

**DEVELOPING ARTIFICIAL NEURAL NETWORKS (ANN) MODELS
FOR PREDICTING E. COLI AT LAKE MICHIGAN BEACHES**

by

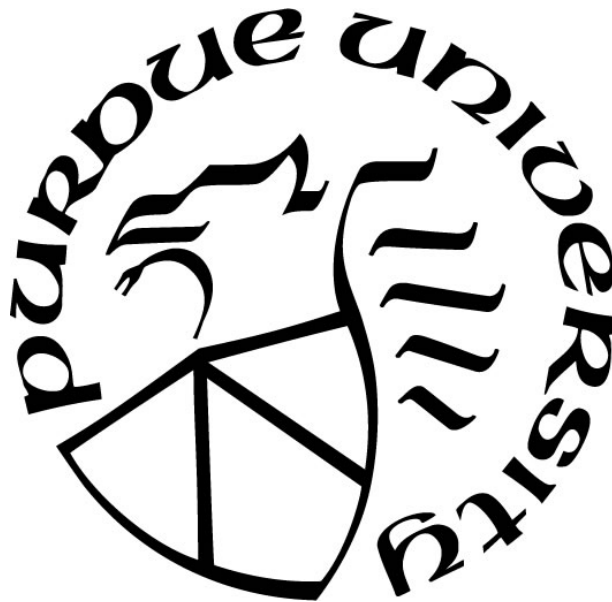
Mitra Khanibaseri

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science in Engineering



Department of Civil and Mechanical Engineering

Hammond, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Chandramouli Viswanatha Chandramouli, Chair

Department of Mechanical and Civil Engineering

Dr. Chien-Chung Chen

Department of Mechanical and Civil Engineering

Dr. Jilian Li

Department of Mechanical and Civil Engineering

Approved by:

Dr. Chenn Q. Zhou

ACKNOWLEDGMENTS

I wish to thank a lot of people that contributed to the success of this research. Their support and encouragement are instrumental in the successful completion of this work.

First and foremost, I would like to thank Professor Chandramouli Viswanatha Chandramouli for providing this opportunity at Purdue University and Purdue Water Institute, respectively. I sincerely appreciate his comprehensive guidance, incisive evaluation, backing, and support throughout this research. I will always be grateful for his belief in me to accomplish the research objectives despite all the challenges encountered. I also want to thank the Indiana department of environmental management for financially supporting my research.

My deep appreciation also goes to Dr. Chien-Chung Chen and Dr. Jiliang Li for serving as my committee members and lending their time, knowledge, and expertise to my work, my colleague Michael Ozeh for his helpful contributions during my research work.

I am deeply grateful to my mother for her prayers, love, and unwavering support; I will always be thankful to my beloved husband, for his love, support, guidance, and prayers; my family and friends and many others for their unreserved love, friendship and immeasurable support, which was critical to my success in this program.

TABLE OF CONTENTS

LIST OF TABLES.....	6
LIST OF FIGURES	8
ABSTRACT.....	9
1. INTRODUCTION	10
1.1 Background.....	10
1.2 Statement of the problem.....	11
2. LITERATURE REVIEW	12
2.1 Introduction.....	12
2.2 Currently Used Predictive Models.....	12
2.2.1 Rapid Analytical Techniques.....	13
2.2.2 Statistical Models.....	13
2.2.3 Artificial Neural Network Models.....	15
3. SYSTEM CONSIDERED	18
3.1 Lake Michigan beaches locations	18
3.1.1 Indiana Dunes State Park Location.....	18
3.1.2 Jeorse Park Beaches Location	19
3.1.3 Whihala Beach Location.....	19
3.2 Field Sampling Training	20
3.3 Data collection	22
3.3.1 Data Collection Details.....	22
3.3.2 Intensive Monitoring Period	23
3.3.3 TDS Collection and Analysis	26
3.3.4 TSS Collection and Analysis.....	26
3.3.5 E. Coli Collection and Analysis.....	26
3.4 Field Data Collection and Recording.....	28
3.5 Chain of Custody/Sample Handling	29
3.6 Collection of Beachgoer, Pet, and Bird Counts	29
3.7 Identification and Use of Other Data	30

3.8	Collection of meteorological data from Gary airport	32
3.9	Data Preparation/pre-processing Attempts	32
3.9.1	Observed normal conditions	33
3.9.2	Closeness to the mean of measured data	35
3.9.3	Binary/ Pseudo-Binary Classifications	35
3.9.4	Miscellaneous	36
3.9.5	E. Coli Classification	37
3.10	Instrumental Analysis	40
3.10.1	Temperature	40
3.10.2	pH:	41
3.10.3	Turbidity:	41
3.10.4	TDS and Electrical conductivity	42
3.10.5	Total Dissolved Solids (TDS) and Temperature	43
4.	ARTIFICIAL NEURAL NETWORK MODEL BUILDING PROCESS	45
4.1	Overview of Artificial Neural Networks	45
4.1.1	Biological Neural Network	45
4.1.2	Artificial Neural Network	46
4.2	Different trials of the model development	54
4.2.1	Initial model with raw data prediction (M1):	54
4.2.2	Model with raw inputs to predict the E. Coli with 4 classes (M2):	56
4.2.3	Model with raw inputs to predict the E. Coli with 3 classes (M3):	59
	Bayesian regularization neural network algorithm:	61
4.2.4	Model with raw inputs to predict the E. Coli with 2 classes (M4):	65
5.	RESULTS AND DISCUSSION	69
5.1	Best Model with 2 E. Coli Classes	71
6.	CONCLUSION	75
	APPENDIX A: MICROBAC INTERNAL CHAIN OF CUSTODY FIELD DATA SHEET FOR MICROBAC-COLLECTED SAMPLES	76
	APPENDIX B: MICROBAC CHAIN OF CUSTODY FORM	77
	REFERENCES	78

LIST OF TABLES

Table 3.1 Sampling location candidates for water quality characterization	23
Table 3.2 Intensive weeks date collection details.....	24
Table 3.3 2019 Beach Program E. Coli Sampling Schedule	27
Table 3.4 Responsible parties for bird, pet, and head counts by beach	30
Table 3.5 Data Sources for Project Secondary (Existing) Data.....	31
Table 3.6 USGS Streamflow Gauges in Project Vicinity	31
Table 3.7 Classification for TSS, Turbidity, pH and Wind Speed	34
Table 4.1 Training algorithm tried during ANN model development	53
Table 4.2 Initial model inputs and their sources	54
Table 4.3 Second model inputs and their sources	56
Table 4.4 Breakdown of number of data per class in the E. Coli dataset	57
Table 4.5 Best prediction accuracy with 4 E. Coli classes (correct predictions in blue and incorrect predictions in red)	57
Table 4.6 Breakdown of the prediction accuracy with 4 E. Coli classes.....	58
Table 4.7 Breakdown of the prediction accuracy with 4 E. Coli classes.....	58
Table 4.8 Third model inputs and its sources	60
Table 4.9 Prediction accuracy (trial 1) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)	63
Table 4.10 Prediction accuracy (trial 3) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)	63
Table 4.11 Prediction accuracy (trial 4) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)	64
Table 4.12 Prediction accuracy (trial 2) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)	64
Table 4.13 Breakdown of the prediction accuracy of the best model with 3 E. Coli classes	65
Table 4.14 Prediction accuracy (trial 1) with 2 E. Coli classes (correct predictions in blue and incorrect predictions in red)	67

Table 4.15 Prediction accuracy (trial 2) with 2 E. Coli classes (correct predictions in blue and incorrect predictions in red)	67
Table 4.16 Prediction accuracy (trial 3) with 2 E. Coli classes (correct predictions in blue and incorrect predictions in red)	68
Table 4.17 Prediction accuracy (trial 4) with 2 E. Coli classes (correct predictions in blue and incorrect predictions in red)	68
Table 5.1 Comparison of the best result for different models' classification	69
Table 5.2 Best prediction accuracy with 2 E. Coli classes	72
Table 5.3 Breakdown of the prediction accuracy with 2 E. Coli classes.....	73
Table 5.4 Breakdown of the prediction accuracy with 2 E. Coli classes.....	73
Table 5.5 The way inputs obtained apply to the ANN model	74

LIST OF FIGURES

Figure 3.1 Indiana Dunes State Park location.....	18
Figure 3.2 Jeorse Park Beach location.....	19
Figure 3.3 Whihala Beach location.....	20
Figure 3.4 Sampling bottles and gloves.....	21
Figure 3.5 Sampling coolers	21
Figure 3.6 Flowchart on overview of the study process	39
Figure 3.7 Oakton pH and Temperature Meter.....	40
Figure 3.8 Hach Turbidity Meter.....	42
Figure 3.9 Extech Conductivity Meter.....	43
Figure 3.10 Apera CE60 TDS and Thermomet	44
Figure 4.1 Biological Neuron and Axon.....	45
Figure 4.2 Simple example of how one single neuron works in ANN	47
Figure 4.3 Original table from [21].....	48
Figure 4.4 Simple neural network with one hidden layer	50
Figure 4.5 Correlation between outputs and targets in the initial model with raw data prediction	55
Figure 4.6 Cross-validation efforts	60
Figure 4.7 Feed forward neural network schematic.....	61

ABSTRACT

A neural network model was developed to predict the E. Coli levels and classes in six (6) select Lake Michigan beaches. Water quality observations at the time of sampling and discharge information from two close tributaries were used as input to predict the E. coli. This research was funded by the Indiana Department of Environmental Management (IDEM). A user-friendly Excel Sheet based tool was developed based on the best model for making future predictions of E. coli classes. This tool will facilitate beach managers to take real-time decisions.

The nowcast model was developed based on historical tributary flows and water quality measurements (physical, chemical and biological). The model uses experimentally available information such as total dissolved solids, total suspended solids, pH, electrical conductivity, and water temperature to estimate whether the E. Coli counts would exceed the acceptable standard. For setting up this model, field data collection was carried out during 2019 beachgoer's season.

IDEM recommends posting an advisory at the beach indicating swimming and wading are not recommended when E. coli counts exceed advisory standards. Based on the advisory limit, a single water sample shall not exceed an E. Coli count of 235 colony forming units per 100 milliliters (cfu/100ml). Advisories are removed when bacterial levels fall within the acceptable standard. However, the E. coli results were available after a time lag leading to beach closures from previous day results. Nowcast models allow beach managers to make real-time beach advisory decisions instead of waiting a day or more for laboratory results to become available.

Using the historical data, an extensive experiment was carried out, to obtain the suitable input variables and optimal neural network architecture. The best feed-forward neural network model was developed using Bayesian Regularization Neural Network (BRNN) training algorithm. Developed ANN model showed an average prediction accuracy of around 87% in predicting the E. coli classes.

1. INTRODUCTION

1.1 Background

Environmental pollution has catastrophically threatened life on Earth. Comprehensive population growth and urban development have always associated with increased waste, industrial effluent, municipal sludge and, agricultural wastewater, which has increased the need for seas and oceans to dispose of treated or untreated wastewater and effluent. Most of the world's pollution offshore is caused by human activity on land. In the meantime, a particular risk factor for public health is contamination caused by human fecal matter (e.g., through sewage), which contains a wide range of pathogens, including human-specific viruses.

Recreational use of beaches such as swimming has put large numbers of people at risk for viral and bacterial diseases, especially on beaches in crowded centers. One of these risks is the risk of intestinal diseases as a result of swimming in sewage-contaminated water. Infectious diseases caused by pathogenic microorganisms due to pollution of coastal wastewater can affect many people and can result in serious economic problems.

Fecal indicator bacteria (FIB) is traditionally used as a surrogate indicator for the presence of pathogenic bacteria in recreational waters. Culture-based methods are used for finding FIB, which requires more time (18–72 h) [1]. According to the USEPA, *Escherichia coli* (*E. Coli*) is considered as the key pathogen (fecal indicator bacteria) in recreational surface water. It has been identified as a major contaminant of water resources in the USA. Also, this agency recommended 235 colony-forming units (cfu) per 100 milliliters (100 ml) of *E. Coli* concentration in the recreational water surface as the safe limit. *E. Coli* contaminations were related to 63,153 cases of illness as well as caused 20 deaths in the United States. It also resulted in \$255 million in losses each year [1]. To alert the public, beach managers use FIB standards developed by USEPA to post warnings or close beaches by following the state's recommendations [2].

Several studies have shown that the *E. Coli* counts in the surface water are influenced by physical (e.g., temperature), chemical (e.g., pH), and biological (e.g., Chlorophyll) factors [3]. By

developing an artificial neural network model which learns the relationship between E. Coli and the observed physicochemical and biological parameters, E. Coli counts could be determined rapidly. It results in real-time decision making and can help one to take quick decisions. Regular E. Coli sampling and lab testing takes longer time (around 24 hours). This procedure is not sufficient to make beach closing decisions on a real-time basis.

This study was formulated to assess the variability of E. Coli concentrations in Lake Michigan waters at select locations. Popular swimming beaches at Indiana Dunes State Park, East Chicago beaches (Jeorse park beach), and Whihala beach were considered for this purpose. In this work, flow measurements in creeks which are draining to Lake Michigan, as well as other hydro-meteorological factors, were used to predict the E. Coli clauses in the beach's swimming zone.

1.2 Statement of the problem

This study is proposed to develop an Artificial Neural Network model to assess the microbial contamination at select locations of Indiana's Lake Michigan shoreline. Developing this model will help the beach managers in taking real-time decision and reduce the cost and effort associated with beach monitoring and public notification process. A significant problem facing beach managers is that the traditional E. Coli analysis generally takes 24 hours to complete. So, beach closures are based on day-old information. Backdate inspections had shown that there had been numerous instances where the decisions made were not satisfactory. Beaches were closed when the E.Coli concentrations were low and were kept open when the concentrations were high. Finding the inter relationship between E.Coli concentration and other water quality parameters will be of great help to the field decision making. It clearly indicates the need for other decision-making tools that do not require a substantial expense.

2. LITERATURE REVIEW

2.1 Introduction

A major concern in surface water bodies is fecal contamination. Identifying the origin of the source of the pollutants of water is an easy task. Bacteria associated gastrointestinal illness is the most widely studied and the diseases were usually caused by unsafe recreational water. Since 1990s, viral and protozoan pathogens have gained attention as areas of potential concern. Contamination due to fecal matters is a threat to human health and is a global problem. E. Coli is a large and diverse group of bacteria. E. Coli are found in the intestines of warm-blooded animals [4].

Clearly identifying the goals is the first step in designing a time-relevant beach water quality and public notification model. A new predictive model will help in providing timely warning to public and protect them from potential health risks. This literature review first presents a brief summary of health concerns and beach water quality monitoring. Later, the factors to be considered while designing a predictive model were presented. Uses of Artificial Neural Network (ANN) techniques in the field of environmental management were also reviewed in this chapter.

2.2 Currently Used Predictive Models

According to the current practice based on the traditional analysis method, 18-24 hours of time is required before the E. Coli concentration can be reported. The persistence model i.e., using last available value to manage beaches, is therefore unsatisfactory because of its lag period. To address this time relevance of water quality assessment issues, a number of strategies have been proposed. Present laboratory-based testing procedures were time-consuming and resulted in difficulties in implementing real-time decisions. Efforts were taken by researchers to develop real-time or near real-time predictive tools for beach managers to take suitable decisions. Rapid analytical techniques, deterministic models, regression models, and artificial neural network-based models are being some of them.

2.2.1 Rapid Analytical Techniques

Rapid analytical techniques of indicator organism quantification, such as amperometric culture-based method, currently take less than 10 hours to complete of low concentrations of viable E. Coli, e.g. 100 cfu/ 100 ml, in environmental water.

Pérez et al., developed this model for the rapid detection of viable Escherichia Coli in environmental samples and cultivated E. Coli in the laboratory. In this method, 4-AP (4-aminophenol) was produced after hydrolysis of 4-APGal (4-aminophenyl- β -d-galactopyranoside) by the enzyme β -galactosidase. Using amperometry, 4-AP was measured and was detected at a considered concentration of E. Coli. This method reduced the time required for finding E.Coli concentration. With initial E. Coli concentrations of 1.0 and 2.0×10^3 cfu ml⁻¹, the new process detected after 10 and 6.6 hours [5].

2.2.2 Statistical Models

The statistical model is a general term for any statistical modeling approach to predict a particular entity for various applications. Linear regression models assume a linear relationship between factors or combinations of factors and indicator organisms [2], [6]. The most highly developed statistical model approach is a multiple linear regression (MLR) relationship between an indicator organism and several independent variables. Many water quality variables are easy and quick to measure. Turbidity, pH, electrical conductivity, hydrodynamic conditions such as flows of nearby tributaries, magnitude, and direction of water currents, wave height and other factors such as a number of birds or pets are good examples. Other meteorological conditions such as air temperature, precipitation, dew point, wind speed and direction are usually available from nearby meteorological observatories. MLR models were usually formulated to find the concentrations or the probability of exceeding the water quality standard limits [7].

Nevers and Whitman monitored five effluent dominated beaches in southern Lake Michigan. They developed regression modeling to nowcast the Escherichia Coli concentrations to assist beach management on an experimental basis. The researchers found out that the swage was present in the river and bathing beaches following heavy rain due to coliphage's positive tests. This study

indicated a positive correlation with mean log E. Coli densities with turbidity, color, Burns Ditch gage height, wave height. It also correlated positively with wind speed, wind gust, and specific conductance but negatively correlated with pH and Dissolved Oxygen. This model predicted E.Coli concentrations more than 235 cfu/100 ml correctly for six out of eleven events and proved to be more reliable [2].

Gonzalez et al. used empirical predictive modeling in eastern North Carolina waters. They developed statistical models which used antecedent rainfall, climate variables and environmental variables for predicting E.Coli and enterococci and validated them. This study indicated 5-day antecedent rainfall, dissolved oxygen, and salinity as important variables. They concluded that these models were very useful in predicting E.Coli and enterococci during the modeling process but did not give satisfactory results with the validation set. But it helped in understanding the important variables involved in the process [8].

Olyphant et al. used a statistical model to predict E.Coli concentration in streamflow from two Lake Michigan sub-watersheds. Precipitation, stream discharge, soil temperature, and water depth in the Great Marsh contributed to 70 percent of the variability in E.Coli concentration for the Derby Ditch watershed. Using regularly observed water quality time series, a time series regression model was also developed to find E.Coli concentration in storm flow. This analysis showed nitrate and ammonia as the most influencing variables. Both these models could be used for real-time prediction [9].

Avila et al. developed many statistical models such as naive model, multiple linear regression, dynamic regression, regression tree, Markov chain, classification tree, random forests, multinomial logistic regression, discriminant analysis, and Bayesian network models with an objective to predict fecal coliform levels in the Oreti River, Wallacetown, New Zealand. The bayesian network was identified as a suitable model in that research. This model had greater flexibility to handle missing data and outliers. By seeing the promise, researchers recommended to use Bayesian network models for modeling other sites [10].

2.2.3 Artificial Neural Network Models

ANN is a mathematical model replicating human brain cells called ‘neurons’ to some extent. They were used as statistical modeling tool in the recent past successfully. As an effective functional approximator, they can establish the inter relationships between inputs and outputs during training. So, they are often referred as functional approximator because for the datasets for which the relationship between inputs and output are not clearly understood, they can identify it. Training of ANN is done with the help of a learning algorithm [11].

Dogan et al. used Artificial neural networks (ANN) to investigate the ability of this model to increase the accuracy of the measuring biological oxygen demand (BOD) of the Melen River in Turkey. They developed a three-layer feed-forward neural network modeling technique which is popular in water resources applications. They tried different architecture by varying hidden layer neurons. They recommended 8 input-3 hidden layers and one output neural network. Chemical oxygen demand (COD), ammonia (NH₃-N), Chlorophyll a (CL-A), nitrite (NO₂-N), temperature (T), nitrate (NO₃-N), dissolved oxygen (DO), and water flow (Q_w) were used as input. This model trained for 1000 iteration. ANN showed reasonable results in predicting BOD [12].

Motamarri and Boccelli used the learning vector quantization (LVQ), MLR, and ANN approach in Charles River Basin, Massachusetts, to provide a quick prediction of microbial concentrations for classification purposes using meteorological, hydrologic and microbial explanatory variables. All the models predicted non- violations very well (> 90%). MLR performed poorly in classifying violations. Current and previous day(s) discharge, rainfall in the last hour(s), and storm intensity were used as inputs in that study. ANN and LVQ models performed very closely and better than MLR model when five or more inputs were used [13].

Gosukonda et al. developed an artificial neural network (ANN) model predicting Escherichia Coli (E. Coli) inactivation due to low-voltage electric current on beef surfaces. This case study's objective was to compare ANN with statistical models such as the polynomial regression for checking its suitability as a tool for online processing by the meat industry. To develop the network, they used current, duty cycles, frequency, and time as inputs and E. Coli as an output. Back-propagation (BP) and Kalman filter (KF) learning algorithms were used in ANN training.

For selecting the best model, many statistical indices, including R^2 were used. The results illustrated that both learning algorithms based ANN performed better than polynomial regression models, especially in interpolating unseen patterns [1].

Yu et al. developed an ANN model to predict the survival/death and growth/no-growth rate of *E. Coli* in a mayonnaise model system. They used a three-layer back-propagation neural network with temperature, pH, acetic acid, sucrose, and salt, as the input variables. They used controlled experiment results as input to the model. The model was able to accurately predict the growth/no-growth by 99.5% and survival/death by 99.1%. Also, in the validation, the ANN model predicted 8 out of 15 correctly. This ANN model was recommended as an alternative tool for the classification of survival and growth conditions in predictive microbiology [14].

Brion et al. carried out the research work to predict the number of viruses in shellfish. Their research compared the performance of the ANN and MLR (multivariate logistic regression) to predict the presence/absence of three kinds of viral pathogens and their indicators. They used a feed-forward ANN model with back-propagation training developed using the software Neurosort VerII, at the University of Kentucky. This research accomplished to predict the presence and absence of PCR-identified human adenovirus (ADV), Norwalk-like virus (NLV), and enterovirus (EV) in shellfish harvested from four different European countries [17]. ANNs were marginally better than the simpler MLR models and they better captured the extreme values [15].

Dwivedi et al. attempted to use BNN model to predict *E. Coli* load in surface water. For this purpose, they compared the results of the BNN model, which utilized thirteen variables to estimate *E. Coli* loads with the comprehensive feature selection technique called LOADEST. They identified 6 out of 13 factors as essential factors for determining *E. Coli* loads. *E. Coli* loads were also predicted using a traditional model called load estimator (LOADEST), developed by the U.S. Geological Survey. In terms of the model efficiency, overall *E. Coli* load predictions by the BNN model were better than the *E. Coli* load predictions by the LOADEST model on all the three occasions (threefold cross-validation). Research indicated the advantages of using LOADEST model in the smaller ranges and BNN model in the higher ranges. Advantages of using BNN as a tool for decision makers and environment managers were presented [3].

Garcia-Gimeno et al. developed an ANN with five input variables: pH, sodium chloride, nitrate concentrations, temperature, and aerobic/anaerobic conditions. They compared it with Response Surface Model (RSM) to estimate the growth response data for E. Coli. The results highlighted ANN to be a useful tool for estimating E. Coli kinetic parameters, including growth rate and lag-time, with less estimation error than RSM (%18 against %27), for a similar complexity. Researchers indicated that having both kinetic parameters in one model as an advantage [16].

Maier and Dandy reviewed 43 papers involving the use of neural network models developed for prediction and forecasting in water resources and environmental research. This review presented a good description of basic ANN theory, limitations and advantages of the ANN modeling [17].

De Vito et al. developed feed-forward neural networks in a regression scheme to predict CO, NO₂, and NO_x in urban pollutant concentration. They used standard station output and used MATLAB environment as a neural network tool for training and simulation. They selected hyperbolic tangent as a hidden neuron transfer function and used the Levenberg–Marquardt training algorithm for the ANN training. They also used early stopping and automatic Bayesian regularization (ABR), a neural network capacity control technique for training to avoid overtraining issues. This research indicated the potential of ANN to capture the cyclic behavior of the pollutant concentration with a short training set [18].

3. SYSTEM CONSIDERED

3.1 Lake Michigan beaches locations

3.1.1 Indiana Dunes State Park Location

Indiana Dunes State Park is located in the Porter County, Indiana, United States. It is located 47 miles (75.6 km) east of Chicago downtown. This beach is located in the southern tip of Lake Michigan. It is surrounded all four sides by Indiana Dunes National Park service. Every summer, many beach users, including day users, campgrounds, and Hispanic social gatherings, visit the Park. The location of the Park's beach is the main attraction for the public, aside from dunes. The park's strategic location with easy access to large population makes Indiana Dunes as a popular stop for visitors from around the world. US Interstates I-80 and I-94 are just few miles away from this park. Citizens from 28 different countries signed in at the Park visiting Center.

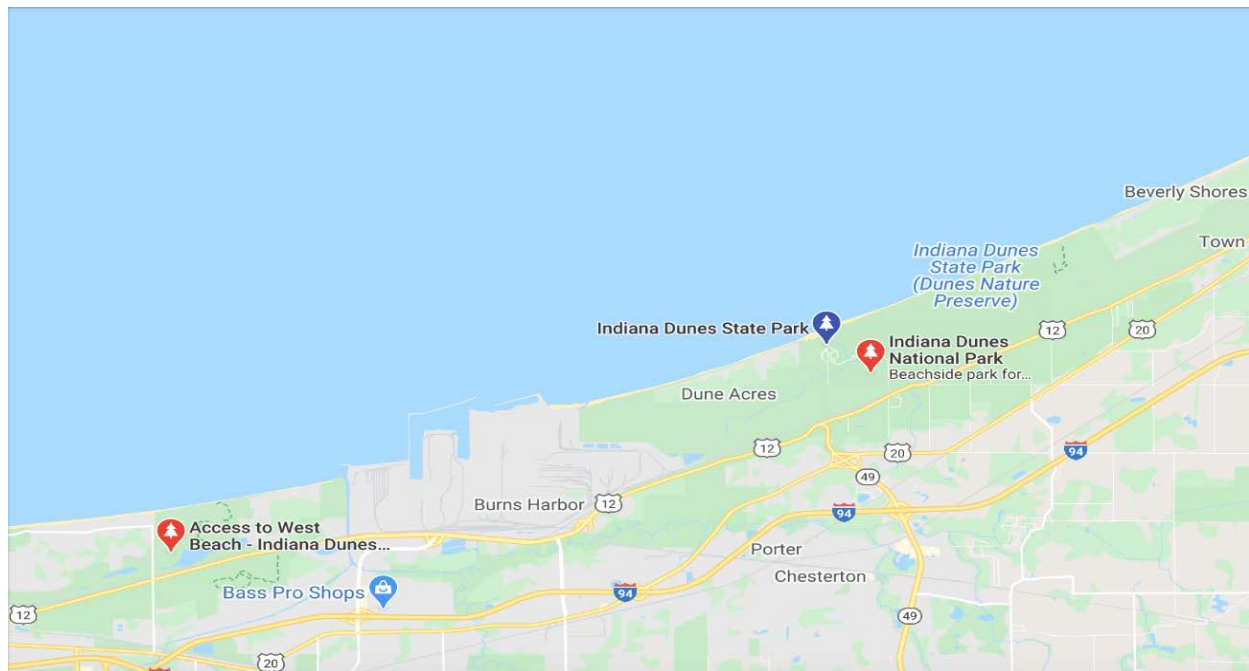


Figure 3.1 Indiana Dunes State Park location

3.1.2 Jeorse Park Beaches Location

Jeorse Park Beach (Jeorse 1, Jeorse 2, and Buffington Harbor) is located in City of East Chicago. This beach is located southeast of Indiana Harbor Canal. From this beach, Chicago Downtown skyline is visible. This beach is surrounded by Casinos. In the north, south and west of the beach, we have Ameristar Casio, Majestic Star Casino and the Cline Avenue respectively.

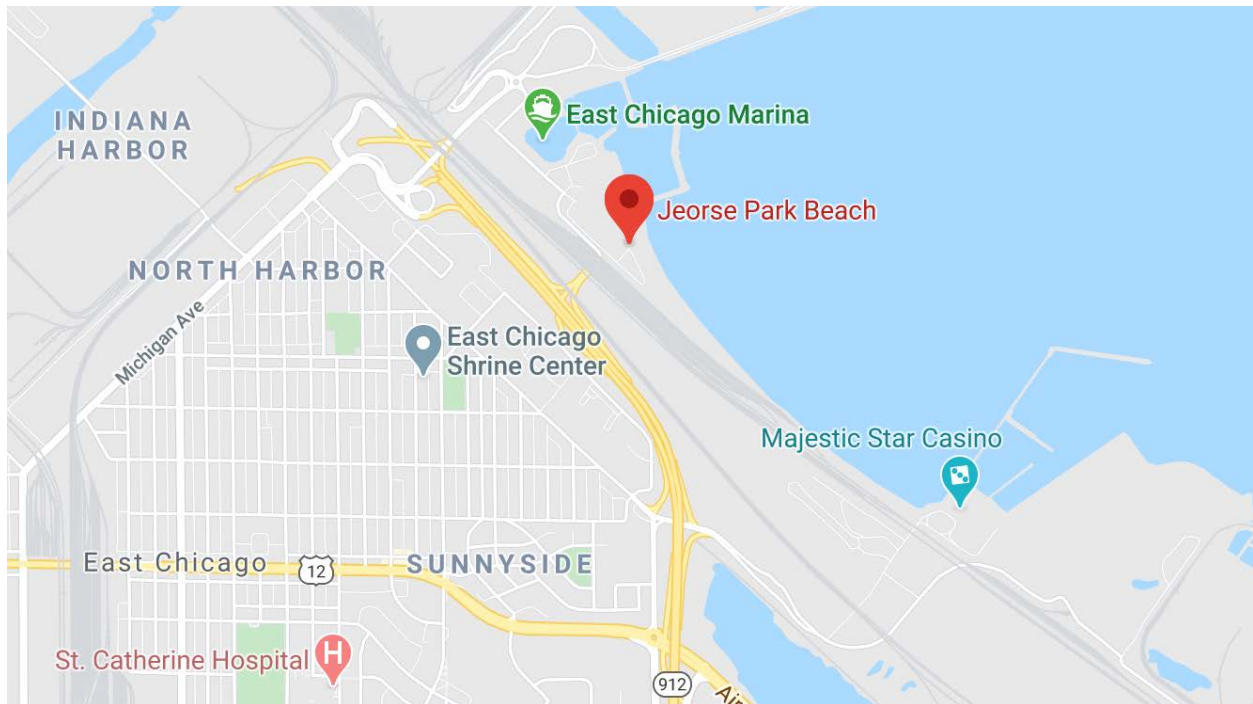


Figure 3.2 Jeorse Park Beach location

3.1.3 Whihala Beach Location

Whihala Beach is also located in the southern tip of Lake Michigan. It is located in Whiting, Indiana, United States. It is located 18 miles east of Chicago downtown. It is a public beach, and the management provides lifeguards during beach seasons. The beach was known as Whiting Beach earlier.

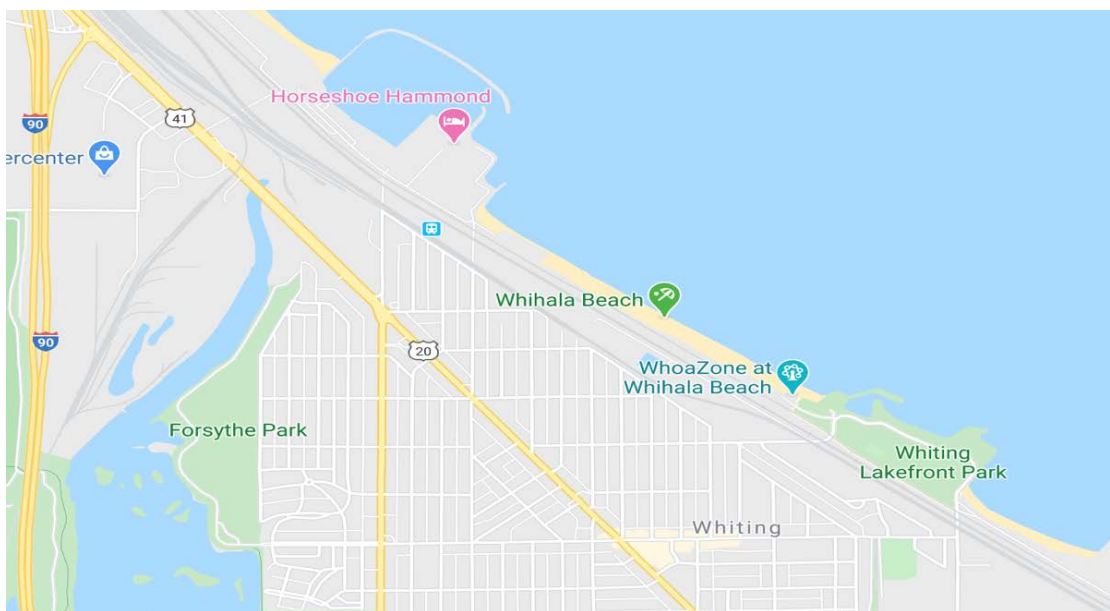


Figure 3.3 Whihala Beach location

3.2 Field Sampling Training

Sampling efforts were planned and implemented from July 24th, 2019 to September 15th, 2019. Mitra Khanibaseri, Michael Ozeh, David Okposio, Katrina Lynn Cook, and Neil Thompson participated had contribution in the sampling work. Mitra Khanibaseri led the Indiana Dunes data collection and Michael Ozeh handled the east beaches. On July 8th and 9th Mitra Khanibaseri and Michael Ozeh got trained by Microbac laboratory field technicians Jim Deter and Darrin Ferris.

Jim Deter met students in the parking lot at the East Chicago Beach between 6:00 – 6:30 a.m. on Monday, July 8th. Jim had the written information and reviewed the standard sampling techniques for the E. Coli beach monitoring project following the Quality Assurance Project Plan provided by the Indiana department of environmental management.

Darrin Ferris met students in the Pavilion parking lot, right side, at the Indiana Dunes State Park Beach between 7:00 – 7:15 a.m. on Tuesday, July 9th. He carried sample bottles, chain of custody forms and a cooler to pass along to students. During the sampling collection period, students strictly followed the standard methods and protocols provided by the Indiana department of environmental management.



Figure 3.4 Sampling bottles and gloves



Figure 3.5 Sampling coolers

3.3 Data collection

3.3.1 Data Collection Details

Data collection for the model development occurred daily, from July 24th through September 15, 2019. Data collection was consisting of the following components:

- Daily collection and analysis of water samples for Total Dissolved Solids (TDS)
- Daily collection and analysis of water samples for Total Suspended Solids (TSS)
- Daily collection and analysis of water samples for E. Coli numeration
- An intensive monitoring period was conducted during one consecutive seven-day period for every month
- Field data collection and recording
- Chain of custody
- Beachgoer counts, pet counts, and bird counts
- Color, odor and algae growth if any were noted.
- Identification of other data relevant for model development

Samples were collected at each of the six different beach locations listed in Table 1 and Table 2. Sampling was performed from July 24, 2019, to September 15, 2019, to collect the necessary data for artificial neural network model development. On July 24, the first day of sampling collection, students were not able to handover the afternoon and evening water samples to the Microbac laboratory for analysis due to the traffic delays and lack of time management. So, only the morning water samples result was available for model development on July 24.

Table 3.1 Sampling location candidates for water quality characterization

Sample Location Description	Latitude (degrees N)	Longitude (degrees W)
Indiana Dunes State Park Beach East	41.663649	87.062202
Indiana Dunes State Park Beach West	41.662465	87.065276
Jeorse Park Beach I	41.650810	87.433422
Jeorse Park Beach II	41.649641	87.432987
Buffington Harbor Beach	41.649083	87.432595
Whihala East Beach	41.685088	87.491932

3.3.2 Intensive Monitoring Period

Intensive sampling was planned for each month. For a continuous seven-day period, apart from morning sampling, two additional samplings were conducted. First afternoon sampling was done between 12 noon to 12.30 pm and the second one was done around 2.00 pm to 2.30 pm. During these periods, PNW students collected, transported, and analyzed two additional water grab samples per day per beach for all the considered variables. All TSS and E. Coli grab samples were delivered to Microbac's Chicagoland Laboratory before 4:30 pm CDT. All Total Dissolved Solids (TDS) water samples were taken to PNW's Water Institute Lab for analysis.

Table 3.2 Intensive weeks date collection details

Beaches	Intensive weeks sampling period		Notes/Challenges
Indiana Dunes East	July	24 to 31	07/24 (afternoon and evening sampling. Missed Microbac laboratory the first day.)
	Aug	9 to 15	
	Sep	5 to 8	
		13 to 15	
Indiana Dunes West	July	24 to 31	07/24 (afternoon and evening sampling. Missed Microbac laboratory the first day.)
	Aug	9 to 15	
	Sep	5 to 8	
		13 to 15	
Jeorse 1	July	24 to 31	07/24 (Missed Microbac laboratory the first day) 07/27 (access to beach barricaded due to private event); 08/02 to 08/05 morning (access to beach barricaded due to private event) 08/07 evening (access to beach barricaded due to private event) 09/15 evening (access to beach barricaded due to private event)
	Aug	9 to 15	
	Sep	5 to 8	
		13 to 15	

Table 3.2 Continued

Jeorse 2	July	24 to 31	07/24 (Missed Microbac laboratory the first day)
	Aug	9 to 15	07/27 (access to beach barricaded due to private event);
	Sep	5 to 8	08/02 to 08/05 morning (access to beach barricaded due to private event)
		13 to 15	08/07 evening (access to beach barricaded due to private event) 09/15 evening (access to beach barricaded due to private event)
Buffington Harbor	July	24 to 31	07/24 (Missed Microbac laboratory the first day)
	Aug	9 to 15	07/27 (access to beach barricaded due to private event);
	Sep	5 to 8	08/02 to 08/05 morning (access to beach barricaded due to private event)
		13 to 15	08/07 evening (access to beach barricaded due to private event) 09/15 evening (access to beach barricaded due to private event)
Whihala	July	24 to 31	07/24 (Missed Microbac laboratory the first day);
	Aug	9 to 15	07/25 (Out of time due to excessive vehicle and train traffic)
	Sep	5 to 8	
		13 to 15	

3.3.3 TDS Collection and Analysis

For TDS analysis, students from Purdue University Northwest, daily collected single grab water samples from July 24, 2019, through September 15, 2019, between 7:00 and 8:00 a.m. for the selected beaches. They collected two additional grab water samples (one between 12 to 12.30 p.m. and another between 2 to 2.30 p.m.) on intensive weeks from each of the six project beaches. Samples were labeled and transported to the Purdue Water Institute in Hammond, Indiana, for analysis within 2 hours.

3.3.4 TSS Collection and Analysis

Particles larger than 2 microns in water is called Total suspended solids (TSS). Particles smaller than that are usually referred as dissolved solids and measured as Total Dissolved solids (TDS). Most of the TSS are inorganic materials. On the other hand, bacteria, algae and organic particles from decomposing materials contribute to TDS concentration.

Trained field staff from Microbac Laboratory, located in Merrillville, Indiana, collected single water samples for purposes of TSS analysis from each of the six project beaches, between 7:00 am and 8:00 a.m., daily from July 24, 2019, through September 15, 2019. Samples were labeled and transported to the laboratory for analysis.

Research students from Purdue University Northwest, collected two water samples during intensive weeks from each of the six project beaches to be analyzed for Total Suspended Solids (TSS) and E. Coli. The sampling was done between 12:00 and 12:30 pm and 2:00 to 2:30 pm during intensive weeks. Samples were labeled and transported to Microbac Laboratory for analysis. Intensive weeks for July, Aug, Sep months were given in Table 3.2.

3.3.5 E. Coli Collection and Analysis

Trained field staff from Microbac Laboratory, located in Merrillville, Indiana, collected single grab water samples for purposes of E. Coli count from each of the six project beaches, between 7:00 am and 8:00 am, daily. They were thereafter preserved, transported, and analyzed according to the 2019-2023 Lake Michigan Beaches Monitoring and Notification Program QAPP (IDEM

2019). The analysis utilized the IDEXX Quanti-Tray/2000 technique, satisfying Standard Method 9223B. Microbac Laboratories (Microbac) possesses a number of accreditations and certifications for E. Coli analysis.

Purdue University Northwest research students collected samples from the selected sites for E. Coli counts at approximately 12:00 to 12:30 p.m., and 2:00 to 2:30 p.m. on intensive weeks. Samples were labeled and transported to Microbac Laboratory for analysis. They also filled out a routine form indicating beachgoer count, bird count and pet count every morning on non-intensive weeks and morning, afternoon and evening on intensive weeks. Other things observed were trash on the beach, odor, discoloration of the water and dead aquatic animals. These forms were shared with Microbac Laboratories while the students retained a copy. During regular sampling days, students visited all six beaches to observe pets, beach visitors and bird counts in the afternoon.

Table 3.3 2019 Beach Program E. Coli Sampling Schedule

Beach	2019 Beach Program Sampling Begins	2019 Beach Program Sampling Ends (last sample)	Number of Beach Program Samples Collected
IN Dunes State Park West	July 24, 2019	September 15, 2019	98
IN Dunes State Park East	July 24, 2019	September 15, 2019	98
Jeorse1	July 24, 2019	September 15, 2019	96
Jeorse2	July 24, 2019	September 15, 2019	91
Buffington Harbor	July 24, 2019	September 15, 2019	91
Whihala	July 24, 2019	September 15, 2019	91
Total	July 24, 2019	September 15, 2019	565

3.4 Field Data Collection and Recording

Most reported models predict the concentration of E. Coli as a function of environmental factors. So, Purdue University Northwest students observed the following during sample collection:

- Wind (calm, light breeze, moderate breeze, windy)
- Wind direction
- Lake Character: water surface (calm, shoreline breakers (waves))
- Lake Character: watercolor (clear medium brown, dark brown, red-brown, green-brown, other)
- Lake Character: smell (none, sewage, oily, rotten eggs, fishy)
- Lake Character: other (dead fish, algal bloom, litter/trash)
- Lake Character: beach pets (visual observations/counts), beach wildlife (visual observation), birds around the sampling site (recorded as low (<10 counts), medium (10 to 20) or large (>20), beach visitors (visual observations)
- Rainfall (precipitation)
- Air and water temperature (°C)
- Weather in the past 24 hours: storm (heavy rain), rain (steady rain), showers (intermittent rain), overcast, clear/sunny
- Weather now: storm (heavy rain), rain (steady rain), showers (intermittent rain), overcast, and clear/sunny.
- pH
- Total Dissolved Solids (TDS)
- Electrical Conductivity (EC)
- Turbidity

Rainfall observations were obtained from the iclimate (<https://iclimate.org/>), provided by the Indiana State Climate Office at Purdue University. Wind categories were recorded as calm (Beaufort Scale wind force 0 or 1), light breeze (Beaufort Scale wind force 2 or 3), moderate breeze (Beaufort Scale wind force 4 or 5), or windy (greater than Beaufort Scale wind force 5).

3.5 Chain of Custody/Sample Handling

Both PNW student samplers and samplers from the Microbac laboratory completed a chain of custody form (i.e., those found in [Appendix A](#), and [Appendix B](#)) for each sampling event. The chain of custody form also documented the laboratory control number for each sample collected during the sampling event. The laboratory control number on the chain of custody form will match the identification number on each sample's container.

The field staff then signed and date the chain of custody form verifying that they collected or viewed the collection of the identified sample(s). All samples collected by the sampler were labeled to identify the sample for database records. The labels included location, date, time, sample's field identification number and other information documents. Labels printed and affixed to the outside wall of the sample container in the lab, before going to the field.

All laboratory identified isolates (with abnormal data values) had labeled for the database records. Record of the chain of custody had maintained for each sample. All samples at PNW are considered non-hazardous and had disposed of down the sink. Microbac had autoclave E. Coli samples before disposal.

3.6 Collection of Beachgoer, Pet, and Bird Counts

Beachgoer, pet, and bird counts had taken during the project to assess E. Coli inputs. Observations had taken by Purdue University Northwest or the Microbac staff during the time of collection of E. Coli and TSS samples and associated field data. Also, observations of beachgoers, pets, and bird counts had collected outside of the intensive monitoring periods between 12:00 pm and 4:00 pm CDT as follows:

Table 3.4 Responsible parties for bird, pet, and head counts by beach

Beach	Count Collection (Beach Season)	Count Collection (Post-Season)
Indiana Dunes State Park West	IDNR Staff	IDNR Staff /Mitra Khanibaseri/Nile Thompson
Indiana Dunes State Park East	IDNR Staff	IDNR Staff/Mitra Khanibaseri/ Nile Thompson
Jeorse1	PNW Students	Michael Ozeh/David Okposio/Katrina Lynn Cook
Jeorse2	PNW Students	Michael Ozeh/David Okposio/ Katrina Lynn Cook
Buffington Harbor	PNW Students	Michael Ozeh/David Okposio/ Katrina Lynn Cook
Whihala	Whiting Parks Dept.	Michael Ozeh/David Okposio/ Katrina Lynn Cook

Bird counts had consisted of total birds only; a breakdown by species or other characteristics was not be required. It is assumed that the pet counts were anticipated to be dogs only; the presence of other pet species or non-avian wildlife observed had noted as a comment on the field datasheet. Every day at approximately 3 pm, a student had visited the Whiting and East Chicago beaches to observe the beachgoer count.

For the Indiana Dunes State Park, entrance fees charge utilized as a proxy for beachgoer counts. Note that entrants possessing valid annual permits would not be charged a separate entrance fee and would not be counted; nor would not the count include each person in a single non-commercial vehicle.

3.7 Identification and Use of Other Data

Other data such as rip current data, lake level, precipitation data, and flow observations in a nearby creek obtained from the appropriate sources, as shown in the table below.

Table 3.5 Data Sources for Project Secondary (Existing) Data

Data	Source
Precipitation	Iclimate
Flow Observations	USGS flow observations at Hart Ditch, Portage Burns Waterway, Burns Ditch, Indiana Harbor Canal, Little Calumet River East Arm and Grand Calumet River
Beach Goer Counts – Whihala East	Whiting Parks Department/Purdue University Observation
Beach Goer Counts – East Chicago Beaches	Purdue University Observation
Beach Goer Counts – INDSP	IDNR/INDSP Gate Office

Table 3.6 USGS Streamflow Gauges in Project Vicinity

Gauge No.	Name	City	Years of Record (As of July 2019)
04092750	Indiana Harbor Canal	East Chicago, IN	22
04092677	Grand Calumet River at Industrial Highway	Gary, IN	21
04093176	Little Calumet River at Grant Street	Gary, IN	N/A
04093250	Little Calumet River Near Lake Station, IN	Lake Station, IN	N/A
04093503	Burns Ditch at US Highway 20	Lake Station, IN	N/A
04095090	Portage-Burns Waterway at Portage, IN	Portage, IN	23
04094000	Little Calumet River at Porter, IN	Porter, IN	73

3.8 Collection of meteorological data from Gary airport

Some meteorological data that were not a part of the data collection, but could influence the detection of E. Coli, such as mean hourly temperature (MHT), relative humidity (RH), mean wind speed (MWS), mean wind direction (MWD), and precipitation were also added to the model's inputs. These meteorological data were collected from the link below.

<https://mrcc.illinois.edu/CLIMATE/welcome.jsp>

3.9 Data Preparation/pre-processing Attempts

On September 15th, the data collection works were completed. After the data collection, the field data were consolidated initially and were cross checked to avoid any mistakes in the documentation. This cross checking was done by the PI and two students who participated in the sampling effort.

The next step after data collection in this project was data pre-processing, usually referred to as data cleanup. This practice converts the data into a sequence with which the ANN training algorithm may make better sense out of it during training without compromising the integrity of the data.

In this practice, the data is usually broken down into classes that capture desired contexts, and these classifications are the products of past research and/or experience. This does not negate the use of the raw data but presents a different way, such that the algorithm can digest the information and lead to new insights. In many cases, the best results come from such pre-processed data used exclusively or as a hybrid with the raw data (for example, some input data are raw data and some input data are classified data with classified data output target).

Some of the preprocessing methods used are listed below.

3.9.1 Observed normal conditions

Studies were carried out to ascertain the normal range of values for a typical lake at the time the data was observed, then the data was cleaned to reflect normal range as 1, less than normal as 2 and greater than normal as 3.

An example is the temperature data. The normal temperature range for a lake shore (which forms the beach) was found from several sources to be 66°F to 76°F (about 18.8°C to 24.5°C). So, this range was classified as 1, then temperatures below 18.8 C were classified as 2 and temperatures above 24.5 C were classified as 3. Others are shown in Table 3.7.

Table 3.7 Classification for TSS, Turbidity, pH and Wind Speed

Parameter	Condition	Classification
TSS	Observed normal range for a calm shore is usually less than or equal to 5mg/L.	1
	Slight breakers, around 6mg/L and 20mg/L (human activity is considered in this range)	2
	Breakers, around 21mg/L and 50mg/L	3
	Turbulent, usually above 50mg/L	4
Turbidity	Observed normal range for a calm shore is usually less than 5NTUs	1
	Slight breakers, around 6NTUs to 15NTUs	2
	Average breakers with human activity, around 15NTUs to 25NTUs	3
	Large breakers, 25NTUs to 50NTUs	4
	Turbulent, more than 50NTUs	5
pH	Less than normal range for a lake shore at the period of data collection is usually less than or equal to 7.55	1
	Lower end of normal range for a lake shore at the period of data collection is usually about 7.56 to 8.34	2
	Higher end of normal range for a lake shore at the period of data collection is usually about 8.35 to 8.65	3
	Higher than normal range for a lake shore at the period of data collection is usually above 8.65	4
Wind speed	Average range for a lake shore at the period of data collection is usually about 6.5mph to 8.4mph	1
	Less than normal range for a lake shore at the period of data collection is usually less than 6.5mph, meaning it was calmer than average	2
	Higher than normal range for a lake shore at the period of data collection is usually more than 8.4 mph, meaning it was windier than average	3

3.9.2 Closeness to the mean of measured data

Some of the data were classified using standard deviation to separate the data into values that were quite close to the mean, and those that behaved like outliers with respect to their location on either side of the mean. The mean and the standard deviation of the data was found, then a bracket of +25% and -25% deviations around the mean were calculated. The data in this bracket is classified as 1, while the remaining lower and upper 25% ends are classified as 2 and 3, respectively. Classifications could be refined as needed as model analysis progressed.

For example, the mean of the TDS data is 198, while the standard deviation is 26. From this, the bracket became $[198 + 13]$ and $[198 - 13]$, classified as 1. Values further from the mean and less than the lowest value in the bracket were classified as 2, while those higher than the maximum value in the bracket, as 3. This resulted in:

$$185 - 211 = 1$$

$$<185 = 2$$

$$>211 = 3$$

Other parameters classified this way are Electrical Conductivity, Discharge data from Little Calumet, Grand Calumet, Portage-Burns Waterway, Indiana Harbor Canal, East Chicago and Hart-Ditch River.

3.9.3 Binary/ Pseudo-Binary Classifications

There were few data in this research which were binary in nature. An example is odor. When there was a perceived odor, it was classified as 1 and if there was no odor, it was classified as 0. Another example is the presence of trash in the vicinity: “Yes” for substantial litter, “Minimal” for little litter and “No” for a clean beach.

Others in this category include Algal bloom, Color, Average Relative Humidity and Precipitation.

All zero precipitation were classified as 0, trace precipitation (those at or under 0.1 in) were classified as 1 and those above 0.1 were classified as 2. A one-day lag was adopted for the

precipitation data because we hypothesize the impact of the measured precipitation would be a more significant factor for the succeeding day.

Average Relative Humidity of the sampling environment was also classified in a binary manner, where <70 was classified as 1 and 70+ classified as 2.

3.9.4 Miscellaneous

Some data were classified based on the population density classification used in the Microbac Laboratory data sheet for number of birds. This was extended to number of people and pets. For example:

Birds:

0 birds	= 0
1--3 birds	= 1
4--10 birds	= 2
11--30 birds	= 3
31--100 birds	= 4
>100 birds	= 5

People:

0 people	= 0
1--10 people	= 1
11--30 people	= 2
31--99 people	= 3
100--499 people	= 4
500+ people	= 5

Pets:

0 pets	= 0
1--3 pets	= 1
4--10 pets	= 2
11--30 pets	= 3
31--100 pets	= 4
>100 pets	= 5

The classification for wind direction is simply a numerical transformation of the directions into 4 quadrants:

0 to 90	= 1
91--180	= 2
181--270	= 3
271--360	= 4

3.9.5 E. Coli Classification

E. Coli classification was based on the safe limit for recreational waters. The classification was done for four groups at first. During modeling process based on suggestions, we adopted to three groups and two groups subsequently. Two groups were finally chosen as it was the grouping that reduced the heavy bias of the data set towards safe limit numbers during training. The number classes below refer to colony forming units per 100 ml:

4 classes:

0 to 125	= safe, classified as 1
126--235	= advisory (but still safe), classified as 2
236--799	= unsafe, classified as 3
800+	= highly unsafe, classified as 4

3 classes:

0 to 235	= safe, classified as 1
236--799	= unsafe, classified as 2
800+	= highly unsafe, classified as 3

2 classes:

0 to 235	= safe, classified as 1
236+	= unsafe, classified as 2

During model building process, several combinations of inputs involving raw data as well as classified data inputs were tried. Lagged input data were also attempted. Likewise, raw E. Coli counts as well as classified ones were tried as outputs.

Some additional data that were not a part of the data collection, but could influence the detection of E. Coli, were also added. These are the discharge data, wind speed and wind direction. All these were consolidated into an input series for the output – E. Coli. An outline of the full modeling process is presented in Figure 3.6.

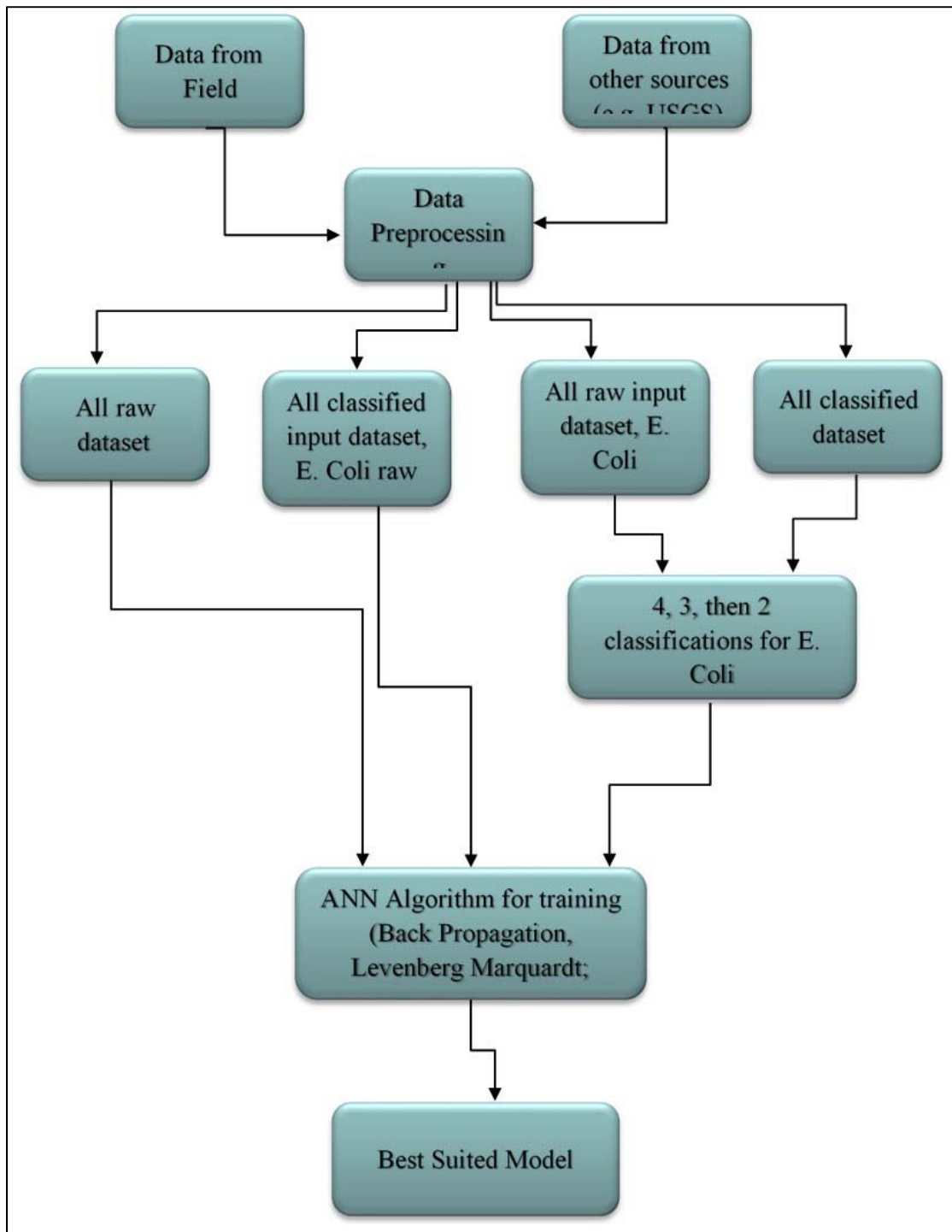


Figure 3.6 Flowchart on overview of the study process

3.10 Instrumental Analysis

3.10.1 Temperature

The temperature of the sample is measured with an Oakton pH/mV/°C pH510 series temperature probe with a 0.1°C resolution. It has a range of 0 °C to 100 °C and an accuracy of ± 0.3 °C.

Method: The temperature probe is dipped into the sample and the temperature output, which fluctuates till it stabilizes, is displayed on the LCD screen. When the temperature reading stabilizes, “ready” is displayed on the screen.

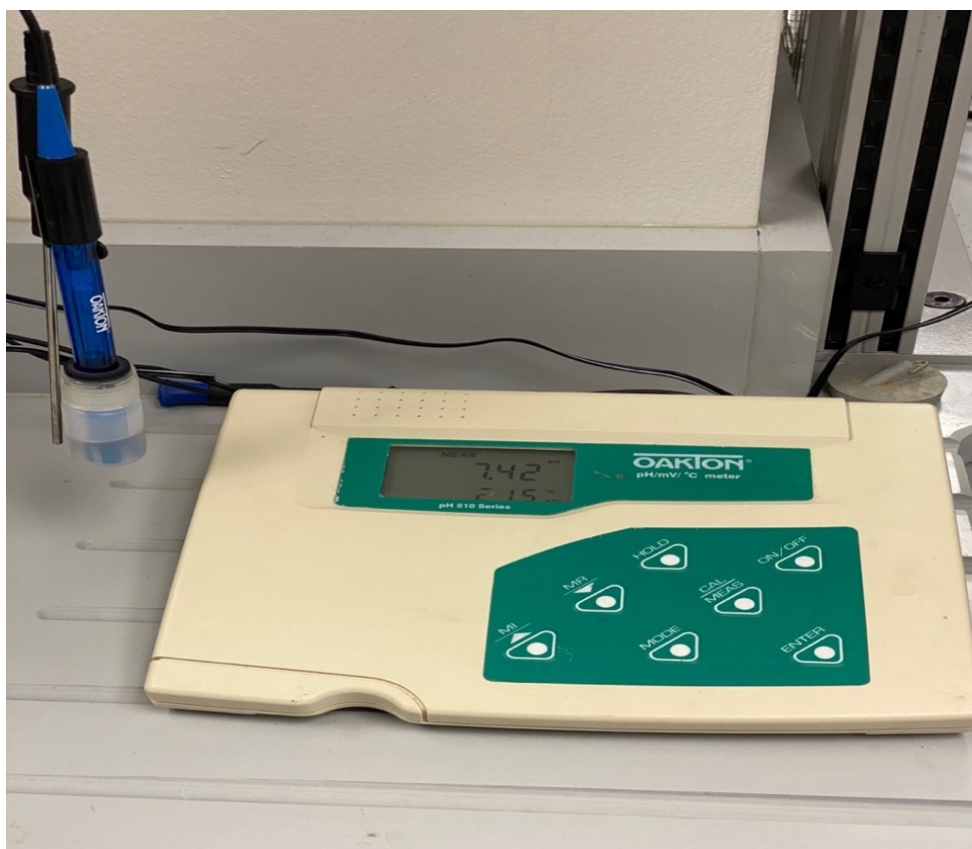


Figure 3.7 Oakton pH and Temperature Meter

3.10.2 pH:

The pH of the sample is also measured with an OakTon pH/mV/°C pH510 series electrode probe with a 0.01°C resolution. It has a range of 0.00pH to 14.00pH and an accuracy of $\pm 0.01 \text{ pH} + 1$ count.

Method: Any electrode soaker bottle or protective rubber cap from the electrode is removed and both the pH electrode and temperature probe are dipped into the sample. The temperature probe is necessary for the pH measurement so as to automatically compensate for the sample temperature if needed by adjusting the pH output accordingly. The pH output, which fluctuates till it stabilizes, is displayed on the LCD screen. When the pH reading stabilizes, [Ready] is displayed on the screen. When the instrument shows “READY mode”, it indicates that the readings are stable within a range of $\pm 0.01 \text{ pH}$. At this mode, reading is observed.

3.10.3 Turbidity:

The Hach 2100N Turbidimeter measures the turbidity of the sample within a range of 0 to 4000 NTU. It has an accuracy of $\pm 2\%$ of reading plus 0.01 NTU from 0 to 1000 NTU and $\pm 5\%$ of reading from 1000 NTU to 4000 NTU. The stability time for taking readings is 30 minutes (with the [Ratio] functionality on) to 60 minutes (with the [Ratio] functionality off). Readings expected to be $> 40 \text{ NTU}$ will need the [Ratio] functionality on. The repeatability of the measurement is either $\pm 1\%$ of the reading or $\pm 0.01 \text{ NTU}$, whichever is greater.

Method: A representative sample is collected in a clean container. The sample is filled in the sample cell to the line marking. The capacity is approximately 30 mL.

The sample cell is capped. From the top, a thin bead of silicone oil is applied to the bottom of the cell. It coats the cell with a thin layer of oil. For spreading the oil uniformly, provided oil cloth is used. After the oil is spread uniformly, any excess oil is wiped.

In the instrument cell compartment, prepared sample cell is placed. After that, the cell cover was closed.

Manual or automatic ranging is selected by pressing the RANGE key (usually automatic), the appropriate SIGNAL AVERAGING setting (on or off; usually on) is selected by pressing the SIGNAL AVG key and the appropriate RATIO setting (on or off) is selected by pressing the RATIO key. (Values >40 NTU require Ratio on).

The appropriate measurement unit (NTU, EBC or NEPH) is selected by pressing the UNITS/EXIT key (usually NTU). The readings are thereafter read and recorded after stabilization.



Figure 3.8 Hach Turbidity Meter

3.10.4 TDS and Electrical conductivity

The Total dissolved solids are measured with an Extech EC600 meter. The measurement range is 0 to 100 g/L with an accuracy of $\pm 2\%$ and a resolution of 0.01g/L. The equipment is ISO9001, CE and CMC Quality/Safety certified.

Method: The electrode is cleaned with deionized water and air-dried, then immersed into the sample solution, gently stirred and allowed to stand till the reading stabilizes. The reading is then taken after using the ENTER key to select the TDS measurement mode.



Figure 3.9 Extech Conductivity Meter

3.10.5 Total Dissolved Solids (TDS) and Temperature

The Total dissolved solids and temperature are also measured with an Apera EC60 meter. The measurement range for TDS is 0 to 100 ppm with an accuracy of $\pm 1\%$ and a resolution of 0.1 ppm. The measurement range for temperature is 0 to 50 °C with an accuracy of $\pm 0.5\%$ and a resolution of 0.1°C.

Method: The temperature probe is dipped into the sample and the temperature output, which fluctuates till it stabilizes, is displayed on the LCD screen. When the temperature reading stabilizes, “laugh emoji” is displayed on the screen. To measure TDS, the probe is cleaned with deionized water and air-dried, then immersed into the sample solution, gently stirred and allowed

to stand till the reading stabilizes. The reading is then taken after using the ENTER key to select the TDS measurement mode.



Figure 3.10 Apera CE60 TDS and Thermometer

4. ARTIFICIAL NEURAL NETWORK MODEL BUILDING PROCESS

4.1 Overview of Artificial Neural Networks

This chapter reviews the theoretical background of ANN, including its learning algorithms, limitations, explains the mathematical foundations and biological inspirations behind ANN.

4.1.1 Biological Neural Network

Artificial Neural Network (ANN) is the method that inspired by brain neurons. Neural networks theory introduced by Warren McCulloch and Walter Pitts [19] in 1943, the method did not apply to an application until 2011 because of the lack of processing and computational power.

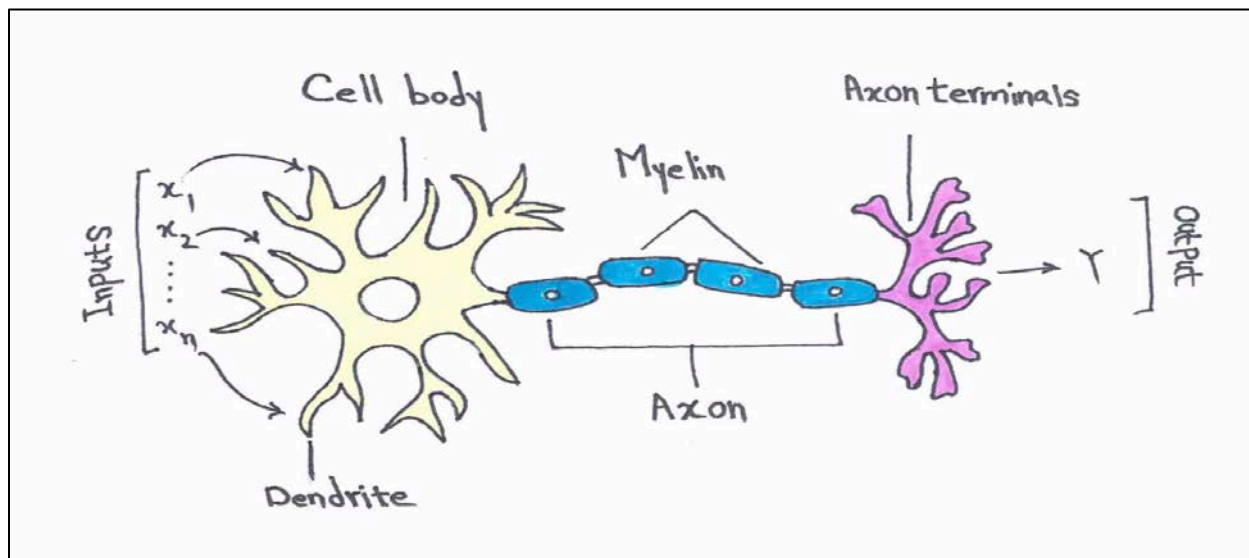


Figure 4.1 Biological Neuron and Axon.

Figure 4.1 shows the neurons in our brains. As shown, the neuron receives the signals from the input and process them through the cell body and send them out throughout the axon to another cell input. In figure 4.1 [19], the x_1 to x_n are the inputs, with raw values like 0 or 1.

4.1.2 Artificial Neural Network

An artificial neural network (ANN) is the part of a computing system that originated to mimic the way the human brain neurons interconnected and process the information. Each neural network has three significant components:

1. Node character
2. Network architecture
3. Learning rules

From the input layer, information flows to each neuron in the hidden layer through interconnecting weights. The number of input nodes connected to the hidden layer neuron is processed through the activation function used in the hidden layer neuron. From there, the information flows to the output node through interconnecting weights. Network architecture outlines the adopted neurons in the input, hidden, and output layers and their inter-connectivities. Learning rules define inter-connectivities and weight initialization [20].

Node Character

As shown in figure 4.2, assume the output of the model is pre-known as one. If, in the first iteration, the output of the sum term is less than the threshold, the output of the model gives zero, which is different than the actual output, which is one. In the process of finding the best weights, all three weights should be changed until the sum term of equation 4.1 becomes greater than the threshold. After the output reaches the threshold, all the weights are stored and remain fixed, and the network considered trained. When the neural network model trained by all the known inputs ($x_1, x_2 \dots x_n$) the model can use for predicting the unknown inputs. Figure 4.2 shows the complete ANN workflow diagram. Equation 4.1, in this figure, illustrated the demonstrated activation function.

$$Output = \begin{cases} 0, & \sum_{j=1}^m w_j x_j < threshold \\ 1, & \sum_{j=1}^m w_j x_j \geq threshold \end{cases} \quad (4.1)$$

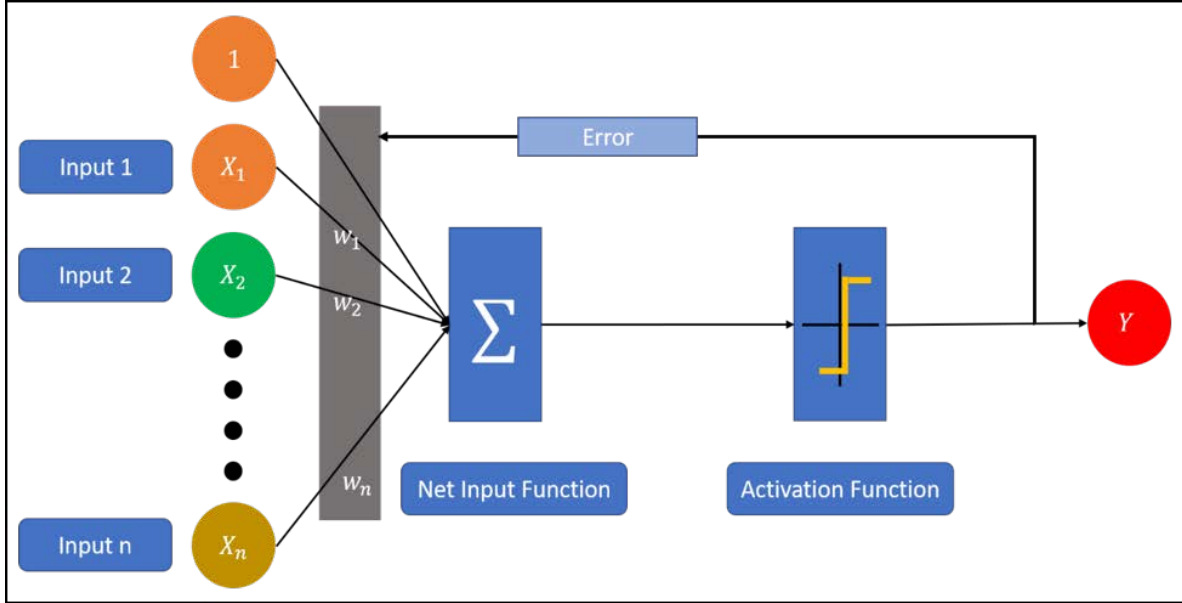


Figure 4.2 Simple example of how one single neuron works in ANN

For simplicity the equation 4.1 is written as

$$Output = \begin{cases} 0, & w \cdot x + b < 0 \\ 1, & w \cdot x + b \geq 0 \end{cases} \quad (4.2)$$

Where w and x are vectors whose components are the weight and inputs, respectively. Bias can be a measure of how easy it is to get the one on model output. For a model with tremendous bias, it's straightforward for the model to output 1. But if the bias is very negative, then it's difficult for the perceptron to output 1. The activation function that is used within the model can be changed according to the various applications. The above example uses a unit step activation function.

Linear and unit step is not really practical for most of the applications. The sigmoid function is often used as the activation function. Figure 4.3 [21] shows more activations functions with their equations.


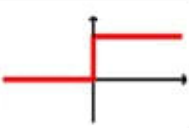
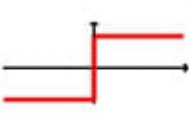
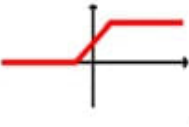

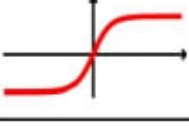

Activation Function	Equation	Example	1D Graph
Linear	$\phi(z) = z$	Adaline, linear regression	
Unit Step (Heaviside Function)	$\phi(z) = \begin{cases} 0 & z < 0 \\ 0.5 & z = 0 \\ 1 & z > 0 \end{cases}$	Perceptron variant	
Sign (signum)	$\phi(z) = \begin{cases} -1 & z < 0 \\ 0 & z = 0 \\ 1 & z > 0 \end{cases}$	Perceptron variant	
Piece-wise Linear	$\phi(z) = \begin{cases} 0 & z \leq -1/2 \\ z + 1/2 & -1/2 \leq z \leq 1/2 \\ 1 & z \geq 1/2 \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multilayer NN	
Hyperbolic Tangent (tanh)	$\phi(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$	Multilayer NN, RNNs	
ReLU	$\phi(z) = \begin{cases} 0 & z < 0 \\ z & z > 0 \end{cases}$	Multilayer NN, CNNs	

Figure 4.3 Original table from [21]

Network Architecture

In Artificial Neural Network, neurons are organized in layers. Typically, there are three essential layers.

- An input layer: this layer gets the direct input from the data. In our case, Temperature, TSS, pH, and TDS are some of the inputs received in this layer.
- Hidden layer(s): this layer or layers is consisting of all summation and multiplication and activation functions. The input of this layer(s) is from the input layer. In our case, the model was consisting of one hidden layer.
- Output layers: this layer usually has a classifier or regression function to make the last decision. In our case, the predicted E. Coli was the model's output.

The multiple-layer perceptron (MLP), is very popular. To make the neural network to learn, supervised training using back-propagation algorithm is very prevalent. In this feed-forward type model, information flow from input layer to output layer. Calculated output value is compared with actual output to find the error. Based on that, back-propagation algorithm, adjust the inter-connecting weights to minimize the error. This is done using the steepest gradient descent method. It needs the activation functions to be differentiable.

The neural network can have as many hidden layers as required. Having a more hidden layer requires more powerful computational systems. For simplicity of describing the hidden layers figure, 4.4 shows a simple perceptron with just one hidden layer between input and output. Each circular node represents one single artificial neuron, and each arrow represents a connection from the output of one node to the input of another. Each neuron connection has carried a different weigh (w_k) for the next nodes. A 3-4-1 architecture is shown in Figure 4.4.

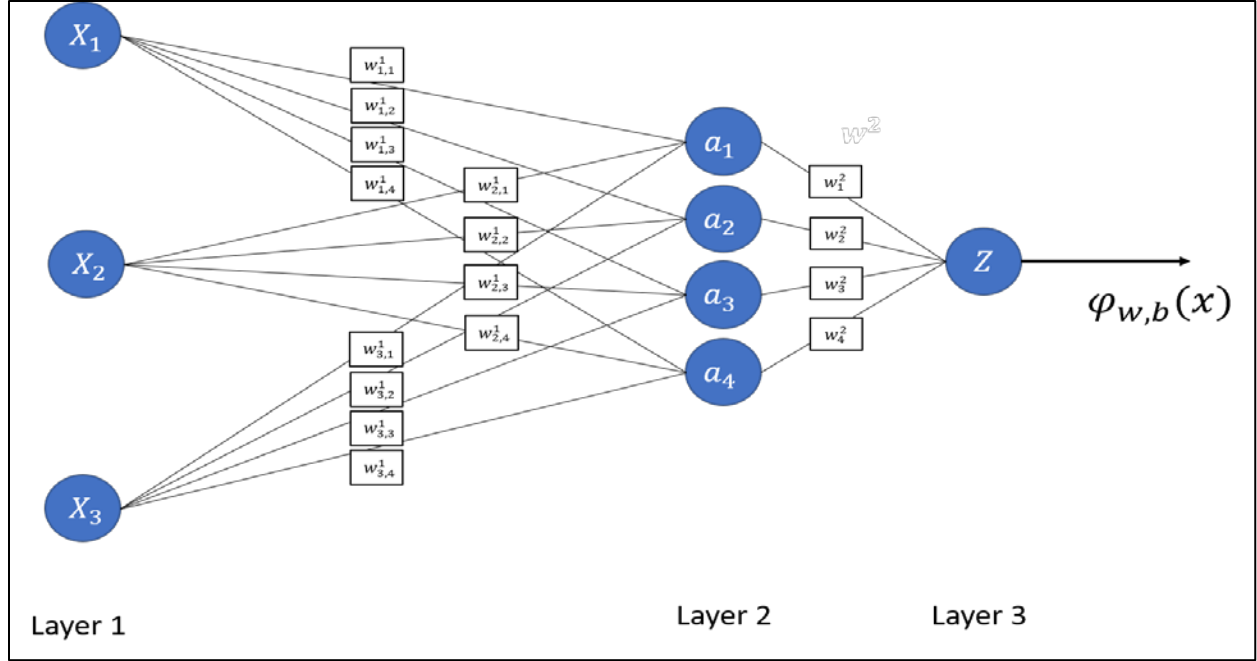


Figure 4.4 Simple neural network with one hidden layer

From equation 4.2, the output of layer 2 calculated as

$$A = W^1 * X + B \quad (4.3)$$

Where A is the output of a hidden layer with a dimension of 4 x 1, W^1 is the 4 x 3 weight matrix between layer one and layer 2, X is 3 x 1 input matrix, and B is the 4 x 1 bias matrix. The output of layer three is calculated by

$$Output = \varphi (W^2 * A + B^2) \quad (4.4)$$

where W^2 is the 1 x 4 weight matrix between layer 2 and layer 3, φ is the activation function, and B^2 is the scalar bias of the last layer threshold.

The dimension of the weight matrix of each layer is achieved by a number of the second layers' node, time to the number of the first layer. For instance, if the network has “n” nodes, in layer j and “m” nodes in layer j+1 the W^j has $m \times n$, which is, in our case, is 4×3 for W^1 . Components of W^1 and W^2 are

$$W^1 = \begin{bmatrix} w_{10}^1 & w_{11}^1 & w_{12}^1 & w_{13}^1 \\ w_{20}^1 & w_{21}^1 & w_{22}^1 & w_{23}^1 \\ w_{30}^1 & w_{31}^1 & w_{32}^1 & w_{33}^1 \end{bmatrix} \quad W^2 = [w_0^2 \quad w_1^2 \quad w_2^2 \quad w_3^2] \quad (4.5)$$

Therefore, equation 4.3 can be expanded as

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} w_{11}^1 & w_{21}^1 & w_{31}^1 \\ w_{12}^1 & w_{22}^1 & w_{32}^1 \\ w_{13}^1 & w_{23}^1 & w_{33}^1 \\ w_{14}^1 & w_{24}^1 & w_{34}^1 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \quad (4.6)$$

$$a_1 = (w_{11}^1 x_1 + w_{21}^1 x_2 + w_{31}^1 x_3) + b_1$$

$$a_2 = (w_{12}^1 x_1 + w_{22}^1 x_2 + w_{32}^1 x_3) + b_2$$

$$a_3 = (w_{13}^1 x_1 + w_{23}^1 x_2 + w_{33}^1 x_3) + b_3$$

$$a_4 = (w_{14}^1 x_1 + w_{24}^1 x_2 + w_{34}^1 x_3) + b_4$$

(4.7)

Substituting the result of equation 4.6 to equation 4.4 follows

$$Output = \varphi (W^2 = [w_1^2 \quad w_2^2 \quad w_3^2 \quad w_4^2] * A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} + B^2) \quad (4.8)$$

$$Output = \varphi(w_1^2 a_1 + w_2^2 a_2 + w_3^2 a_3 + w_4^2 a_4 + B^2) \quad (4.9)$$

For avoiding the complexity of equation 4.9, a_1, a_2, a_3 didn't substitute with the values from equation 4.7. From equation 4.2, the network output can be described as

$$Output = \begin{cases} 0, & w_1^2 a_1 + w_2^2 a_2 + w_3^2 a_3 + w_4^2 a_4 + B^2 < 0 \\ 1, & w_1^2 a_1 + w_2^2 a_2 + w_3^2 a_3 + w_4^2 a_4 + B^2 \geq 0 \end{cases} \quad (4.10)$$

As described earlier, for training the network, the output of equation 4.10 compares with the actual value, and if the mismatch happened, all the weights and biases change in order to get the correct output. The activation function in a neural network identifies if the output of the weighted sum value is above the defined threshold or not.

The sigmoid function is one of the most popular activation functions that use in many applications, as shown in equation 4.10.

$$f(x) = \frac{1}{1 + e^{-x}} = \frac{e^x}{1 + e^x} \quad (4.10)$$

The sigmoid function has a lot of properties that make it perfect for many applications [22].

- Non-linear function
- Output values range between (0,1) make it perfect for probabilistic problems
- Has a vast input range can be any large number or any small negative number

Learning

The Artificial Neural Network learns by training. During training, the training algorithm adjusts the weights to get desired output. The learning in general is classified into two major categories:

1. Supervised Learning
2. Unsupervised Learning

In supervised learning, the dataset used for training has input data and expected target output data. The weights are modified to minimize the error. The training dataset is usually formulated with different ranges to make the model learn better. When one adopts supervised training, the initial model is trained using a training algorithm. After satisfactory training, the weights are fixed. A validation dataset is used to validate the model performance. In unsupervised training, only input

data is provided to the network during training. Using the information provided, ANN identifies the underlying patterns to group them to different clusters [20].

To achieve different learning goals, many learning algorithms were developed to train Artificial Neural Networks. Bayesian regularization and Levenberg-Marquardt are two popular learning schemes for supervised training used during ANN model development.

In this research work two software was used for developing the Artificial Neural Network models. The first one was the Neurosort ANN program that was developed at the University of Kentucky Environmental Research and Training Laboratory (ERTL) lab by Prof. Gail Brion and Prof. Srin Lingireddy with the help of USEPA Star grant. Two postdoctoral researchers Dr. T. R. Neelakandan and Dr. C.V.Chandramouli were involved in developing the software. This software uses a back-propagation algorithm and helps the users to develop neural network models. The second software was Matlab. The neural network toolbox of the Matlab 2019 package was utilized for this research analysis too.

Table 4.1 Training algorithm tried during ANN model development

Training function	Brief explanation
<i>trainlm</i>	Levenberg-Marquardt: It is often the fastest backpropagation algorithm in the Matlab toolbox and usually a first choice of supervised algorithm.
<i>trainbr</i>	Bayesian Regularization: It is another backpropagation algorithm in the Matlab toolbox. This function minimizes a combination of squared error and weights, and then determines the correct combination. the correct combination that will lead to a good generalization for the network.

4.2 Different trials of the model development

4.2.1 Initial model with raw data prediction (M1):

The first model had none of the inputs pre-processed, and all the input (or features) and output (or target) variables were the actual values obtained from laboratory measurements or from designated water quality monitoring websites. In this practice, 23 inputs were applied to the model. Some of them come directly from the field observations and the rest is collected from the Gary airport meteorological website and USGS flow observation website.

Table 4.2 Initial model inputs and their sources

Source of model's inputs	Model's Inputs
Field Observations	Temperature, Total Dissolved Solids (TDS), pH, Turbidity, Electrical Conductivity (EC), Total Suspended Solids (TSS), Color, Odor, Algae, Birds, Trash, People and Pets
Gary airport meteorological website	Mean Hourly Temperature, Relative Humidity, Mean Wind Speed, Mean Wind Direction and Precipitation
USGS flow observation website	Hart Ditch at Munster, Little Calumet at Porter, Grand Calumet at Gary, Portage Burns Waterway and Indiana Harbor Canal at East Chicago

The data set was randomized by rows to negate any effect the dataset's time-series nature might have on the model, then subdivided into three groups: 85% training and 15% for validation and testing.

A three-layer feed-forward network was utilized for all the modeling works. The number of neurons in the hidden layer was chosen depending on the performance of the network. The

activation function of hidden layer neurons was used as tanh-sigmoidal transfer function. A linear transfer function was used in the output layer. Levenberg- Marquardt training algorithm was taken for training which is very widely used in the recent past.

This algorithm typically requires more memory but less training time. Training is terminated using optimal training termination by monitoring the testing dataset. However, after several iterations with different feature combinations and hidden layer neurons, the best results obtained were not very promising. The prediction resulted in 24% R^2 value is shown in figure 4.5.

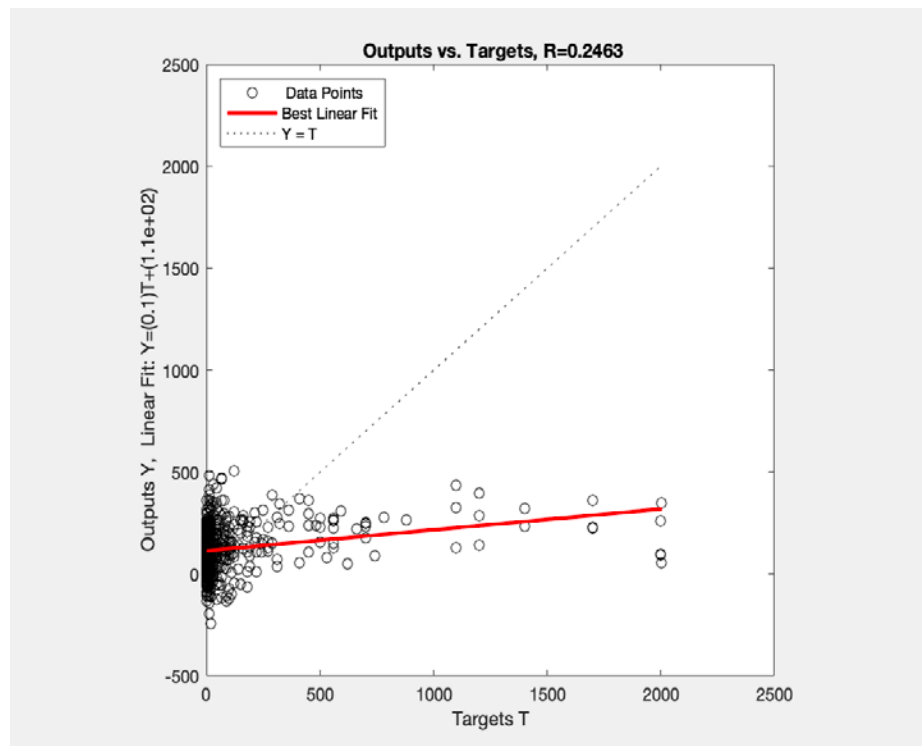


Figure 4.5 Correlation between outputs and targets in the initial model with raw data prediction

The unsatisfactory results led to an experiment to determine which input mode would yield the best result. The model was prepared in three (3) distinct formats:

- (a) Classified input and classified output
- (b) Raw input and classified output
- (c) Classified input and raw output

More than 250 combinations of models were developed and examined. Of the three groups, the second format with raw inputs and classified output showed the most promise and was adopted for further modeling. This decision was taken after experimenting trial models with other formats.

4.2.2 Model with raw inputs to predict the E. Coli with 4 classes (M2):

In this model (M2), none of the input's variables pre-processed and were the actual values obtained from field observations and different agencies, but the output of the model was divided into 4 E. Coli classes as explained in section 3.9.5. In this approach, like previous model, the network structure was consisting of three layers, one input layer, one hidden layer and an output layer.

Table 4.3 Second model inputs and their sources

Source of model inputs	Model Inputs
Field Observations	Temperature, Total dissolved solids (TDS), pH, Turbidity, Electrical Conductivity (EC), Total Dissolved Solids (TSS) lagged, Color, Odor, Algae, Trash, Birds, People, pets
USGS flow observation website	Indiana Harbor Canal at East Chicago and Portage Burns Waterway

Eighty-five percent of the data set was used for the training, and fifteen percent were used for validation and testing. In this approach, a three-layer network, with tanh-sigmoid transfer function in the hidden layer and a linear transfer function (Purelin) in the output layer was applied. Also, Levenberg-Marquardt algorithm was used for training of the data set. The number of neurons in the hidden layer was selected based on the performance of the network.

Best model results were presented here in this category. Results of this model to predict E. Coli classes was not satisfactory too, because the model was unable to adequately distinguish between

those classes. The hypothesis was the algorithm could be susceptible to some form of bias due to the fact that a lot of the data had values less than 236, hence there were much more dataset in class 1 than 2, 3 or 4 (see Table 4.4). Without enough data for 2, 3 and 4, the neural network did not yield a satisfactory result.

Table 4.4 Breakdown of number of data per class in the E. Coli dataset

Class number	Number of samples with E. Coli counts in a select class
1	476
2	29
3	43
4	17

Table 4.5 Best prediction accuracy with 4 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			Class 1			Class 2			Class 3			Class 4		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Trainin g	302	173	64 %	270	132	67 %	19	4	83 %	10	27	27 %	3	10	23 %
Testing	55	35	61 %	47	27	64 %	5	1	83 %	1	5	17 %	2	2	50 %

Table 4.6 Breakdown of the prediction accuracy with 4 E. Coli classes

TRAINING	Total number of 1s predicted as 2s:	130
	Total number of 1s predicted as 3s:	2
	Total number of 1s predicted as 4s:	0
	Total number of 2s predicted as 1s:	3
	Total number of 2s predicted as 3s:	1
	Total number of 2s predicted as 4s:	0
	Total number of 3s predicted as 1s:	1
	Total number of 3s predicted as 2s:	25
	Total number of 3s predicted as 4s:	1
	Total number of 4s predicted as 1s:	0
	Total number of 4s predicted as 2s:	5
	Total number of 4s predicted as 3s:	5

Table 4.7 Breakdown of the prediction accuracy with 4 E. Coli classes

TESTING	Total number of 1s predicted as 2s:	26
	Total number of 1s predicted as 3s:	1
	Total number of 1s predicted as 4s:	0
	Total number of 2s predicted as 1s:	1
	Total number of 2s predicted as 3s:	0
	Total number of 2s predicted as 4s:	0
	Total number of 3s predicted as 1s:	0
	Total number of 3s predicted as 2s:	5
	Total number of 3s predicted as 4s:	0
	Total number of 4s predicted as 1s:	0
	Total number of 4s predicted as 2s:	1
	Total number of 4s predicted as 3s:	1

After this experiment, the results were presented to the IDEM review committee. The committee recommended to try 3 classification output schemes instead of 4 classes. This led to a reclassification of the dataset to 3 classes for the output. In this modeling effort, raw data inputs were used to predict E. Coli 3 classifications (1, 2 and 3), (see section 3.9.5). The results obtained showed more promise than previous model.

4.2.3 Model with raw inputs to predict the E. Coli with 3 classes (M3):

The only difference between model three and two is the organization of output data classes. Since the second model had difficulties to capture highly unsafe values, class 3 and class 2 were merged with the previous classes to get the better result. The network structure was consisting of three layers, one input layer, one hidden layer and an output layer like the earlier model.

Data segregation for training, testing and validation, as well as the network architecture and neuron activation functions were adopted as the same in this modeling work. In this class, the best model presented had 13 inputs and 8 hidden layer neurons. These model trials were started with 23 input combinations and then by eliminating low contributing inputs through relative strength effect [22], 13 best inputs were identified in stages.

Apart from modeling using randomized, to verify the consistency of the model's performance, cross validation of the dataset was carried out. Data were segmented to 4. Each segment was made as the testing data and the other 3 segments were used for training. In this way 4 trials were created to examine the best model. The results were compared afterwards, and it was found that all correctly predicted variables were correctly predicted no matter which subset of the data set was used for testing and validation. The same applies to all incorrectly predicted variables. This shows that the model's ability to correctly predict the observed class.

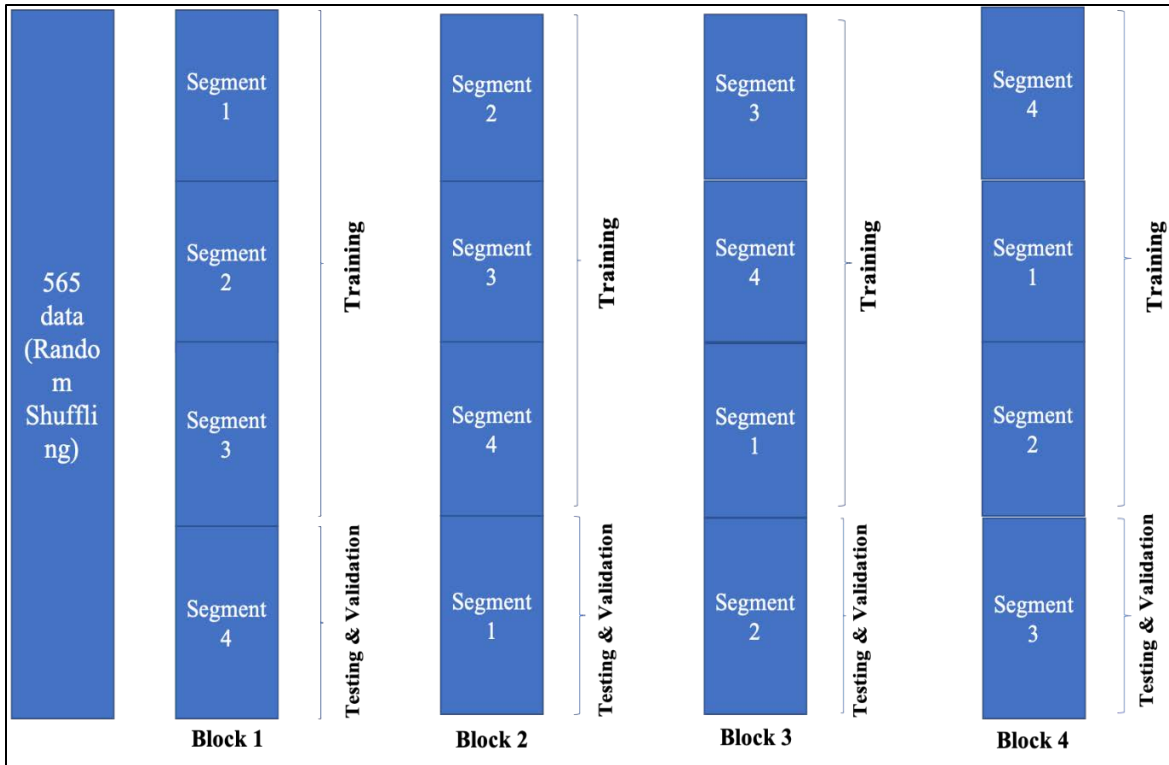


Figure 4.6 Cross-validation efforts

During cross-validation analysis too, consistently similar results were achieved. For the three-class model, Bayesian Regularization Neural Network (BRNN) training algorithm was used for training of data set because it performed superior. The number of neurons in the hidden layer were selected based on the performance of the network. This model showed improved performance.

Table 4.8 Third model inputs and its sources

Source of model inputs	Model Inputs
Field Observations	Temperature, Total Dissolved Solids (TDS), pH lagged, pH, Turbidity lagged, Turbidity, Electrical Conductivity (EC), Total Suspended Solids (TSS) lagged, Color, Algae, Trash
USGS flow observation website	Indiana Harbor Canal at East Chicago and Portage Burns Waterway

Bayesian regularization neural network algorithm:

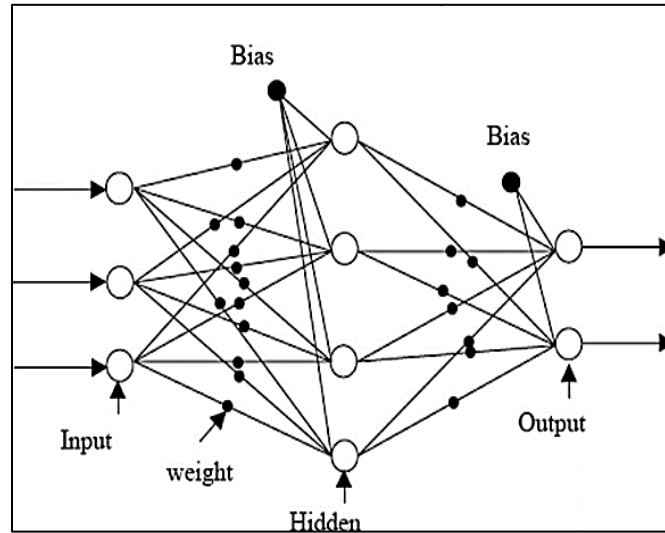


Figure 4.7 Feed forward neural network schematic

It is a feed forward ANN model which uses Bayesian Regularization algorithm for training. It is a supervised learning algorithm.

There are three main layers to the model: input, hidden and output. The input layer contains the input parameters and they are all connected to the hidden layer via a link quantified as weights, like the relative strength of the contribution of any particular input parameter to any particular hidden neuron at that point in the network. Taken together, these dependencies and inter-dependencies make up the trained network.

During training, these weights are adjusted by the algorithm in a bid to reduce to a minimum the chosen error function (usually the mean squared error) so that the network can be reliably used to predict unknown samples. The advantage of this algorithm is that it is much less susceptible to overfitting/overtraining – a challenge that leads to trivial data being given undue importance in a model, leading to a generalization that does not truly reflect the network.

Another advantage is that Bayesian regularization assigns probability values to weights as it learns the during training.

Normally, a validation set is provided for models so as to prevent overfitting. In Bayesian regularized networks, the probability values operate to detect and penalize trivial weights, so they are eventually driven to zero. As such, they can no longer be a factor that the system will try to fit to, which prevents overfitting. Eventually, the network only evaluates and trains only the effective number of parameters, which ultimately converges to a constant.

This frees up more data for training since there is no need to extract separate data for validation out of the training data. The validation process becomes an integral behavior of the network during training.

The disadvantage with using this algorithm is its speed of execution. It is slower than Levenberg-Marquardt and takes more computing memory to perform the same neural training tasks. However, where accuracy is desired and preferred to speed, and when the dataset is a relatively small and noisy one, then Bayesian Regularization is the best algorithm to use, as shown by the result obtained for the best model.

To verify the consistency of the model's performance, cross validation of the dataset was carried out. Data were segmented to 4. Each segment was made as the testing data and the other 3 segments were used for training. In this way 4 trials were created to examine the best model (segments 1,2,3 for training, 4 for testing and validation in trial 1, segment 2,3,4 for training and segment 1 for testing and validation in trial 2 etc).

The results were compared afterwards, and it was found that all correctly predicted variables were correctly predicted no matter which subset of the data set was used for testing and validation. The same applies to all incorrectly predicted variables. This shows that the model's ability to correctly predict the observed class and generalize it well.

Tables 4.9, 4.10, 4.11, and 4.12 illustrate the prediction accuracy with 3 E. Coli classes after cross-validation was applied to the dataset of the model. Among these, Table 4.12 demonstrates the best prediction accuracy of the three-class model.

Table 4.9 Prediction accuracy (trial 1) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			Class 1			Class 2			Class 3		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	417	58	88%	401	31	93%	14	18	44%	2	9	18%
Testing	75	15	83%	69	4	95%	5	6	45%	1	5	17%

Table 4.10 Prediction accuracy (trial 3) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			Class 1			Class 2			Class 3		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	399	76	84%	378	54	88%	17	13	57%	4	9	31%
Testing	74	16	82%	65	8	89%	9	4	69%	0	4	0%

Table 4.11 Prediction accuracy (trial 4) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			Class 1			Class 2			Class 3		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	422	53	89%	420	12	97%	2	28	7%	0	13	0%
Testing	66	24	73%	66	7	90%	0	13	0%	0	4	0%

Table 4.12 Prediction accuracy (trial 2) with 3 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			Class 1			Class 2			Class 3		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	422	53	89%	399	26	94%	21	16	57%	2	11	15%
Testing	79	11	88%	74	6	93%	3	3	50%	2	2	50%

Table 4.13 Breakdown of the prediction accuracy of the best model with 3 E. Coli classes

TRAINING	Total number of 1s predicted as 2s:	26
	Total number of 1s predicted as 3s:	0
	Total number of 2s predicted as 1s:	16
	Total number of 2s predicted as 3s:	0
	Total number of 3s predicted as 1s:	4
	Total number of 3s predicted as 2s:	7
TESTING		
	Total number of 1s predicted as 2s:	6
	Total number of 1s predicted as 3s:	0
	Total number of 2s predicted as 1s:	3
	Total number of 2s predicted as 3s:	0
	Total number of 3s predicted as 1s:	0
	Total number of 3s predicted as 2s:	2

Upon closer analysis, it was discovered that a lot of the class 3s were predicted as class 2s in this model validation (Table 4.13). This discovery revealed that the neural network understood that those were not safe values. However, from our earlier hypothesis, it seemed the neural network was still struggling to adequately learn the model given the relative paucity of data on those higher classes, compared to class 1. For example, in this data set, there are 505 class 1 data, 43 class 2 data and 17 class 3 data. It was hypothesized to try a two-classification scheme of E. Coli output in the next level. Several training algorithms were tried in the modeling effort.

4.2.4 Model with raw inputs to predict the E. Coli with 2 classes (M4):

The previously described neural network design procedure was applied to develop the E. Coli 2 class model prediction. In this approach, 13 inputs variables (Temperature, Total Dissolved Solids (TDS), pH lagged, pH, Turbidity lagged, Turbidity, Electrical Conductivity (EC), Total Suspended

Solids (TSS) lagged, Color, Algae, Trash, Indiana Harbor Canal at East Chicago (HC), and Portage Burns Waterway (BD)) were used to predict the output variable.

The input-output data were grouped in fourteen variables (thirteen inputs and one output) for this approach. Some of the data variables were pre-processed to minimize the difference between the predicted model's output and the actual value of E. Coli. In this model, besides pH, TSS and turbidity that lagged by one day, the Indiana harbor canal and Portage-Burns waterway flow lagged by 10 hours. Also, the output classified into two classes. It was essential for increasing the efficiency of network training. All of the data set variables were subdivided into three groups: training, validation, and testing. Eighty-five percent of the data were used for the training and the remaining fifteen percent was used for the validation and testing.

This network structure was selected after different trial combinations. Like previous models, the three-layers network was used (one input layer, one hidden layer, and an output layer), to keep the model as simple as possible. The number of neurons in the hidden layer were selected after testing the performance of the network at different combinations. It was noticed that 13 neurons are the best number of neurons, in the hidden layer, which converged to a final solution. Activation functions for neurons, were taken to be tanh-sigmoidal and linear respectively for hidden and output layers. This is a good choice for functional approximation neural network [23].

The algorithm chosen was Bayesian Regularization Neural network training algorithm due to its ability to handle small, noisy datasets.

The tables below show the prediction accuracy with 2 E. Coli classes. Table 4.15 illustrates the best prediction accuracy in comparison to other ones.

Table 4.16 Prediction accuracy (trial 3) with 2 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			1's			2's		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	390	85	82%	351	81	81%	39	4	91%
Testing	81	9	90%	66	7	90%	15	2	88%

Table 4.17 Prediction accuracy (trial 4) with 2 E. Coli classes (correct predictions in blue and incorrect predictions in red)

	ALL			1's			2's		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	413	62	87%	375	45	89%	38	17	69%
Testing	76	14	84%	72	13	85%	4	1	80%

5. RESULTS AND DISCUSSION

This chapter presents the results generated by the ANN nowcasting models for E. Coli concentrations in recreational water at the six Lake Michigan beaches. The best result generated from ANN models were introduced.

Different types of input groups, transfer function, hidden neuron, and training algorithm were examined during model development. As it might be recalled, while running the ANN models, the input data were divided into three sets as training, validation, and testing. As a result, for this particular research work, training data set were used for screening and prediction performance, the threshold for screening out best model among all developed models was selected based on Mean Square Error (MSE), Regression R values (R^2). Error value reveals underlying relationships between the output data (actual values) and target data, and Regression R values reveals the correlation between actual and predicted outputs.

The performance statistics presented in Table 5.1 provide an overall assessment of the best ANN models investigated for the effect of different training algorithms, training functions and the number of hidden neurons.

Table 5.1 Comparison of the best result for different models' classification

No	Model	Transfer function (for hidden layer neuron, output neuron)	Training algorithm	No. of neurons	Input normalization	Cross-validation
1	M1	tansig purelin	Trainlm	12	Yes	Yes
2	M2	tansig purelin	Trainlm	12	Yes	Yes
3	M3	tansig purelin	Trainbr	8	Yes	Yes
4	M4	tansig purelin	Trainbr	13	Yes	Yes

Input Normalization: Normalization of the data set usually helps the neural network's training because it helps obtain a mean close to 0. According to [24], [25], normalization of input variables plays a vital role during ANN model development as all inputs have different units. So, during this research work, all models developed by applying a normalization function. In this approach, the mapminmax function was used for the normalization of input datasets. This function scales inputs so that they fall in the range $[-1,1]$. In this way, better predictions can be made; hence all input data are linearly normalized into a particular range before applying transfer functions.

Training Algorithm: The algorithm used in the Matlab were discussed in previous chapter (Table 4.8). Trainlm is the most popular Levenberg-Marquardt training algorithm for all the feedforward ANN models. But in order to verify and to understand the difference in terms of performance, some other training algorithms were tried keeping other criteria's same. From the results, it was evident that trainbr function which is BRNN training algorithm has shown better results than trainlm for all different models in terms of MSE and R^2 value. Trainlm algorithm is generally the fastest training function among others and is the default training function for feedforward networks. Trainbr takes more time to converge but for small or noisy datasets in can provide better generalization. For that reason, trainbr was used for the final model development.

Transfer Function: The default transfer function of Neural Network Toolbox for Levenberg-Marquardt algorithm (trainlm) and Bayesian Regularization (trainbr) are a Hyperbolic Tangent (tansig) in the input to hidden layer and a Linear transfer function for the hidden to the output layer (purelin). In comparison to the Levenberg-Marquardt algorithm Bayesian regularization typically requires more time but can result in good generalization for small or noisy datasets.

In the Bayesian regularization algorithm, training stops according to adaptive weight minimization. To understand each training function's influence, different training functions were employed during model development from input to hidden layer. The performance of those models (Table 4.15) indicated that Bayesian regularization algorithm produced better results for E. Coli prediction.

Numbers of Neurons: During ANN model's development different hidden neurons were assessed using trial and error method. For each model, the best results in terms of the MSE and R^2

value of the training and testing data sets was calculated to determine the appropriate number of hidden neurons to provide adequate generalization while avoiding overfitting. Table 5.1 demonstrated the best performance of each model with the best number of neurons.

5.1 Best Model with 2 E. Coli Classes

Datasets with 2 E. Coli classes showed much more improvement over the previous datasets that had more classes. After several iterations with number of hidden layer neurons, feature selections and choice of algorithm for the neural network, it was found that a 13-input dataset comprising of Temperature, TDS, pH, Turbidity Electrical Conductivity, TSS (lagged by a day), Water Color, Algae Bloom and Presence of Trash on the beach. The dataset also included one-day-lagged data for pH and turbidity as preliminary analysis showed that pH and turbidity had a lot of influence on the model prediction.

The logic behind introducing the one-day-lagged values as part of the inputs is to provide the model with a little bit of time series information on any data that it considers an important factor in its learning process. The discharge data for Indiana Harbor Canal and Portage-Burns Waterway, two close tributaries with respect to the beaches under study, were also part of the inputs, bringing the total number of inputs to 13. Harbor canal and Portage Burns Waterway flow were average of 10 hours of discharge from the time of sample collection.

The number of hidden layer neurons that yielded the best result for the dataset described in the preceding paragraph was 13 neurons. The algorithm chosen was Bayesian Regularization.

A careful comparison between Table 5.2, 5.3, and 5.4 shows that with 2 E. Coli classes, there was even less mispredictions of class 2 as class 1 or vice versa. Table 5.3 and 5.4 breaks down the accuracy of the best model. This was the best performance out of all the other models obtained in this study. As a part of this research, a user friendly excel sheet was created based on the best model. This tool will be helpful to the beach managers to find the E. Coli class instantaneously when they input the 13 data to the sheet. All these data used here are available from yesterday's observations. Few observations are needed from the same day, but they can be measured instantaneously. So, this tool can be used on a real time basis.

Table 5.2 Best prediction accuracy with 2 E. Coli classes

	ALL			1's			2's		
	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct	Correctly Predicted	Wrongly Predicted	Percent predicted correct
Training	442	33	93%	400	25	94%	42	8	84%
Testing	82	8	91%	74	6	93%	8	2	80%

Table 5.3 Breakdown of the prediction accuracy with 2 E. Coli classes

TRAINING	Total number of 1s predicted as 2s:	25
	Total number of 1s predicted as 3s:	0
	Total number of 2s predicted as 1s:	8
	Total number of 2s predicted as 3s:	0
	Total number of 3s predicted as 1s:	0
	Total number of 3s predicted as 2s:	0

Table 5.4 Breakdown of the prediction accuracy with 2 E. Coli classes

TESTING	Total number of 1s predicted as 2s:	6
	Total number of 1s predicted as 3s:	0
	Total number of 2s predicted as 1s:	2
	Total number of 2s predicted as 3s:	0
	Total number of 3s predicted as 1s:	0
	Total number of 3s predicted as 2s:	0

The 2-way classification model proposed as the best model would be used in the Lake Michigan beaches to predict E. Coli classes in real-time. For this purpose, every day, a field technician would be taking a water sample and test the level of the water temperature, Total dissolved solids (TDS), Total suspended solids (TSS), pH, Turbidity, and Electrical conductivity (EC). Water discoloration, algae bloom, and trash observations should be also made simultaneously. After gathering all the other variables needed from web, together with the observed values and lagged pH, turbidity and TSS values, on a real time basis, one can find the E.Coli class (Table 5.5).

Table 5.5 The way inputs obtained apply to the ANN model

Model Inputs	Temperature	pH	pH lagged	Turbidity	Turbidity lagged	TDS	TSS	EC	Color	Algae	Trash	Indiana Harbor Canal	Portage Burns Waterway
Instantaneous observation	✓	✓		✓		✓		✓	✓	✓	✓		
one day lag			✓		✓		✓						
Average of 10 hours lag												✓	✓

6. CONCLUSION

An ANN model has been developed that predicts the E. Coli safe/unsafe concentration limit with an average prediction accuracy of around 87%. The results of trained ANN model have been made into an easy to use excel tool that will assist beach managers to make future predictions based on the trained model. Even though the prediction accuracy is not 100%, 87% is a very good start on a quest to be able to obtain absolute real time knowledge on E. Coli classes in recreational waters. The final model uses 13 inputs, namely: Temperature (°C), TDS (mg/L), pH-lagged, pH, Turbidity-lagged (NTU), Turbidity (NTU), Electrical Conductivity (µs/cm), TSS (mg/L) lagged, Color, Algae, Trash, Indiana Harbor Canal discharge rate and Portage-Burns Waterway discharge rate. As a part of this research, a user friendly excel sheet was created based on the best model. This tool will be helpful to the beach managers to find the E.Coli class instantaneously when they input the 13 input data to the sheet. All these data used here can be observed instantaneously and few are from yesterday's observations. So, this tool can be used on a real-time basis.

To improve the model, the researcher team recommends a second data gathering phase over another beachgoers season and fine tune the existing model. With more data available to train the model, the model becomes more robust and can handle wider ranges of data fluctuations.

APPENDIX A: MICROBAC INTERNAL CHAIN OF CUSTODY FIELD DATA SHEET FOR MICROBAC-COLLECTED SAMPLES.

PURDUE NORTH WEST Lake Michigan Beach XXX
E.coli Field Sheet

Beach

Date of Collection _____
Air Temp °F _____

By _____

Station: Some Fine BEACH Sample# - 01 Time Collected: _____

Wind Direction (from) _____

Water Surface ☐ Calm ☐ Breakers

Odor ☐ None ☐ Sewage ☐ Oily ☐ Rotten Eggs ☐ Fishy ☐ Other _____

☐ Algal Bloom ☐ Litter Trash Dead Fish # _____ Dogs # _____

Birds# (circle) None 1-3 4-10 10-31 32-100 >100

Field pH: _____ Field Temperature (C): _____

Station: Some Better BEACH Sample# - 02 Time Collected: _____

Wind Direction (from) _____

Water Surface ☐ Calm ☐ Breakers

Odor ☐ None ☐ Sewage ☐ Oily ☐ Rotten Eggs ☐ Fishy ☐ Other _____

☐ Algal Bloom ☐ Litter Trash Dead Fish # _____ Dogs # _____

Birds# (circle) None 1-3 4-10 10-31 32-100 >100

Field pH: _____ Field Temperature (C): _____

Weather Conditions Past 24 Hours

☐ Clear Sunny ☐ Stormy ☐ Steady Rain ☐ Cloudy ☐ Hail
☐ Light Showers ☐ Overcast ☐ Partly Cloudy

Relinquished by _____ Date _____ Time _____

Rec'd at Lab by _____ Date _____ Time _____

Sample Temperature Rec'd at LAB °C _____ Trip Blank Lot# _____

Lab Work Order ID: _____

(Left - click to open)

77

REFERENCES

- [1] R. Gosukonda, A. K. Mahapatra, X. Liu, and G. Kannan, “Application of artificial neural network to predict Escherichia coli O157:H7 inactivation on beef surfaces,” *Food Control*, vol. 47, pp. 606–614, 2015, doi: 10.1016/j.foodcont.2014.08.002.
- [2] M. B. Nevers and R. L. Whitman, “Nowcast modeling of Escherichia coli concentrations at multiple urban beaches of southern Lake Michigan,” *Water Res.*, vol. 39, no. 20, pp. 5250–5260, 2005, doi: 10.1016/j.watres.2005.10.012.
- [3] D. Dwivedi, B. P. Mohanty, and B. J. Lesikar, “Estimating Escherichia coli loads in streams based on various physical, chemical, and biological factors,” *Water Resour. Res.*, vol. 49, no. 5, pp. 2896–2906, 2013, doi: 10.1002/wrcr.20265.
- [4] L. M. (Lee) He and Z. L. He, “Water quality prediction of marine recreational beaches receiving watershed baseflow and stormwater runoff in southern California, USA,” *Water Res.*, vol. 42, no. 10–11, pp. 2563–2573, 2008, doi: 10.1016/j.watres.2008.01.002.
- [5] F. Pérez, I. Tryland, M. Mascini, and L. Fiksdal, “Rapid detection of Escherichia coli in water by a culture-based amperometric method,” *Anal. Chim. Acta*, vol. 427, no. 2, pp. 149–154, 2001, doi: 10.1016/S0003-2670(00)00984-3.
- [6] G. A. Olyphant and R. L. Whitman, “Elements of a predictive model for determining beach closures on a real time basis: The case of 63rd Street Beach Chicago,” *Environ. Monit. Assess.*, 2004, doi: 10.1023/B:EMAS.0000038185.79137.b9.
- [7] O. of R. & Development, “2010 U.S. Environmental Protection Agency (EPA) Decontamination Research and Development Conference.”
- [8] R. A. Gonzalez, K. E. Conn, J. R. Crosswell, and R. T. Noble, “Application of empirical predictive modeling using conventional and alternative fecal indicator bacteria in eastern North Carolina waters,” *Water Res.*, vol. 46, no. 18, pp. 5871–5882, 2012, doi: 10.1016/j.watres.2012.07.050.
- [9] G. A. Olyphant, J. Thomas, R. L. Whitman, and D. Harper, “olyphant et al 2003 Characterization and Statistical Modeling of Bacterial,” *Environ. Monit. Assess.*, 2003, doi: 10.1023/A:1021345512203.

- [10] R. Avila, B. Horn, E. Moriarty, R. Hodson, and E. Moltchanova, "Evaluating statistical model performance in water quality prediction," *J. Environ. Manage.*, vol. 206, pp. 910–919, 2018, doi: 10.1016/j.jenvman.2017.11.049.
- [11] "Transactions of the American Society of Civil Engineers, Vol. 175 (2010)."
<https://sp360.asce.org/PersonifyEbusiness/Merchandise/Product-Details/productId/18286>
(accessed May 04, 2020).
- [12] E. Dogan, B. Sengorur, and R. Koklu, "Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique," *J. Environ. Manage.*, vol. 90, no. 2, pp. 1229–1235, 2009, doi: 10.1016/j.jenvman.2008.06.004.
- [13] S. Motamarri and D. L. Boccelli, "Development of a neural-based forecasting tool to classify recreational water quality using fecal indicator organisms," *Water Res.*, vol. 46, no. 14, pp. 4508–4520, 2012, doi: 10.1016/j.watres.2012.05.023.
- [14] C. Yu, V. J. Davidson, and S. X. Yang, "A neural network approach to predict survival/death and growth/no-growth interfaces for Escherichia coli O157:H7," *Food Microbiol.*, vol. 23, no. 6, pp. 552–560, 2006, doi: 10.1016/j.fm.2005.09.008.
- [15] G. Brion *et al.*, "Artificial neural network prediction of viruses in shellfish," *Appl. Environ. Microbiol.*, vol. 71, no. 9, pp. 5244–5253, 2005, doi: 10.1128/AEM.71.9.5244-5253.2005.
- [16] R. M. Garciaa-Gimeno, C. Hervas-Martianez, E. Barco-Alcala, G. Zurera-Cosano, and E. Sanz-Tapia, "An Artificial Neural Network Approach to Escherichia Coli O157:H7 Growth Estimation," *J. Food Sci.*, vol. 68, no. 2, pp. 639–645, 2003, doi: 10.1111/j.1365-2621.2003.tb05723.x.
- [17] H. R. Maier and G. C. Dandy, "Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications," *Environ. Model. Softw.*, vol. 15, no. 1, pp. 101–124, 2000, doi: 10.1016/S1364-8152(99)00007-9.
- [18] S. De Vito, M. Piga, L. Martinotto, and G. Di Francia, "CO, NO₂ and NO_x urban pollution monitoring with on-field calibrated electronic nose by automatic bayesian regularization," *Sensors Actuators, B Chem.*, vol. 143, no. 1, pp. 182–191, 2009, doi: 10.1016/j.snb.2009.08.041.

- [19] “Artificial neural network - Wikipedia.”
https://en.wikipedia.org/wiki/Artificial_neural_network (accessed Apr. 06, 2020).
- [20] D. J. Livingstone, “Artificial neural networks. Methods and applications. Preface.,”
Methods in molecular biology (Clifton, N.J.). 2008.
- [21] “What is Perceptron | Simplilearn.” <https://www.simplilearn.com/what-is-perceptron-tutorial> (accessed Apr. 06, 2020).
- [22] “Everything you need to know about Neural Networks and Backpropagation — Machine Learning Easy and Fun.” <https://towardsdatascience.com/everything-you-need-to-know-about-neural-networks-and-backpropagation-machine-learning-made-easy-e5285bc2be3a> (accessed Apr. 15, 2020).
- [23] F. S. Mjalli, S. Al-Asheh, and H. E. Alfadala, “Use of artificial neural network black-box modeling for the prediction of wastewater treatment plants performance,” *J. Environ. Manage.*, vol. 83, no. 3, pp. 329–338, 2007, doi: 10.1016/j.jenvman.2006.03.004.
- [24] M. Heydari, E. Olyaie, H. Mohebzadeh, and Ö. Kisi, “Development of a neural network technique for prediction of water quality parameters in the Delaware River, Pennsylvania,” *Middle East J. Sci. Res.*, 2013, doi: 10.5829/idosi.mejsr.2013.13.10.1238.
- [25] W. Thoe, S. H. C. Wong, K. W. Choi, and J. H. W. Lee, “Daily prediction of marine beach water quality in Hong Kong,” *J. Hydro-Environment Res.*, 2012, doi: 10.1016/j.jher.2012.05.003.