Water quality data with a greater amount of spatial and temporal variations for 20 years sampling period collected from a public data source system were used for the analysis. Living microorganism's behavior is harder to predict than physical and chemical processes. In this study, living microorganisms' behavior (*E. coli* concentration MPN/100 mL) were predicted using ML algorithms that can explain the greater amount of unexplainable variation in the data through the independent physical and chemical water quality parameters. These algorithms analyzed the real-time data of source water, found the underlying pattern in a large volume of data using a mapping function, and identified complex relationships between the output and inputs, which are unattainable by the traditional or process-based methods used for water quality analysis. The significance of the study is the application of the developed ensemble-hybrid models for prediction of total organic matter pollution and *E. coli* contamination in surface water. The developed ensemble-hybrid methods were not previously used for TOC and *E. coli* prediction that can provide a reliable and direct approach to complement existing monitoring techniques in the Milwaukee River system with satisfactory

prediction accuracies. Also, limited studies have been conducted so far for developing TOC prediction models using ML in natural streams. The research purpose is to develop prediction models that would provide a more accurate, reliable, and direct measure of organic matter pollution and FIB contamination in the Milwaukee River system. More interestingly, using the same dataset used for TOC prediction, prediction models for *E. coli* were developed and tried to explain the variability in living microorganisms' behavior based on the specific physicochemical parameters, such as total solids (TS), total suspended solids (TSS), volatile suspended solids (VSS), chlorophyll a, turbidity, pH, electrical conductivity (EC), temperature, dissolved oxygen (DO), nitrate, alkalinity, total phosphorous (Total P), chloride, BOD_5, TOC, dissolved organic carbon (DOC), and identify the most influential physical and chemical water quality parameters in predicting *E. coli*. The goal is to make a comprehensive assessment of the differences in prediction performances of the models between TOC and *E. coli* with data collected from a specific study area.

The main objective of this study is to predict TOC level and *E. coli* concentration in three major rivers: Milwaukee River, Menomonee River, and Kinnickinnic River within the Milwaukee River basin during a sampling period of 2000-2020 using ML methods. The efficiencies of different standalone regression ML models: Artificial Neural Network (ANN), Support Vector Machine (SVM), Gradient Boosting Machine (GBM), Random Forest (RF) and ensemble-hybrid models: RF-SVM, ANN-SVM, GBM-SVM, RF-ANN, GBM-ANN, and RF-GBM were developed and evaluated for predicting TOC and *E. coli* of the river system using the specific water quality parameters. The prediction performances of the models were significantly improved by integrating different standalone models and developing the ensemble-hybrid models. Also, identifying the most influential physicochemical parameters in predicting both

TOC and *E. coli* is one of the objectives of the study. In addition, a comprehensive assessment of the employed ML techniques was conducted and the differences in model performances were evaluated for predicting two different outputs (TOC and *E. coli*) using the specific dataset.

6.2 Materials and Methods

6.2.1 Study area and data collection

Water quality monitoring data of the three primary rivers (Milwaukee River, Menomonee River, and Kinnickinnic River) in Wisconsin were collected from the Milwaukee Metropolitan Sewerage District (MMSD). There was a total number of 32 monitoring sites that are located at the Milwaukee, Waukesha, Ozaukee, and Washington county in Wisconsin (Figure 6.1). After data cleaning and processing, the complete dataset used for the analysis was composed of 5976 sample observations with 18 water quality parameters: total solids (TS), total suspended solids (TSS), volatile suspended solids (VSS), chlorophyll a, turbidity, pH, electrical conductivity (EC), temperature, dissolved oxygen (DO), sampling depth, nitrate, alkalinity, total phosphorous (Total P), chloride, biochemical oxygen demand (BOD_5), total organic carbon (TOC), dissolved organic carbon (DOC), and *E. coli*. The data collection period was during different seasonal months of 2000-2020 so that the generalized ML models could be developed and trained with water quality data with possible local and seasonal variations during the 20 years sampling periods.

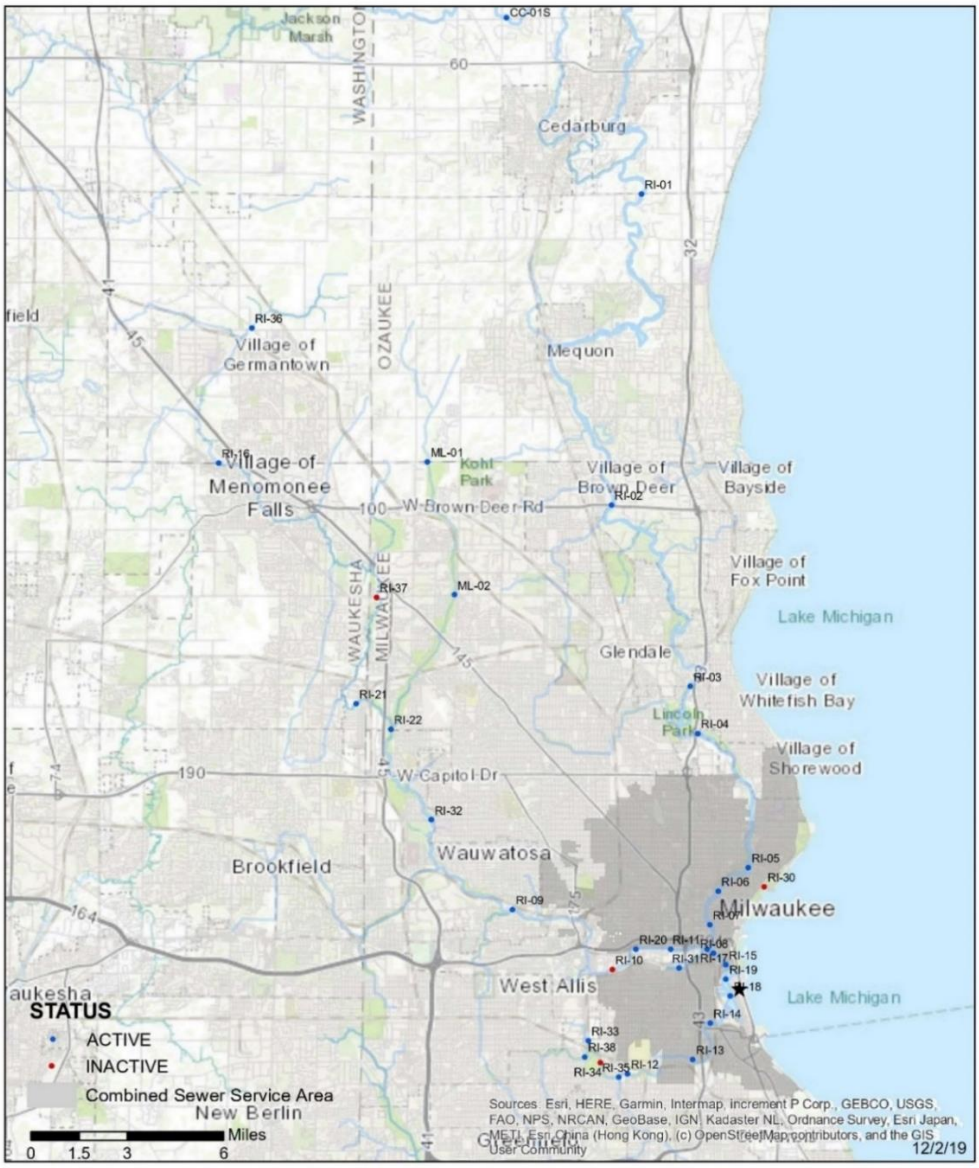


Figure 6.1. Water quality monitoring sites (total active monitoring sites: 32) of Milwaukee River, Menomonee River, and Kinnickinnic River in Wisconsin (source: MMSD)

6.2.2 Data preprocessing for ML models

ML models were developed to predict TOC and *E. coli* in the Milwaukee River system using the water quality dataset. Before applying the model for prediction to a new dataset, the

model accuracy should be ensured. In the model development process, data splitting was done into a training set for model training and validation and a test set for model performance evaluation. In this analysis, 75% of the dataset (4482 samples) with target value were considered as the training and 25% (1494 samples) as the testing set. Again, 75% of the training set was considered for training (3362 samples) and 25% for validation (1120 samples). Another important step for data preparation is data normalization that transforms the data into a standard range. Machine learning models such as ANN and SVM are sensitive to data scaling, and they require all the features to be on the same scale to perform well. We used the '*RobustScaler*' in *scikit learn* python library that used median and quartiles (25^{th} and 75^{th} quartile) to standardize the features. Same scaling and transformation were applied to the training set and testing set for the models to perform efficiently on the test set.

6.2.3 Machine learning tools

Several supervised ML algorithms such as ANN, SVM, RF, and GBM were used to develop regression models for predicting water quality parameters such as TOC and *E. coli* of the natural streams in Wisconsin. In regression-based algorithm, the models learned from the input data with the corresponding output value, and based on that learning, predicted the output for new observations. The analysis was performed using a python program with a built-in machine learning toolkit *scikit-learn* (version 0.21) (Pedregosa et. al., 2011).

Artificial Neural Network (ANN) is a good approach for regression problems with complex datasets. The model consists of one input layer, one or several hidden layers, and one output layer. The hidden layers include many interconnected units (neurons) arranged with the input vectors to convert them into output using an activation function. In a feed-forward network such

as Multilayer Perceptron (MLP), each unit feeds its output to all the units on the next layer. In this analysis, we used the MLP Neural Network with two hidden layers and five units in each layer that resulted in the best model performances for predicting both TOC and *E. coli*. The ANN model architecture used for prediction of TOC and *E. coli* are shown in Figure 6.2.

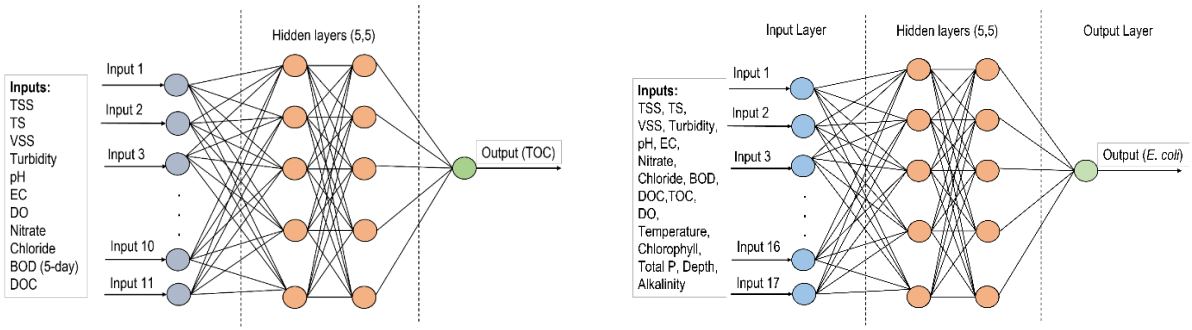


Figure 6.2. ANN model architecture for TOC prediction (left) and *E. coli* prediction (right)

Support vector machine (SVM) is used as 'support vector regressor' (SVR) in regression problems which finds a decision boundary or hyperplane to classify data points appropriately. SVM uses a kernel method that converts the original input 2-dimensional data space into higher dimensional feature space. Different kernel functions such as radial basis function (RBF), sigmoid kernel, linear kernel, and polynomial kernel are available to transform data into higher dimensional space (Muller and Guido, 2016). We used the RBF kernel function and optimized the two key parameters: regularization parameter (C) and kernel width (gamma).

Random Forest (RF) is an ensemble ML model that combines multiple decision trees to build an effective prediction model (Figure 6.3). The model makes different random choices to develop several independent trees. The trees are randomized by selecting the data points to build trees and the maximum features in each split test. Each tree in the forest predicts the output, and the final output is determined by averaging the outputs from all the decision trees.

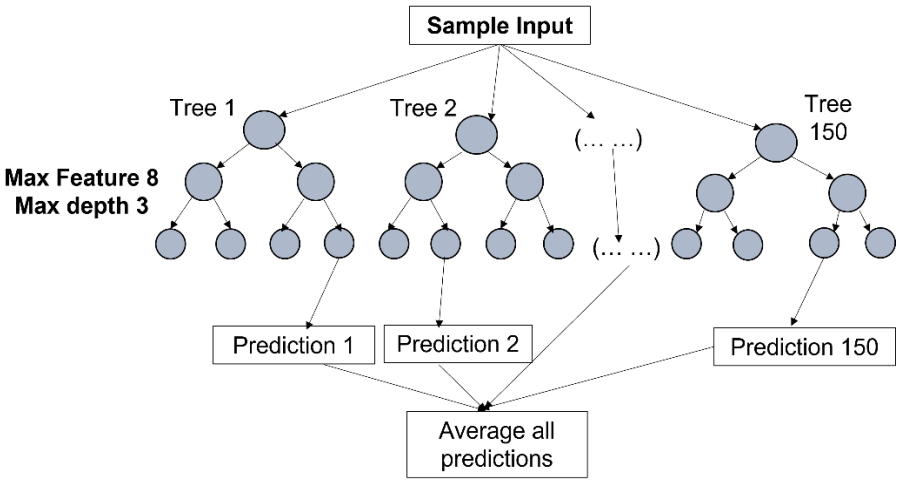


Figure 6.3. Random Forest model structure

Gradient boosting machine (GBM) works by building multiple models or decision trees sequentially and reducing the errors from the previous model. Each decision tree takes a portion of the input data and make predictions. The new models are built over the errors or residuals of the previous predictions. Several decision trees are added iteratively to improve the prediction performance. The degree to which each model is allowed to correct the errors from the previous tree is controlled by the key parameters: number of decision trees and learning rate.

In addition, several ensemble-hybrid models: RF-SVM, ANN-SVM, GBM-SVM, RF-ANN, GBM-ANN, and RF-GBM were developed by integrating the standalone traditional ML algorithms. An ensemble meta estimator ‘voting regressor’ was used to fit the dataset on each standalone models in this hybridization process. The final prediction of the hybrid model was determined by averaging the individual prediction of each standalone model. The contributing models were optimized to achieve the best performance of the ensemble-hybrid model. The details of the ensemble-hybrid ML algorithms are provided in chapter 5, section 5.2.3.

6.2.4 Performance evaluation metrics

Evaluation is a key part of the model development process of supervised machine learning. Evaluation is performed on a specific test dataset, and the model with the 'best' evaluation metric is selected. In this analysis, the statistical metrics: coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) were used to evaluate the performances of different models.

The coefficient of determination (R^2) or 'goodness-of-fit' describes the variability in the dependent variables that can be explained by the independent variables. The range of value is from 0.0 to 1.0, which means 0% to 100% variation in the dependent variable can be explained by independent variables. Mean absolute error (MAE) represents the average of the absolute difference between the actual and predicted value. It measures the average of the residuals in the dataset. The standard deviation of the residuals is indicated by root mean squared error (RMSE). The RMSE score measures the distance between the regression fit line and the actual data. The amount of error in regression models is also determined by mean squared error (MSE) which measures the average squared difference between the actual and predicted value. A high value of R^2 and low values of errors indicate satisfactory model performance that can predict the output accurately. The equations for calculating the performance metrics are provided in chapter 5, section 5.2.4.

6.3 Results and Discussion

6.3.1 Statistical analysis of water quality data

Statistical analysis was performed to visualize water quality data of the Milwaukee River, Menomonee River, and Kinnickinnic River during the sampling period of 2000-2020. The basic statistical parameters such as minimum, maximum, mean, standard deviation (SD), and coefficient of variation (CV) of the water quality data are presented in Table 6.1. The coefficient of variations of all the parameters were ranging from 4.73% to 451.82%. During the sampling period, water quality of the Milwaukee River system varied significantly because of *E. coli*, BOD_5, and TSS with high values of CV. The higher value of CV indicates relatively high variability in the dataset. Among the parameters, *E. coli* dataset had the highest variability during the 20 years sampling period with a CV of 452.82%. *E. coli* level varied within a range of 0.00-250000 MPN/100 mL with a mean value of 3643 MPN/100 mL. The level of TOC in the river system were found within a range of 0.67 mg/L-190 mg/L with a mean value of 8.25 mg/L and CV of 88.13%. Also, BOD_5 and TSS had high variability with %CV greater than or about 300%. Other parameters such as turbidity, VSS, chlorophyll a, and chloride resulted in greater than 100% CV.

Table 6.1. Statistical analysis of water quality parameters during the sampling period of 2000-2020

Parameter	**Unit**	**Minimum**	**Maximum**	**Mean**	**Standard deviation**	**Coefficient of variation %**
TS	mg/L	100.00	8400.00	633.57	412.17	65.05
TSS	mg/L	0.80	2700.00	20.18	60.37	299.20
VSS	mg/L	0.30	260.00	5.07	8.49	169.71
Chlorophyll a	mg/m^3	0.08	280.00	9.83	15.68	159.45
Turbidity	NTU	0.60	378.00	14.19	23.49	165.52
pH	-	5.50	9.93	8.05	0.38	4.73
EC	µS/cm	101.00	15600	1045.18	765.44	73.23
Temperature	^oC	-0.50	33.44	14.12	7.92	56.12
DO	mg/L	0.00	25.70	9.88	3.36	33.97
Nitrate	mg/L	0.00	4.30	0.81	0.56	69.68

Alkalinity	mg/L	4.50	440.00	223.38	71.04	31.80
Total P	mg/L	0.00	2.60	0.12	0.11	88.38
Chloride	mg/L	5.00	3100.00	174.84	210.51	120.40
BOD_5	mg/L	0.00	310.00	3.98	12.00	301.03
TOC	mg/L	0.67	190.00	8.25	7.27	88.13
DOC	mg/L	0.52	190.00	7.89	7.04	89.27
E. coli	MPN/100 mL	0.00	250000	3643.51	16461.95	451.82

Pearson's correlation analysis was performed at 0.05 level of significance to identify the input parameters that could impact the output variables such as TOC and *E. coli*. The results in Table 6.2 indicated significant correlations (p-value<0.05) between TOC and input variables. Among the parameters, DOC was strongly positively correlated with TOC (R-value=0.975, p<0.05). BOD_5 had a moderately strong linear correlation with TOC (R-value=0.725, p<0.05). Total solids, chloride, and EC had weak linear correlations with an R-value of 0.423, 0.408, and 0.405, respectively. Temperature and pH were negatively correlated with TOC. Oher parameters had a very weak or almost no linear correlation with TOC with an R-value of less than 0.1. Similarly, Table 6.3 shows the Pearson's correlation coefficient between *E. coli* and other input variables. The results indicate that the physicochemical parameters had a very weak (R-value < 0.3) or about no linear correlation with *E. coli*.

Table 6.2. Pearson's correlation coefficient (R-value) between TOC and other parameters at 0.05 level of significance

Parameter	**Correlation coefficient**	**P-value**	**Parameter**	**Correlation coefficient**	**P-value**
TS	0.423	0.000	Temperature	-0.147	0.000
TSS	0.036	0.005	DO	-0.018	0.003
VSS	0.091	0.000	Nitrate	0.055	0.000
Chlorophyll	-0.010	0.032	Alkalinity	0.048	0.000

Turbidity	0.013	0.039	Chloride	0.408	0.000
pH	-0.105	0.000	BOD_5	0.725	0.000
EC	0.405	0.000	DOC	0.975	0.000
Depth	-0.054	0.000	*E. coli*	0.058	0.000
Total P	0.089	0.000			

Table 6.3. Pearson's correlation coefficient (R-value) between *E. coli* and other parameters at 0.05 level of significance

Parameter	**Correlation coefficient**	**P-value**	**Parameter**	**Correlation coefficient**	**P-value**
TS	-0.056	0.000	Temperature	0.108	0.000
TSS	0.110	0.000	DO	-0.136	0.000
VSS	0.178	0.000	Nitrate	-0.049	0.000
Chlorophyll	-0.001	0.027	Alkalinity	-0.247	0.000
Turbidity	0.151	0.000	Chloride	-0.034	0.009
pH	-0.131	0.000	BOD_5	0.079	0.000
EC	-0.076	0.000	DOC	0.047	0.000
Depth	0.001	0.048	TOC	0.058	0.000
Total P	0.258	0.000			

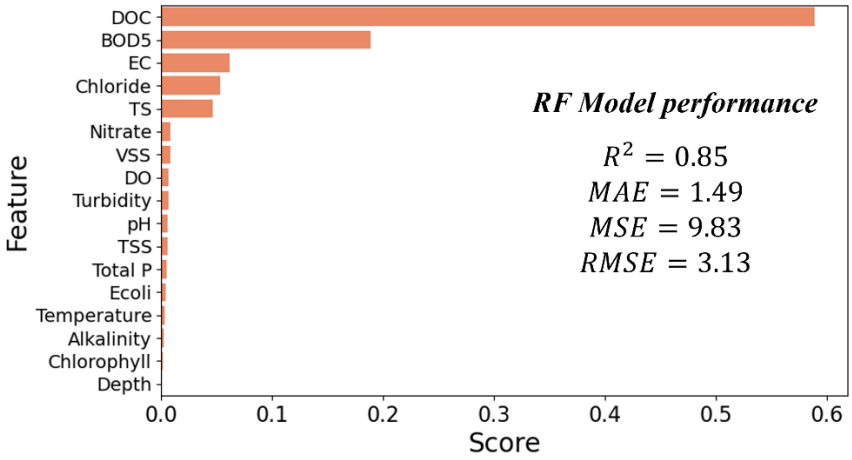
6.3.2 Feature importance

Although Pearson's correlation analysis indicated the statistical relationship between the independent and dependent variables, it can only explain linear relationships without considering the non-linearity of a complex dataset. In this analysis, the decision-tree based RF algorithm was

used to identify the relative importance of each feature in predicting the output. The feature importance assigns a score typically numbered between 0 and 1 to individual features and normalized to sum to 1. The higher the score, the more relevant the feature is for predicting the output. Figure 6.4 and Figure 6.5 indicate the feature importance chart for prediction of TOC and *E. coli*, respectively. For the particular test-train split of the dataset, DOC had the largest feature importance score of 0.58 in prediction of TOC. BOD_5 was the second most important feature, with a feature score of 0.18. EC, chloride, and TS had feature scores of 0.07, 0.06, and 0.05, respectively. The results indicated that DOC and BOD_5 were the two most important variables for predicting TOC with comparatively higher feature importance scores (greater than 0.1). TOC, DOC, and BOD_5 indicate the organic matter pollution in water and wastewater. TOC is a measure of organic carbon that can present in water in different forms. TOC consists of dissolved organic carbon (DOC) and non-dissolved organic carbon (NDOC). DOC is considered as the particulate TOC that can pass through a 0.45 µm filter, while larger size of TOC is known as NDOC. Because of the direct association of TOC with DOC, Pearson's analysis and feature importance analysis by RF algorithm showed significant correlation between the parameters. Also, TOC and BOD_5 are correlated as they both indicate the presence of organic matter in water. TOC provides a direct measure of organic carbon while BOD_5 measures the amount of oxygen consumed by microorganisms to oxidize soluble organic matter. A high content of organic carbon increases the growth of microorganisms and as a result consumption of DO increases which eventually increases BOD_5.

When considering all the input features, the RF model performance accuracy were found as 85%. However, the important features were extracted from the feature importance chart, eliminated the features with lower scores, and developed models with only the feature

importance. The results indicated that the model performance was significantly improved with higher accuracy (accuracy 90%) and lower values of error with the input variable combination of BOD_5, DOC, EC, chloride, TS, nitrate, VSS, DO, turbidity, pH, and TSS. Based on the analysis, the input combination of the 11 water quality parameters out of the 18 parameters was selected that had comparatively higher feature importance scores to develop TOC prediction models. The performances of other developed ML models are shown in section 6.3.4.



DOC: Dissolved organic carbon; BOD5: 5-day Biochemical oxygen demand; EC: Electrical conductivity; TS: Total solids; VSS: Volatile suspended solids; DO: Dissolved oxygen; TSS: Total suspended solids; Total P: Total Phosphorous; *E. coli: Escherichia coli*

Figure 6.4. Relative importance of input features for prediction of TOC

For prediction of *E. coli*, although Pearson's analysis indicated poor correlation between *E. coli* and other parameters, the feature importance analysis computed from RF algorithm were able to capture non-linear relationships between the input and output. The result indicated that BOD_5 was the most important variable for predicting bacteria with a feature importance score of 0.13. Other influential variables were DO, Total P, temperature, turbidity, and nitrate with feature scores of 0.12, 0.10, 0.09, 0.08, and 0.07 respectively. The feature importance scores were comparatively lower (less than 0.14) for *E. coli* prediction than TOC as shown in Figure

6.5. For developing prediction models of *E. coli* only the features that had relatively higher feature importance score were selected. The selected input variables are BOD_5, DO, total phosphorous, temperature, turbidity, nitrate, and alkalinity.

The result indicates that *E. coli* concentration was associated with the level of BOD_5 and DO in surface water. With an increasing growth of microorganisms, the rate of decomposition of organic matter also increases, which results in dissolved oxygen level depletion and increased BOD level in water. For survival and growth of microorganisms, enough supply of nutrients such as nitrogen and phosphorous and appropriate temperature are required. However, the requirements of temperature vary with different species. For example, *E. coli* requires temperature within a range of 25^{o}-40^{o}C. Turbidity can also affect the microbial growth in water. Bacteria has the potential to attach to the surface of particulate turbidity causing material (TCM) influencing the inactivation of microorganisms (Farrell et al., 2018). Alkalinity also contributes to some extent in prediction of bacteria level. Previous study (Tan et al., 2018) shows that an appropriate alkaline environment can be effectively inhibit growth of microorganism through inactivation of ATP synthesis. Although the influence of the physical and chemical parameters on prediction of *E. coli* was poor, the decision tree-based ML algorithm were able to extract data-driven information about the non-linear relationships that could exist between the inputs and output.

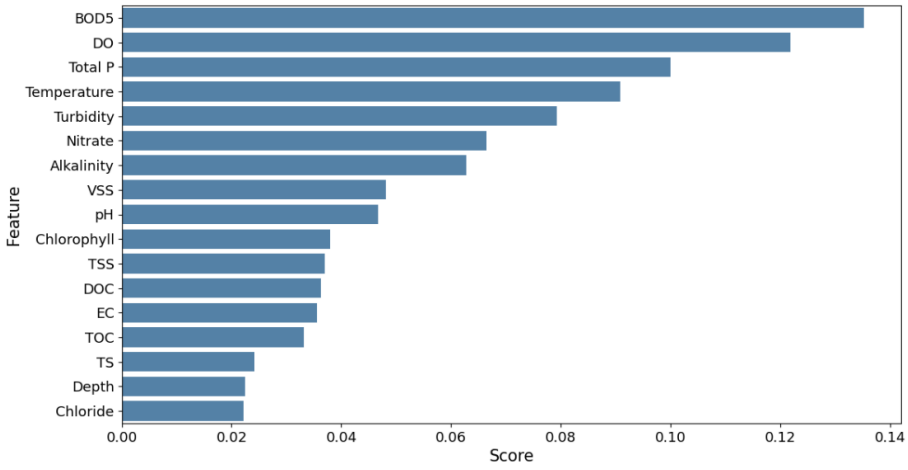


Figure 6.5. Relative importance of input features for prediction of *E. coli*

6.3.3 Optimization of model parameters

The generalization performances of ML models were improved by the model's parameter tuning. The grid search and five-fold cross validation method was used and all possible combinations of the hyperparameters that control the learning process were examined. For the kernelized SVM, the two strongly correlated key parameters are the regularization parameter C and kernel-specific parameter gamma. For ANN, the key model parameters are the regularization parameter alpha and activation function for the hidden layers. The grid search method found the best parameters for SVM, ANN, RF, and GBM models, and with the parameters, fit the models on the whole training set that yielded the best cross-validation performance. Also, the models were evaluated using the test set to identify how well the best-found parameters were generalized. The optimized key parameters used for the models are summarized in Table 6.4 and Table 6.5.

Table 6.4. Model key parameter selection for prediction of TOC

Model's key parameters	Optimal values									
	ANN	**SVM**	**RF**	**GBM**	**RF-SVM**	**GBM-SVM**	**RF-GBM**	**RF-ANN**	**GBM-ANN**	**ANN-SVM**
Hidden layer	(5,5)							(4,4)	(4,4)	(5,5)
activation	'relu'							'relu'	'relu'	'relu'
alpha	0.1							0.005	0.01	0.22
solver	'lbfgs'							'lbfgs'	'lbfgs'	'lbfgs'
kernel		rbf			'rbf'	'rbf'				'rbf'
C		500			500	700				650
gamma		0.001			0.001	0.0001				0.0001
n-estimators			150	100	150	500	100	50	250	
max-features			8		8		5	8		
max-depth			3	2	6	3	6	5	3	
learning rate				0.08		0.085	0.08		0.085	

Table 6.5. Model key parameter selection for prediction of *E. coli*

Model's key parameters	Optimal values							
	ANN	**SVM**	**RF**	**GBM**	**RF-GBM**	**RF-ANN**	**GBM-ANN**	**ANN-SVM**
Hidden layer	(5,5)					(4,4)	(4,4)	(5,5)
activation	'relu'					'relu'	'relu'	'relu'
alpha	0.025					0.02	1.1	0.6
solver	'lbfgs'					'lbfgs'	'lbfgs'	'lbfgs'
kernel		rbf						'rbf'
C		500						500
gamma		0.05						0.06
n-estimators			200		10	12	100	
max-features			8		4	4		
max-depth			9	2	9	11	5	
learning rate				0.08	0.06		0.08	

6.3.4 Model performance evaluation for TOC prediction

The performances of the developed ML models were evaluated using statistical measures such as coefficient of variation (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE). In regression analysis, the most important statistical measure is the R^2 -value (goodness-of-fit) which represents the proportion of variance in the output that the model can explain based on the features. The range of R^2 -value is from 0 to 1, where 0 indicates the model is unable to explain the variability in the data resulting in high errors while an R^2 -value of 1 represents the model is a good fit to explain the variability accurately. Based on the R^2, the regression models were evaluated whether they resulted in very satisfactory ($R^2 \geq 0.90$), fairly satisfactory ($0.5< R^2 < 0.90$), or unsatisfactory model execution ($R^2 \leq 0.50$). For the nonlinear ML models, besides R^2 value, other statistical measures such as standard errors (MAE, MSE. And RMSE) were also determined to evaluate the models' performance efficiencies. A good-fit model with higher R^2-value also results in lower errors.

Table 6.6 shows the performances of the developed four standalone and six hybrid ML models for TOC prediction based on the coefficient of determination (R^2-value), RMSE, MSE, and MAE. The results indicated that the standalone ML models had prediction accuracies within a range of 89.9%-95.8% which indicate the models perform satisfactorily in predicting TOC and the models are considered as good-fit for the dataset. The ensemble-hybrid models were also developed by integrating two standalone ML algorithms that further improved the prediction performances of the models. The prediction accuracy of the hybrid models ranged from 94.8%-97.0%. Based on the R^2-value and errors, the developed models were ranked from the best performed to the lowest performed model as shown in Table 6.6. Among the developed TOC prediction models, the best four performing models were the ensemble-hybrid models: ANN-

GBM, SVM-GBM, ANN-SVM, and ANN-RF with prediction accuracies higher than 96%. The hybrid model ANN-GBM outperformed others with an R^2 value of 0.97 when using the selected input features computed from the feature importance analysis. The best performing hybrid model ANN-GBM exhibited the lowest values of MAE (0.664), MSE (2.334), and RMSE (1.528). The performance metrics indicated that the employed regression models can efficiently predict TOC based on the input combination of the features: BOD_5, DOC, EC, chloride, TS, nitrate, VSS, DO, turbidity, pH, and TSS. The linear correlation between the actual and predicted TOC for the employed ML models are presented in Figure 6.6-Figure 6.8. From the scatter plots it can be observed that most of the data points were best fit with the regression line that explained the percent of variance of the output through the input variables. The deviation of few data points from the regression line were indicated by the degree of errors. The time variation graphs in Figure 6.6, Figure 6.7, and Figure 6.8 indicate that the prediction models exhibited small deviation between the predicted value and actual value for each sample observation of the test set, verifying good generalization capacity of the models with satisfactory R^2 values.

In the hybridization process of model development, the two base algorithms were integrated together to develop a model with greater flexibility and higher prediction accuracy than the standalone models. For example, the standalone model ANN and GBM had TOC prediction accuracies of 95.8% and 93.6%, respectively. Although both standalone models performed satisfactorily, the performances of the models were further improved by hybridizing the GBM with ANN algorithm with a TOC prediction accuracy of 97.0%. When developing the ANN-GBM model, the hyperparameters for the ANN (hidden layer size with the number of nodes in each layer, activation function, solver, and alpha) and GBM (learning rate, n_estimators, max_depth) were defined. The hyperparameters were optimized to achieve the best

performance of the hybrid model. An ensemble meta-estimator 'VotingRegressor' was applied to fit the standalone algorithms each on the dataset. The final prediction was determined by averaging the individual predictions of the regression models.

Among the standalone models, ANN could better model data with high volatility and non-constant variance and learn hidden relationships without imposing any fixed relationships in the data. For the specific dataset with high variance, ANN resulted in the highest prediction accuracy of 95.8% among the standalone models. The prediction accuracy was improved (R^2 = 0.97) when integrating the ANN model with another standalone regressor, GBM. The ensemble-hybrid model ANN-GBM outperformed others because of the significant advantages of ANN over other regression models, such as ANN's ability to learn and model complex non-linear relationships between the dependent and independent variables and establish all possible interactions between the dependent variables without requiring the need for making assumptions about data properties, data distribution (such as normality assumption), and specific hypothesis for testing. The ANN model benefitted from the mathematical functions of hidden layers consisting of neurons that assigned weights to the inputs, directed them to an activation function, and performed specific non-linear transformations of the input data. The activation function allowed complex functional mapping of the network's input and output with the dataset of non-linearity. In addition, the boosting mechanism of GBM with properly optimized hyperparameters allowed to build individual decision trees at a time and learn from the mistakes of previous trees to improve the overall performance sequentially with each iteration. GBM overcame the errors of decision trees by using gradients in the loss function and optimizing the model's coefficients to fit the underlying data. The incorporation of boosting mechanism along with the non-linear transformation of the input data using an activation function allowed extraction of specific

patterns from the data and minimized the difference between the actual and the predicted output, resulting in a more powerful ensemble-hybrid model ANN-GBM.

Table 6.6. Model performances for 4 standalone and 6 hybrid algorithms for prediction of TOC

Algorithms	MAE	MSE	RMSE	R^2	Rank order
ANN	0.750	2.788	1.669	0.958	5
GBM	0.718	4.315	2.077	0.936	8
SVM	0.807	5.276	2.297	0.921	9
RF	1.177	6.739	2.596	0.899	10
ANN-GBM	**0.664**	**2.334**	**1.528**	**0.970**	**1**
SVM-GBM	0.652	2.366	1.538	0.965	2
ANN-SVM	0.672	2.394	1.547	0.964	3
ANN-RF	0.703	2.626	1.620	0.961	4
SVM-RF	0.722	2.888	1.699	0.957	6
RF-GBM	0.738	3.514	1.875	0.948	7

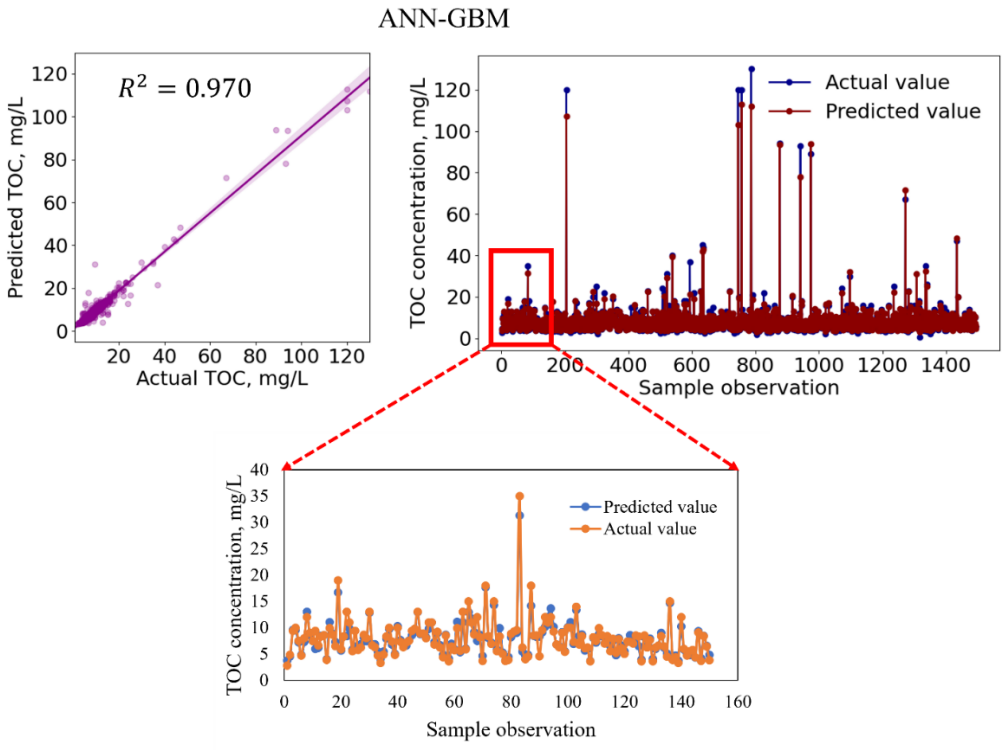


Figure 6.6. Regression analysis plot (top left) and time variation graph comparing the actual and predicted TOC concentration for ANN-GBM hybrid model with all testing data (1494

observations) (top right) and with smaller test data (150 observations) for better visualization (bottom).

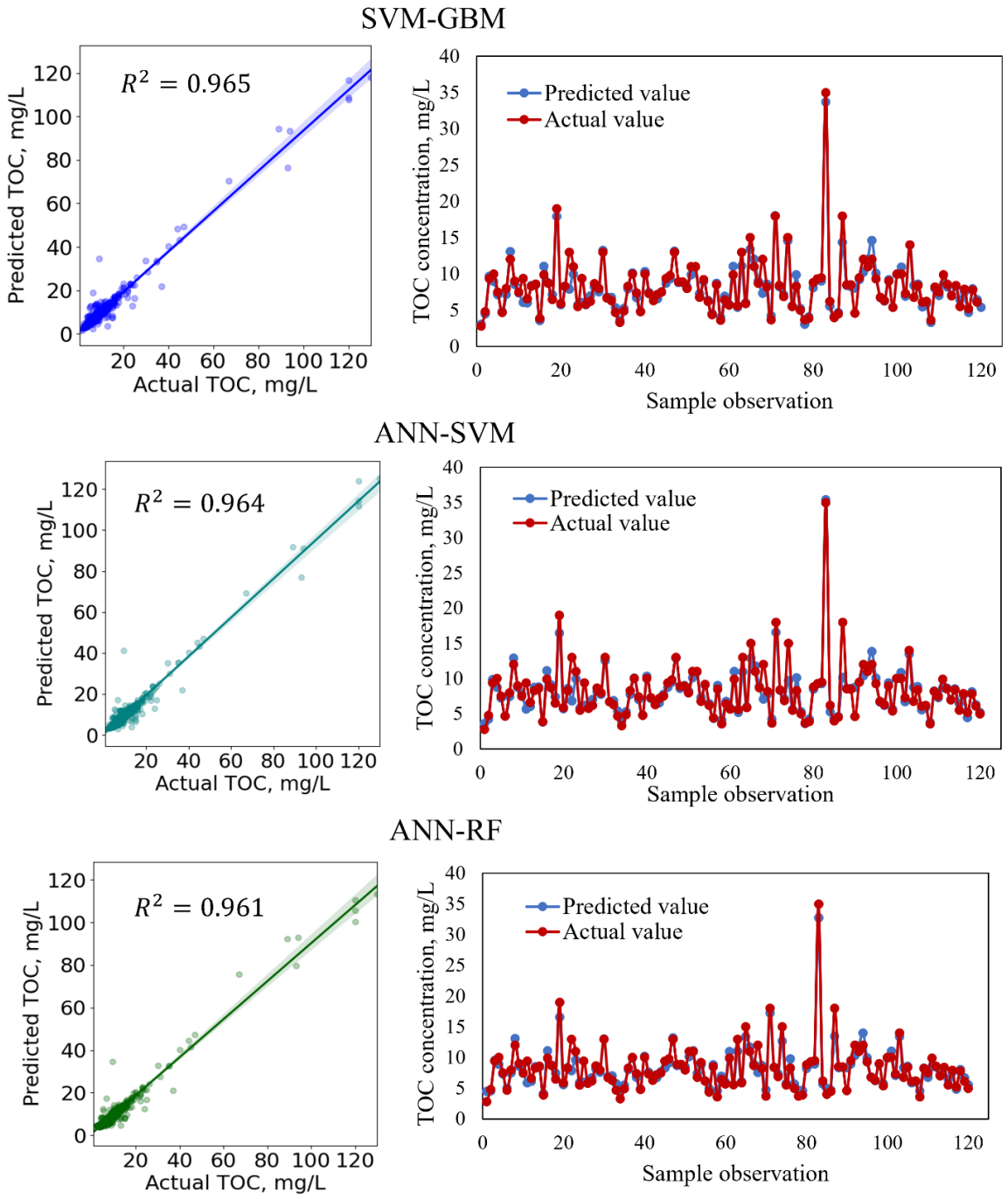


Figure 6.7. Regression analysis plots (left) and time variation graphs (right) comparing the actual and predicted TOC concentration for the hybrid models SVM-GBM, ANN-SVM, and ANN-RF with a small portion (120 sample observations) of the test dataset.

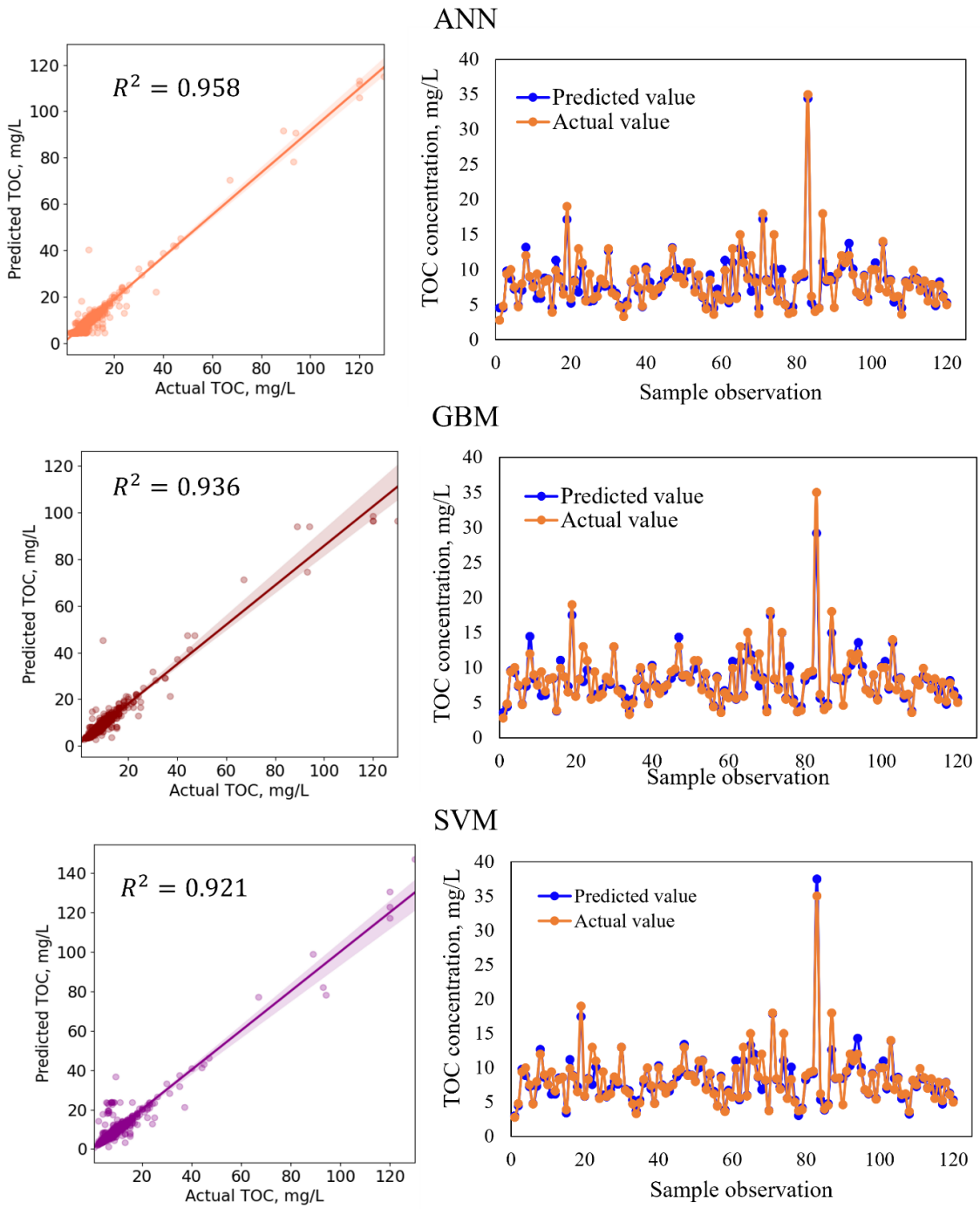


Figure 6.8. Regression analysis plots (left) and time variation graphs (right) comparing the actual and predicted TOC concentration for standalone ML models with a small portion (120 sample observations) of the test dataset.

Also, learning curves were used to indicate whether the models were a good fit, underfit, or overfit, based on the performance of the training and validation set. The plot of learning curves in Figure 6.9 indicates the learning and generalization performance of the standalone ML models: RF, GBM, and ANN over experience. A good fit model is represented by a training and validation score that approaches towards a point of stability with a minimal gap between these two scores. The score (MSE score) should be lower on the training set than the validation set, creating a generalization gap between the two curves. Figure 6.9 represents the learning curves for RF, GBM, and ANN that indicates the relationship between the training score and validation score with varying training sizes. The MSE score was used to evaluate the performances of the models for the specific training size. The results indicated that the models performed satisfactorily with a training size of 4482 sample observations with lower validation scores. For RF, the training score and validation score moved toward a stable point with a small gap between the curves. Also, with the increasing training size, the validation score decreased. For RF and GBM models, when the training set size increased to 3000, the training MSE remained constant, while the validation MSE started decreasing significantly. For RF model, with a training size of 3000, the MSE of training set was found as 3.48 while the MSE score for validation set was 10.68. Beyond the point of training size of 3000, the validation score further decreased to 6.74. For GBM, the final validation score was 4.32 with training size 4482 and a minimum gap was found between the two curves. The gap between the two curves decreased with the increase in training size and continued to reach a point of stability.

Similarly, the learning curves for ANN model indicate that the model generalized well with the specific training size, and the training score and cross validation score converged at the point of stability with a minimum training size of 4000. When the training size was 1, the MSE

for the training set was 0 while the validation score was about 300. The high value of validation score is expected because the model had no problem fitting perfectly a single data point. And the model that trained on a single data point cannot generalize accurately to new sample observations in the validation set, resulting in a high prediction error. When the training size increased to 3000, the training MSE increased sharply, while the validation MSE decreased. At a training size of 4000, the training and validation curves converged and reached at the point of stability. The learning curves indicate that the models generalized well on the validation set with a training size of 4482 sample observations and are considered as good fit models with lower MSE scores.

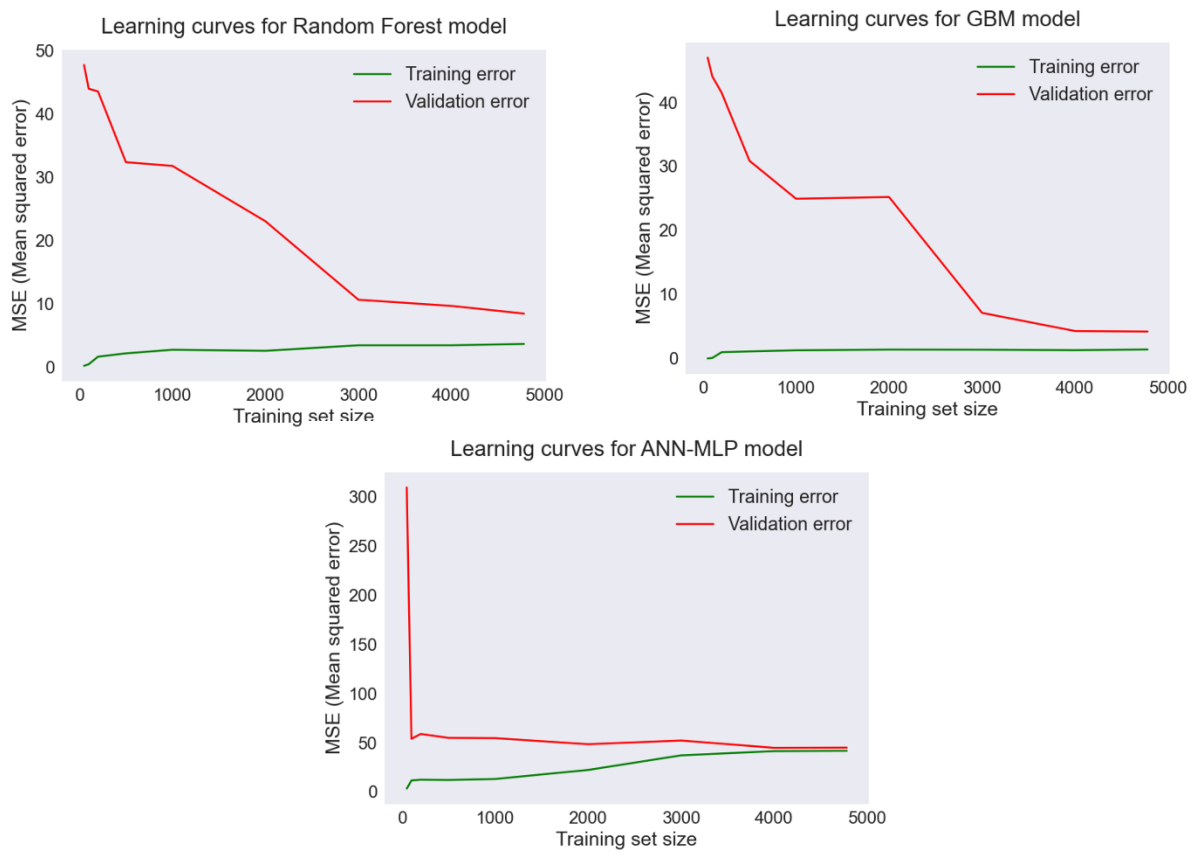


Figure 6.9. Learning curves indicating performances of RF, GBM, and ANN-MLP models for TOC prediction based on MSE score with varying training size

6.3.5 Model performance evaluation for E. coli prediction

Using the specific dataset of different water quality parameters, ML prediction models were developed for *E. coli* to analyze the microbial data, find specific patters in the data, and establish complex non-linear relationships that might exist between the physiochemical and microbiological parameter during the sampling period. In this study, an attempt has been made to predict living microorganisms' behavior that have greater amount of unexplainable variation in the data. And living objects such as microbial community data are harder to predict than physical and chemical processes. The goal was to apply machine learning to develop and evaluate models using algorithms that can predict bacteria concentrations (MPN/100 mL) from the data and try to explain the variability in the dependent variable through the independent physical and chemical water quality parameters. Also, based on the established relationship between microbial population and surface water environment, the models could provide favorable support to understand disease outbreak and human health associated with exposure to water with *E. coli* contamination.

Prediction models for *E. coli* were developed using four standalone and four hybrid ML algorithms. Initially, the models were developed using all input features. To reduce the number of redundant features and improve the performance efficiency, the important variables with relatively higher feature importance scores computed from the feature importance analysis were considered. The prediction performances of the eight developed ML models with all the input features and the feature importance (BOD_5, DO, total phosphorous, temperature, turbidity, nitrate, and alkalinity) are shown in Table 6.7. The results indicated that the R^2 value ranged from 0.26-0.40 when using all the input variables while with the feature importance the prediction performances of the models were improved within a range of 0.29-0.42. The mean

absolute error (MAE) scores were found higher that also indicated unsatisfactory model execution for prediction of *E. coli* with the specific dataset. Among the developed ML models, the hybrid model ANN-GBM exhibited the highest prediction accuracy of 42%. It appears that prediction performances of the models for *E. coli* were not satisfactory, even with the feature importance.

Table 6.7. Comparison of model performances for 4 standalone and 4 hybrid algorithms for prediction of *E. coli* between all input features and feature importance

ML Algorithms	**R^2 (all features)**	**Only Feature Importance**	
		R^2	**MAE**
ANN	0.36	0.38	2062.44
RF	0.29	0.32	3226.22
SVM	0.27	0.29	1861.43
GBM	0.26	0.30	3244.35
ANN-RF	0.40	0.41	3095.05
ANN-SVM	0.37	0.32	3051.09
ANN-GBM	0.34	**0.42**	2994.89
RF-GBM	0.30	0.37	3023.69

From the statistical analysis of microbial data, it can be observed that *E. coli* concentrations varied within a range of 0-250000 MPN/mL with a standard deviation of 16461.95 MPN/mL during the 20 years sampling period. There was about 452% dispersion of the data around the mean value of 3643.51 MPN/100 mL. The reason that the ML models performed unsatisfactorily was because of the high variation in bacteria data and it was difficult to explain such variability based on the input variables of physicochemical water quality parameters. Also, from the statistical analysis, no significant and strong correlation was found between the output and input variables. Although the prediction performances of the models were relatively poor, the ML algorithms were able to explain some percentage of variability in

the data by extracting useful data-driven information about the existing hidden non-linear relationships between the output and input variables. However, as ML models are used as black boxes in predicting the output, there was little understanding of how the models explained such variability and arrived at the prediction with prediction accuracies within a range of 29%-42%. In addition, it can be observed that higher values of MAE for the prediction models. MAE indicates the average of absolute error (difference between the actual and predicted value). As the measuring values were found within a high range with a maximum value of 250000 MPN/mL, the difference between the actual and predicted value was also found higher, more than 1000 MPN/100 mL. The models' performances could be improved if besides the physicochemical parameters, other hydrometeorological variables such as air temperature, air humidity, atmospheric pressure, precipitation level, stormwater runoff flow were also available during the sampling period to consider as inputs in model development process. Because of the unavailability of the meteorological data for the corresponding bacteria concentration, only the available physical and chemical parameters measured by MMSD were used for the specific rivers as inputs to the models.

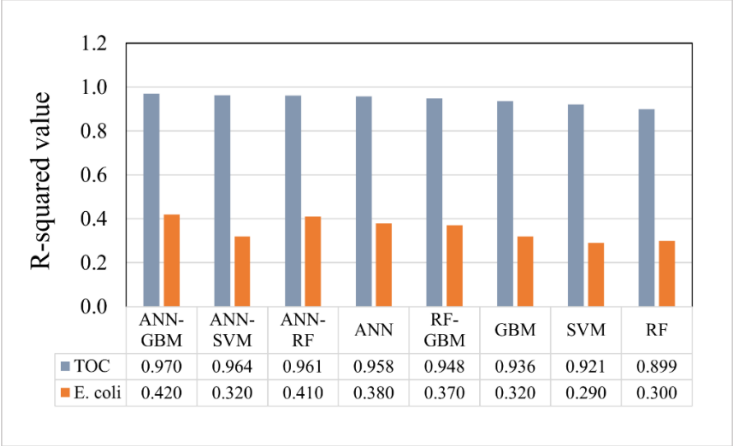


Figure 6.10. Comparison between TOC and *E. coli* prediction models' performances based on R^2 value

In this study, several standalone and hybrid ML models were developed and evaluated for predicting of TOC and *E. coli* in the major rivers of the Milwaukee River basin. In addition, using the ML algorithm, the most influential parameters were identified in predicting TOC and *E. coli* of the river system by interpreting a large water quality dataset. For TOC prediction, the most influential variables were identified as BOD_5, DOC, EC, chloride, TS, nitrate, VSS, DO, turbidity, pH, and TSS, while for *E. coli* prediction the decision-tree-based RF algorithm indicated BOD_5, DO, total phosphorous, temperature, turbidity, nitrate, and alkalinity as the relatively important features. TOC, DOC, and BOD_5 are directly associated with organic pollution in the rivers. The growth of microorganism increases with high level of organic carbon in water and as a result, oxygen consumption by the microorganism increases, which eventually increases BOD_5. The feature importance scores of the input variables for *E. coli* prediction were relatively less than TOC prediction. However, the ML algorithm was able to extract useful data-driven information about the hidden complex non-linear relationships between bacteria concentration and other physicochemical water quality parameters and indicated BOD_5, DO, and total phosphorous as the most influential parameters for predicting *E. coli*. The input variables with lower feature importance scores did not indicate that these features were uninformative for the prediction models. The specific features were not chosen since they might be profoundly related to another informative feature and did not provide any new additional signal for prediction. With the specific dataset, the ML models performed very satisfactorily for TOC prediction with high prediction accuracies of greater than 96%. However, for *E. coli* prediction, because of the high variability of bacteria data, it was difficult to explain such greater amount of unexplained variation in the dataset based on the physicochemical water quality parameters, resulting in relatively lower R^2 values of the models. The results also indicated that for both TOC

and *E. coli* prediction with the specific dataset, ANN-GBM outperformed others with prediction accuracies of 97% and 42%, respectively (Figure 6.10). The reason is that the hybrid model benefitted from advantages of the specific activation function of ANN in performing non-linear transformations of the input data and allowing complex functional mapping of the network's input and output. In addition, the errors in prediction of individual trees developed by GBM algorithm were overcome by boosting mechanism and optimizing the coefficients to fit the underlying data. The incorporation of boosting mechanism along with the non-linear transformation of the input data using an activation function allowed extraction of specific patterns from the data and minimized the difference between the actual and the predicted output, resulting in a more powerful ensemble-hybrid model ANN-GBM.

The application of ML methods in river water quality monitoring ensures a reliable and cost-effective environmental monitoring and assessment program. The direct measure of water quality parameters using AI techniques can potentially eliminate the longer computational time in traditional methods for measuring TOC and *E. coli*. The ability of the employed ML algorithms to be re-trained with the newly available sample observations provides a way to improve model performance efficiency. The application of such ML algorithms in river water quality monitoring is useful in forecasting water quality in future time-steps that will allow timely information and alert the water operators about the water quality levels associated with possible future pollution events. As future work, ML regression models can be developed for the prediction of *E. coli* considering both the hydrometeorological variables and physicochemical parameters measured in a controlled laboratory environment that would explain the variability in microbial data successfully.

6.4 Conclusion

In this study, several regression ML models were developed to predict TOC and *E. coli* in the major rivers (Milwaukee River, Menomonee River, and Kinnickinnic River) within the Milwaukee River basin. Water quality data of 18 physical and chemical parameters were collected from the 32 active monitoring sites of the Milwaukee River basin during the sampling period of 2000-2020. The employed standalone ML models accurately and directly measured TOC with prediction accuracies ranging from 89.9%-95.8%. The prediction performances were further improved ($R^2 > 0.96$) by developing ensemble-hybrid models such as ANN-GBM, SVM-GBM, ANN-SVM, and ANN-RF using the selected input features with relatively higher feature importance scores as computed by the decision-tree based algorithm. The ensemble-hybrid model ANN-GBM achieved the highest prediction accuracy of 97% and lowest error values (MAE =0.664, MSE = 2.334, and RMSE =1.528) in predicting TOC of the river system. The developed ensemble-hybrid models for TOC prediction performed very satisfactorily and were able to successfully explain most of the variability in the dataset based on the combination of input variables: DOC, BOD_5, EC, Chloride, TS, Nitrate, VSS, DO, Turbidity, pH, TSS. The developed ensemble-hybrid methods were not previously used for TOC and *E. coli* prediction that can provide a reliable and direct approach to complement existing monitoring techniques in the Milwaukee River system with satisfactory prediction accuracies. However, for *E. coli* prediction it was difficult to explain the greater amount of unexplained variation in bacteria data based on the physicochemical water quality parameters, resulting in R^2 values within a range of 0.29-0.42; the hybrid model ANN-GBM outperformed others with a prediction accuracy of 42%. Although the statistical analysis didn't identify any significant correlation between bacteria concentrations and physicochemical parameters, the ML models provided data-driven decisions

by extracting predictive information from the dataset and established hidden non-linear relationships between the output and input variables that could explain some percentages of variability in the data. The model performances could be improved if other hydrometeorological variables such as air temperature, air humidity, atmospheric pressure, precipitation level were also available for the corresponding *E. coli* data to consider as inputs in the model development process. This study on the application of AI/ML techniques to river water quality monitoring can be useful for efficient water management and control policies that will allow timely information on river water contamination by interpreting historical water quality data and provide a more reliable and cost-effective environmental monitoring and assessment program. Properly tested and optimized ML models can potentially be used in forecasting the river water quality parameters in future time-steps based on the historical water quality dataset. This will alert the river water operators about the water quality associated with possible future organic matter pollution and microbiological contamination in rivers.

6.5 References

1. Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, *578*, 124084.
2. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31.
3. Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019a). Efficient water quality prediction using supervised machine learning. *Water*, *11*(11), 2210.
4. Babbar, R., & Babbar, S. (2017). Predicting river water quality index using data mining techniques. *Environmental Earth Sciences*, *76*(14), 1-15.
5. Bourel, M., Segura, A. M., Crisci, C., López, G., Sampognaro, L., Vidal, V., ... & Perera, G. (2021). Machine learning methods for imbalanced data set for prediction of faecal contamination in beach waters. *Water Research*, *202*, 117450.
6. Bui, D. T., Khosravi, K., Tiefenbacher, J., Nguyen, H., & Kazakis, N. (2020). Improving prediction of water quality indices using novel hybrid machine-learning algorithms. *Science of the Total Environment*, *721*, 137612.
7. Burzynski, M. (2001). The State of the Milwaukee River Basin. Department of Natural Resources.
8. Chen, K., Chen, H., Zhou, C., Huang, Y., Qi, X., Shen, R., ... & Ren, H. (2020). Comparative analysis of surface water quality prediction performance and identification of key water parameters using different machine learning models based on big data. *Water research*, *171*, 115454.

9. Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020a). A review of the artificial neural network models for water quality prediction. *Applied Sciences*, *10*(17), 5776.
10. David, M. M., & Haggard, B. E. (2011). Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois River Watershed, Arkansas and Oklahoma, USA. *Water, Air, & Soil Pollution*, *215*(1), 525-547.
11. Dogo, E. M., Nwulu, N. I., Twala, B., & Aigbavboa, C. (2019). A survey of machine learning methods applied to anomaly detection on drinking-water quality data. *Urban Water Journal*, *16*(3), 235-248.
12. Emamgholizadeh, S., Kashi, H., Marofpoor, I., & Zalaghi, E. (2014). Prediction of water quality parameters of Karoon River (Iran) by artificial intelligence-based models. *International Journal of Environmental Science and Technology*, *11*(3), 645-656.
13. Farrell, C., Hassard, F., Jefferson, B., Leziart, T., Nocker, A., & Jarvis, P. (2018). Turbidity composition and the relationship with microbial attachment and UV inactivation efficacy. Science of the total environment, 624, 638-647.
14. Goz, E., Yuceer, M., Karadurmus, E. (2019). Total Organic Carbon Prediction with Artificial Intelligence Techniques. *In Computer Aided Chemical Engineering* (Vol. 46, pp. 889– 894). Elsevier
15. Gullick, R. W., Grayman, W. M., Deininger, R. A., & Males, R. M. (2003). Design of early warning monitoring systems for source waters. *Journal-American Water Works Association*, *95*(11), 58-72.
16. Haghiabi, A. H., Nasrolahi, A. H., & Parsaie, A. (2018). Water quality prediction using machine learning methods. *Water Quality Research Journal*, *53*(1), 3-13.

17. Hameed, M., Sharqi, S. S., Yaseen, Z. M., Afan, H. A., Hussain, A., & Elshafie, A. (2017). Application of artificial intelligence (AI) techniques in water quality index prediction: a case study in tropical region, Malaysia. *Neural Computing and Applications*, *28*(1), 893-905.
18. Hayder, G., Kurniawan, I., & Mustafa, H. M. (2020). Implementation of machine learning methods for monitoring and predicting water quality parameters. *Biointerface Res. Appl. Chem*, *11*, 9285-9295.
19. He, L. M. L., & He, Z. L. (2008). Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA. *Water research*, *42*(10-11), 2563-2573.
20. Herrig, I. M., Böer, S. I., Brennholt, N., & Manz, W. (2015). Development of multiple linear regression models as predictive tools for fecal indicator concentrations in a stretch of the lower Lahn River, Germany. Water research, 85, 148-157.
21. Khan, F. M., Gupta, R., & Sekhri, S. (2021). Superposition learning-based model for prediction of E. coli in groundwater using physico-chemical water quality parameters. *Groundwater for Sustainable Development*, *13*, 100580.
22. Kim, S., Maleki, N., Rezaie-Balf, M., Singh, V. P., Alizamir, M., Kim, N. W., ... & Kisi, O. (2021). Assessment of the total organic carbon employing the different nature-inspired approaches in the Nakdong River, South Korea. *Environmental Monitoring and Assessment*, *193*(7), 1-22.
23. Lawal, L. O., Mahmoud, M., Alade, O. S., & Abdulraheem, A. (2019). Total Organic Carbon Characterization Using Neural-Network Analysis of XRF Data. *Petrophysics*, *60*(04), 480–493.

24. Mandal, P. P., Rezaee, R., & Emelyanova, I. (2021). Ensemble Learning for Predicting TOC from Well-Logs of the Unconventional Goldwyer Shale. *Energies*, *15*(1), 216.

25. Mohammed, H., Longva, A., & Seidu, R. (2018). Predictive analysis of microbial water quality using machine-learning algorithms. *Environmental Research, Engineering and Management*, *74*(1), 7-20.

26. Muharemi, F., Logofătu, D., Andersson, C., & Leon, F. (2018). Approaches to Building a Detection Model for Water Quality: A Case Study. In: Modern Approaches for Intelligent Information and Database Systems. Springer, Cham, Switzerland, pp 173-183.

27. Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294-307.

28. Muller, A.C., & Guido, S. (2016) Introduction to Machine learning with Python: A Guide for Data Scientists. O'Reilly Media, Inc.

29. Nafsin, N., & Li, J. (2021). Using CANARY event detection software for water quality analysis in the Milwaukee River. *Journal of Hydro-environment Research*, *38*, 117-128.

30. Nafsin, N., Bevers, B., Schruender, R., Liao, Q., & Li, J. (2022). Escherichia coli and Enterococci Bacteria in Lake Michigan Beach Sand. *Environmental Engineering Science*, *39*(1), 3-14.

31. Najah, A., El-Shafie, A., Karim, O. A., & El-Shafie, A. H. (2013). Application of artificial neural networks for water quality prediction. *Neural Computing and Applications*, *22*(1), 187-201.

32. Ouadfeul, S. A., & Aliouane, L. (2015). Total organic carbon prediction in shale gas reservoirs from well logs data using the multilayer perceptron neural network with

Levenberg Marquardt training algorithm: Application to Barnett shale. *Arabian Journal for Science and Engineering*, *40*(11), 3345–3349.

33. Paule-Mercado, M. A., Ventura, J. S., Memon, S. A., Jahng, D., Kang, J. H., & Lee, C. H. (2016). Monitoring and predicting the fecal indicator bacteria concentrations from agricultural, mixed land use and urban stormwater runoff. *Science of the Total Environment*, *550*, 1171-1181.
34. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.
35. Perelman, L., Arad, J., Housh, M., & Ostfeld, A. (2012). Event detection in water distribution systems from multivariate water quality time series. *Environmental science & technology*, *46*(15), 8212-8219.
36. Rong, J., Zheng, Z., Luo, X., Li, C., Li, Y., Wei, X., ... & Lei, Y. (2021). Machine Learning Method for TOC Prediction: Taking Wufeng and Longmaxi Shales in the Sichuan Basin, Southwest China as an Example. *Geofluids*, *2021*.
37. Sakizadeh, M. (2016). Artificial intelligence for the prediction of water quality index in groundwater systems. *Modeling Earth Systems and Environment*, *2*(1), 1-9.
38. Sarkar, A., & Pandey, P. (2015). River water quality modelling using artificial neural network technique. *Aquatic procedia*, *4*, 1070-1077.
39. Tinelli, S., & Juran, I. (2019). Artificial intelligence-based monitoring system of water quality parameters for early detection of non-specific bio-contamination in water distribution systems. *Water Supply*, *19*(6), 1785-1792.

40. Southeastern Wisconsin Regional Planning Commission. (2007). Water quality conditions and sources of pollution in the Greater Milwaukee watersheds: Southeastern Wisconsin Regional Planning Commission, Technical Report, 39, 141.
41. Tan, J., Wang, D., Cao, H., Qiao, Y., Zhu, H., & Liu, X. (2018). Effect of local alkaline microenvironment on the behaviors of bacteria and osteogenic cells. *ACS Applied Materials & Interfaces*, 10(49), 42018-42029.
42. Wang, X., Zhang, F., & Ding, J. (2017). Evaluation of water quality based on a machine learning algorithm and water quality index for the Ebinur Lake Watershed, China. *Scientific reports*, *7*(1), 1-18.
43. Wang, X., Xie, R., Wang, T., Liu, R., & Shao, L. (2021). Total organic carbon content prediction of source rocks with conventional well log data based on regression committee machine. *Arabian Journal of Geosciences*, *14*(15), 1-11.
44. Yeon, I. S., Kim, J. H., & Jun, K. W. (2008). Application of artificial intelligence models in water quality forecasting. *Environmental Technology*, *29*(6), 625–631.
45. Zou, X. Y., Lin, Y. L., Xu, B., Guo, Z. B., Xia, S. J., Zhang, T. Y., ... & Gao, N. Y. (2019). A novel event detection model for water distribution systems based on data-driven estimation and support vector machine classification. *Water Resources Management*, *33*(13), 4569-4581.

CHAPTER 7: CONCLUSION

Urban water sources are negatively impacted by rapid urbanization and population growth. The local hydrologic system is affected when a rural area is turned into an area full of housing developments, industries, and roads. Urban areas have the potential to pollute natural source water. Urban water takes on a large number of contaminants from industrial discharges, residential and commercial wastewater, mobile sources (e.g., trucks/cars), and stormwater runoff carrying oil, rubber, heavy metals from automobiles and the urban landscape. Untreated or poorly treated sewage discharged into surface water bodies can deplete the dissolved oxygen level in the water and increase the level of fecal indicator bacteria contamination, organic matter and nutrient pollution. Pollution of urban water sources creates public and environmental health hazards by degrading the quality of drinking water and surface water bodies that are unsafe for people to swim in or use for other recreational purposes. Solutions to these problems involve sustainable urban water management by improving pollution control and prevention strategies. Continuous monitoring of surface water allows timely information to the water operators about any contamination event with significant changes in water quality parameters. This dissertation is directed toward the application of advanced water quality monitoring technologies to urban water sources (e.g., rivers and lakes) using CANARY Event Detection System and AI/ML techniques.

In Chapter 1, the motivation and goals of the dissertation are discussed.

In chapter 2, an overview of the application of smart sensor technology and machine learning (ML) methods to urban water systems is provided. Real-time monitoring of water quality and quantity parameters from multiple locations and remote operations with low power

consumption is essential for improving water management and developing pollution control strategies for urban water systems. The large amount of data collected by online water sensor networks provide valuable information for the deployment of data-driven approaches, such as AI-based machine learning methods in urban water management. The current implementation of advanced monitoring technologies in areas such as water distribution networks, natural source water, water treatment plants, pipe infrastructure, filtration efficacy in treatment processes, and early flood warnings are reported in this chapter.

In chapter 3, the concentration and interaction of fecal indicator bacteria in beach sand and water were analyzed, and methods for more accurate predictions of public health outcomes from the measured bacterial concentration were developed. The sampling location (Bradford Beach) was one of the top urban beaches and the most visited place for public recreational activities in Milwaukee, Wisconsin, where several contamination sources such as stormwater discharge, sewage overflow along with a large population of shorebirds cause bacterial pollution at the beach. Analysis of bacterial concentrations at this location could help to develop methods for accurate prediction of public health outcomes as a result of increased contamination with fecal indicator bacteria. Also, this study provides an insight into the first successful application of the CANARY event detection system to identify abnormal conditions, i.e., anomalies in bacterial concentration at a recreational beach.

In chapter 4, analysis of anomalous water quality events was performed for the Milwaukee River using CANARY based on the available monitoring water quality data of pH, conductivity, and turbidity. The study provided an insight into the effectiveness of the application of CANARY statistical software to river water system for detection of anomaly or

‘events’ in water quality signals that can be useful for real-time monitoring of surface water quality in future.

Chapter 5 provides an efficient water data management system using AI/ML techniques for developing improved pollution control strategies to protect natural source water from domestic and industrial pollutants in developing countries such as Bangladesh. The Buriganga River system of Bangladesh is highly polluted with significantly higher organic matter pollution (BOD_5) and lower levels of dissolved oxygen than the standard limits. Because of the limitations of the traditional lab analysis methods for measuring BOD_5 and rapid seasonal and local fluctuations of surface water quality, machine learning (ML) techniques have been developed.

In this study, four standalone ML models: Artificial Neural Network (ANN), Support Vector Machine (SVM), Random Forest (RF), and Gradient Boosting Machine (GBM) and six novel ensemble hybrid models: RF-SVM, ANN-SVM, GBM-SVM, RF-ANN, GBM-ANN, and RF-GBM were developed and evaluated for predicting BOD_5 in a highly polluted river of Bangladesh. The significance of this study is the application of the novel ensemble hybrid models that resulted in higher prediction accuracies. The best-performing prediction model was RF-SVM with the highest prediction accuracy of 91%. Additionally, the most influential water quality parameters in predicting BOD_5 of the river system were identified using the decision-tree based ML algorithm. The prediction performance of the models was verified by comparing the predicted BOD_5 with the measured BOD_5, and a small deviation found between the two datasets indicated good model executions. Properly tested and optimized ML hybrid models can potentially be used in forecasting BOD_5 in future time-steps based on the historical water quality dataset that will allow timely information to river water operators about the BOD levels associated with possible future organic matter pollution.

Chapter 6 focuses on developing ML regression prediction models (both standalone and ensemble-hybrid models) that would provide a more accurate, reliable, and direct measure of organic matter pollution (TOC levels) and FIB contamination (*E. coli* bacteria) in the three primary rivers within the Milwaukee River basin. The prediction models developed for *E. coli* explained the greater amount of unexplainable variation in living microorganisms' behavior based on the specific physicochemical parameters and identified the most influential physical and chemical water quality parameters in predicting *E. coli*. A comprehensive assessment of the differences in prediction performances of the ML models were performed between TOC and *E. coli* with data collected from the specific study area.

The developed ensemble-hybrid methods were not previously used for TOC and *E. coli* prediction and can be applied as a reliable and direct approach to complement existing monitoring techniques in the Milwaukee River system with satisfactory prediction accuracies. The employed ensemble-hybrid ML models accurately and directly measured TOC with prediction accuracies greater than 96%. However, for *E. coli* prediction it was difficult to explain the greater amount of unexplained variation in bacteria data based on the physicochemical water quality parameters, resulting in R^2 values within a range of 0.29-0.42; the hybrid model ANN-GBM outperformed others with a prediction accuracy of 42%. Although the statistical analysis didn't identify any significant correlation between bacteria concentrations and physicochemical parameters, the employed ML models provided data-driven decisions by extracting predictive information from the dataset and established hidden non-linear relationships between the output and input variables that could explain some percentages of variability in the microbial data.

In summary, this dissertation provides an insight into the effectiveness of advanced monitoring technologies, including event detection systems (EDS) and AI/ML algorithms for

monitoring physical, chemical, and microbiological parameters of urban water systems. The significance of this dissertation is the first successful application of CANARY EDS to natural source water and the development of novel ensemble-hybrid ML models in predicting different water quality parameters of rivers. The employed advanced and real-time water monitoring technologies can be used as a direct approach to complement the existing conventional water quality analysis methods and can provide a reliable and cost-effective solution to urban water management.

CURRICULUM VITAE

RESEARCH INTEREST

Environmental Engineering, Water quality monitoring, Smart water sensor technology, Machine learning technologies

EDUCATION

Ph.D., Civil and Environmental Engineering **May 2022**

University of Wisconsin-Milwaukee, Milwaukee, WI

GPA: 3.935/4.000

B.Sc., Civil Engineering **March 2016**

Bangladesh University of Engineering and Technology, Dhaka, Bangladesh

PROFESSIONAL CERTIFICATION

FE (Civil)

Board: Wisconsin

PUBLICATIONS

- Nafsin, N. and Li, J. (2021). Using CANARY event detection software for water quality analysis in the Milwaukee River. Journal of Hydro-environment Research, 38, 117-128.
- Nafsin, N., Bevers, B., Schruender, R., Liao, Q., and Li, J. (2022). *Escherichia coli* and Enterococci bacteria in Lake Michigan beach sand. Environmental Engineering Science, 39, 3-14.

- Nafsin, N. and Li, J. (2022). Prediction of 5-day Biochemical Oxygen Demand in the Buriganga River of Bangladesh using novel hybrid machine learning algorithms. Water Environment Research, 94 (5), 10718.
- Nafsin, N. and Li, J. (2022). Prediction of Total Organic Carbon and *E. coli* in rivers within the Milwaukee River basin using machine learning methods. (Submitted to Water Research, Submission ID: WR69139)
- Jean-Pierre, G., Akbarihaghighat, H., Zhao, T., Berger, A., Nafsin, N., Nasir, F. B., ... & Nowak, M. (2021, June). Development of a Data Analytics Platform for an Electrical/Water Microgrid. In 2021 IEEE 12th International Symposium on Power Electronics for Distributed Generation Systems (PEDG) (pp. 1-7). IEEE.

CONFERENCES

- Nafsin, N. and Li, J. The Use of CANARY Event Detection Software for Urban Water Quality Analysis. Presented in International Symposium on Sustainable Urban Drainage, 2019, Ningbo, China.
- Nafsin, N., Bevers, B., Liao, Q., and Li, J. Monitoring of *E. coli* and Enterococci in Lake Michigan Beach Sand. Presented in 22nd EGU General Assembly Conference, 2020 (ID 2459).
- Nafsin, N., Nasir, F., and Li, J. Smart Sensor Networks and Machine Learning Technologies in Urban Water Systems. Presented in International Workshop on Sustainable Urban Drainage, 2021, Ningbo University, Ningbo, Zhejiang, China.

POSTER PRESENTATIONS

- Nafsin, N and Li, J. Biochemical Oxygen Demand prediction in the Buriganga River of Bangladesh Using Novel Hybrid Machine Learning Methods, 2022 Climate Change Conference, Loyola University Chicago.
- Nafsin, N and Li, J. Monitoring Fecal Indicator Bacteria in shoreline water and beach sand of Lake Michigan, UWM Student Research Poster Competition, CEAS 2021.

AWARDS

- Chancellor's Graduate Student Award (2017-2021), University of Wisconsin-Milwaukee
- University Technical Scholarship, Bangladesh University of Engineering & Technology
- Education Board Scholarship, Bangladesh

PROFESSIONAL MEMBERSHIP

- American Society of Civil Engineers (ASCE)
- Water Environment Federation (WEF)