Physico-Chemical Parameters Based Artificial Intelligence Model for Prediction of Faecal Coliform Bacteria in Ground Water

THESIS

Submitted in partial fulfilment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

by

FARHAN MOHAMMAD KHAN

ID No. 2018PHXF0411P

Under the Supervision of

Prof. Rajiv Gupta

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

2022

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE

PILANI – 333 031 (RAJASTHAN) INDIA

CERTIFICATE

This is to certify that the thesis entitled "**Physico-Chemical Parameters Based Artificial Intelligence Model for Prediction of Faecal Coliform Bacteria in Ground Water**" submitted by **Farhan Mohammad Khan**, ID No. 2018PHXF0411P for award of Ph.D. of the Institute embodies original work done by him under my supervision.

Signature of the Supervisor

Name: Prof. Rajiv Gupta

Designation: Senior Professor

Date: 30/07/2022

Acknowledgement

I express sincere gratitude and indebtedness to my guide, Prof. Rajiv Gupta, Senior Professor, Department of Civil Engineering, BITS Pilani, Pilani Campus. He has been a constant source of inspiration throughout my work. The enthusiasm, moral support and advice that he has given me will stimulate me to be the best in my endeavors. I sincerely acknowledge the help rendered by him at all times.

I express my profound sense of gratitude and indebtedness to the Co-Principal Investigator of BITS-UVA Joint Research Project, Dr. Sheetal Sekhri, Associate Professor, Department of Economics, University of Virginia, USA, for her valuable guidance and suggestions to do my research and the help she provided to me in the formulation of ideas for my thesis. I gratefully acknowledge the University of Virginia, USA, for financial support in the form of Junior Research Fellowship.

I express my deep appreciation to BITS Pilani for providing all the necessary facilities and support to complete the research work. My special thanks to Prof. S. Bhattacharyya, Vice-Chancellor of the University, Prof. S. K. Barai, Director, BITS-Pilani, Pilani Campus for allowing me to pursue my research work successfully. I also express my gratitude to Prof. S. K. Verma, Dean, Academic Research and other faculty members of ARD for providing valuable support throughout the programme. I would like to take this opportunity to express my gratefulness to Prof. Anupam Singhal and Dr. Lavika Goel, who are members of the Doctoral Advisory Committee (DAC), for their kind suggestions, motivation and moral support throughout the program both technically and personally. I would like to thank Dr. Anshuman, Ex-HoD, Civil Engineering and Prof. Anupam Singhal, HoD, Civil Engineering, for their motivation during difficult times.

I am also thankful to all my professors for their support and motivation throughout my studies at BITS-Pilani, Pilani Campus. Thanks are also due to all the non-teaching staff of the Department of Civil Engineering Kamalji, Shivpalji, Sureshji, Shivratanji, Rameshji and Sultanji for their constant co-operation throughout my work. I also thank all my friends and colleagues, especially Dr. Gaurav Kumar, Mr. Abheek Gupta, Ms. Soumya Kar, Mr. Raya Raghavendra Kumar, Mr. Danish Khaki, Mr. Manigandan R, Mr. Dhanesh Kumar and all others whose name are not here and have helped me during my Ph.D. Thanks are also due to all faculty and staff members of BITS-Pilani, Pilani Campus, for helping me out at various times.

I express my hearty gratitude to my beloved parents, my brother, for their care and affection shown towards me to undertake this journey. I owe thanks to my family for their love, encouragement and moral support, without which this work would have been an impossible task. My parents have always been supportive and their encouragement in all my endeavors from the beginning has always been a source of inspiration for me. I am grateful to Almighty God, with whose blessings this thesis has seen the light of the day. May, I would cherish happy moments spent at BITS Pilani campus for my whole life.

Farhan Mohammad Khan

Abstract

Only 3% of the earth's surface is covered by fresh water, with the remaining (97%) made up of saltwater water from the ocean. Groundwater is critical to our environment and economies since it accounts for 30% of fresh water. Groundwater, being a valuable resource, requires regulations to maintain and protect its quality and quantity. *E. coli* bacteria are associated with the coliform group and are a more precise indicator of faecal contamination than other coliform bacteria; its existence indicates the possible presence of harmful disease-causing bacteria. Physicochemical parameters are a significant element that influences bacteriological water quality and, as a result, the characterization of water and groundwater.

In this research study, groundwater analysis of the Rajasthan region has been carried out. The groundwater samples were collected from eight districts of the state of Rajasthan, India, under the BITS-UVA (University of Virginia) groundwater contamination project, containing 1302 water samples. These samples were collected from 348 villages and cities during the years 2019–2021. Eight physico-chemical parameters such as pH, total dissolved solids (TDS, mg/l), oxidation-reduction potential (ORP, mg/l), dissolved oxygen (DO, mg/l), electrical conductivity (EC, s/m), turbidity (NTU), fluoride (mg/l), and nitrate (mg/l) were determined in the laboratory using the titration and spectroscopy method.

Microbiological water quality analysis was performed to identify the bacteria present in water samples. The viable count analysis of the water samples showed *E. coli* bacterial strains with minimum cell counts of $4{\times}10^7$ CFU/100 mL and maximum cell counts of 132×10^7 CFU/100 mL. A total of 99 groundwater samples were found positive for *E. coli*. The prediction of waterborne bacteria is crucial to prevent health risks. Hence, *E. coli* have been chosen for detailed studies as many diseases are associated with their presence. A superposition-based learning algorithm (SLA) was proposed to observe the patterns of ANN-based sensitivity analysis to determine the importance of each water quality parameter resulting in the prediction of *E. coli* in groundwater. Mean Square Error (MSE) and the Coefficient of determination (R^2) were calculated using MATLAB (R2019b) software for model performance evaluation. The highest correlation was observed between *E. coli* and the pH values, whereas the lowest correlation was observed with Dissolved Oxygen.

The detection of *Escherichia coli* bacteria is essential to prevent health diseases. According to the laboratory-based methods, 12– 48 h is required to detect bacteria in water. The drawback of depending on laboratory-based methods for the detection of *E. coli* bacteria can be prone to human errors. Hence, the bacterial detection process must be automated to reduce error. We implement an automated *E. coli* bacteria detection process using a convolutional neural network (CNN) to address this issue. We have also proposed a mobile application to rapidly detect *E. coli* bacteria in water that uses CNN. The developed CNN model achieved an accuracy of 96% and an error (loss) of 0.10, predicting each sample in only 458ms. The performance of the model was validated using the F-score, precision, sensitivity, and accuracy statistical measures, which shows that the model is reliable and effective in detecting *E. coli*.

Manual counting of the viable bacterial colony on agar plates is time-consuming and can be prone to human error. The method requires experts to identify and count colonies on agar plates using a microscope. Hence, the bacterial counting procedure must be automated to decrease error. We automated the process of *E. coli* bacteria identification using a convolutional neural network (CNN). We developed a smartphone application for the rapid detection of *E. coli* bacteria on agar plates using CNN. We also automated the process of bacteria colony counting using a faster region-based convolutional neural network (R-CNN) to overcome manual cell counting process limitations. A graphical user interface (GUI) application was created to rapidly count bacteria colony-forming units on agar plates using faster R-CNN. The developed faster R-CNN model achieved an overall accuracy of 97% and an error (loss) of 0.10. The performance of the CNN and faster R-CNN models were validated various statistical measures. The comparative analysis showed that the faster R-CNN model is reliable and effective in *E. coli* cell counting. The study developed a system for identifying and counting viable cells of *E. coli* bacteria in water that can be used to forecast hotspots of water contamination.

Keywords: *Escherichia coli*, Bacteria, Detection, Viable but Nonculturable, Groundwater, Water quality, Prediction, Physico-chemical parameters, Artificial neural network, Superposition, Sensitivity analysis, Machine learning, Convolutional neural network (CNN), Agar plates, Colony counting, Faster R-CNN.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI	Artificial Intelligence
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Network
AP	All parameters
BITS-UVA	Birla Institute of Technology and Science- University of Virginia
BOD	Biochemical Oxygen Demand
BR	Bayesian regularization
BW	Black and white
CFU	Colony-forming units
CNN	Convolutional Neural Network
COD	Chemical Oxygen Demand
DBP	Disinfection by-products
DNN	Deep Neural Network
DO	Dissolved Oxygen
EC	Electrical Conductivity
EMB	Eosin methylene blue
Escherichia coli	*E. coli*
FA	Firefly Algorithm
FC	Fully connected
FCB	Faecal coliform bacteria
FIS	Fuzzy Inference System
FN	False negative
FP	False positive
GA	Genetic Algorithm
GUI	Graphical user interface
HM	Hybrid model

in	input
IOT	Internet of Things
LM	Levenberg Marquardt
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
Mf	Membership function
MF	Membrane Filter
MLP	Multi-Layer Perceptron
MPN	Most Probable Number
MSE	Mean Square Error
NTU	Nephelometric Turbidity Units
ORP	Oxidation-Reduction Potential
PA	Present/Absent
PC	Principal Components
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
R	Regression
R-CNN	Region-based CNN
ReLU	Rectified Linear Unit
RGB	Red Green Blue
RMSE	Root Mean Square Error
ROI	Region of interest
RPN	Region proposal network
RSIMCA	Robust Soft Independent Modelling of Class Analogies
SGDM	Stochastic Gradient Descent with Momentum
SI	Sensitivity Index
SLA	Superposition-based Learning Algorithm
SRB	Sulphate Reducing Bacteria
STP	Sewage Treatment Plant
SVM	Support Vector Machines

T	Temperature
TC	Total coliform
TDS	Total Dissolved Solids
TN	True negative
TP	True positive
TSS	Total Suspended Solids
USB	Universal Serial Bus
UTI	Urinary Tract Infection
VBNC	Viable but Non-Cultural
VGG	Visual Geometry Group
WQI	Water Quality Indexing
WSN	Wireless Sensor Networks
.tflite	TensorFlow Lite

1. Introduction

Water is more essential to human life. Adequate, accessible, and secure supply for consumers is needed. Water helps to maintain the moisture in the internal organs of the body (Gerald, 2011). It also protects the volume and uniformity of blood and lymph fluids (Dooge, 2001), controls body temperature, and eliminates toxins from the body through urine, sweat, and respiration (Molden, 2013), which is essential for maintaining skin functions (Burton et al., 1987). Water pollution is one of the most critical challenges for sustainable development. Water pollution can lead to kidney failure and can cause death (David et al., 2011). Improving access to clean drinking water would carry essential health benefits. Water quality measurement is an important stepping stone towards finding a solution to this problem. In the present situation, people are struggling to obtain access to water. Generally, the most harmful infectious diseases are caused due to bacterial contamination in water.

This chapter deals with various aspects of water, including its quality parameters, sources, classification and health effects of bacterial contamination on human health. The importance and application of different modeling techniques employed in water quality monitoring are also discussed in detail. The significance of the prediction of bacterial contamination in groundwater has been discussed and problems associated with physico-chemical water quality parameters are also provided here.

1.1 Background

Water is a transparent fluid which we get from rain, also found in seas, streams and lakes across the world and it is a significant component of organism's fluids having the chemical formula H_2O. A water molecule is a chemical substance that comprises covalent bonding between one oxygen and two hydrogen atoms. Water is a liquid at standard pressure and temperature. On Earth, it often coexists with its solid-state (ice) and gaseous state (steam). It can also be found as snow, fog, mist, and cloud (Gleick, 1993; Clark et al., 2001). Water comprises 71% of the Earth's surface and is necessary for all known kinds of life (Gleick, 1993). Seas and oceans contain 96.5% of the world's water. Groundwater contains 1.7%. The glaciers and ice caps of Antarctica, Greenland contain the remaining 1.7%. A minor fraction in other significant bodies of water, and 0.001% in the atmosphere as clouds, precipitation and vapors. Freshwater constitutes only 2.5% of the world's water, and 98.8% of this freshwater is in the form of icecaps and groundwater. Atmosphere, rivers and lakes contain less than 0.3% of total freshwater. An even smaller quantity (0.003%) is found in biological organisms and manufactured goods (Gleick, 1993; WHO, 1996; Clark et al., 2001). Figure 1.1 is a schematic representation of the global distribution of water (Clark et al., 2001).

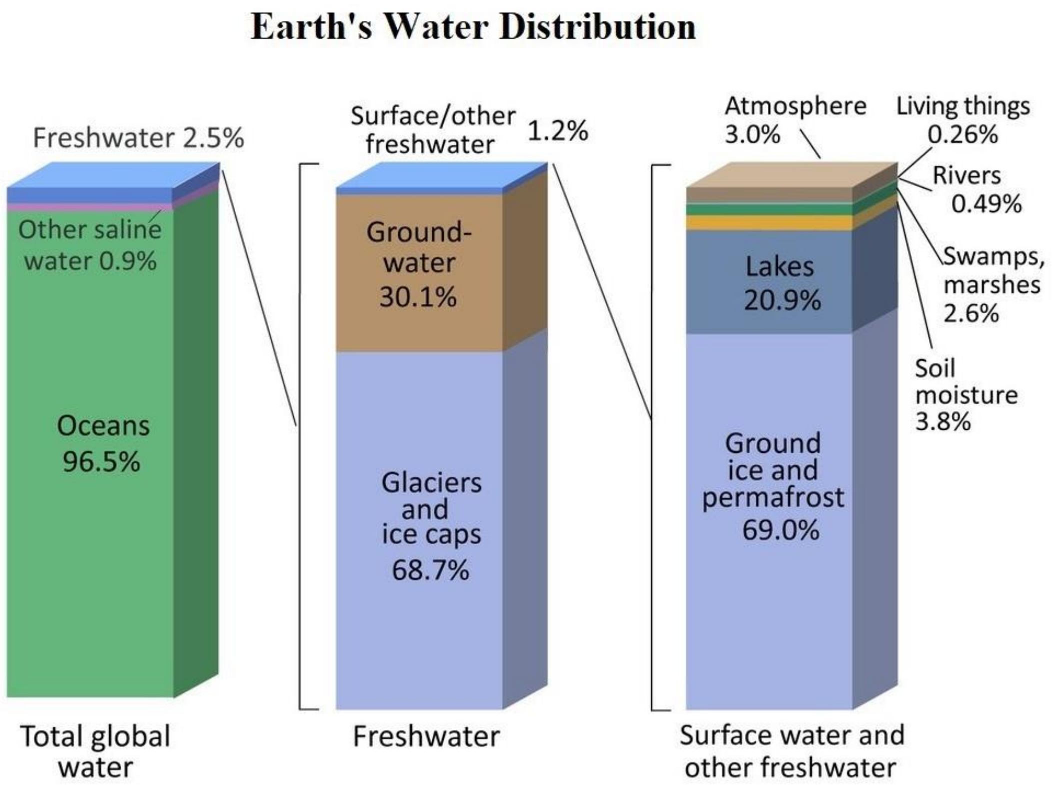


Figure 1.1: A graphical distribution of the water on Earth (Clark et al., 2001).

Water is more essential to human life. Adequate, accessible, and secure supply for consumers is needed. Improving access to clean drinking water would carry essential health benefits. Efforts should be made to achieve the cleanest possible water quality for drinking (WHO, 2011). In the present situation, people are struggling to obtain access to water. The infectious diseases associated with water consumption are usually polluted with human or animal faeces. Some of the significant health disorders are caused by micro-organisms such as bacteria, pathogens, viruses etc. because they can live, reproduce and disperse in water systems (Payus et al., 2018).

The hydrological cycle includes groundwater as a significant constituent. Groundwater is a significant source of water and an essential component of our water supply. Groundwater normally flows inside aquifers and remains below the water table. Water from extremely deep sources has been maintained for a long time, and these aquifers were produced a few million years ago, and this water has been held in these aquifers for a few million to hundreds of years. Groundwater enters rivers at periods of low river flow, maintaining river flow. Occasionally, groundwater comes back to the surface in the form of springs or, more spectacularly, hot springs and geysers (Ben et al., 2011). As a result, deep aquifer water is sometimes referred to as fossil water (US EPA, 2015).

Water resources have significant environmental, social, and economic significance, and if water quality deteriorates, this resource will lose economic value. The physical, chemical and biological parameters of water are often used to determine its quality. Measurements of these water quality parameters can be used to determine and monitor changes in water quality. Conventional water quality measurement techniques include on-site sampling and subsequent laboratory-based tests. Water quality is defined as the composition of elements dissolved in water throughout the operation of natural processes and human activities. Chemical parameters are often used in determining water quality (UNEP, 2015). In general, water quality parameters can be classified as follows (IAEA, 2007):

- Physical Parameters: Turbidity, Temperature, Color, Taste and Odour, Total Dissolved Solids (TDS), Total Suspended Solids (TSS), Electrical Conductivity (EC).
- Chemical Parameters: pH, Chloride, Fluoride, Iron, Hardness, Dissolved Oxygen (DO), Biochemical Oxygen Demand (BOD), Chemical Oxygen Demand (COD), Nitrate.
- Biological Parameters: Bacteria, Indicator organisms, Protozoa, Viruses, Algae.

Groundwater is constantly prone to pollution as a result of the water cycle, and data on groundwater quality is scarce. Water quality is determined by several parameters rather than a single parameter (Eddy-Miller et al., 2007). Water quality is inextricably connected to the surrounding environment and land usage (Nicolau, 2001). Dams and weirs can also alter natural stream flows, which can have an impact on water quality. Weather can also significantly influence water quality, especially in a dry region where drought conditions are usually present (Nicolau, 2001). Water quality frequently deteriorates when rivers pass through areas with extensive land and water usage, and pollution from intensive agriculture, major towns, industry, and recreation areas rises.

According to Korngold et al. (1966), Mohamed et al. (2011), NFIS (2012), water quality can increase downstream, behind dams and weirs, at locations where tributaries or better quality groundwater meet the main stream and wetlands. Approximately 1.7 billion children under the age of five in developing countries died from diarrhea, primarily from consuming contaminated water reported by the World Health Organization in 2011. 525,000 children die worldwide annually in 2018 due to poor water quality, sanitation, and hygiene, mainly due to infectious diarrhoea. Worldwide, 1.9 billion people use contaminated water (WHO, 2017). Around 37.7 million people in India suffer from waterborne diseases every year and 1.5 million children have died from diarrhea. Faecal matter is the primary source of water-borne bacteria causing the disease.

As a result, in many regions of the world, water quality regulation is a high-priority policy agenda (WHO, 2011). Many factors impact water quality and have intricate nonlinear interactions with the parameters in water quality analysis. Water quality parameters must be assessed in order to establish optimal water resource planning and management (Liu et al., 2009; Najah et al., 2014). Many water quality models have been created and widely used to tackle water quality problems due to increased computer capabilities (Dogan et al., 2009; He et al., 2011; Lindim et al., 2011). Although deterministic models have been used to estimate water quality, the outcomes of these models need input data, model parameters, and a lot of information (Vieira et al., 2013).

To guarantee that water is of excellent quality, different factors must be examined for all purposes of water quality evaluation. Physical, chemical, and biological parameters are the three types of factors that make up water quality. Colour, Temperature, pH, Taste, Odour, Salinity, Hardness, Turbidity, Electrical Conductivity (EC), Total Dissolved Solids (TDS) and Total Suspended Solids (TSS) are some of the physical parameters. Fluorides, metal irons, organics, nutrients, alkalinity, pesticides, dissolved oxygen (DO), biological oxygen demand (BOD), chemical oxygen demand (COD), and disinfection by-products (DBP) are some of the chemical parameters that may be found in water. Bacteria, viruses, and protozoa are all biological parameters of water. All three water quality parameters must be tested in accordance with the APHA standard for water quality (WHO, 2004; Baird et al., 2017).

Contamination from anthropogenic or natural activities increases the sensitivity to groundwater consumption and well digging in unsuitable settings. The depth and type of well affect groundwater degradation and pollution (Hynds, 2014). Due to a lack of groundwater quality monitoring and rules controlling well digging, the public consumes water that has not been properly treated. Water consumption without proper quality control is a public health problem since it is frequently used as a vehicle for disease transmission. Water may flow fast in aquifers, allowing microorganisms to be transported with little contact between them and the host rock. Permeability is determined by the pore space between grains in porous aquifers such as gravel or coarse sand aquifers (Azizullah, 2011).

1.2 *Escherichia coli* bacteria

Escherichia coli (*E. coli)* bacteria include gram-negative, non-spore, rod-shaped pathogenic bacteria that generate gas in prescribed growth media after fermentation within 48 hours at 35℃. In 1982, *E. coli* was first recognized as a human pathogen. *E.coli* bacteria are found in the intestines of men and animals that released into the atmosphere as fecal material. *E.coli* is commonly used as an indicator of pollution impacting rivers, sea beaches, reservoirs, groundwater, surface water and recreational water. In the last five years since 2017, India has caused 10,738 deaths. The highest deaths from diarrhea were recorded in Uttar Pradesh, followed by Assam, West Bengal, Delhi and Madhya Pradesh (CBHI, 2018).

In India, 19% of the population washes hands with soap and water associated with excreta. 26% drink water that is generally polluted with *E.coli* (WHO, 2017). 44% of people have access to piped water, of which only 32% is treated. People have no access to water, thereby increasing the possibility of infection (India Water Portal, 2019). To facilitate the removal and control of water pollution, WHO, US EPA, and IS 10500: 2012 have established microbiological water quality. According to drinking water quality standards, *E. coli* bacteria shall not be detectable in 100 ml of water sample. The standards of water quality are summarized in Table 1.1 (IS 10500: 2012).

Table 1.1: Bacteriological drinking water quality (IS 10500: 2012)

S. No	Organisms	Requirements
1	*E. coli* or thermotolerant coliform bacteria	Shall not be detectable in 100 ml water sample
2	Total coliform bacteria	Shall not be detectable in 100 ml water sample

1.3 Classification of *E. coli* bacteria

Faecal coliform bacteria (FCB) can be categorized into three classes of commensal, diarrheagenic, and extraintestinal groups. The FCB is *Citrobacter, Enterobacter, Hafnia, Klebsiella, and Escherichia coli*, where *Escherichia coli* are the most common bacteria that usually survive in the gastrointestinal tract of warm-blooded animals. Some bacterial strains are harmless, like the commensal classes, but there are some infectious as well. Diarrheagenic strains can cause diseases such as diarrhea, hemorrhagic colitis, hemolytic uremic syndrome, inflammatory colitis, and dysentery. The extraintestinal strains can cause urinary tract infections, septicemia, and neonatal meningitis.

E. coli is a non-spore forming and rod-shaped bacteria with a diameter of around 0.5 μm and a length between 1.0 and 3.0 μm. *E. coli* bacteria are capable of surviving 4 to 12 weeks in water. Faecal matter is the primary source of disease-causing agents in water, and *E. coli* bacteria are commonly used as an indicator of water contamination (Atlas, 1998). The bacteria can be exhibited to be undergoing different stresses, and they are well known to be able to live below freezing temperatures (Nevers et al., 2011). Various classifications have been established for coliform bacteria. The MacConkay (1909) identified 128 different types of coliform in 1909, and 256 types of coliform were identified in 1908 by Bergey and Deehan. However, in the 1920s, coliform variation indicated reactions from indole and Voges-Proskauer that are among the most significant tests used to identify faecal contamination (Hendricks, 1978). These advancements resulted in the IMViC (Indole, Methyl red, Voges-Proskauer, and Citrate) tests for distinguishing faecal coliforms, *E. coli* and soil coliforms. Figure 1.2 (Monk et al., 2013) shows the characterization of *E. coli* pathotypes based on conditions that support growth.

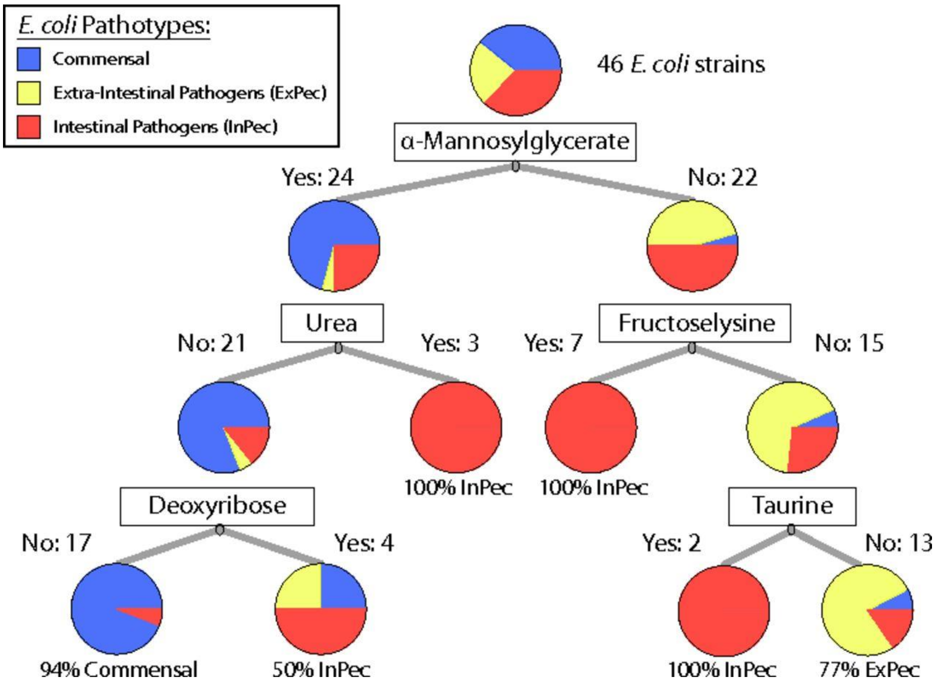


Figure 1.2: Characterizing of *E. coli* pathotypes based on conditions that support growth (Monk, 2013)

1.4 Sources of *E. coli* bacteria

Sewage discharges are usually linked to the sources of *E. coli* bacteria, classified into three general categories: human, animal and plant. Human sources include failed septic systems, urban landfills, and sewage sludge land applications. *E. coli* also originates from diverse animal sources, including domestic animals, wildlife, poultry and manure land use, pasture and feedlots. Drinking water originates from groundwater and surface streams. Surface water sources consist of rivers, ponds, and lakes, where groundwater from wells or boreholes are drained and then drilled into aquifers. Safe water availability is almost inaccessible due to bacterial and chemical contamination (Cabral, 2010). Contamination is caused by water draining from the soil, animals and birds, as well as by waste leakage, sewer overflow due to storm events and contaminated water release into the water sources (Cornejova et al., 2015; Pandey et al., 2014). Sewage treatment plants (STP) are among the sources of pathogenic *E. coli* bacteria introduced into the river systems (Eichmiller et al., 2013; Anastasi et al., 2012). Low contact rates with contaminated water in rivers (Madoux-Humery et al., 2016) or beaches (Boehm et al., 2014) can result in gastrointestinal disorders. *E. coli* bacteria are widely used as an indicator of contamination, affecting rivers, sea, beaches, lakes, groundwater, surface water, recreational water, and the actions associated with waterborne diseases (Rompré et al., 2012).

Although *E. coli* bacteria usually do not cause serious diseases but they should be used to indicate the possible presence of pathogenic bacteria and viruses (Dorevitch et al., 2012).

1.5 Health Effects of *E. coli* bacteria

Water helps to maintain the moisture in the internal organs of the body (Gerald, 2011). It also protects the volume and uniformity of blood and lymph fluids (Dooge, 2001), controls body temperature, and eliminates toxins from the body through urine, sweat, and respiration (Molden, 2013), which is essential for maintaining skin functions (Burton et al., 1987). Water pollution can lead to kidney failure and can cause death (David et al., 2011). In the present situation, people are struggling to obtain access to clean water. Generally, the most important infectious diseases are caused due to the presence of human or animal waste in groundwater. Some primary health diseases are caused by micro-organisms, including bacteria, pathogens, viruses, etc., because they can survive, reproduce, and spread in water (Payus et al., 2018). About 37.7 million people in India are affected by waterborne diseases annually, and 1.5 million children have died from diarrhea (WHO, 2017).

1.6 *E. coli* as an indicator of faecal contamination

The presence of indicator bacteria shows the occurrence of contamination. They also show the extent and nature of the pollutants. Indicator bacteria do not cause diseases associated with pathogens. *E. coli* bacteria can be used as an indicator bacteria. It is a micro-organism whose presence indicates faecal contamination. *E. coli* bacteria can survive in water for 4 – 12 weeks. At present, due to the availability of accessible, cheap, fast, sensitive, and accurate detection methods, it is an active bacterial pollution indicator. Total coliform (TC) is not considered to be an indicator organism. The ideal indicator organism has the following characteristics (Tallon et al., 2005):

- Indicator organisms should exist when there are pathogenic strains.
- The number of indicator organism counts correlates with the extent of the pollution.
- The number of indicator organism counts should be higher than that for pathogenic strains. Also, the indicator organism should not grow in water.
- Indicator organisms should have a survival time greater than or equal to pathogenic strains.
- The laboratory tests should detect indicator organisms easily and quickly.
- The indicator organism should be harmless to humans.

1.7 Detection of *E. coli* bacteria in water

1.7.1 Laboratory Methods of Bacteriological Examination of Water:

E. coli bacteria can be identified in the laboratory using Conventional Methods (Co-ordination Action Food, 2007), Enzymatic Methods (Co-ordination Action Food, 2007), Molecular Methods (Tamerat et al., 2016; Saxena et al., 2015), and Biosensor based methods (Maas et al., 2017). Basically, there are three methods of bacteriological analysis of water [57]:

(a) Multiple Tube or Most Probable Number (MPN) method.

(b) Membrane Filter (MF) method.

(c) EC-MUG Test for confirmation of *E. coli*.

(a) Multiple Tube/ Most Probable Number (MPN) method:

MPN is a method used for predicting the density of viable cells of microorganisms in a test sample. It is based on applying probability theory to the number of reported positive growth samples to a standard dilution series of sample inoculums put into a predetermined number of culture medium tubes. Observations of gas generation in fermentation tubes or apparent turbidity in broth tubes depend on the type of medium used. It denotes a good growth response after incubation.

(b) Membrane Filter (MF) Method:

Unlike the multiple-tube (MT) approach, the membrane filter method provides a direct count of total coliforms and faecal coliforms in a given water sample. The procedure involved filtering a known amount of water through a membrane filter made of a cellulose composite with a consistent pore width of 0.45 µm. The bacteria are retained on the membrane filter's surface. When the bacteria-containing membrane is cultured in a sterile container with a selective differential culture medium at a suitable temperature, typical colonies of coliforms and faecal coliforms develop, which can be counted immediately. This technique proved ineffective in turbid waters.

(c) EC-MUG Test for confirmation of E. coli

This is a standard test for confirming the presence of *E. coli* bacteria in water samples and can be used as a confirmatory test into the Multiple Tube Fermentation (MTF) process. If the lab chose to employ the EC-MUG test, the BGLB and tryptone broth (indole test) would be replaced at 44.5°C.

1.7.2 Identification of *E. coli* bacteria

The most significant bacteriological task is to classify water-borne pathogens. Generally, bacteria display three basic shapes: round, rod-shaped, and spiral. After water samples are collected, bacteria must be grown on culture media to be identified. Gram staining is the first step toward identifying bacteria (Tripathi et al. 2020). Staining is a method used to differentiate bacteria in the cell wall based on their different constituents. By coloring these cells violet or red, the gram staining method categorizes bacteria into two classes: gram-positive and gram-negative.

Eosin methylene blue (EMB) agar is a selective and differential medium used to isolate fecal coliform bacteria. It provides a rapid and accurate method of differentiating *E. coli* from other gram-negative pathogens. *E. coli* bacteria are the indicator of faecal contamination in water. The presence of *E. coli* bacteria indicates the possibility of the existence of pathogenic bacteria and viruses (Khan et al. 2020). Nobody can ferment lactose except *E. coli*. If *E. coli* bacteria are present in water. In this case, a colony will appear on an agar plate with a metallic sheen with a dark center. Gram-positive bacteria growth is typically hindered on EMB agar because of the toxicity of the methylene blue dye. Therefore, only colonies of *E. coli* will appear on agar plates. If no colony appears on the agar plates, it indicates that *E. coli* bacteria are absent in water. Consequently, it can be concluded that only *E. coli* bacteria will grow on agar plates; gram-positive bacteria will not grow on agar plates, so this method is only valid for *E. coli* bacteria. Figure 1.3 shows a Petri dish containing *E. coli* bacteria.

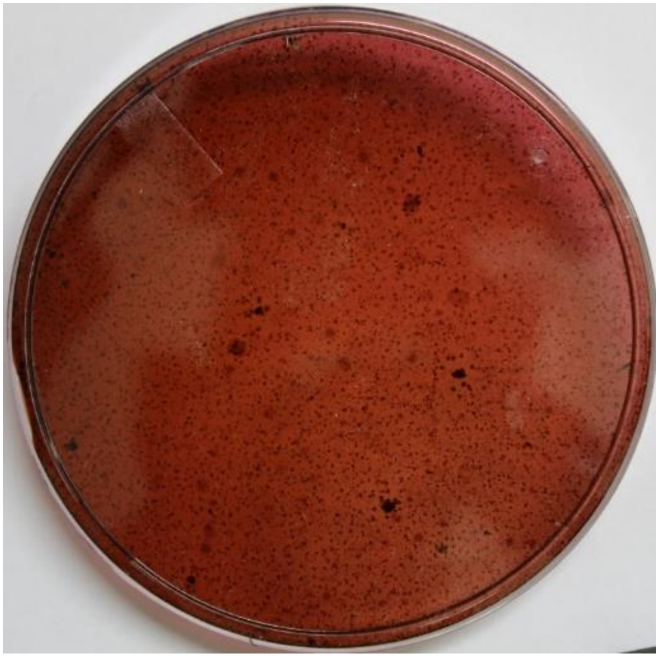


Figure 1.3: Petri dish containing *E. coli* bacteria.

1.7.3 Viable count

The USEPA approved gold standard methods for detecting *E. coli* and counting viable cells are based on culturing the water samples on solid agar plates or in liquid media. Viable cell count can be done by the plate count method (USEPA, 2010). In the plate count technique, serial dilution is made by aliquoting a certain volume of liquid culture and plating numerous serial dilutions onto culture plates. A glass spreader is used to spread the volume of culture over the surface of an agar plate, which is then incubated to develop colonies. The bacterial concentration in water sample can then be calculated assuming that each viable cell would form a single colony (Harrigan et al., 2014). The number of colonies is counted manually using the bacterial colony counter (Rompré et al., 2002).

After the identification of bacteria, a viable count of *E. coli* bacteria can be performed to count the number of actively growing bacterial cells in terms of colony-forming units (CFU). It is a microscopically visible grouping of millions of bacteria from one single bacterial cell. The plate count method can be used in which serial dilution of the water sample is done to count the number of bacterial cells present in water. Colonies of faecal coliform bacteria appear on agar plates with a metallic sheen with a dark center. This metallic sheen either covers the whole colony or appears solely in the colony's core. Other types of colonies should not be counted. The digital colony counter is used to count the number of colonies of bacteria on a petri dish.

The number of cells of *E. coli* bacteria present per 1 ml of the water sample is then given by equation (1.1);

$$\underset{\text{(CFU/ 100 mL)}}{\textit{E. coli}} = \frac{\text{Number of coliform colonies counted x 100}}{\text{Amount of water sample filtered (mL)}} \tag{1.1}$$

1.8 Fluoride

Freshwater reserves comprise of fluoride in varying concentrations, from trace amounts to some mg/L and even toxic concentrations (Schmedt et al. 2012, Celinski et al. 2016, O'Donnel 1973, Álvarez et al. 2011, Wenzel et al. 1992). High levels of fluoride are generally found at the foot of high mountains and in geological regions with marine deposits (Koblar et al., 2011). Fluoride is known to have beneficial effects on dental health within permissible limits. On the other hand, extreme fluoride ingestion above the allowable limit can lead to detrimental effects, including the accumulation of dental fluorosis or skeletal fluorosis in both adults and children. The acceptable consumption has been set at 0.05 mg/day/kg weight based on experimental observations. The frequency and intensity of this clinical incidence can differ between persons and communities because of the effects of environmental and physiological influences, the volume of fluoride absorbed, and the duration of exposure (Carvalho et al. 2011, Buzalaf et al. 2006, Khairnar et al. 2015, Ando et al. 1998).

Fluoride toxicity awareness remains relatively low (Ando et al. 1998). Millions of people worldwide are affected by adverse health effects with exposure to a high concentration of naturally occurring fluoride in potable water supplies (Moseley et al. 2003, Yi et al. 2008). Thus, fluoride has been called one of the top ten public health concern chemicals (WHO, 2006). A recent study of the US National Research Council has reported a range of possible health issues linked to elevated exposure to fluoride, including disrupted biochemical and physiological processes, cardiovascular, reproductive, endocrine, gastrointestinal, neurological, and bone fractures (Beir 2005).

1.9 Biomarkers

In order to get relevant results in a large population, a fluoride exposure biomarker should be easily collectable without donor objections, and there should be an accurate, reliable, and legitimate fluoride estimation tool. Samples of the nails and hair can be used as biomarkers to monitor fluoride contamination. Nails have been proposed as appropriate biomarkers for fluoride intake (Pessan et al. 2011, Buzalaf et al., 2006). They can help to detect chronic and sub-chronic exposures to fluoride. The use of nails as fluoride markers is appealing, provided that the samples are easy to obtain (Fukushima et al., 2009), as nails can be collected non-invasively. The user-friendly methodology for assessing nail fluoride and its fast use in an essential laboratory condition exhibits strong ability as a biomarker for epidemiological surveys. The fluoride concentration in nails reflects the total concentration of fluoride absorption and plasma during the processing of nail samples. The fluoride concentration in the nail samples is thus directly correlated to the average fluoride consumption that happened around three months ago (Whitford et al., 2005).

1.10 Water Quality Measurement and Classification

Water pollution is one of the most critical challenges for sustainable development. Water quality measurement is an important stepping stone towards finding a solution to this problem. Water quality parameters are currently measured using laboratory testing methods, where the standard laboratory sensors are stationary and water samples are brought in from the field for analysis. In this way, the current water quality monitoring system is a repetitive manual system. It is extremely tedious with the time-consuming procedure. The test sensor can be mounted in the water sample, and pollution detection can be performed remotely to improve device performance. There are some field usable devices, but those devices are large and cumbersome and way too costly. Basic water quality parameters like pH, Temperature, Turbidity, and TDS are taken as references, as the variations in the value of these parameters indicate the extent of water pollution.

1.11 Organization of thesis

The thesis is organized as follows:

Chapter 2 discusses the research carried out by many scientists and researchers from India and abroad to detect and predict fecal coliform bacteria.

Chapter 3 discusses the water sampling and laboratory testing for the study. A total of 1301 groundwater samples were collected from 348 villages and cities in the pre- and post-monsoon season. Water samples were analyzed for various physical, chemical, and microbiological water quality tests in the laboratory. These parameters are as follows; pH, total dissolved solids (TDS, mg/l), oxidation-reduction potential (ORP, mg/l), dissolved oxygen (DO, mg/l), electrical conductivity (EC, s/m), turbidity (NTU), fluoride (mg/l), nitrate (mg/l) and *E. coli* concentration (MPN/100 mL) were determined in the laboratory using the titration and spectroscopy method. Microbiological water quality analysis was done to identify the bacteria present in water using the gram staining culturing method. After identification, a viable count of bacteria was done to count the number of actively growing bacterial cells in terms of colony-forming units (CFU).

Chapter 4 discusses the different implementation of Artificial intelligence-based methods for the prediction of *E. coli* bacteria in groundwater. Artificial Neural Networks have been widely used for the classification of data where the boundaries are not clearly defined for the parameters involved. The study implements the Artificial Neural Networks and Superposition learning-based algorithm was proposed for the prediction of *E. coli* bacteria in groundwater using physico-chemical water quality parameters. Sensitivity analysis was also performed to study the importance of different water quality parameters on bacterial concentration.

Chapter 5 discusses the implementation of Machine Learning for the prediction of *E. coli* bacteria in groundwater. We implement an automated *E. coli* bacteria detection process using Convolutional Neural Network (CNN). We have also proposed a mobile application to rapidly detect *E. coli* bacteria in water that uses CNN.

Chapter 6 discusses the implementation of Machine Learning methods to develop an automatic system of bacterial colony counting on agar plates. We have automated the process of *E. coli* bacteria identification using Convolutional Neural Network (CNN). We have proposed a mobile application for the rapid detection of *E. coli* bacteria on agar plates using CNN. We have also automated the process of bacteria colony counting using Faster Region-based Convolutional Neural Network (Faster R-CNN). A Graphical user interface (GUI) application was created to rapidly count bacteria colony-forming units on agar plates using Faster R-CNN.

Chapter 7 discusses the identification of significant water quality parameters and other factors that affect the fluoride content in nail samples. Apart from laboratory tests, different Artificial Intelligence (AI) methods were used to predict fluoride in nails, which will help identify the degree of fluoride exposure to children, females, and males. The proposed Hybrid Model (HM) combines Principal Component Analysis (PCA), Firefly Algorithm (FA), and Artificial Neural Network (ANN) to predict fluoride concentration in nails.

Chapter 8 discusses the implementation of the artificial neural network to measure water quality. The data has been collected and prepared for the artificial neural network. The data collection process is very important for the implementation of Artificial Neural Networks. The data was collected from various sources in and around the BITS Pilani campus area. A prototype was designed and calibrated using standard solutions. The collected water samples were tested, and the prototype readings were tested against the readings of YSI Sonde and/or standard laboratory procedures. This technology has been patented.

Chapter 9 discusses the main finding of this study, conclusions and scope for further research.

References:

- Álvarez-Ayuso, E., Giménez, A., & Ballesteros, J. C. (2011). Fluoride accumulation by plants grown in acid soils amended with flue gas desulphurisation gypsum. *Journal of hazardous materials*, *192*(3), 1659-1666.

- Anastasi, E. M., Matthews, B., Stratton, H. M., & Katouli, M. (2012). Pathogenic Escherichia coli found in sewage treatment plants and environmental waters. *Applied and environmental microbiology*, *78*(16), 5536-5541.

- Ando, M., Tadano, M., Asanuma, S., Tamura, K., Matsushima, S., Watanabe, T., ... & Cao, S. (1998). Health effects of indoor fluoride pollution from coal burning in China. *Environmental Health Perspectives*, *106*(5), 239-244.

- Atlas, R. M. (1998). *Microbial ecology: fundamentals and applications*. Pearson Education India.

- Azizullah, A., Khattak, M. N. K., Richter, P., & Häder, D. P. (2011). Water pollution in Pakistan and its impact on public health—a review. *Environment international*, *37*(2), 479-497.

- Baird, R. B., Eaton, A. D., Rice, E. W., & Bridgewater, L. (Eds.). (2017). *Standard methods for the examination of water and wastewater* (Vol. 23). Washington, DC: American Public Health Association.

- Beir, V. I. I. (2005). Health risks from exposure to low levels of ionizing radiation. *The National Academies report in brief*.

- Ben-Naim, A., & Ben-Naim, R. (2011). *Alice's Adventures in Water-land*. World Scientific.
- Bergey, D. H., & Deehan, S. J. (1908). The colon-aerogenes group of bacteria. *The Journal of medical research*, *19*(1), 175.

- Boehm, A., & Sassoubre, L. M. (2014). Enterococci as indicators of environmental contamination. *Commensals to leading to causes of drug resistant infections. Gilmore MS, Clewell DB, Ike Y.(editors), Boston, Mass: Eye and Ear Infirmary*.

- Breed, R. S., & Dotterrer, W. D. (1916). The number of colonies allowable on satisfactory agar plates. *Journal of bacteriology*, *1*(3), 321-331.

- Burton Jr, G. A., Gunnison, D., & Lanza, G. R. (1987). Survival of pathogenic bacteria in various freshwater sediments. *Applied and Environmental Microbiology*, *53*(4), 633-638.
- Buzalaf, M. A. R., Pessan, J. P., & Alves, K. M. R. P. (2006). Influence of growth rate and length on fluoride detection in human nails. *Caries research*, *40*(3), 231-238.

- Cabral, J. P. (2010). Water microbiology. Bacterial pathogens and water. *International journal of environmental research and public health*, *7*(10), 3657-3703.

- Carvalho, R. B. D., Medeiros, U. V. D., Santos, K. T. D., & Pacheco Filho, A. C. (2011). Influence of different concentrations of fluoride in the water on epidemiologic indicators of oral health/disease. *Ciencia & saude coletiva*, *16*(8), 3509-3518.

- Celinski, V. R., Ditter, M., Kraus, F., Fujara, F., & Schmedt auf der Günne, J. (2016). Trace determination and pressure estimation of fluorine F2 caused by irradiation damage in minerals and synthetic fluorides. *Chemistry—A European Journal*, *22*, 18-388.

- Central Bureau of Health Intelligence (2018) National Health Profile 2018. Ministry of Health and Family Welfare. Government of India, New Delhi
- Clark D. W. and Briar D. W. (2001). Open-File Report 93-643. Department of the Interior, U.S. Geological Survey. [Online] Available: http://pubs.usgs.gov/of/1993/ofr93-643/pdf/ofr93-643.pdf

- Clarkson, J. J., & McLoughlin, J. (2000). Role of fluoride in oral health promotion. *International dental journal*, *50*(3), 119-128.

- Co-ordination Action Food (CAF) (2007). Methods for detection and molecular characterisation of pathogenic Escherichia coli. In: O'Sullivan J, Bolton DJ, Duffy G, Baylis C, Tozzoli R, Wasteson Y, Lofdahl S (eds.)

- Cornejova, T., Venglovsky, J., Gregova, G., Kmetova, M., & Kmet, V. (2015). Extended spectrum beta-lactamases in Escherichia coli from municipal wastewater. *Annals of Agricultural and Environmental Medicine*, *22*(3).

- David, M. M., & Haggard, B. E. (2011). Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois River Watershed, Arkansas and Oklahoma, USA. *Water, Air, & Soil Pollution*, *215*(1), 525-547.

- Diarrhoeal disease. World Health Organization (2017). Available Online at http://www.who.int/mediacentre/factsheets/fs330/en

- Dogan, E., Sengorur, B., & Koklu, R. (2009). Modeling biological oxygen demand of the Melen River in Turkey using an artificial neural network technique. *Journal of Environmental Management*, *90*(2), 1229-1235.

- Dooge, J. C. (2001). Integrated management of water resources. In *Understanding the Earth System* (pp. 115-123). Springer, Berlin, Heidelberg.

- Dorevitch, S., Pratap, P., Wroblewski, M., Hryhorczuk, D. O., Li, H., Liu, L. C., & Scheff, P. A. (2012). Health risks of limited-contact water recreation. *Environmental health perspectives*, *120*(2), 192-197.
- Eddy-Miller, C. A., Peterson, D. A., Wheeler, J. D., & Leemon, D. J. (2010). *Characterization of water quality and biological communities, Fish Creek, Teton County, Wyoming, 2007-08*.

- Eichmiller, J. J., Hicks, R. E., & Sadowsky, M. J. (2013). Distribution of genetic markers of fecal pollution on a freshwater sandy shoreline in proximity to wastewater effluent. *Environmental science & technology*, *47*(7), 3395-3402.
- Fukushima, R., Rigolizzo, D. S., Maia, L. P., Sampaio, F. C., Lauris, J. R. P., & Buzalaf, M. A. R. (2009). Environmental and individual factors associated with nail fluoride concentration. *Caries research*, *43*(2), 147-154.
- Gautam, S. P., Reeta, K., Suniti, P., Basu, D. D., & Kamyotra, J. S. (2011). Guide manual: Water and wastewater analysis. *Central Pollution Control Board, Ministry of Environment & Forests, Government of India*.

- Gerald, P. (2011). Water science. University of Washington. *PMC [serial on the Internet]*.
- Gleick, P.H. (1993). Water in Crisis: A Guide to the World's Freshwater Resources. Oxford University Press, pp. 13.
- Gourmelon, M., Lazure, P., Hervio-Heath, D., Le Saux, J. C., Caprais, M. P., Le Guyader, F. S., ... & Pommepuy, M. (2010). Microbial modelling in coastal environments and early warning systems: useful tools to limit shellfish microbial contamination. *Safe management of shellfish and harvest waters*, 297-318.
- Harrigan, W. F., & McCance, M. E. (2014). *Laboratory methods in microbiology*. Academic press.
- He, G., Fang, H., Bai, S., Liu, X., Chen, M., & Bai, J. (2011). Application of a three-dimensional eutrophication model for the Beijing Guanting Reservoir, China. *Ecological Modelling*, *222*(8), 1491-1501.
- Hendricks, C. W. (1978). Exceptions to the coliform and the fecal coliform tests. *Indicators of viruses in water and food*, *99*.
- Hynds, P., Misstear, B. D., Gill, L. W., & Murphy, H. M. (2014). Groundwater source contamination mechanisms: physicochemical profile clustering, risk factor analysis and multivariate modelling. *Journal of contaminant hydrology*, *159*, 47-56.
- India Water Portal. (2019). Available Online at https://www.indiawaterportal.org/
- International Atomic Energy Agency (IAEA). (2007). IAEA Safety Glossary: Terminology Used in Nuclear Safety and Radiation Protection (IAEA. ISBN 92-0-100707-8). Vienna, pp.5-175.
- IS10500, B. I. S. (2012). Indian standard drinking water–specification (second revision). *Bureau of Indian Standards (BIS), New Delhi*.
- Khairnar, M. R., Dodamani, A. S., Jadhav, H. C., Naik, R. G., & Deshmukh, M. A. (2015). Mitigation of fluorosis-a review. *Journal of clinical and diagnostic research: JCDR*, *9*(6), ZE05.

- Khan A, Sohail A, Zahoora U, Qureshi AS (2020) A survey of the recent architectures of deep convolutional neural networks. Artif Intell Rev 53(8):5455–5516

- Koblar, A., Tavčar, G., & Ponikvar-Svet, M. (2011). Effects of airborne fluoride on soil and vegetation. *Journal of Fluorine Chemistry*, *132*(10), 755-759.

- Korngold E., S. Belfer, C. Urtizberea. (1966). Removal of heavy metals from tap water by a cation exchanger. Desalination. 104, pp. 197-201.

- Lindim, C., Pinho, J. L., & Vieira, J. M. P. (2011). Analysis of spatial and temporal patterns in a large reservoir using water quality and hydrodynamic modeling. *Ecological Modelling*, *222*(14), 2485-2494.

- Liu, W. C., Chen, W. B., & Kimura, N. (2009). Impact of phosphorus load reduction on water quality in a stratified reservoir-eutrophication modeling study. *Environmental monitoring and assessment*, *159*(1), 393-406.

- MacConkey, A. (1909). Further observations on the differentiation of lactose-fermenting bacilli, with special reference to those of intestinal origin. *Epidemiology & Infection*, *9*(1), 86-103.

- Madoux-Humery, A. S., Dorner, S., Sauvé, S., Aboulfadl, K., Galarneau, M., Servais, P., & Prévost, M. (2016). The effects of combined sewer overflow events on riverine sources of drinking water. *Water research*, *92*, 218-227.

- Mohamed E. Mahmoud, Sawsan S. Haggag. (2011). Implementation of layer-by-layer chemical deposition technique for static removal of magnesium from various matrices. Chemical Engineering Journal. 171, pp. 181–189.

- Molden, D. (Ed.). (2013). *Water for food water for life: A comprehensive assessment of water management in agriculture*. Routledge.

- Monk, J. M., Charusanti, P., Aziz, R. K., Lerman, J. A., Premyodhin, N., Orth, J. D., ... & Palsson, B. Ø. (2013). Genome-scale metabolic reconstructions of multiple Escherichia coli strains highlight strain-specific adaptations to nutritional environments. *Proceedings of the National Academy of Sciences*, *110*(50), 20338-20343.

- Moseley, R., Waddington, R. J., Sloan, A. J., Smith, A. J., Hall, R. C., & Embery, G. (2003). The influence of fluoride exposure on dentin mineralization using an in vitro organ culture model. *Calcified tissue international*, *73*(5), 470-475.

- Najah, A., El-Shafie, A., Karim, O. A., & El-Shafie, A. H. (2014). Performance of ANFIS versus MLP-NN dissolved oxygen prediction models in water quality monitoring. *Environmental Science and Pollution Research*, *21*(3), 1658-1670.

- Nevers, M. B., & Boehm, A. B. (2010). Modeling fate and transport of fecal bacteria in surface water. *The fecal bacteria*, 165-188.

- NFIS: National Fluoridation Information Service. (2012). Household water treatment system for fluoride removal. National fluoridation information service advisory, Wellington, New Zealand, pp. 11-135.

- Nicolau, B. A. (2001). Water Quality and Biological Characterization of Oso Creek & Oso Bay, Corpus Christi, Texas. *Center for Coastal Studies, Texas A&M University Corpus Christi*, 1-124.

- O'Donnel, T. A. (1973). The chemistry of fluorine.

- Organisation mondiale de la santé, Światowa Organizacja Zdrowia, World Health Organization, & World Health Organisation Staff. (2004). *Guidelines for drinking-water quality* (Vol. 1). World health organization.

- Pandey, P. K., Kass, P. H., Soupir, M. L., Biswas, S., & Singh, V. P. (2014). Contamination of water resources by pathogenic bacteria. *Amb Express*, *4*(1), 1-16.

- Payus, C., Haziqah, N., Basri, N., & Wan, V. L. (2018). Faecal bacteria contaminations in untreated drinking water (Groundwater well and hill water) from rural community areas.

- Pessan, J. P., Pin, M. L. G., Martinhon, C. C. R., Silva, S. M. B. D., Granjeiro, J. M., & Buzalaf, M. A. R. (2005). Analysis of fingernails and urine as biomarkers of fluoride exposure from dentifrice and varnish in 4-to 7-year-old children. *Caries research*, *39*(5), 363-370.

- Rompré, A., Servais, P., Baudart, J., De-Roubin, M. R., & Laurent, P. (2002). Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. *Journal of microbiological methods*, *49*(1), 31-54.

- Saxena T, Kaushik P, MohanMK (2015) Prevalence of *E. coli* O157: H7 in water sources: an overview on associated diseases, outbreaks and detection methods. Diagn Microbiol Infect Dis 82(3):249–264.

- Schmedt auf der Günne, J., Mangstl, M., & Kraus, F. (2012). Occurrence of difluorine F2 in nature—in situ proof and Quantification by NMR spectroscopy. *Angewandte Chemie International Edition*, *51*(31), 7847-7849.

- Tallon, P., Magajna, B., Lofranco, C., & Leung, K. T. (2005). Microbial indicators of faecal contamination in water: a current perspective. *Water, air, and soil pollution*, *166*(1), 139-166.

- Tamerat N, Muktar Y, Shiferaw D (2016) Application of molecular diagnostic techniques for the detection of *E. coli* O157: H7: a review. J Vet Sci Technol 7(362):1–9.

- UNEP. (2015). WMO: Terrestrial Observation Panel for Climate (TOPC). [Online] Available: http://www.wmo.int/pages/prog/gcos /?name=TOPC.

- United States Environmental Protection Agency (US EPA). (2010). EPA Microbiological Alternate Test Procedure (ATP) Protocol for Drinking Water, Ambient Water, Wastewater, and Sewage Sludge Monitoring Methods.

- US EPA. (2009). Drinking water standards and health advisories table.

- US EPA. (2015). Drinking water from household wells. [Online] Available: water.epa.gov/learn/resources/ground water.cfm.

- Vieira, J. M. P., Pinho, J. L., Dias, N., Schwanenberg, D., & Van Den Boogaard, H. F. P. (2013). Parameter estimation for eutrophication models in reservoirs. *Water science and technology*, *68*(2), 319-327.

- Wenzel, W. W., & Blum, W. E. H. (1992). Fluoride speciation and mobility in fluoride concentration soil and minerals. *Soil Sci*, *153*, 357-364.
- Whitford, G. M. (2005). Monitoring fluoride exposure with fingernail clippings. *Schweizer Monatsschrift fur Zahnmedizin*, *115*(8), 685.
- WHO. (1996). Guidelines for drinking-water quality (2nd ed. Vol. 2.), Report no-WHO/SDE/WSH/ 03.04/16, Health criteria and other supporting information, World Health Organization, Geneva, pp. 443-504.
- WHO. Guidelines for Drinking-water Quality. 4th ed.; WHO (2011) Geneva, Switzerland.
- World Health Organization, 2017. World Health Statistics. https://www.who.int/gho/publications/world_health_statistics/2017/EN_WHS2017_TOC.pdf (accessed 06th August 2021).
- World Health Organization. (2006). *The world health report 2006: working together for health*. World Health Organization.
- World Health Statistics. WHO (2017).
- Yi, J., & Cao, J. (2008). Tea and fluorosis. *Journal of Fluorine Chemistry*, *129*(2), 76-81.

2. Literature Review

This chapter deals with the review of work from other authors globally from the year 1979 to the year 2021. The review is divided into 10 main components viz. (1) Groundwater quality, (2) Relationship between physico-chemical water quality parameters and E. coli growth, (3) Viable but Non-Cultural cells of E. coli, (4) Artificial Neural Networks, (5) Superposition-based learning algorithm, (6) Convolutional Neural Networks, (7) Principal Component Analysis, (8) Firefly Algorithm, (9) Prediction of Fluoride, (10) Water Quality Measurement and Classification, (11) Identified research gaps, (12) Objectives of the study.

2.1 Groundwater quality

Water pollution is one of the most critical challenges for sustainable development. Water quality measurement is an important stepping stone towards finding a solution to this problem. Water quality parameters are currently measured using laboratory testing methods, where the standard laboratory sensors are stationary and water samples are brought in from the field for analysis. The monitoring of bacteriological drinking water quality relies mainly on the study of indicator bacteria. *E. coli* is a more precise indicator of water contamination than other fecal coliform bacteria due to the advancement in testing methods. *E. coli* bacteria can be identified in the laboratory using conventional methods (Co-ordination Action Food, 2007), enzymatic methods (Co-ordination Action Food, 2007), molecular methods (Tamerat et al., 2016; Saxena et al., 2015), and biosensor-based methods (Maas et al., 2017). According to the method based on laboratory experiments, it takes 12–48h for the concentration of bacterial cells to be recorded. The limitation of relying solely on sensor-based water quality analysis for identification is that it can lead to errors. Therefore, there is a need for real-time monitoring.

2.2 Relationship between physicochemical water quality parameters and *E. coli growth*

Recent studies have been reported significant relationships between various water quality parameters and *E. coli* bacteria. In a study reported by Doran and Linn (1979) in eastern Nebraska for three years, runoff from a cow-calf pasture was observed. The number of fecal streptococci was higher in runoff from the ungrazed region, exposing the wildlife contributions. Baudišová (1997) performed a comparative study on the survival of fecal coliforms, total coliforms, and *E. coli* in polluted and unpolluted river water and found that all bacteria lived for several months under polluted water conditions. Still, the elimination of all types of bacteria was significantly faster under unpolluted water conditions. Total coliforms lasted the most prolonged and *E. coli* the shortest. The existence of bacteria in water sources usually increases with decreased temperature. The properties of electrolyzed oxidizing water and chemically modified solutions for the inactivation of *E. coli O157: H7* bacteria were studied by Kim et al. (2000). Inactivation of *E. coli* occurred within 30 seconds after electrolyzed oxidizing water was added with solutions containing 1% of bromine and chlorine. Residual chlorine was added to reduce oxidation-reduction potential (ORP). Iron has been found to be the only effective solution for inactivating *E. coli* and then having high residual ORP readings. The study recommended that electrolyzed oxidizing water might be simulated by chemical modification of deionized water, whereas ORP of the solution was the critical factor affecting bacterial inactivation.

In a comparative study on the growth of 10 different bacterial strains. *E. coli, Citrobacter freundii, Klebsiella pneumonia,* and *Enterobacter cloacae subsp* were identified by Boualam et al. (2002). After 96 hours of incubation, only cloacae remained cultivable. In a previous study, Boualam et al. (2003) found that only *Citrobacter freundii* and *Enterobacter cloacae subsp*. *Cloacas* were found alive after 28 days. Hughes (2003) studied the impacts of temperature, water salinity, solar radiation, sea ice conditions, and fecal contamination on the *E. coli* count around Rothera Point, Adelaide Island, and the Antarctic Peninsula from February 1999 to September 1999. In summer, i.e., February, due to the effects of solar radiation and high station population, presumptive *E. coli* counts were low, the daily amount of solar radiation was high, and the estimated *E. coli* counts were low.

In winter, i.e., April, *E. coli* counts were high because migrant wildlife had increased fecal matter and the intensity of solar radiation dropped by 95%. By late winter, i.e., September near the station sewage outfall, *E. coli* counts were high. However, the *E. coli* counts in North Cove were high as compared to February. Solar radiation was found to be the leading factor in the estimation of *E. coli* counts at sea. Water depth, temperature, and salinity also affect fecal bacterial viability by increasing cell inactivation. Some factors that affect the existence of *E. coli* include dissolved organic carbon content, the intensity of the sunlight (Medema et al., 2003). Table 2.1 (Medema et al., 2003) shows the reduction times for *E. coli* in surface water.

Table 2.1: Times for reduction of fecal coliform in surface water (Medema et al., 2003)

Bacterial group	Time for 50% reduction in concentration (days)
Total coliforms	0.9
E. coli	1.5 – 3
Enterococci	0.9 – 4
Clostridium perfringens	60 - > 300
Salmonella	0.1 – 0.67
Shigella	1

The existence of bacterial strains in groundwater is affected by various factors that are linked with soil. Bacteria have to infiltrate through the soil to enter the groundwater with low temperature, high soil humidity, acidic or alkaline soil pH, and organic carbon (Medema et al., 2003). The pollutants carried in runoff originate from urban and sub-urban non-point sources (US EPA, 2001). Table 2.2 (Medema et al., 2003) shows the disappearance rates of *E. coli* in groundwater sources.

Table 2.2: Disappearance rates of fecal coliform in groundwater sources (Medema et al., 2003)

Bacterial group	Disappearance rate (per day)
E. coli	0.063 – 0.36
Fecal streptococci	0.03 - 0.24
Clostridium bifermentans spores	0.00
Salmonella enterica subsp. enterica serovar Typhimurium	0.23 – 0.22

The effect of pH and chlorine on *E. coli O157: H7* and *Listeria monocytogenes* was explained by Park et al. (2004). The results revealed that both *Escherichia coli* and *Listeria monocytogenes* were sensitive to the residual chlorine and chlorine level of electrolyzed water. Electrolyzed water bactericidal activity increased with decreased water pH for both *Escherichia coli* and *Listeria monocytogenes*. The study recommended the application of electrolyzed water with residual chlorine greater than 2 mg/l to achieve complete inactivation of *E. coli* and *Listeria monocytogenes* within a pH range between 2.6 and 7.0. Roslev et al.

(2004) studied the effect of oxygen on the survival of *E. coli* bacteria in drinking water that is not disinfected. *E. coli ATCC 25922* bacteria shown a decline in growth, both reduced and biphasic. The survival of *fecal enterococci*, *somatic coliphages*, and coliforms were also seen to be reduced in aerobic conditions, and oxygen was found to be the main factor for *E .coli* growth in drinking water which is not disinfected. Juhna et al. (2007) studied the effects of phosphoric addition on *E. coli* bacteria survival. Higher concentrations of phosphorus increased the life of cultivable *E. coli* bacteria in water and biofilms. The study found that higher concentrations of phosphorus in water increased the cultivability of *E. coli* in the water distribution system. Ellie et al. (2007) showed a direct correlation with an R^2 value of between 0.6 and 0.8. Turbidity from the six sites ranged from 5.7 to 120 NTU, with an average of 12-17 NTU. The *E. coli* ranged from 20 to 25000 CFU/100mL with 180 to 340 CFU/100mL as geometric mean. A direct correlation between *E. coli* and turbidity was observed. Higher standards of turbidity can be used to predict increased levels of *E. coli* bacteria.

Kreske et al. (2008) identified *E. coli O157: H7* ability to grow in acidified vegetable products at pH 3.2 and 3.7, with specific dissolved oxygen content and a range of ionic strengths between 0.086 and 1.14. The study revealed that in acid solutions under anaerobic conditions, *E .coli* survived significantly better than under aerobic conditions. *The E. coli* strain decreased by 1.55 log CFU/ml for all acid solutions that were evaluated without oxygen. Kalantari et al. (2008) investigated the effects of iron, cadmium, and chromium on *E. coli* bacteria growth. In the series of experiments, *E. coli* was cultivated for 5 hours at 37 °C in a nutrient broth added with Fe^{+2}, Fe^{+3}, Cr^{+3}, Cd^{+2}. After every half hour, the bacterial growth was measured using a spectrophotometer. Results indicated that bacterial growth decreased with a concentration of 1 mM/L of Fe^{+3} and 0.5 mM/L Fe^{+2}. However, the growth was completely affected by 1 mM/L concentration of iron (II). Chromium also exhibited growth effects, while cadmium exhibited poisonous effects. Cr^{+3} and Cd^{+} showed antagonism to the growth of bacteria with iron. Than (2011) reported the growth of *E. coli* in water under different temperatures ranges from 0-70ºC at the laboratory of microbiology in the Department of Zoology, University of Yangon. The bacteria cell growths were recorded as $1.28x10^8$ CFU/ml at 20ºC, $3.25x10^8$ CFU/ml at 30ºC and $4.85x10^8$ CFU/ml at 40ºC on nutrient agar. The bacterial count at 37ºC was $4.98x10^8$ CFU/ml. Bacterial colonies were not observed under the temperatures of 50ºC, 60ºC, and 70ºC. The results revealed that *E. coli* was found to grow at temperatures between 20ºC and 40ºC.

Sinaga et al. (2016) observed the counts of *E. coli* in well water sources and the factors correlated with bacterial growth. Water samples from 5 wells were collected to test total E. coli bacteria concentrations, mercury inorganic nitrogen compounds, total phosphorus, dissolved oxygen, pH, and salinity. Results showed that *E. coli* and mercury contaminated the drinking water resources at the Sekotong regency and mercury and salinity showed an inverse correlation with *E. coli* growth. Whereas pH supports the *E. coli* survival at the range of 6.05–6.50, but no correlation to the growth of *E. coli* was found between total phosphorus and inorganic nitrogen compounds. However, the growth of *E. coli* was positively related to phosphorus concentration in water but negative to nitrate concentration. Kim et al. (2018) observed the growth features of foodborne pathogens in a laboratory medium incubated at a range of temperatures 25°C to 45°C and pH levels 3 to 10. Results showed that when subjected to pH 3 and 4 at any temperature measured, the concentration of all bacteria was restricted to about 3 log CFU/ml and all pH at 45 °C. The results showed that at pH 6, the growth rates of *E. coli* and *Salmonella* were approximately three and a half to four times faster than that of *Listeria* and at pH 7, the growing rates of *Bacillus*, *E. coli* and *Salmonella* were significantly higher than those of *Listeria* and *Staphylococcus*. At pH 8, the growth rate of *Bacillus* was the highest as compare to *Salmonella*, *E. coli*, *Listeria* and *Staphylococcus*. *E. coli* and *Salmonella* were less prone than other bacterial classes to acidic environments at pH 5-6, while Bacillus was the least prone to alkaline environments at pH 8-9.

2.3 Viable but Non-Cultural cells of *E. coli*

Laboratory-grown bacteria constitute only a minor part of the bacteria found in nature. It is found that on standard laboratory media, less than 1% of environmental bacteria can grow (Davey et al., 2011). The survival of microbial organisms depends mainly on their ability to exist in intimidating environments (Barcina et al., 2009; Keep et al., 2006). Bacteria should be able to withstand stress when environmental conditions are unfavorable and follow strategies that allow them to survive until sufficient conditions for growth are restored (Barcina et al., 2009). Clinical laboratories often grow enriched-media bacteria, developed to upkeep the growth of specific pathogens. It is achieved by certain bacterial genera, for example, by evolving resistant structures such as endospores. While many bacterial cells enter a condition of deficient metabolic activity, it is generally called the viable but non-cultural condition, i.e., VBNC (Barcina et al., 2009; Keep et al., 2006).

Colwell and colleagues first described the VBNC condition in 1982 (Xu et al., 1982). When bacterial cells can grow and form colonies on conventional culture media, they are said to be 'culturable', whereas if they are metabolically or physiologically active, they are 'viable' (Fakruddin et al., 2013). According to Oliver (1995), VBNC can be defined as a metabolically active bacterial cell that has crossed a threshold due to known or unknown causes and became unable to multiply in or on a medium that would normally support its growth. Under different stress conditions, various bacteria, including *E. coli*, are known to enter a viable but non-cultural state (VBNC). Cells lose colony-forming units on Petri dishes during VBNC state while retaining the signs of viability. Various environmental stresses like starvation stimulate the VBNC state. Bacteria in the VBNC condition cannot be grown on conventional media, usually escape plate count detection and pose a severe risk to drinking water safety and public health (Oliver, 2005); nevertheless, they maintain metabolic activity, respiration, membrane integrity and slow transcription of genes (Chowdhury et al., 1995; Huq et al., 1996; Kinjo et al., 2011; Oliver, 2005; Roszak et al., 1987). Despite the low metabolic rate of bacteria in this state, they may become culturable once again after specific resuscitation processes (Oliver, 2010). When exposed to adverse environmental conditions, many bacterial species use these conditions for long-term existence. Hence it can be recommended as a unique adaptation technique (Li et al., 2014; Ramamurthy et al., 2014).

The VBNC condition is defined as a state of dormancy in which certain bacterial strains may enter when encountered with severe environmental conditions (Besnard et al., 2002; Oliver et al., 2005). Recent studies have been shown that *E. coli* and certain bacteria may become viable but nonculturable (VBNC) under sublethal stress, such as extreme temperature changes (Riley et al., 1983; Roszak et al., 1987; Barer et al., 1993; Chowdhury et al., 1995; Oliver et al., 1995; Ravel et al., 1995; Besnard et al., 2002; Fakruddin et al., 2013; Ramamurthy et al., 2014), starvation (Chowdhury et al., 1995; Mascher et al., 2000; Nitta et al., 2000; Besnard et al., 2002; Fakruddin et al., 2013), high osmotic pressure (Barer et al., 1993; Chowdhury et al., 1995; Besnard et al., 2002; Oliver et al., 2010; Fakruddin et al., 2013), chlorine exposure (Chowdhury et al., 1995; Besnard et al., 2002; Mason et al., 2015), changes in pH (Chowdhury et al., 1995; Nevers et al., 2010; Li et al., 2014), oxygen availability (Lonsane et al., 1967; Noor et al., 2009; Liu et al., 2010; Munna et al., 2014), heavy metals (Murata Li et al., 2012) or exposure to white light (Chowdhury et al., 1995; Na et al., 2006; Fakruddin et al., 2013). However, bacteria can resuscitate a culturable state under suitable conditions (Gourmelon et al., 1994; Grey et al., 2001; Greenwood et al., 2003; Cook et al., 2007; BIS, 2012; Harmel et al., 2016). Apart from starvation, various severe environmental conditions such as changes in temperatures (Mary et al., 2002; Leclair et al., 2009; Lee et al., 2012), salinity (Linder et al., 1989; Noor et al., 2009), nutrient scarcity (McDaniels et al., 1985), incubation outside the normal growth temperature range (McKay et al., 1992; Mizunoe et al., 1999; Motion, 2009), osmotic pressure (Muela et al., 2008; Munna et al., 2013), UV radiation in combination with high salinity (Nilsson et al., 1991), low water availability (Noor et al., 2013), high concentration of copper (Noor et al., 2013), and severe environmental conditions (Dolezalova et al., 2015) induced the VBNC state. Earlier studies showed that temperature upshift with oxidative stress generation hinders the count of viable and culturable bacterial cells (Makino et al., 2000; Maalej et al., 2004; Maier et al., 2015; Lundquist, 2020). Such conditions could be lethal unless the organism has reached a VBNC state (Fakruddin et al., 2013).

The non-culturability related to the VBNC state poses a possible problem to public health because the methods commonly used to identify and count *E. coli* depend on culturing (Ghezzi et al., 1999). All non-pathogenic and pathogenic strains of *E. coli* have been shown to persist in sublethal conditions of environmental stress by entering the VBNC state (Huq et al., 1996; Jones et al., 2004; Keep et al., 2006; Kana et al., 2008; Fakruddin et al., 2013). Infectious bacteria, for example, pathogenic *E. coli*, is a crucial public health concern capable of entering a VBNC state (Kinjo et al., 2011). Studies indicate that many pathogenic bacteria can persist and remain in pasteurized milk, processed food, and drinking water, as well as in the environment (Kinjo et al., 2011). There are various significant concerns regarding the involvement of cells in water in the VBNC environment. An example is that *E. coli* cannot be used as an indicator of fecal contamination when the cells are in VBNC state (Kolling et al., 2001). However, except for *E. coli* and *V. cholerae*, other pathogens such as *Aeromonashydrophil* (Oliver, 2000; Oliver, 2005), *Listeria monocytogenes* (McKay et al., 1992; Oliver, 2012) and *Vibrio vulnificus* (Pawlowski et al., 2011) are reported to have entered VBNC state (Dutka et al., 1980). Such pathogens present in the VBNC condition can easily evade testing by conventional plating methods while retaining or recovering toxic effects after achieving suitable conditions (Pommepuy et al., 1996; Mizunoe et al., 1999; Rahman et al., 2001; Pope et al., 2003). So monitoring of *E. coli* VBNC cells is essential in drinking water due to the possible transmission of pathogens in water distribution.

2.4 Artificial Neural Networks

Artificial Neural Networks (ANN) are a model of the Biological Neural Network. Biological Neural Networks help living beings interpret, classify, and learn from their environment patterns for future applications. Humans use these patterns and prior knowledge to process any information and thus come to an output (Fausett, 2006). ANNs lend this property to machines. ANNs provide the machines with a general and functional system of learning from examples and improvising their functioning. ANNs have proved especially useful in areas where a fixed mapping algorithm between input and output does not specify the output of the system (Gupta et al., 2018). Most commonly, ANNs are used when the mapping between the inputs and the outputs are not linear. An ANN neuron is modeled, as shown in Figure 2.1.

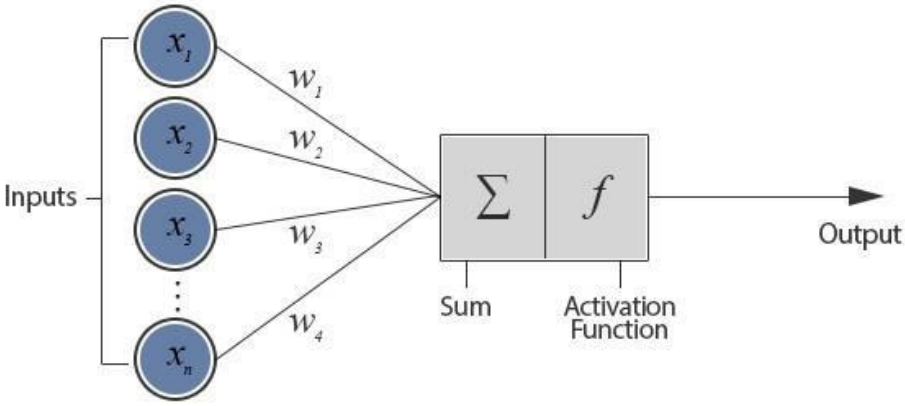


Figure 2.1: Model of a neuron of an ANN.

The input points in the above figure are analogous to synaptic connections on a nerve cell. For an n-dimension input vector ($x1$, $x2$, xn), each input is multiplied by a synaptic weight ($w1$, 2,…wn). These products are hereafter summed up in the nerve center, and the final sum is passed through an activation function that defines the final output of the neuron. Many such neurons form a layer of parallel processing centers that can work on a vast range of inputs. The outputs of one such layer of neurons serve as the input to many such subsequent layers of neurons, thus implementing a hugely parallel processing capability to the system. The outputs are then compared with the expected outputs, and the errors are measured. The weights of the synapses are thus altered in accordance with the error. This is the basic learning process of a neuron. On most neural networks, each neuron has a hidden layer connected to each unit in the previous (input) layer and the subsequent (output) layers. ANNs can be implemented in a number of architectures.

In 1943, McCulloch and Pitts presented the first-ever model of an artificial neuron, called the perceptron. A layer of perceptrons can perform some tasks. Thus a single layer of perceptrons can form a network. We term such a network as a Single Layer Perceptron. An arrangement of a series of a Single Layer Perceptron is called a Multi-Layer Perceptron (MLP). MLPs are also called feed-forward networks. Backpropagation is the most common training methodology and is simpler to implement, which reduces the time to market and is also much more capable when it comes to supervised pattern matching. Backpropagation has its limitations, such as problems with convergence, and the time cost of backpropagation hardware is hugely variable. Mathematically, Hornik et al. (1989) proved that a multilayer neural network with finite hidden layers and enough hidden neurons is a universal approximator for any Borel measurable function from one finite-dimensional space to another. Several complicated multilayer neural network models have been proposed and used in different fields (Paliwal and Kumar, 2009). Applications of ANN in the groundwater, ecology, and environmental engineering fields were documented in the early 1990s. However, in recent years ANN has been intensively used for prediction and forecasting in a variety of engineering and water-related areas, including water resource analysis by Liong and Sivapragasam, 2002; Muttil and Chau, 2006; El-Shafie et al. 2008; El-Shafie et al. 2011; Noureldin et al. 2011; Najah et al. 2009; oceanography by Makarynska et al. 2008 and environmental engineering by Grubert, 2003.

Krishnamurti et al. (1951) studied simultaneous electrical conductivity and bacterial concentration measurements in glucose and peptone-containing solution. Results show that there was an appreciable increase in conductivity even before the multiplication of bacteria occurred. In a study reported by Doran and Linn (1979) in eastern Nebraska for three years, runoff from a cow-calf pasture was observed. The cell count of fecal *streptococci* was higher in runoff from the ungrazed region, exposing the wildlife contributions. The effects of temperature, pH, and water activity were studied, out of which temperature was found to be the most crucial parameter for the thermal inactivation of *E. coli*. The limitation of the model was that it could not provide a prediction equation for the inactivation rate of bacteria (Lou and Nakai, 2001). Some studies have reported a significant correlation between various water quality parameters and *E. coli* bacteria. The existence of bacterial strains in groundwater was affected by factors that were linked with soil. Bacteria have to infiltrate through the soil to enter the groundwater with low temperature, high soil humidity, acidic or alkaline soil pH, and organic carbon (Medema et al., 2003).

Initial research on *E. coli* bacteria prediction dates back to 2003, which included the analysis and development of the model to predict compliance of bathing waters along the Firth of Clyde coastline, located in the southwest of Scotland, UK. In this study, rainfall, river discharge, sunlight, and tidal conditions were used as inputs of networks, and *E. coli* was used as an output. River discharge was found to be the most significant parameter affecting bacterial concentration (Lin et al., 2003). A few studies have evaluated the correlation between *E. coli* bacteria and physico-chemical parameters in water samples. Results show that the growth of *E. coli* was correlated with pH, dissolved oxygen (DO), specific conductivity (SC), and water temperature (T). The correlation between the Kosi River's physicochemical water quality parameters was studied in the pre-monsoon, monsoon, and post-monsoon seasons by Bhandari et al. (2007). Results show a positive correlation for chloride with pH, Magnesium, Sodium, Hardness, and Total Suspended Solids with the highest positive correlation of 0.748, 0.821, 0.7442, 0.8121, and 0.8774, respectively. Also, Electrical Conductivity was found to be negatively correlated with Total Suspended Solids, Total dissolved solids, and Hardness with a correlation of -0.8865, -0.9477, and -0.8979, respectively. Tufail et al. (2008) explained the application of AI-based models for the prediction of *E. coli* in surface water based on streamflow and other water quality parameters. Three classes of fecal bacteria were observed. The relationship between total suspended solids (TSS) and turbidity with *E. coli* was studied by Huey et al. (2010). In all four of the watersheds, a significant correlation was observed between turbidity and *E. coli*.

Recent studies in the literature reported the correlation between biological, physical, and chemical water quality using the Pearson correlation coefficient. Jothivenkatachalam et al. (2010) collected groundwater samples from different Coimbatore district locations, Tamil Nadu, India, and studied the relationship between physico-chemical water quality parameters. The study shows a strong positive correlation for Electrical Conductivity with Total dissolved Solids, Total Hardness, Calcium ions, Magnesium ions, and Chloride ions with the highest positive correlation of 0.978, 0.852, 0.835, 0.68, and 0.84559, respectively. A few studies have been done to evaluate the relationship between physico-chemical water quality parameters and E. coli on the site. Chigor et al. (2011) studied the water quality of surface water sources in Zaria, Nigeria, using Pearson correlation analysis. A total of 228 water samples were collected and analyzed to observe the relationship between physico-chemical water quality parameters and *E. coli* bacteria. Results show a positive correlation of $P<0.05$ for *E. coli* with Electrical Conductivity, Total Suspended Solids, Total dissolved Solids, Chloride ions, Potassium ions, Nitrate, and Biochemical Oxygen Demand. No significant correlation was observed between *E. coli* with pH, Temperature, and Turbidity.

An AI-based model on multiple regression analysis was developed to predict coliform bacteria concentrations at the selected sites based on available USGS NWIS data (David and Haggard, 2011). Francy et al. (2013) proposed a model to evaluate the correlation between *E. coli* and other parameters based on water samples collected from eight inland recreational lakes in Ohio. The parameters used were rainfall, wind direction, speed, turbidity, and water temperature, but the model was not developed at sites where the *E. coli* concentration was exceeded. Cheng et al. (2013) studied the correlation between water quality parameters and *E. coli* growth by Pearson's correlation analysis. They observed that the density of bacteria varied negatively with pH and the removal of total suspended solids. The growth of *E. coli* in aeration pond was negatively correlated with the increased dissolved oxygen. Mouna et al. (2014) observed factors that have an impact on the growth of *E. coli* with a negative correlation of higher salinity (R=-0.97) and pH (R=-0.98) against a positive correlation of higher turbidity (R=0.93). The model was used to predict environmental quality along the Penchala River urban catchment area located in Kuala Lumpur, Malaysia, directly affected by human activities (Zamani and Saybani, 2014). The model provided the best training performance, with 70 neurons in the hidden layer.

Shroff et al. (2015) collected groundwater samples from different locations of Valsad district, Gujarat, India. They studied the correlation between physico-chemical water quality parameters using the Pearson correlation coefficient. Results show that Electrical Conductivity was found to be significantly correlated with Total Hardness, Calcium ions, Total Alkalinity, Total dissolved Solids, Chloride ions, Sulphate ions, Sodium ions, and Sodium Absorption Ratio, with the highest correlation of 0.88, 0.94, 0.99, 0.98, 0.94, 0.95 and 0.90 respectively. Also, Electrical Conductivity was found to be the lowest correlated with Fluoride, Manganese, Chemical Oxygen Demand, and Silica. Wu et al. (2015) studied the relationship between microbial biofilm and physico-chemical water quality parameters using the Pearson correlation coefficient. Also, the study has considered the effect of sampling site distance on water quality. Shamsudin et al. (2016) observed *E. coli* growth using Pearson correlation, and the results showed a linear relationship with pH (R=0.971), time (R=0.958), turbidity (R=0.885), dissolved oxygen (R=-0.861), and temperature (R=0.763). Rao et al. (2015) examined the relationship between water quality and turbidity. Increased *E. coli* cell counts were significantly associated with increased turbidity (β = 0.003; $p < 0.0001$) and decreased dissolved oxygen concentrations (β = −0.310; $p < 0.0001$). The water quality prediction model was developed with water samples from 4 different sampling stations on the Panchaganga River for modeling river quality with BOD and DO parameters (Mulla et al., 2016). Bisi-Johnson et al. (2017) studied the physico-chemical and microbial properties of water.

Islam et al. (2017) developed a linear regression model to evaluate the impact of the atmospheric conditions on bacterial counts. Precipitation and temperature of the water showed a positive correlation with the growth of bacteria. Raw water turbidity, colour, and alkalinity were found to have a significant influence on the growth of *E. coli* (Mohammed et al., 2018). In a few studies, statistical models were used for prediction (Katip, 2018). Mohammed et al. (2018) study was based on observed counts of bacteria and measured water quality parameters, including pH, temperature, conductivity, turbidity, colour, and alkalinity. Considerable improvement in the efficiency of the model was achieved when the input data was normalized before training. Vijayashanthar et al. (2018) developed a model for the prediction of bacteria using water temperature and turbidity. Results show that the accuracy of the developed model was 86.5%. The temperature was found to be the most crucial parameter of bacterial regrowth and death rates, and pH also showed effects on the decay rates. In the three perennial watersheds, a correlation was observed between TSS and *E. coli*. Pachepsky et al. (2018) studied the relationship of temperature, pH, dissolved oxygen (DO), turbidity, nitrate, ammonium with *E. coli* concentration, using Spearman's rank correlation coefficient of -0.247, -0.267, -0.246, -0.293, 0.015, -0.220 respectively.

Singh et al. (2019) studied the seasonal effect on potable water sources of the Eastern Himalayan region. The relationship between physico-chemical and microbiological water quality parameters was determined using Pearson correlation analysis. Results show that the growth of *E. coli* was positively correlated with the rainy season with a possibility of water-borne diseases. The microbiological standard of the municipal water supply system in the Osijek-Baranja district of Eastern Croatia was studied by Habuda-Stanić et al. (2013). The study showed that *E. coli* bacterial growth was negatively correlated with free residual chlorine and positively correlated with Turbidity. Seo et al. (2019) studied the relationship between coliform bacteria and physico-chemical water quality parameters in the Nakdong River, South Korea, using Pearson correlation and multiple regression analysis. Results show that the growth of coliform bacteria was affected by Phosphate Phosphorus and Total Suspended Solids. It was also found that the growth of coliform bacteria was inversely correlated with organic matter and directly correlated with Phosphate Phosphorus.

Bouharati et al. (2008) proposed a method for detecting micro bacterial pollution in freshwater using an ANFIS. The model produced instantaneous results by the measurement of the physical and chemical properties of the sensors. ANFIS based methods are based on the concept of Fuzzy set theory, which states that a variable can partially belong to a set and can have a membership value between 0 and 1. The number of rules depicted the number of fuzzy sets created. The author revealed the use of an artificial neural network model of three layers trained and tested on the collected water samples. ANFIS based model was used because it combines the advantages of fuzzy systems with transparent knowledge representation and those neural networks which deal with the implicit knowledge that can be acquired using learning. Kamali and Binesh (2013) used ANN and ANFIS to study the diffusion of water through nanotubes using molecular dynamics data. They concluded that ANFIS outperformed ANN.

Similarly, Azeez et al. (2013) compared the performance of ANN and ANFIS in the triage of emergency patients using various vital signs of patients as input parameters. Chandaran et al. (2012) explained the detection of sulphate-reducing bacteria (SRB) using ANFIS, which can be crucial in curbing the corrosion of iron material in the system. The author used three parameters: Voltage, Temperature, and humidity, for training the model. The membership functions were taken to be trapezoidal and bell-shaped. The ANFIS model used three inputs, which finally gives the output as either 1 or 0. The predicted results were obtained by the input parameters and the number of epochs was taken as 20. Lastly, the model was tested with testing data up to 250 epochs.The author compared the results and the best membership function was given by trapezoidal shape. Keshavarz et al. (2018) explained the application of ANFIS based method in determining the compressive strength of concrete. The model used 150 different concrete specimens with various mix design parameters. Five different concrete mix parameters, i.e., cement, water to cement ratio, gravel, sand, and micro-silica, were considered as the parameters. For results, two of the soft computing methods: ANN and ANFIS, were selected to detect the compressive strength of concrete. The results were computed in MATLAB, where the concrete mix parameters were used as input variables and the compressive strength of concrete was used as an output parameter. In order to compare the ANN and ANFIS based methods, the author used parameters like the R squared coefficient of both models. The higher values of the coefficient of determination would indicate the better capability of the model in predicting the specific studied characteristics. Calp (2019) proposed a hybrid model for the estimation of the regional rainfall amount. The proposed model focused on providing efficient water resources management by estimating the amount of rainfall that can occur in the region. While creating the model, the MATLAB package program was used and regression values (R) or mean squared error (MSE) were taken into account. The error rate was obtained as 0.9920, 0.9840 and 0.0011, respectively, for the model. The author concludes by stating that this hybrid model is an important support tool for estimating the amount of annual rainfall and ensuring the effective management of water resources.

2.5 Superposition-based learning algorithm

The superposition-based learning algorithm (SLA) is based on the search algorithm of Grover (1996). The learning algorithm can be used for training neural networks. Grover's algorithm (1997) is used for searching an unordered data linearly faster than any conventional method. The Grover's learning algorithms for neural networks (Altaisky, 2001; Zhou et al., 2007; Silva et al., 2010; Panella et al., 2011) can be categorized as superposition based (Silva et al., 2010; Panella et al., 2011) or iterative (Altaisky, 2001; Zhou et al., 2007). This algorithm is a supervised learning algorithm for neural networks, where all training set patterns are introduced simultaneously to the network using a superposition state. The iteration number I of the algorithm is calculated from equation (2.1):

$$I = [\ \frac{\pi}{4}\ \sqrt{\frac{N}{M}}\ ] \tag{2.1}$$

Where,

I = iteration number

M = Number of solutions

N = Number of parameters

2.6 Convolutional Neural Network

A Convolutional Neural Network (CNN) is a subset of the neural network mentioned previously. One or two convolutional layers are present in a CNN, always with a subsampling layer, accompanied by one or more fully connected layers (Khan et al. 2020). The conception of a CNN was sparked by the discovery of a sense system in the brain, the visual cortex. The visual cortex comprises many cells that sense light in small. Receptive fields are overlapping sub-regions of the visual field. The more complex cells have wider receptive fields, and they serve as local filters over the input space. The convolution layer in a CNN has the same function as the cells in the visual cortex (Hubel et al. 1968). A hand-designed feature extractor collects essential information from the input. It extracts irrelevant variables in the conventional model of pattern recognition (Fukushima et al. 1983). After the extractor, a trainable classifier is used, which is a regular neural network that divides feature vectors into classes. Convolution layers serve as feature extractors in CNNs. They are not, however, handcrafted. The kernel weights for convolution filters are selected during the training phase. Since the receptive fields of the hidden layers are restricted to be local, convolutional layers can extract local features. The weights of the convolutional and fully connected layers are calculated in CNN during the training phase and used for feature extraction (Brownlee 2019) and classification (Huang et al. 2018). The improved network architectures result in reduced memory and computing complexity.

Machine learning algorithms for object detection are based on autonomous learning and have good detection accuracy. Object detection algorithms based on machine learning have been developed for various applications like face detection (Viola et al. 2004), pedestrian detection (Dollar et al. 2011), medical object detection (Zhu et al. 2016), military object detection (Hua et al. 2018), intelligent transportation systems (Zhang et al. 2011), and intelligent monitoring systems (Chen et al. 2014). An anchor is used in object detection for classification and regression. The algorithm replaces the preceding region proposal network (RPN), feature selection method, and selective search (Kulkarni et al., 2015) with the guided anchor method (Wang et al., 2019). The network module is obtained using function selection (Hu et al. 2018) to elucidate the object compression problem. The skip pooling method (Bell et al. 2016) is used to solve the problem of small object size to improve the detection efficiency of faster region-based CNN (R-CNN) in complex scenes.

The USEPA-approved gold-standard methods for detecting *E. coli* and counting viable cells are based on culturing water samples on solid agar plates or in liquid media. Viable cell counts can be done by the plate count method (USEPA 2010). In the plate count technique, serial dilutions are made by creating aliquots of a certain volume of liquid culture and plating numerous serial dilutions onto culture plates. A glass spreader is used to spread the volume of culture over the surface of an agar plate, which is then incubated to develop colonies. The bacterial concentration in a water sample can then be calculated, assuming that each viable cell forms a single colony (Harrigan et al., 2014). The number of colonies is counted manually using a bacteria colony counter (Rompré et al. 2002). Manual counting of viable bacterial cells on agar plates is time-consuming and can be prone to human error. The method requires experts to identify and count viable cells. Furthermore, due to bacterial overcrowding, high numbers of colony-forming units on a plate will lead to inaccurate results (Breed et al. 1916).

E. coli bacteria can be identified in the laboratory using Conventional Methods (Co-ordination Action Food, 2007), Enzymatic Methods (Co-ordination Action Food, 2007), Molecular Methods (Tamerat et al. 2016; Saxena et al. 2015), and Biosensor based methods (Maas et al. 2017). According to the method based on laboratory experiments, it takes 12-48 hours for the concentration of bacterial cells to be recorded. The limitation of relying solely on sensor-based water quality analysis for identification is that it can lead to errors. Therefore, there is a need for real-time monitoring. Enzymatic methods of detection are color-based methods (Rice et al. 1989). The amount of colour appearance can be used to determine the degree of bacterial contamination. The detection method is based on the concept that only *E. coli* bacteria are fed. No substrate is given for other bacteria. The specified substrate is used as an essential source of nutrients for bacteria. A chromogenic or fluorogenic substance is released from the specified substrate during the substrate utilization period, which indicates the presence of *E. coli*. Manually performing this process is highly time-consuming and difficult. This detection process is analytical. There is always a possibility of human error, which may result in a disastrous decision. The colours of each concentration can be scanned using conventional computer vision methods. It is, however, extremely difficult to determine the colour intensity for each concentration level. This is made simple with deep learning since the algorithm calculates these colour intensities using statistically generated training sets.

A Convolutional Neural Network (CNN) is a subset of the neural network mentioned previously. One or two convolutional layers are present in a CNN, always with a subsampling layer, accompanied by one or more fully connected layers (Khan et al., 2020). The conception of a CNN was sparked by the discovery of a sense system in the brain, the visual cortex. The visual cortex comprises many cells that sense light in small. Receptive fields are overlapping sub-regions of the visual field. The more complex cells have wider receptive fields, and they serve as local filters over the input space. The convolution layer in a CNN has the same function as the cells in the visual cortex (Hubel et al. 1968). A hand-designed feature extractor collects essential information from the input. It extracts irrelevant variables in the conventional model of pattern recognition (Fukushima et al. 1983). After the extractor, a trainable classifier is used, which is a regular neural network that divides feature vectors into classes. Convolution layers serve as feature extractors in CNNs. They are not, however, handcrafted. The kernel weights for convolution filters are selected during the training phase. Since the receptive fields of the hidden layers are restricted to be local, convolutional layers can extract local features. The weights of the convolutional and fully connected layers are calculated in CNN during the training phase and used for feature extraction (Brownlee 2019) and classification (Huang et al., 2018). The improved network architectures result in reduced memory and computing complexity.

Mohanty et al. (2016) used a public dataset containing 54,306 images of healthy and diseased plant leaves. They developed a deep CNN to identify 26 diseases and 14 crop species with an accuracy of 99%. Turra et al. (2017) developed a hyperspectral image based on the acquisition of spectral signs from bacterial colonies growing on blood agar plates and bacteria identification using machine learning methods. PCA+SVM and RSIMCA methods were used to differentiate five selected UTI bacteria. The RSIMCA method outperformed the PCA+SVM method in terms of Sensitivity, Precision and F-Score for the classification of *E. coli* bacteria. The study shows that this method is time-consuming as it requires 16h of incubation compared to the currently available EPA-approved gold-standard analytical methods (USEPA 2010). Arrigoni et al. (2017) developed an HSI processing and classification system to rapidly identify UTI bacteria. The sheep blood agar plate samples were collected from American Type Culture Collection (ATCC) for the study. This method is time-consuming as it requires 18h of incubation compared to EPA-approved gold-standard methods (USEPA 2010). Further research is needed for rapid, accurate identification of bacteria and bacterial cell counting on agar plates using hyperspectral image analysis.

Zieliński et al. (2017) used the deep learning method for bacterial colony classification using publicly available datasets collected from the Chair of Microbiology of the Jagiellonian University in Krakow, Poland. The data collection comprises 33 bacterial species, each with 20 pictures. The developed FC-CNN model achieved an accuracy of 0.82. However, the developed model has not been verified by the authors using other statistical measures. Ferrari et al. (2017) used two different machine learning methods for automatic bacterial colony counting. The authors have developed Support Vector Machines and Convolutional Neural Networks model using publicly available datasets of urine samples. The model performance was validated using Accuracy, Sensitivity, and Precision statistical measures. The Sensitivity and Precision of the CNN model were found to be 0.73 and 0.71. Thus, we cannot rely on this study for bacteria colony counting. Hay et al. (2018) developed a Convolutional Neural Network to identify bacteria in light-sheet fluorescence microscopy images of larval zebrafish intestines. The authors have used Google's open-source Tensorflow to create convolutional neural networks. The accuracy of the developed CNN model was 0.90, but the authors have not validated the model performance using other statistical measures. Huang L et al. (2018) developed a convolutional neural network (CNN) for bacterial colony classification using digital images. The data from Peking University First Hospital was used for the classification of bacterial colonies. Results show that the network was able to classify 18 bacterial colonies with an accuracy of 73%.

Alaslani et al. (2018) extracted the learned features from a pre-trained CNN and Support Vector Machine (SVM) algorithm for image classification. The Alex Net pre-trained CNN model was used for feature extraction, and the SVM algorithm was used for classification. The Iris public images were used for the development of an iris recognition system. Results show that the recognition accuracy of the Iris database was 98.3%. Nehal et al. (2019) developed an AI-based lab-on-chip for the detection of bacterial contamination using the Photonic Crystal-based optical biosensor. These biosensors came up with a few limitations of using separate sensors to measure physical, chemical, and bacteriological parameters of water quality which affect sensitivity and accuracy of the results. The method is cost-intensive and requires maintenance. Gunda et al. (2019) developed an AI-based mobile application for water quality monitoring with an accuracy of 99%. The authors have not validated the proposed model with performance functions. Hence, this model is not reliable for bacterial detection. Wang et al. (2020) developed a Deep Neural Network to rapidly detect bacterial growth and classify the corresponding species. The authors have developed a model that captures coherent microscopy images of bacterial growth inside a 60-mm-diameter agar plate and analyzes these time-lapsed holograms using two different DNNs. The first DNN is used to detect bacterial growth. The second DNN was utilized to categorize bacteria based on spatial and temporal characteristics extracted from incubated agar plate coherent pictures. The dataset used in this study contains 71 images of agar plates. Six agar plate images were used for validation of the classification model with an accuracy of 0.90. Hence, we cannot rely on this study for bacteria colony counting.

Various unsupervised learning methods were also used to automate counting and classification of images in various environmental applications like Wheat ear counting using K-means clustering segmentation and convolutional neural network (Xu et al. 2020), White blood cells segmentation using the K-means algorithm (Sarrafzadeh et al. 2015), Application of the t-SNE method for creating urban microbial fingerprints (Ryan 2019), Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome (Durán et al. 2021). Unsupervised learning methods can be used for Clustering and Association. It works on uncategorized and unlabeled data, which makes it more important. The limitation of unsupervised learning is that it is a more time-consuming procedure than supervised learning because there is no matching output (Karim et al., 2020). Unsupervised learning methods do not predict the result of a new sample as there is no notation of the output along the training process (Caballé-Cervigón et al. 2020). Also, the accuracy of the unsupervised learning algorithm is less as compared to the supervised learning algorithm since input data is not categorized and systems do not know the exact response in advance (Chen et al. 2016; Li et al. 2020). Further research is needed to automate colony counting using unsupervised learning methods.

Recent studies used publicly available datasets and platforms for colony counting. Torelli et al. (2018) used publicly available OpenCV and CellProfiler software platforms for automatic bacterial cell counting. Albaradei et al. (2020) used the CSRNet transfer learning application for cell counting. The training of the model was performed using Python with the Keras library. The model performance was validated using Root mean squared error (RMSE) values. The average RMSE value of the developed model was 22.38, which is very high. We cannot rely on the model for viable cell counting. The aforementioned studies indicate that the prediction models are based on public datasets, research reports that have been published, and testing data that is freely available on the internet, making it impossible to assess the model's accuracy. As a result, we cannot depend entirely on these studies to count bacteria colonies. However, no study has been done to identify and count *E. coli* bacterial cells on agar plates with great accuracy using experimental laboratory data.

2.7 Principal Component Analysis (PCA)

The principal component analysis is a multivariate method used to reduce the dimension of input variables when we have a vast amount of observations and an improved understanding of variables (Lu et al., 2003). The PCA algorithm helps to reduce the dimension of the data into limited numbers of variables for data interpretation and then create basic plots to display essential statistics, including score plot and loading plot, to study the correlation between the broad clustered data set (Stojanovic et al. 2012, Beltran et al. 2006). Such associated variables are known as principal components (Shinde et al., 2009). PCA has its mathematical algorithm in linear algebra, which describes the association between the data containing the variables as columns and the observations or samples as rows. The fundamental purpose is to create a transformed matrix using coefficients of principal components that includes the maximum amount of information and then plot the data using a 2-dimensional plot in MATLAB software (Bell et al. 1997).

2.8 Firefly Algorithm (FA)

The firefly algorithm is used to improve the performance of machine learning models by optimizing the weights and bias between the input layer and the hidden layer of the ANN model. Firefly algorithm is one of the swarm intelligent algorithms developed by Yang. It is a metaheuristic algorithm that is inspired by nature and, based on the flashing behavior of fireflies, used to solve complex problems and non-linear optimization problems (Moazenzadeh et al. 2018). The brightness of the fireflies is the main criterion for the optimization of the fitness function (Gandomi et al. 2011, Yang et al., 2011). Yang developed the algorithm based on the following assumptions:

1. The attraction of firefly is independent of gender due to unisexuality, and it is directly proportional to the brightness of the emitted light, but it is indirectly proportional to the distance between the fireflies(x_i, x_j). The firefly can move in any direction if the brightness of the neighboring firefly is same.
2. The brightness of the light is associated with the optimization of objective function $f(x)$ in the algorithm.

2.9 Prediction of Fluoride

Recent studies examined the concentration of fluoride in water by the usage of urine (Buzalaf et al. 2012, Antonijevic et al. 2016, Akpata et al. 2014) and nail samples (Buzalaf et al. 2012, Lima-Arsati et al. 2010, Amaral et al. 2014, Linhares et al. 2016, Sousa et al. 2018). Still, no study has been done to examine the correlation between physico-chemical water quality parameters and fluoride concentration in nail samples. Fluoride fingernail analysis has been widely used to determine low-level concentrations in water fluoridation, toothpaste, salt, and milk (Whitford et al. 1999, Buzalaf et al., 2012, Lima-Arsati et al. 2010, De Almeida et al. 2007, Buzalaf et al., 2009, Pessan et al. 2005, Levy et al., 2004).

Fukushima et al. 2009 have used nails for investigating the correlation between fluoride exposure biomarkers and total daily intake of fluoride with significant fluoride exposure in drinking water. They studied the impact of age, gender, nail growth rate, and geographic area on the absorption of fluorides in the fingernails and toenails (Elekdag-Turk et al., 2019). They obtained drinking water and nail samples and used an ion-selective electrode to examine fluoride concentration. A comparison mark was created on each nail, and growth levels were calculated. The analysis was done by ANOVA and linear regression. All the factors they considered were directly associated with the fluoride concentration in nail samples. The study recommended that nails should be used as biomarkers of fluoride contamination, with the advantage of being easily obtained. But they do not consider water characteristics. At present, none of the studies on nails as biomarkers of fluoride exposure have examined the impact of age, gender, and factors affecting the bioavailability of fluoride (Clarkson et al., 2000). There is a need to study the effect of age, weight, gender, water fluoride, nitrate, turbidity, dissolved oxygen, electrical conductivity, and pH levels on fluoride concentration in nail samples since water characteristics might also impact fluoride.

2.10 Water Quality Measurement and Classification

Ranković et al. (2010) used ANN to predict the Dissolved Oxygen concentration (DO). The limitation of this study is that it can't be used for real-time monitoring since the parameter involved are chemical parameters and can only be detected in a laboratory setup. Gazzaz et al. (2012) used 23 water quality parameters for the prediction of WQI using ANN. The model cannot be used for real-time monitoring as it turns out to be expensive given the price of the sensors involved. Menon et al. (2012) developed a wireless sensor network-based river water quality monitoring system for continuous and remote monitoring of water quality data in India. The wireless sensor node in the device was intended for water pH monitoring purposes. The device was restricted in that it could not be used to control the contamination of water in a region. Meanwhile, Ali et al. (2013) classified the water quality into three classes using the unsupervised machine learning method. The limitation of this study is that they have not considered the various parameters which are correlated with Water Quality Indexing (WQI).

Faustine et al. (2014) developed a solar-powered system for monitoring water quality in the Lake Victoria Basin using WSN was developed by Sensor nodes used an Arduino core which was further used by sensor nodes to process measured data. Then it was sent to the gateway via ZigBee. The gateway gathered all the information and, using GPRS, transferred it to the application program. The authors demonstrated the proper functionality and implementation of the proposed system in the real world based on field test results. Nevertheless, the device had no local data analysis provision. It will thus be offline any time a mobile network interruption is encountered. These technologies typically work in the free band of the 2.4 GHz ISM license, which is often crowded and susceptible to interference and security attacks. Vijayakumar et al. (2015) developed a low-cost, real-time water quality monitoring system using Internet of Things (IoT) technologies. A Raspberry Pi model B+ microprocessor was operated by the node, with several water quality sensors connected to it. The water quality parameters such as temperature, pH, turbidity, conductivity, and dissolved oxygen can be measured using this system. The Raspberry Pi platform was used as a central controller. The proposed device was able to demonstrate water quality parameters on the Internet from experimental data. This approach incorporated vulnerabilities that can affect the authenticity, credibility, and confidentiality of measurement data due to cyber-attacks. Kalpana et al., 2016 developed a water monitoring device consisted of conductivity, turbidity, and pH sensor. The parameters can be automatically detected under the Raspberry Pi3 Model B single-board computer. The single-board computer receives the data from the three sensors, and the data is sent via the internet to the webserver. This device can be used for commercial as well as domestic purposes. The system can be extended to track hydrology, air pollution, the development of an industrial and agricultural product, etc.

Amruta et al. (2017) proposed a regulated water supply system using a board arranged with the sun. The device includes the center and base station in which the center point is connected to the base station via the Zigbee advance that is operated by the board based on daylight. If the panel located in the sun cannot be charged due to any reason, the mechanism would cease to work. Previous studies used basic water quality parameters like pH, Temperature, Turbidity and TDS as a reference, as the variations in the value of these water quality parameters indicate the extent of water pollution. Therefore, overcoming this restriction leads us to prepare a new system that will make a negligible effort, improvement, and be user-friendly. Gopavanitha et al., 2017 developed a system for the monitoring and control of water quality in real-time using

IoT. The device consists of sensors capable of measuring the physical and chemical parameters of water, including temperature, turbidity, conductivity, pH, and flow. The sensor takes the output value from Raspberry Pi and sends it to the cloud. Using cloud computing, the sensed data is finally visible on the cloud, and IoT controls the water flow in the pipeline. Puneeth et al., 2018 proposed an application using the concept of WSN and IoT to monitor parameters via every node, including pH, turbidity, and temperature, which were stored and made available on the cloud. The system runs on solar power.

Lin et al. (2018) developed a water quality monitoring system using wireless sensor network technology and powered by a solar panel. The prototype device was developed and implemented using one node powered by a solar cell and WSN technology. Data was collected from various node-side sensors, such as pH, turbidity, and oxygen density, and was sent to the base station via WSN. Kumar et al., 2019 proposed a smart sensor interface device for water quality monitoring systems in an IoT environment. The inventors used sensors such as CO_2, temperature sensor, pH sensor, water level sensor, and turbidity sensor. This sensor system controls the entire process and is controlled by wireless communication devices centered on the cloud. The water level sensor is used to sense the water level within the tank and show it. The sensors can automatically monitor the water quality. Amareshwar et al. (2019) developed a sensor-based water quality monitoring system that uses MEMS sensors to assess physical parameters of water quality like temperature, pH, and Water Humidity. The Raspberry Pi model can be used as a controller for the center. The sensor information can finally be viewed on the web using API.

A low-cost, real-time water quality monitoring system was developed by Demetillo et al. (2019). It can be used in remote rivers, reservoirs, coastal areas, and other water bodies. The 6 V/3.5 amp-hour (Ah) lead-acid battery was used in the device to power the nodes. Minu et al. (2019) developed an IOT-based sensor that measures the pH, temperature, conductivity, dissolved oxygen, turbidity, bacteria, etc., in the water sample. Data were obtained by the sensors and sent through a network. The server would then upload the details to the cloud. The remote water station will read the gathered data and assess the water quality. Ahmed et al. (2019) predicted and classified the Water Quality of Rawal Lake, Pakistan, using various Machine Learning algorithms. They have taken 12 parameters – Alkalinity, Appearance, Calcium, Chlorides, Conductance, Fecal Coliforms, Hardness as $CaCO_3$, Nitrate as NO_2^-, pH, Temperature, Total Dissolved Solids, and Turbidity. The maximum accuracy reached by any algorithm in their study was about 85%, and also, they didn’t propose any practical implementation of the algorithms for field usage. Abyaneh, 2006 used ANN and multivariate linear regression to predict Biological Oxygen Demand (BOD) and Chemical Oxygen Demand (COD) using four parameters, viz, pH, temperature, Total Suspended Solids (TSS), and Total Suspended (TS).

Conventional water quality measurement techniques include on-site sampling and subsequent laboratory-based tests; both are labor-intensive and cost-intensive processes. The measurements are not in real-time. Therefore there is a need for real-time monitoring of water quality for drinking applications to reduce labor costs and time usage. With the help of Zigbee boards, recorded data is uploaded to the remote data storage in the traditional system. It requires more hardware to set up this technology and is very expensive. There's also no alert indication in that system when parameters are abnormal. In the Solar Powered Water Quality Monitoring System using remote Sensor Network, the advancement of the water sensing network is controlled using sun board. If the sun board is not charged, then the system will not switch on, which is the restriction associated with this method.

2.11 Research Gaps

According to the laboratory experiment based on the conventional analysis method, 24-48 hours are required before the bacteria concentration get reported (Gautam et al., 2011). As a result of limitations associated with laboratory quantification of microbial water quality, studies have been done to develop real-time or near real-time predictive models to aid in water management decisions. At present, it is not possible to measure bacterial concentrations in water and to obtain an immediate quantitative result to evaluate and prevent human health risks. This study would be supportive of data management and it could be the basis for ground water management in the early warning system studies for public health. Early warning systems, which are based on modeling, aims at obtaining real-time data for risk management. It is also based on real-time observation in which the information is immediately sent to a computer that will analyze the information, but if the parameter exceeds defined values, then the system moves into an alarm mode (Gourmelon et al., 2011). The gaps identified for this study can be listed as follows:

- The majority of the existing techniques are limited to most of the substantial features of water to limit pH, Temperature, Turbidity, Conductivity and Colour of water, but few major parameters were not considered which have direct effects on the growth of faecal coliform bacteria. The parameters such as DO, TDS, ORP, Nitrate and Fluoride are to be considered.

- No study has been carried out to predict faecal coliform bacteria in ground water using various physical and chemical parameters which have a direct effect on the growth of bacteria. So, the aim of this study is focused on this main parameter through the study of the influence of water quality parameters.

- Limited studies are available to evaluate the sensitivities of each input parameter to determine their respective influences on the predictive abilities of the models. It is essential to determine which physical or chemical parameters of water affect the variations in the faecal coliform bacteria concentrations. So the model sensitivity analysis should be done to evaluate the sensitivities of various physico- chemical parameters.

- Previous studies show that the predictive models are based on published research reports, public datasets, and open-source testing data, so it is difficult to check the accuracy of the model. Thus, we cannot rely solely on these studies for bacterial detection.

- Manual counting of viable bacterial cells on agar plates is time-consuming and can be prone to human errors. The method requires experts to identify and count viable cells. Furthermore, due to bacterial overcrowding, high numbers of colony-forming units on a plate will lead to inaccurate results (Breed et al. 1916).

- None of the studies on nails as biomarkers of fluoride exposure have examined the impact of age, gender, and factors affecting the bioavailability of fluoride (Clarkson and McLoughlin 2000). There is a need to study the effect of age, weight, gender, water fluoride, nitrate, turbidity, dissolved oxygen, electrical conductivity, and pH levels on fluoride concentration in nail samples since water characteristics might also impact fluoride.

- Conventional water quality measurement techniques include on-site sampling and subsequent laboratory-based tests; both are labor-intensive and cost-intensive processes. The measurements are not in real-time. Therefore there is a need for real-time monitoring of water quality for drinking applications to reduce labor costs and time usage. Previous studies cannot fulfill the objective of real-time monitoring of water quality parameters. There is a need for a portable system, the output is legible for people with limited or no literacy and it will work in all environmental conditions.

2.12 Objectives of the study

The purpose of this work is to describe the characterization of water quality in Rajasthan state, using the data collected during the year 2019-2021. A total of 1301 groundwater samples are collected from 348 villages and cities in the pre-and post-monsoon seasons. These water samples are tested for various physical, chemical, and microbiological water quality parameters in laboratories at Birla Institute of Technology and Science, Pilani, India. Data from pre and post-monsoon seasons of the study were compared with national water quality standards.

The objective of this study is to develop a model based on laboratory experiments to predict the count of faecal coliform bacteria for cost-effective water quality management. This study would evaluate the accuracy of the modeling approach to predict faecal coliform bacteria concentrations. In recent studies where the process-based models based on laboratory experiments are difficult to develop, the proposed model would provide quick and accurate predictions of bacterial concentrations in water. This work would also include the limited amount of research on the use of data-driven modeling methods for bacteria prediction in water. There are various methods available for fecal coliform identification, which is associated with environmental effects include to the challenges in accurately modeling fecal coliform concentrations in surface waters. According to recent studies, comprehensive process-based models are neither possible nor available for use in site-specific models based on water quality monitoring data. As real-time water quality monitoring and stream gauges become more widely available in watersheds, there is a necessity to find effective models that will be highly beneficial in applications such as analyzing water bodies for current water quality requirements, developing a faecal coliform, giving early warnings to the public, and enforcing groundwater recommendations. It is intended for human contact recreation activities and provides rapid estimations of bacterial contamination. Water infections caused by pathogen exposure in ground water sources continue to be a major public health problem. The ability to swiftly assess bacterial cell quantities will be highly useful in preventing individuals from coming into contact with contaminated water and, therefore, limiting health risks associated with bacterial infection contamination.

The study focuses on developing a laboratory experimental-based model to get the accurate time-based prediction of fecal coliform bacteria in water using the various physical, chemical, and bacteriological parameters. In order to accomplish this objective, the following issues need to be addressed.

- Water Sampling, Testing and Identification of concentration for different physical, chemical and biological water quality parameters in groundwater of Rajasthan.

- To identify the correlations between fecal coliform bacteria (FCB) concentrations and physico-chemical water quality parameters. Sensitivity analysis will be performed to study the importance of different water quality parameters on bacterial concentration.

- To automate the process of bacterial detection using Artificial Intelligence.

- To automate the process of bacterial colony counting using Machine Learning.

- Comparing the experimental laboratory results with the model performance in terms of the predictive ability. So, the error and correlation analysis will be done to get the accurate model and to develop an AI model based on laboratory experimental results to predict fecal coliform bacteria in water.

- Apart from laboratories test, the application of different artificial intelligence (AI) methods will be used to predict fluoride in nails, which will help identify the degree of fluoride exposure to children, females, and males.

- To develop a low-cost system for real-time monitoring of water quality. A system that is portable, output is legible for people with limited or no literacy and it will work in all environmental conditions.

References:

- Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, *12*(1), 1-8.

- Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, *11*(11), 2210.

- Akpata, E. S., Behbehani, J., Akbar, J., Thalib, L., & Mojiminiyi, O. (2014). Fluoride intake from fluids and urinary fluoride excretion by young children in Kuwait: a non-fluoridated community. *Community dentistry and oral epidemiology*, *42*(3), 224-233.

- Alaslani, M. G. (2018). Convolutional neural network based feature extraction for iris recognition. *International Journal of Computer Science & Information Technology (IJCSIT) Vol*, *10*.

- Albaradei, S. A., Napolitano, F., Uludag, M., Thafar, M., Napolitano, S., Essack, M., ... & Gao, X. (2020). Automated counting of colony forming units using deep transfer learning from a model for congested scenes analysis. *IEEE Access*, *8*, 164340-164346.

- Ali, M., & Qamar, A. M. (2013, September). Data analysis, quality indexing and prediction of water quality for the management of rawal watershed in Pakistan. In *Eighth International Conference on Digital Information Management (ICDIM 2013)* (pp. 108-113). IEEE.

- Altaisky, M. V. (2001). Quantum neural network. *arXiv preprint quant-ph/0107012*.

- Amaral, J. G., Freire, I. R., Valle-Neto, E. F., Cunha, R. F., Martinhon, C. C., & Delbem, A. C. (2014). Longitudinal evaluation of fluoride levels in nails of 18–30-month-old children that were using toothpastes with 500 and 1100 μg F/g. *Community dentistry and oral epidemiology*, *42*(5), 412-419.

- Amareshwar, E., & Jahan, S. (2019). Raspberry pi based water quality monitoring and flood alerting system using IoT. *International Journal of Innovative Technology and Exploring Engineering*, *8*(4S2), 237-240.

- Amruta, M. K., & Satish, M. T. (2013, March). Solar powered water quality monitoring system using wireless sensor network. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)* (pp. 281-285). IEEE.

- Antonijevic, E., Mandinic, Z., Curcic, M., Djukic-Cosic, D., Milicevic, N., Ivanovic, M., & Antonijevic, B. (2016). “Borderline” fluorotic region in Serbia: correlations among fluoride in drinking water, biomarkers of exposure and dental fluorosis in schoolchildren. *Environmental geochemistry and health*, *38*(3), 885-896.

- Arrigoni, S., Turra, G., & Signoroni, A. (2017). Hyperspectral image analysis for rapid and accurate discrimination of bacterial infections: A benchmark study. *Computers in biology and medicine*, *88*, 60-71.

- Azeez, D., Ali, M. A. M., Gan, K. B., & Saiboon, I. (2013). Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. *SpringerPlus*, *2*(1), 1-10.

- Barcina, I., & Arana, I. (2009). The viable but nonculturable phenotype: a crossroads in the life-cycle of non-differentiating bacteria?. *Reviews in Environmental Science and Bio/Technology*, *8*(3), 245-255.

- Barer, M. R., Gribbon, L. T., Harwood, C. R., & Nwoguh, C. E. (1993). The viable but non-culturable hypothesis and medical bacteriology. *Reviews in Medical Microbiology*, *4*(4), 183-191.

- Baudišová, D. (1997). Evaluation of Escherichia coli as the main indicator of faecal pollution. *Water Science and Technology*, *35*(11-12), 333-336.

- Bell, A. J., & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision research*, *37*(23), 3327-3338.

- Bell, S., Zitnick, C. L., Bala, K., & Girshick, R. (2016). Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2874-2883).

- Beltran, L. (2006). Nonparametric multivariate statistical process control using principal component analysis and simplicial depth.

- Besnard, V., Federighi, M., Declerq, E., Jugiau, F., & Cappelier, J. M. (2002). Environmental and physico-chemical factors induce VBNC state in Listeria monocytogenes. *Veterinary research*, *33*(4), 359-370.

- Bhandari, N. S., & Nayal, K. (2008). Correlation study on physico-chemical parameters and quality assessment of Kosi river water, Uttarakhand. *Journal of Chemistry*, *5*(2), 342-346.

- BIS, I. (2012). 10500 Indian standard drinking water–specification, second revision. *Bureau of Indian Standards*, *New Delhi*.

- Bisi-Johnson, M. A., Adediran, K. O., Akinola, S. A., Popoola, E. O., & Okoh, A. I. (2017). Comparative physicochemical and microbiological qualities of source and stored household waters in some selected communities in southwestern Nigeria. *Sustainability*, *9*(3), 454.

- Boualam, M., Fass, S., Saby, S., Lahoussine, V., Cavard, J., Gatel, D., & Mathieu, L. (2003). Organic matter quality and survival of coliforms in Low-Nutritive Waters. *Journal-American Water Works Association*, *95*(8), 119-126.

- Boualam, M., Mathieu, L., Fass, S., Cavard, J., & Gatel, D. (2002). Relationship between coliform culturability and organic matter in low nutritive waters. *Water research*, *36*(10), 2618-2626.

- Bouharati, S., Benmahammed, K., Harzallah, D., & El-Assaf, Y. M. (2008). Application of artificial neuro-fuzzy logic inference system for predicting the microbiological pollution in fresh water. *Journal of Applied Sciences*, *8*(2), 309-315.

- Breed, R. S., & Dotterrer, W. D. (1916). The number of colonies allowable on satisfactory agar plates. *Journal of Bacteriology*, *1*(3), 321.

- Brownlee, J. (2019). A gentle introduction to object recognition with deep learning. *Machine Learning Mastery*, *5*.

- Buzalaf, M. A. R., Massaro, C. S., Rodrigues, M. H. C., Fukushima, R., Pessan, J. P., Whitford, G. M., & Sampaio, F. C. (2012). Validation of fingernail fluoride concentration as a predictor of risk for dental fluorosis. *Caries research*, *46*(4), 394-400.
- Buzalaf, M. A. R., Vilhena, F. V., Iano, F. G., Grizzo, L., Pessan, J. P., Sampaio, F. C., & Oliveira, R. C. (2009). The effect of different fluoride concentrations and pH of dentifrices on plaque and nail fluoride levels in young children. *Caries research*, *43*(2), 142-146.

- Caballé-Cervigón, N., Castillo-Sequera, J. L., Gómez-Pulido, J. A., Gómez-Pulido, J. M., & Polo-Luque, M. L. (2020). Machine learning applied to diagnosis of human diseases: A systematic review. *Applied Sciences*, *10*(15), 5135.

- CALP, M. H. (2019). A hybrid ANFIS-GA approach for estimation of regional rainfall amount. *Gazi University Journal of Science*, *32*(1), 145-162.
- Chandaran, U. D., Halim, Z. A., & Sian, L. K. (2012, October). Study on sulfate reducing bacteria detection using Adaptive Neuro-fuzzy Inference System. In *2012 IEEE International Conference on Circuits and Systems (ICCAS)* (pp. 59-64). IEEE.

- Chen, B. H., & Huang, S. C. (2014). An advanced moving object detection algorithm for automatic traffic monitoring in real-world limited bandwidth networks. *IEEE transactions on multimedia*, *16*(3), 837-847.

- Chen, J., Huang, P. S., He, X., Gao, J., & Deng, L. (2016). Unsupervised learning of predictors from unpaired input-output samples. *arXiv preprint arXiv:1606.04646*.

- Cheng, J., Niu, S., & Kim, Y. (2013). Relationship between water quality parameters and the survival of indicator microorganisms–Escherichia coli–in a stormwater wetland. *Water science and technology*, *68*(7), 1650-1656.

- Chigor, V. N., Umoh, V. J., Okuofu, C. A., Ameh, J. B., Igbinosa, E. O., & Okoh, A. I. (2012). Water quality assessment: surface water sources used for drinking and irrigation in Zaria, Nigeria are a public health hazard. *Environmental monitoring and assessment*, *184*(5), 3389-3400.

- Chowdhury, M. A. R., Xu, B., Montilla, R., Hasan, J. A. K., Huq, A., & Colwell, R. R. (1995). A simplified immunofluorescence technique for detection of viable cells of Vibrio cholerae O1 and O139. *Journal of microbiological methods*, *24*(2), 165-170.

- Clarkson, J. J., & McLoughlin, J. (2000). Role of fluoride in oral health promotion. *International dental journal*, *50*(3), 119-128.

- Cook, K. L., & Bolster, C. H. (2007). Survival of Campylobacter jejuni and Escherichia coli in groundwater during prolonged starvation at low temperatures. *Journal of applied microbiology*, *103*(3), 573-583.

- Co-ordination Action Food (CAF) (2007). Methods for detection and molecular characterisation of pathogenic Escherichia coli. In: O'Sullivan J, Bolton DJ, Duffy G, Baylis C, Tozzoli R, Wasteson Y, Lofdahl S (eds.)

- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499-510.

- David, M. M., & Haggard, B. E. (2011). Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois River Watershed, Arkansas and Oklahoma, USA. *Water, Air, & Soil Pollution*, *215*(1), 525-547.

- De Almeida, B. S., da Silva Cardoso, V. E., & Buzalaf, M. A. R. (2007). Fluoride ingestion from toothpaste and diet in 1-to 3-year-old Brazilian children. *Community dentistry and oral epidemiology*, *35*(1), 53-63.

- Demetillo, A. T., Japitana, M. V., & Taboada, E. B. (2019). A system for monitoring water quality in a large aquatic area using wireless sensor network technology. *Sustainable Environment Research*, *29*(1), 1-9.

- Dolezalova, E., & Lukes, P. (2015). Membrane damage and active but nonculturable state in liquid cultures of Escherichia coli treated with an atmospheric pressure plasma jet. *Bioelectrochemistry*, *103*, 7-14.

- Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*(4), 743-761.

- Doran, J. W., & Linn, D. M. (1979). Bacteriological quality of runoff water from pastureland. *Applied and environmental microbiology*, *37*(5), 985-991.

- Durán, C., Ciucci, S., Palladini, A., Ijaz, U. Z., Zippo, A. G., Sterbini, F. P., ... & Cannistraci, C. V. (2021). Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome. *Nature communications*, *12*(1), 1-22.

- Dutka, B. J., & El-Shaarawi, A. (1980). Microbiological water and effluent sample preservation. *Canadian journal of microbiology*, *26*(8), 921-929.

- Elekdag-Turk, S., Almuzian, M., Turk, T., Buzalaf, M. A. R., Alnuaimi, A., Dalci, O., & Darendeliler, M. A. (2019). Big toenail and hair samples as biomarkers for fluoride exposure–a pilot study. *BMC oral health*, *19*(1), 82.

- Ellie, L. B. (2007, October). The correlation of fecal coliform and turbidity of the little Tallapoosa River in the West Georgia Region. In *GSA Denver annual meeting* (pp. 28-31).

- El-Shafie, A., Mukhlisin, M., Najah, A. A., & Taha, M. R. (2011). Performance of artificial neural network and regression techniques for rainfall-runoff prediction. *International Journal of Physical Sciences*, *6*(8), 1997-2003.

- El-Shafie, A., Noureldin, A. E., Taha, M. R., & Basri, H. (2008). Neural network model for Nile river inflow forecasting based on correlation analysis of historical inflow data.

- Fakruddin, M., Mannan, K. S. B., & Andrews, S. (2013). Viable but nonculturable bacteria: food safety and public health perspective. *International Scholarly Research Notices*, *2013*.

- Fausett, L. V. (2006). *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.

- Faustine, A., Mvuma, A. N., Mongi, H. J., Gabriel, M. C., Tenge, A. J., & Kucel, S. B. (2014). Wireless sensor networks for water quality monitoring and control within lake victoria basin: prototype development.

- Ferrari, A., Lombardi, S., & Signoroni, A. (2017). Bacterial colony counting with convolutional neural networks in digital microbiology imaging. *Pattern Recognition*, *61*, 629-640.

- Francy, D. S., Stelzer, E. A., Duris, J. W., Brady, A. M., Harrison, J. H., Johnson, H. E., & Ware, M. W. (2013). Predictive models for Escherichia coli concentrations at inland lake beaches and relationship of model variables to pathogen detection. *Applied and environmental microbiology*, *79*(5), 1676-1688.

- Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5), 826-834.

- Gandomi, A. H., & Yang, X. S. (2011). Benchmark problems in structural optimization. In *Computational optimization, methods and algorithms* (pp. 259-281). Springer, Berlin, Heidelberg.

- Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine pollution bulletin*, *64*(11), 2409-2420.

- Ghezzi, J. I., & Steck, T. R. (1999). Induction of the viable but non-culturable condition in Xanthomonas campestris pv. campestris in liquid microcosms and sterile soil. *FEMS Microbiology Ecology*, *30*(3), 203-208.

- Gopavanitha, K., & Nagaraju, S. (2017, August). A low cost system for real time water quality monitoring and controlling using IoT. In *2017 International conference on energy, communication, data analytics and soft computing (ICECDS)* (pp. 3227-3229). IEEE.

- Gourmelon, M., Cillard, J., & Pommepuy, M. (1994). Visible light damage to Escherichia coli in seawater: oxidative stress hypothesis. *Journal of Applied Bacteriology*, *77*(1), 105-112.

- Greenwood, D., Slack, R., & Peutherer, J. (2003). Escherichia. *Medical Microbiology. 16th ed. Edinburgh: Churchill Livingstone*, 265-273.

- Grey, B., & Steck, T. R. (2001). Concentrations of copper thought to be toxic to Escherichia coli can induce the viable but nonculturable condition. *Applied and Environmental Microbiology*, *67*(11), 5325-5327.

- Grover, L. K. (1996, July). A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* (pp. 212-219).

- Grover, L. K. (1997). Quantum mechanics helps in searching for a needle in a haystack. *Physical review letters*, *79*(2), 325.

- Grubert, J. P. (2003). Acid deposition in the eastern United States and neural network predictions for the future. *Journal of Environmental Engineering and Science*, *2*(2), 99-109.

- Gunda, N. S. K., Gautam, S. H., & Mitra, S. K. (2019). Artificial intelligence based mobile application for water quality monitoring. *Journal of The Electrochemical Society*, *166*(9), B3031.
- Gupta, A., Gupta, A., & Gupta, R. (2018). Power and Area Efficient Intelligent Hardware Design for Water Quality Applications. *Sensors & Transducers*, *227*(11), 67-72.
- Habuda-Stanić, M., Santo, V., Sikora, M., & Benkotić, S. (2013). Microbiological quality of drinking water in public and municipal drinking water supply systems in Osijek-Baranja County, Croatia. *Croatian journal of food science and technology*, *5*(2), 61-69.

- Harmel, D., Wagner, K., Martin, E., Smith, D., Wanjugi, P., Gentry, T., ... & Hendon, T. (2016). Effects of field storage method on E. coli concentrations measured in storm water runoff. *Environmental monitoring and assessment*, *188*(3), 170.

- Harrigan, W. F., & McCance, M. E. (2014). *Laboratory methods in microbiology*. Academic press.

- Hay, E. A., & Parthasarathy, R. (2018). Performance of convolutional neural networks for identification of bacteria in 3D microscopy datasets. *PLoS computational biology*, *14*(12), e1006628.

- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132-7141).

- Huang, L., & Wu, T. (2018). Novel neural network application for bacterial colony classification. *Theoretical Biology and Medical Modelling*, *15*(1), 1-16.

- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215-243.

- Huey, G. M., & Meyer, M. L. (2010). Turbidity as an indicator of water quality in diverse watersheds of the Upper Pecos River Basin. *Water*, *2*(2), 273-284.

- Hughes, A.A.: Influence of seasonal environmental variables on the distribution of pre-sumptive fecal coliforms around an antarctic research station. Applied and Environmental Microbiology. 69, 4884-4891 (2003).

- Huq, A., & Colwell, R. R. (1996). A microbiological paradox: viable but nonculturable bacteria with special reference to Vibrio cholerae. *Journal of food protection*, *59*(1), 96-101.

- Islam, M. M., Hofstra, N., & Islam, M. A. (2017). The impact of environmental variables on faecal indicator bacteria in the Betna river basin, Bangladesh. *Environmental Processes*, *4*(2), 319-332.

- Jones, T., Gill, C. O., & McMullen, L. M. (2004). The behaviour of log phase Escherichia coli at temperatures that fluctuate about the minimum for growth. *Letters in applied microbiology*, *39*(3), 296-300.

- Jothivenkatachalam, K., Nithya, A., & Chandra Mohan, S. (2010). Correlation analysis of drinking water quality in and around Perur block of Coimbatore District, Tamil Nadu, India. *Rasayan Journal of Chemistry*, *3*(4), 649-654.

- Juhna, T., Birzniece, D., & Rubulis, J. (2007). Effect of phosphorus on survival of Escherichia coli in drinking water biofilms. *Applied and environmental microbiology*, *73*(11), 3755-3758.

- Kalantari, N., & Ghafari, S. (2008). Evaluation of toxicity of heavy metals for Escherichia coli growth.

- Kalpana, M. B., & Student, M. T. (2016). Online monitoring of water quality using raspberry Pi3 model B. *International Journal of Innovative Technology and Research*, *4*(6), 4790-4795.

- Kamali, R., & Binesh, A. R. (2013). A comparison of neural networks and adaptive neuro-fuzzy inference systems for the prediction of water diffusion through carbon nanotubes. *Microfluidics and nanofluidics*, *14*(3-4), 575-581.

- Kana, B. D., Gordhan, B. G., Downing, K. J., Sung, N., Vostroktunova, G., Machowski, E. E., ... & Mizrahi, V. (2008). The resuscitation-promoting factors of Mycobacterium tuberculosis are required for virulence and resuscitation from dormancy but are collectively dispensable for growth in vitro. *Molecular microbiology*, *67*(3), 672-684.

- Karim, S., Zhang, Y., Yin, S., Bibi, I., & Brohi, A. A. (2020). A brief review and challenges of object detection in optical remote sensing imagery. *Multiagent and Grid Systems*, *16*(3), 227-243.

- Katip, A. (2018). The usage of artificial neural networks in microbial water quality modeling: a case study from the lake Iznik. *Applied Ecology and Environmental Research*, *16*(4), 3897-3917.

- Keep, N. H., Ward, J. M., Robertson, G., Cohen-Gonsaud, M., & Henderson, B. (2006). Bacterial resuscitation factors: revival of viable but non-culturable bacteria. *Cellular and molecular life sciences*, *63*(22), 2555.

- Keshavarz, Z., & Torkian, H. (2018). Application of ANN and ANFIS models in determining compressive strength of concrete. *Journal of Soft Computing in Civil Engineering*, *2*(1), 62-70.

- Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455-5516.

- Kim, C., & Ndegwa, E. (2018). Influence of pH and temperature on growth characteristics of leading foodborne pathogens in a laboratory medium and select food beverages.

- Kim, C., Hung, Y. C., & Brackett, R. E. (2000). Roles of oxidation–reduction potential in electrolyzed oxidizing and chemically modified water for the inactivation of food-related pathogens. *Journal of food protection*, *63*(1), 19-24.

- Kinjo, Y., & Ueno, K. (2011). iNKT cells in microbial immunity: recognition of microbial glycolipids. *Microbiology and immunology*, *55*(7), 472-482.

- Kolling, G. L., & Matthews, K. R. (2001). Examination of recovery in vitro and in vivo of nonculturable Escherichia coli O157: H7. *Applied and environmental microbiology*, *67*(9), 3928-3933.

- Kreske, A. C., Bjornsdottir, K., Breidt Jr, F., & Hassan, H. (2008). Effects of pH, dissolved oxygen, and ionic strength on the survival of Escherichia coli O157: H7 in organic acid solutions. *Journal of food protection*, *71*(12), 2404-2409.

- Krishnamurti, K., & Kate, S. R. (1951). Changes in electrical conductivity during bacterial growth. *Nature*, *168*(4265), 170.

- Kulkarni, A., & Callan, J. (2015). Selective search: Efficient and effective search of large textual collections. *ACM Transactions on Information Systems (TOIS)*, *33*(4), 1-33.

- Kumar, M. J. V., & Samalla, K. (2019). Design and development of water quality monitoring system in IoT. *International Journal of Recent Technology and Engineering (IJRTE)*, *7*, 527-533.

- Leclair, R. M., McLean, S. K., Dunn, L. A., Meyer, D., & Palombo, E. A. (2019). Investigating the effects of time and temperature on the growth of Escherichia coli O157: H7 and Listeria monocytogenes in raw cow's milk based on simulated consumer food handling practices. *International journal of environmental research and public health*, *16*(15), 2691.

- Lee, D. W., Gwack, J., & Youn, S. K. (2012). Enteropathogenic Escherichia coli Outbreak and its Incubation Period: Is it Short or Long?. *Osong public health and research perspectives*, *3*(1), 43-47.

- Levy, F. M., Bastos, J. R. D. M., & Buzalaf, M. A. R. (2004). Nails as biomarkers of fluoride in children of fluoridated communities. *Journal of dentistry for children*, *71*(2), 121-125.

- Li, L., Mendis, N., Trigui, H., Oliver, J. D., & Faucher, S. P. (2014). The importance of the viable but non-culturable state in human bacterial pathogens. *Frontiers in microbiology*, *5*, 258.

- Li, Y., Li, W., Xiong, J., Xia, J., & Xie, Y. (2020). Comparison of Supervised and Unsupervised Deep Learning Methods for Medical Image Synthesis between Computed Tomography and Magnetic Resonance Images. *BioMed Research International*, *2020*.

- Lima-Arsati, Y. B. O., Martins, C. C., Rocha, L. A., & Cury, J. A. (2010). Fingernail may not be a reliable biomarker of fluoride body burden from dentifrice. *Brazilian dental journal*, *21*(2), 91-97.

- Lin, B., Kashefipour, S. M., & Falconer, R. A. (2003). Predicting near-shore coliform bacteria concentrations using ANNS. *Water Science and technology*, *48*(10), 225-232.

- Lin, J. S., & Liu, C. Z. (2008, November). A monitoring system based on wireless sensor network and an SoC platform in precision agriculture. In *2008 11th IEEE International Conference on Communication Technology* (pp. 101-104). IEEE.

- Linder, K. A. T. H. E. R. I. N. E., & Oliver, J. D. (1989). Membrane fatty acid and virulence changes in the viable but nonculturable state of Vibrio vulnificus. *Applied and Environmental Microbiology*, *55*(11), 2837-2842.

- Linhares, D. P. S., Garcia, P. V., Amaral, L., Ferreira, T., Cury, J. A., Vieira, W., & dos Santos Rodrigues, A. (2016). Sensitivity of two biomarkers for biomonitoring exposure to fluoride in children and women: A study in a volcanic area. *Chemosphere*, *155*, 614-620.

- Liong, S. Y., & Sivapragasam, C. (2002). Flood stage forecasting with support vector machines 1. *JAWRA Journal of the American Water Resources Association*, *38*(1), 173-186.

- Liu, Y., Wang, C., Tyrrell, G., & Li, X. F. (2010). Production of Shiga-like toxins in viable but nonculturable Escherichia coli O157: H7. *Water Research*, *44*(3), 711-718.

- Lonsane, B. K., Parhad, N. M., & Rao, N. U. (1967). Effect of storage temperature and time on the coliforms in water samples. *Water Research*, *1*(4), 309-316.

- Lou, W., & Nakai, S. (2001). Application of artificial neural networks for predicting the thermal inactivation of bacteria: a combined effect of temperature, pH and water activity. *Food Research International*, *34*(7), 573-579.

- Lu WZ, Wang WJ, Wang XK, Xu ZB, Leung AY. 2003. Using improved neural network model to analyze RSP, NO x and NO 2 levels in urban air in Mong Kok, Hong Kong. Environmental monitoring and assessment. 87(3):235-54.

- Lundquist C.J. (2020). National Institute of Water and Atmospheric Research, New Zealand.

- Maalej, S., Gdoura, R., Dukan, S., Hammami, A., & Bouain, A. (2004). Maintenance of pathogenicity during entry into and resuscitation from viable but nonculturable state in Aeromonas hydrophila exposed to natural seawater at low temperature. *Journal of Applied Microbiology*, *97*(3), 557-565.

- Maas, M. B., Perold, W. J., & Dicks, L. M. T. (2017). Biosensors for the detection of Escherichia coli. *Water Sa*, *43*(4), 707-721.

- Maier, A., Krolik, J., Fan, S., Quintin, P., McGolrick, D., Joyce, A., & Majury, A. (2015). Evaluating appropriate maximum holding times for private well water samples. *Environmental Health Review*, *58*(2), 35-40.

- Makarynska, D., & Makarynskyy, O. (2008). Predicting sea-level variations at the Cocos (Keeling) Islands with artificial neural networks. *Computers & Geosciences*, *34*(12), 1910-1917.

- Makino, S. I., Kii, T., Asakura, H., Shirahata, T., Ikeda, T., Takeshi, K., & Itoh, K. (2000). Does enterohemorrhagic Escherichia coli O157: H7 enter the viable but nonculturable state in salted salmon roe?. *Applied and Environmental Microbiology*, *66*(12), 5536-5539.

- Mary, P., Chihib, N. E., Charafeddine, O., Defives, C., & Hornez, J. P. (2002). Starvation survival and viable but nonculturable states in Aeromonas hydrophila. *Microbial ecology*, 250-258.

- Mascher, F., Hase, C., Moënne-Loccoz, Y., & Défago, G. (2000). The viable-but-nonculturable state induced by abiotic stress in the biocontrol agent Pseudomonas fluorescens CHA0 does not promote strain persistence in soil. *Applied and environmental microbiology*, *66*(4), 1662-1667.

- Mason, D., Cooper, K., Drysdale, J., Dyer, A., Meehan, E., Stanley, M. and Lorentz, P. (2015). Effects of holding time and temperature on *E. Coli* and total coliforms in surface water samples, 28 th Annual FLRC Workshop at Massey University on the 10th, 11th and 12th February 2015.

- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133.

- McDaniels, A. E., Bordner, R. H., Gartside, P. S., Haines, J. R., Brenner, K. P., & Rankin, C. C. (1985). Holding effects on coliform enumeration in drinking water samples. *Applied and Environmental Microbiology*, *50*(4), 755-762.

- McKay, A. M. (1992). Viable but non-culturable forms of potentially pathogenic bacteria in water. *Letters in Applied Microbiology*, *14*(4), 129-135.

- Medema, G. J., Shaw, S., Waite, M., Snozzi, M., Morreau, A., & Grabow, W. (2003). Catchment characterisation and source water quality. *Assessing Microbial Safety of Drinking Water*, *4*, 111-158.

- Menon, K. U., Divya, P., & Ramesh, M. V. (2012, July). Wireless sensor network for river water quality monitoring in India. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)* (pp. 1-7). IEEE.

- Minu, M. S., Kumari, P., Singh, A. K., & Singh, A. (2019). Wired Sensor Systems for Water Quality Monitoring. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(4), 847-852.

- Mizunoe, Y., Wai, S. N., Takade, A., & Yoshida, S. I. (1999). Restoration of culturability of starvation-stressed and low-temperature-stressed Escherichia coli O157 cells by using H2O2-degrading compounds. *Archives of Microbiology*, *172*(1), 63-67.

- Moazenzadeh, R., Mohammadi, B., Shamshirband, S., & Chau, K. W. (2018). Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran. *Engineering Applications of Computational Fluid Mechanics*, *12*(1), 584-597.

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

- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in plant science*, *7*, 1419.

- Motion, E. Q. C. (2009). Department of Environmental Quality Memorandum.

- Mouna, H., Ahmed, A., & Omar, A. (2014). An evaluation of environmental factors affecting the survival of Escherichia coli in coastal area, Oualidia Lagoon. *International Journal of Current Microbiology and Applied Science*, *3*(10), 710-721.

- Muela, A., Seco, C., Camafeita, E., Arana, I., Orruño, M., López, J. A., & Barcina, I. (2008). Changes in Escherichia coli outer membrane subproteome under environmental conditions inducing the viable but nonculturable state. *FEMS microbiology ecology*, *64*(1), 28-36.

- Mulla, K. R., & Bhosale, M. S. (2016). Water quality Analysis and simulation of Panchaganga River using Matlab. *International Journal of Engineering Sciences & Research Technology*, *5*(8), 613-620.

- Munna, M. S., Nur, I. T., Rahman, T., & Noor, R. (2013). Influence of exogenous oxidative stress on Escherichia coli cell growth, viability and morphology. *Am J Biosci*, *1*(4), 59.

- Munna, M. S., Tamanna, S., Afrin, M. R., Sharif, G. A., Mazumder, C., Kana, K. S., ... & Noor, R. (2014). Influence of aeration speed on bacterial colony forming unit (CFU) formation capacity. *Am J Microbiol Res*, *2*(1), 47-51.

- Murata, M., Noor, R., Nagamitsu, H., Tanaka, S., & Yamada, M. (2012). Novel pathway directed by σE to cause cell lysis in Escherichia coli. *Genes to Cells*, *17*(3), 234-247.

- Muttil, N., & Chau, K. W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, *28*(3-4), 223-238.

- Na, S. H., Miyanaga, K., Unno, H., & Tanji, Y. (2006). The survival response of Escherichia coli K12 in a natural environment. *Applied microbiology and biotechnology*, *72*(2), 386-392.

- Najah, A., Elshafie, A., Karim, O. A., & Jaffar, O. (2009). Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of scientific research*, *28*(3), 422-435.

- Nehal, S. A., Roy, D., Devi, M., & Srinivas, T. (2019). Highly sensitive lab-on-chip with deep learning AI for detection of bacteria in water. *International Journal of Information Technology*, 1-7.

- Nevers, M. B., & Boehm, A. B. (2010). Modeling fate and transport of fecal bacteria in surface water. *The fecal bacteria*, 165-188.

- Nilsson, L., Oliver, J. D., & Kjelleberg, S. (1991). Resuscitation of Vibrio vulnificus from the viable but nonculturable state. *Journal of bacteriology*, *173*(16), 5054-5059.

- Nitta, T., Nagamitsu, H., Murata, M., Izu, H., & Yamada, M. (2000). Function of the ςE regulon in dead-cell lysis in stationary-phase Escherichia coli. *Journal of Bacteriology*, *182*(18), 5231-5237.

- Noor, R., Islam, Z., Munshi, S. K., & Rahman, F. (2013). Influence of temperature on Escherichia coli growth in different culture media. *J Pure Appl Microbiol*, *7*(2), 899-904.

- Noor, R., Murata, M., & Yamada, M. (2009). Oxidative stress as a trigger for growth phase-specific σE-dependent cell lysis in Escherichia coli. *Journal of molecular microbiology and biotechnology*, *17*(4), 177-187.

- Noureldin, A., El-Shafie, A., & Bayoumi, M. (2011). GPS/INS integration utilizing dynamic neural networks for vehicular navigation. *Information fusion*, *12*(1), 48-57.

- Oliver, J. D. (1995). The viable but non-culturable state in the human pathogen Vibrio vulnificus. *FEMS microbiology letters*, *133*(3), 203-208.

- Oliver, J. D. (2000). The viable but nonculturable state and cellular resuscitation. *Microbial biosystems: new frontiers. Atlantic Canada Society for Microbial Ecology, Halifax, Canada*, 723-730.

- Oliver, J. D. (2005). The viable but nonculturable state in bacteria. *Journal of microbiology*, *43*(spc1), 93-100.

- Oliver, J. D. (2010). Recent findings on the viable but nonculturable state in pathogenic bacteria. *FEMS microbiology reviews*, *34*(4), 415-425.

- Oliver, J. D., Dagher, M., & Linden, K. (2005). Induction of Escherichia coli and Salmonella typhimurium into the viable but nonculturable state following chlorination of wastewater. *Journal of water and health, 3*(3), 249-257.

- Oliver, J. D., Hite, F., McDougald, D., Andon, N. L., & Simpson, L. M. (1995). Entry into, and resuscitation from, the viable but nonculturable state by Vibrio vulnificus in an estuarine environment. *Applied and Environmental Microbiology, 61*(7), 2624-2630.

- Pachepsky, Y., Kierzewski, R., Stocker, M., Sellner, K., Mulbry, W., Lee, H., & Kim, M. (2018). Temporal stability of Escherichia coli concentrations in waters of two irrigation ponds in Maryland. *Appl. Environ. Microbiol.*, *84*(3), e01876-17.

- Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, *36*(1), 2-17.

- Panella, M., & Martinelli, G. (2011). Neural networks with quantum architecture and quantum learning. *International Journal of Circuit Theory and Applications*, *39*(1), 61-77.

- Park, H., Hung, Y. C., & Chung, D. (2004). Effects of chlorine and pH on efficacy of electrolyzed water for inactivating Escherichia coli O157: H7 and Listeria monocytogenes. *International journal of food microbiology*, *91*(1), 13-18.

- Pawlowski, D. R., Metzger, D. J., Raslawsky, A., Howlett, A., Siebert, G., Karalus, R. J., ... & Whitehouse, C. A. (2011). Entry of Yersinia pestis into the viable but nonculturable state in a low-temperature tap water microcosm. *PLoS One*, *6*(3), e17585.

- Pessan, J. P., Pin, M. L. G., Martinhon, C. C. R., Silva, S. M. B. D., Granjeiro, J. M., & Buzalaf, M. A. R. (2005). Analysis of fingernails and urine as biomarkers of fluoride exposure from dentifrice and varnish in 4-to 7-year-old children. *Caries research*, *39*(5), 363-370.

- Pommepuy, M., Butin, M., Derrien, A., Gourmelon, M., Colwell, R. R., & Cormier, M. (1996). Retention of enteropathogenicity by viable but nonculturable Escherichia coli exposed to seawater and sunlight. *Applied and Environmental Microbiology*, *62*(12), 4621-4626.

- Pope, M. L., Bussen, M., Feige, M. A., Shadix, L., Gonder, S., Rodgers, C., ... & Standridge, J. (2003). Assessment of the effects of holding time and temperature on Escherichia coli densities in surface water samples. *Applied and Environmental Microbiology*, *69*(10), 6201-6207.

- Puneeth, K. M., Bipin, S., Prasad, C., Kumar, R. J., & Urs, M. K. (2018, May). Real-time Water Quality Monitoring using WSN. In *2018 3rd IEEE International*

Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) (pp. 1152-1156). IEEE.

- Rahman, M. H., Suzuki, S., & Kawai, K. (2001). Formation of viable but non-culturable state (VBNC) of Aeromonas hydrophila and its virulence in goldfish, Carassius auratus. *Microbiological research*, *156*(1), 103-106.

- Ramamurthy, T., Ghosh, A., Pazhani, G. P., & Shinoda, S. (2014). Current perspectives on viable but non-culturable (VBNC) pathogenic bacteria. *Frontiers in public health*, *2*, 103.

- Ranković, V., Radulović, J., Radojević, I., Ostojić, A., & Čomić, L. (2010). Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. *Ecological Modelling*, *221*(8), 1239-1244.

- Rao, G., Eisenberg, J. N., Kleinbaum, D. G., Cevallos, W., Trueba, G., & Levy, K. (2015). Spatial variability of Escherichia coli in rivers of northern coastal Ecuador. *Water*, *7*(2), 818-832.

- Ravel, J., Knight, I. T., Monahan, C. E., Hill, R. T., & Colwell, R. R. (1995). Temperature-induced recovery of Vibrio cholerae from the viable but nonculturable state: growth or resuscitation?. *Microbiology*, *141*(2), 377-383.

- Rice, E. W., Geldreich, E. E., & Read, E. J. (1989). The presence-absence coliform test for monitoring drinking water quality. *Public Health Reports*, *104*(1), 54.

- Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., ... & Cohen, M. L. (1983). Hemorrhagic colitis associated with a rare Escherichia coli serotype. *New England Journal of Medicine*, *308*(12), 681-685.

- Rompré, A., Servais, P., Baudart, J., De-Roubin, M. R., & Laurent, P. (2002). Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. *Journal of microbiological methods*, *49*(1), 31-54.

- Roslev, P., Bjergbæk, L., & Hesselsoe, M. (2004). Effect of oxygen on survival of faecal pollution indicators in drinking water. *Journal of applied microbiology*, *96*(5), 938-945.

- Roszak, D. B., & Colwell, R. R. (1987). Survival strategies of bacteria in the natural environment. *Microbiological reviews*, *51*(3), 365-379.

- Ryan, F. J. (2019). Application of machine learning techniques for creating urban microbial fingerprints. *Biology direct*, *14*(1), 1-13.

- Sarrafzadeh, O., Dehnavi, A. M., Rabbani, H., & Talebi, A. (2015, October). A simple and accurate method for white blood cells segmentation using K-means algorithm. In *2015 IEEE Workshop on Signal Processing Systems (SiPS)* (pp. 1-6). IEEE.

- Saxena, T., Kaushik, P., & Mohan, M. K. (2015). Prevalence of E. coli O157: H7 in water sources: an overview on associated diseases, outbreaks and detection methods. *Diagnostic microbiology and infectious disease*, *82*(3), 249-264.

- Seo, M., Lee, H., & Kim, Y. (2019). Relationship between Coliform Bacteria and Water Quality Factors at Weir Stations in the Nakdong River, South Korea. *Water*, *11*(6), 1171.

- Shamsudin, S. N., Rahman, M. H. F., Taib, M. N., Razak, W. R. W. A., Ahmad, A. H., & Zain, M. M. (2016, August). Analysis between Escherichia Coli growth and physical parameters in water using Pearson correlation. In *2016 7th IEEE Control and System Graduate Research Colloquium (ICSGRC)* (pp. 131-136). IEEE.

- Shinde, R. L., & Khadse, K. G. (2009). Multivariate process capability using principal component analysis. *Quality and Reliability Engineering International*, *25*(1), 69-77.

- Shroff, P., Vashi, R. T., Champaneri, V. A., & Patel, K. K. (2015). Correlation study among water quality parameters of groundwater of Valsad district of south Gujarat (India). *Journal of Fundamental and Applied Sciences*, *7*(3), 340-349.

- Silva, A., de Oliveira, W., & Ludermir, T. (2010, October). A weightless neural node based on a probabilistic quantum memory. In *2010 Eleventh Brazilian Symposium on Neural Networks* (pp. 259-264). IEEE.

- Sinaga, D. M., Robson, M. G., Gasong, B. T., Halel, A. G., & Pertiwi, D. (2016). Fecal coliform bacteria and factors related to its growth at the Sekotong shallow wells (West Nusa Tenggara, Indonesia). *Public Health of Indonesia*, *2*(2), 47-54.

- Singh, A. K., Das, S., Singh, S., Pradhan, N., Gajamer, V. R., Kumar, S., ... & Tiwari, H. K. (2019). Physicochemical Parameters and Alarming Coliform Count of the Potable Water of Eastern Himalayan State Sikkim: An Indication of Severe Fecal Contamination and Immediate Health Risk. *Frontiers in public health*, *7*, 174.

- Sousa, E. T. D., Alves, V. F., Maia, F. B. M., Nobre-dos-Santos, M., Forte, F. D. S., & Sampaio, F. C. (2018). Influence of fluoridated groundwater and 1,100 ppm fluoride dentifrice on biomarkers of exposure to fluoride. *Brazilian dental journal*, *29*(5), 475-482.

- Stojanovic, B., & Neskovic, A. (2012, November). Impact of PCA based fingerprint compression on matching performance. In *2012 20th Telecommunications Forum (TELFOR)* (pp. 693-696). IEEE.

- Tamerat N, Muktar Y, Shiferaw D (2016) Application of molecular diagnostic techniques for the detection of E. coli O157: H7: a review. J Vet Sci Technol 7(362):1–9.

- Than, A. A. (2011). Effect of temperatures on the growth of Escherichia coli from water. *Universities research journal*, *4*(2), 163-171.

- Torelli, A., Wolf, I., & Gretz, N. (2018). AutoCellSeg: robust automatic colony forming unit (CFU)/cell analysis using adaptive image segmentation and easy-to-use post-editing techniques. *Scientific reports*, *8*(1), 1-10.

- Tufail, M., Ormsbee, L., & Teegavarapu, R. (2008). Artificial intelligence-based inductive models for prediction and classification of fecal coliform in surface waters. *Journal of Environmental Engineering*, *134*(9), 789-799.

- Turra, G., Conti, N., & Signoroni, A. (2015, August). Hyperspectral image acquisition and analysis of cultured bacteria for the discrimination of urinary tract infections. In *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 759-762). IEEE.

- United States Environmental Protection Agency (US EPA). (2010). EPA Microbiological Alternate Test Procedure (ATP) Protocol for Drinking Water, Ambient Water, Wastewater, and Sewage Sludge Monitoring Methods.

- US EPA, Office of Water. (2001). Source water protection practices bulletin managing septic systems to prevent contamination of drinking water. EPA 816-F-01-021.

- Vijayakumar, N., & Ramya, A. R. (2015, March). The real time monitoring of water quality in IoT environment. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1-5). IEEE.

- Vijayashanthar, V., Qiao, J., Zhu, Z., Entwistle, P., & Yu, G. (2018). Modeling fecal indicator bacteria in urban waterways using artificial neural networks. *Journal of Environmental Engineering*, *144*(6), 05018003.

- Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*(2), 137-154.

- Wang, H., Koydemir, H. C., Qiu, Y., Bai, B., Zhang, Y., Jin, Y., & Ozcan, A. (2020). Early detection and classification of live bacteria using time-lapse coherent imaging and deep learning. *Light: Science & Applications*, *9*(1), 1-17.

- Wang, J., Chen, K., Yang, S., Loy, C. C., & Lin, D. (2019). Region proposal by guided anchoring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2965-2974).

- Whitford, G. M., Sampaio, F. C., Arneberg, P., & Von der Fehr, F. R. (1999). Fingernail fluoride: a method for monitoring fluoride exposure. *Caries research*, *33*(6), 462-467.

- Wu, H., Zhang, J., Mi, Z., Xie, S., Chen, C., & Zhang, X. (2015). Biofilm bacterial communities in urban drinking water distribution systems transporting waters with different purification strategies. *Applied microbiology and biotechnology*, *99*(4), 1947-1955.

- Xu, H. S., Roberts, N., Singleton, F. L., Attwell, R. W., Grimes, D. J., & Colwell, R. R. (1982). Survival and viability of nonculturable Escherichia coli and Vibrio cholerae in the estuarine and marine environment. *Microbial ecology*, *8*(4), 313-323.

- Xu, X., Li, H., Yin, F., Xi, L., Qiao, H., Ma, Z., ... & Ma, X. (2020). Wheat ear counting using K-means clustering segmentation and convolutional neural network. *Plant Methods*, *16*(1), 1-13.

- Yang, X. S. (2011). Metaheuristic optimization. *Scholarpedia*, *6*(8), 11472.

- Zamani, M. A. T., & Saybani, M. ARTIFICIAL NEURAL NETWORK MODEL FOR PREDICTION OF ENVIRONMENTAL STATUS OF URBAN CATCHMENT OF PENCHALA RIVER, KUALA LUMPUR, MALAYSIA.

- Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1624-1639.

- Zhou, R., & Ding, Q. (2007). Quantum mp neural network. *International Journal of Theoretical Physics*, *46*(12), 3209-3215.

- Zhu, W., Huang, W., Lin, Z., Yang, Y., Huang, S., & Zhou, J. (2016). Data and feature mixed ensemble based extreme learning machine for medical object detection and segmentation. *Multimedia Tools and Applications*, *75*(5), 2815-2837.

- Zieliński, B., Plichta, A., Misztal, K., Spurek, P., Brzychczy-Włoch, M., & Ochońska, D. (2017). Deep learning approach to bacterial colony classification. *PloS one*, *12*(9), e0184554.

3. Analysis of increase in cell counts of *E. coli* in groundwater of Rajasthan: Possible presence of VBNC cells

E. coli bacteria are associated with the coliform group and is a more precise indicator of faecal contamination than other coliform bacteria; its existence indicates the possible presence of harmful disease-causing bacteria. This chapter discusses the water sampling and laboratory testing for the study. A study is carried out to analyze the groundwater quality of the Rajasthan region. The experimental laboratory results are synthesized to test the physical, chemical, and microbiological parameters of water. Detection of E. coli bacteria in groundwater samples is performed. The detection of waterborne bacteria is crucial to prevent health risks. We developed an automated process of waterborne bacterial detection using AI.

3.1 Introduction

Water pollution is one of the most critical challenges for sustainable development. Water quality measurement is an important stepping stone towards finding a solution to this problem. Water quality parameters are currently measured using laboratory testing methods. The standard laboratory sensors are stationary and water samples are brought in from the field for analysis. The monitoring of bacteriological drinking water quality relies mainly on the study of indicator bacteria. *E. coli* is a more precise indicator of water contamination than other fecal coliform bacteria due to the advancement in testing methods. *E. coli* bacteria can be identified in the laboratory using conventional methods (Co-ordination Action Food, 2007), enzymatic methods (Co-ordination Action Food, 2007), molecular methods (Tamerat et al., 2016; Saxena et al., 2015), and biosensor-based methods (Maas et al., 2017).

3.2 Study Area

The study was carried out in Rajasthan, India (Figure 3.1). This study focused on costs and remediation of groundwater contamination in India, with an emphasis on Rajasthan. During the years 2019–2021, 1,301 groundwater samples were collected from 348 villages and towns. The water samples were examined in the laboratory with different physical, chemical, and microbiological water quality tests utilizing titration and spectroscopy. Laboratory testing was carried out at the environmental engineering laboratory in the department of civil engineering at BITS Pilani, Rajasthan. The water quality parameters are as follows; pH, total dissolved solids (TDS, mg/l), oxidation-reduction potential (ORP, mg/l), dissolved oxygen (DO, mg/l), electrical conductivity (EC, s/m), turbidity (NTU), fluoride (mg/l), and nitrate (mg/l) are measured in the laboratory using the titration and spectroscopy method.

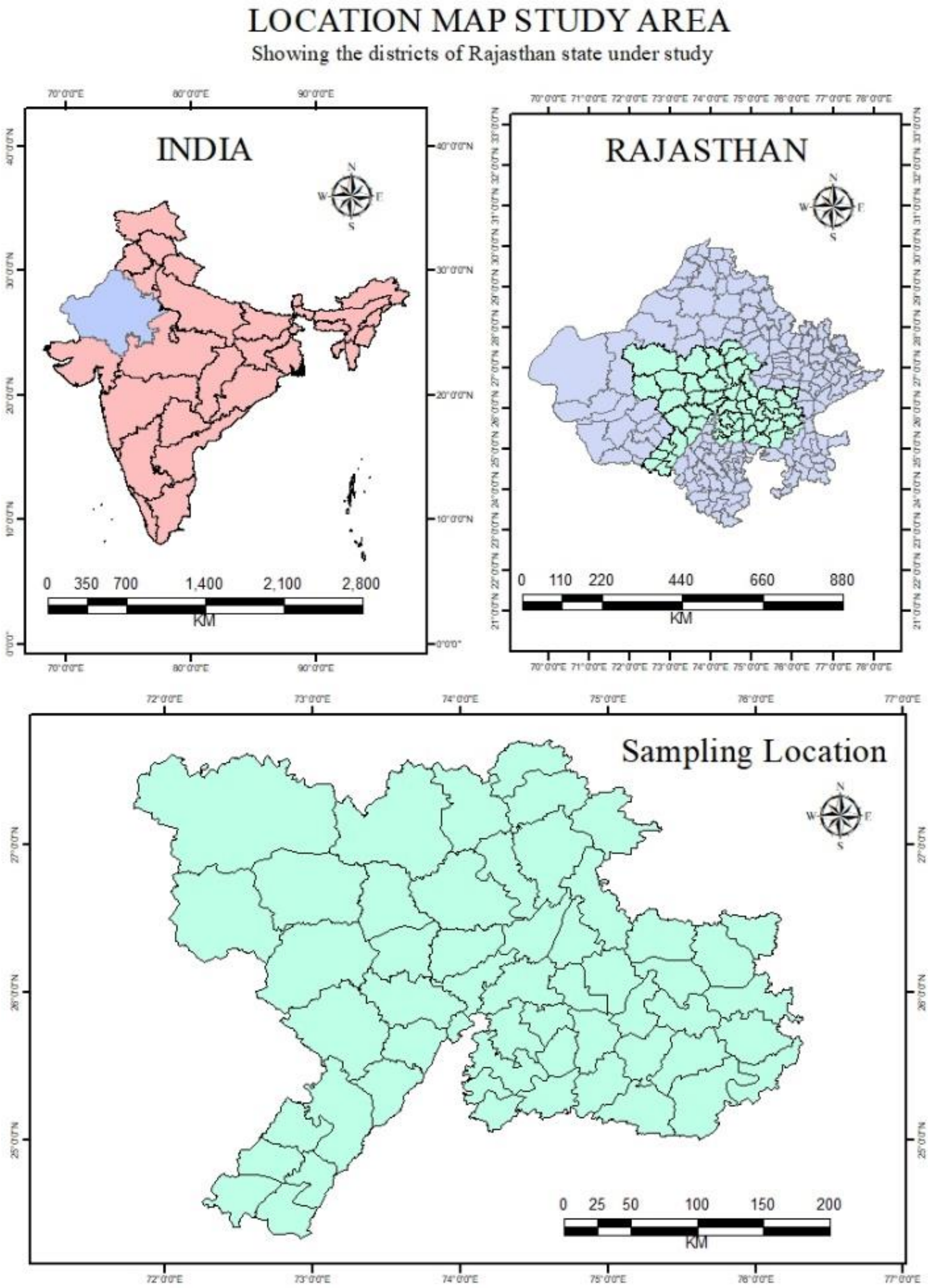


Figure 3.1: Location of the study area.

3.3 Methodology

The following methodology has been adopted to characterize the groundwater:

- Physico-chemical analysis of water samples for identification pH, total dissolved solids, oxidation-reduction potential, dissolved oxygen, electrical conductivity, turbidity, fluoride and nitrate concentration in groundwater of Rajasthan.
- Detection of the presence or absence of bacteria in groundwater samples is performed.
- Identification of bacteria present in groundwater samples.
- Viable count.
- Detection of *E. coli* bacteria using Adaptive Neuro-Fuzzy Inference System.

3.4 Groundwater quality

The groundwater samples were collected from eight districts of Rajasthan, India, under the BITS-UVA (University of Virginia) groundwater contamination project, containing 1302 water samples used in this study. Microbiological water quality analysis was performed to identify the bacteria present in water using the gram staining culturing method. After identification, a viable count of *E. coli* bacteria was performed to count the number of actively growing bacterial cells in terms of colony-forming units (CFU). The laboratory testing is carried out at Environmental Engineering Lab, Department of Civil Engineering, BITS Pilani, Rajasthan. Table 3.1 provides the summary of the water quality parameters of the groundwater samples.

Table 3.1: Summary of water quality parameters of the groundwater samples.

Parameter	Minimum	Maximum
pH	5.5	9.65
Electrical conductivity (µs/cm)	0	8.81
Total dissolved solids (TDS, mg/l)	42	3820
Dissolved oxygen (DO, mg/l)	16	1.7
Oxidation-reduction potential (ORP, mg/l)	4	760
Turbidity (NTU)	0	62
Fluoride (mg/l)	0.007	4.696
Nitrate (mg/l)	0.71	357.678

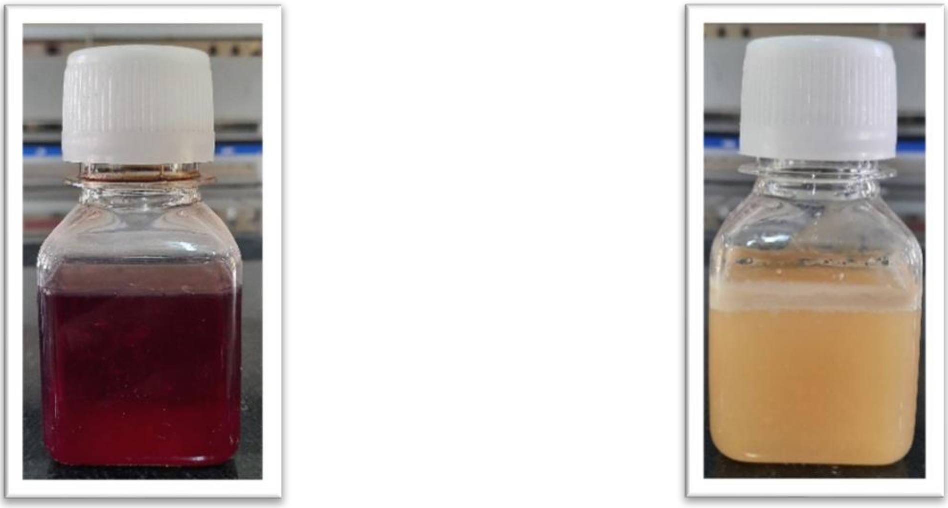
The standards established by the Bureau of Indian standard (IS 10500: 2012) were used to characterize the groundwater of the study area. The various standards adopted in India are given in Table 3.2 (IS 10500: 2012).

Table 3.2: Standard limits for physico-chemical parameters under study.

Parameter	Acceptable Limit
pH	6.5-8.5
Electrical conductivity (s/cm)	0.005-0.05
Total dissolved solids (TDS, mg/l)	500-2000
Dissolved oxygen (DO, mg/l)	4-6
Oxidation-reduction potential (ORP, mg/l)	200-600
Turbidity (NTU)	5-10
Nitrate	0-45
Fluoride	0-1.0

3.4.1 Present/Absent test (PA test)

The enzymatic method is used in this study to determine the presence or absence of *E. coli* bacteria in groundwater samples (Olstadt et al., 2007). Present/Absent test is a substrate method developed to overcome some constraints of the multiple tube fermentation method (Oshiro 2002) and membrane filter method (Jagals et al. 2000). The detection method is based on the concept that only *E. coli* bacteria are fed. No substrate is given for other bacteria. Firstly 100 ml water sample is added to the sterile disposable bottle. The powder medium (PA broth) is then swirled into water so that it gets dissolved completely. Once dissolved, water samples can be incubated for 24-48 hours at 35 °C. After the incubation period, the transition in the colour of the medium from reddish-purple (Figure 3.2a) to yellow (Figure 3.2b) indicates the presence of *E. coli*. Figure 3.2 shows the change in colour of the culture medium due to the presence of bacteria.



a) The initial state of culture medium b) Final state of culture medium (if bacteria is present)

Figure 3.2: Colour change in culture medium due to the presence of bacteria.

3.4.2 Identification of bacteria

The most significant bacteriological task is to classify water-borne pathogens. Generally, bacteria display three basic shapes: round, rod-shaped, and spiral. After water samples are collected, bacteria must be grown on culture media for identification. Gram staining is the first step towards identifying bacteria. Staining is a method used for the differentiation of bacteria in the cell wall based on their different constituents. By coloring these cells violet or red, the gram staining method categorizes bacteria into two classes: gram-positive and gram-negative.

Eosin methylene blue (EMB) agar is a selective and differential medium used to isolate fecal coliform bacteria. It provides a rapid and accurate method of differentiating *E. coli* from other gram-negative pathogens. *E. coli* bacteria is an indicator of fecal contamination in water. The presence of *E. coli* bacteria indicates the possibility of the presence of pathogenic bacteria and viruses (Khan et al. 2020). Nobody can ferment lactose except *E. coli*. If *E. coli* bacteria are present in water. In this case, a colony will appear on an agar plate with a metallic sheen with a dark center. Gram-positive bacteria growth is typically hindered on EMB agar because of the toxicity of the methylene blue dye. Therefore, only colonies of *E. coli* will appear on agar plates. If no colony appears on the agar plates, it indicates that *E. coli* bacteria are absent in water. Consequently, it can be concluded that only *E. coli* bacteria will grow on agar plates; gram-positive bacteria will not grow on agar plates, so this method is only valid for *E. coli* bacteria. Figure 3.3 shows a petri dish containing *E. coli* bacteria.

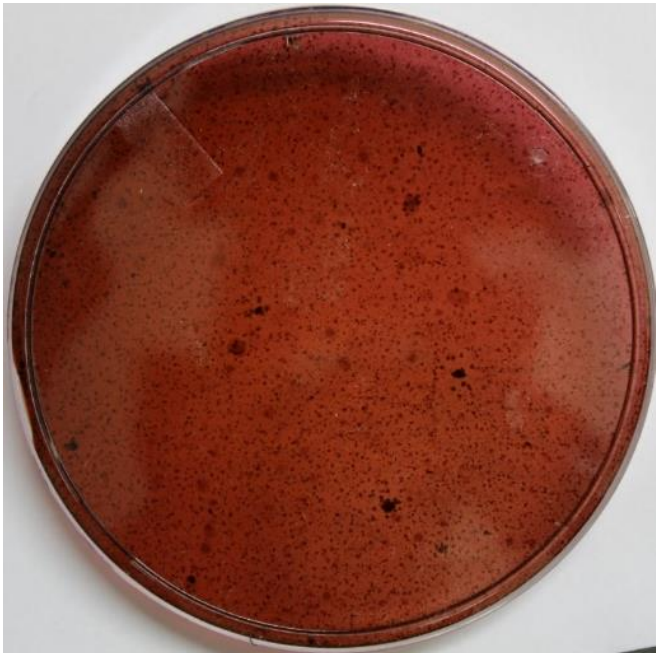


Figure 3.3: Petri dish containing *E. coli* bacteria.

3.4.3 Viable count

Viable cell counts were performed using the plate count method (USEPA 2002). EMB agar (Leininger et al. 2001) was used as a growth media for the identification of *E. coli*. Using 1-mL water samples, serial dilution was performed so that dilution two had a concentration one-tenth that of dilution one and one hundredth that of the water sample. Next, 20 mL of molten cooled agar solution and diluted water samples were mixed well and poured into a sterile petri dish with a diameter of 90 mm. The agar plates were placed in an incubator at 35°C for 24–48 h to distribute the colonies throughout the depth of the medium. Colony-forming units present in the petri dish were counted using a microscope at 10× magnification. The colony-forming units present in a water sample can be determined by multiplying the number of colonies present on the agar plate by the sample's dilution factor (Bartram et al. 1996), as shown in Equation (3.1)

CFU/mL = number of colonies * dilution factor (3.1)

The viable count analysis of the water samples showed *E. coli* bacterial strains with minimum cell counts of $4{\times}10^7$ CFU/100 mL and maximum cell counts of 132×10^7 CFU/100 mL, as shown in Figure 3.4. A total of 99 groundwater samples were found positive for *E. coli*.

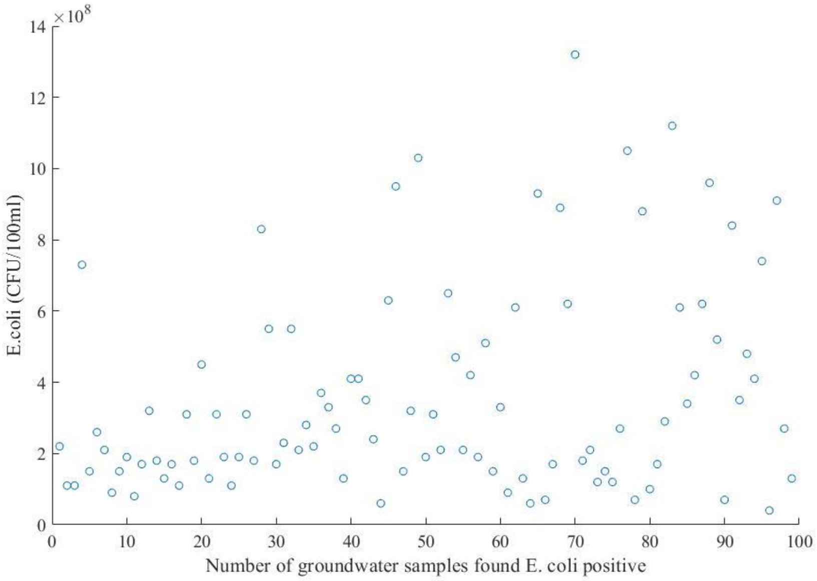


Figure 3.4: Viable cell count of *E. coli* in groundwater samples.

According to laboratory-based culture methods, 12-48 hours are required for bacteria to be reported. After plating, water samples are kept in an incubator at 35°C for 48 hours. The time of incubation depends on the organism and medium of growth. However, every viable cell that has been spread on the petri dish containing agar must grow and divide several times during the incubation to form a detectable colony of microorganisms. The growth of the bacteria is observed after 48 hours of incubation. No changes are observed in any of the water samples after 48 hours of incubation. The sterile disposable bottles are stored in the laboratory at room temperature for preservation after analysis. After 30 days of water testing for detection of presence or absence of *E. coli* bacteria, the presence of bacteria was observed in 99 samples due to a change in colour of the medium from reddish-purple to yellow, indicating the presence of *E. coli.* According to WHO, USEPA and IS 10500: 2012, *E. coli* bacteria shall not be detectable in 100 mL of the water sample. The viable count analysis of water samples showed the presence of *E. coli* with minimum cell counts of 4x10^2 CFU/100mL and maximum cell counts of 132x10^2 CFU/100mL. It indicates that there may be a possible presence of viable but not culturable (VBNC) cells of *E. coli* induced by diverse environmental stresses that restricted the growth of bacteria under controlled laboratory conditions. When samples were kept at room temperature under anaerobic conditions, the bacterial cells become culturable once again after specific resuscitation protocols.

The limitation of only relying on sensor-based water quality analysis for detection can be prone to human errors. Hence, there is a need to automate real-time bacterial monitoring to minimize the error, as mentioned above. To address this issue, we implement an automated process of water-borne bacterial detection using a hybrid technique called Adaptive Neuro-fuzzy Inference System (ANFIS) that integrates the advantage of learning in an ANN and a set of fuzzy if-then rules with appropriate membership functions.

3.5 Detection of *E. coli* bacteria using Adaptive Neuro-Fuzzy Inference System

Artificial neural networks (ANN) mimic the biological neural network of the visual cortex in the brain. The brain consists of a densely interconnected set of information-processing units called neurons. Information is stored and processed in the brain by the involvement of each neuron, which subsequently helps in human learning. Similarly, an ANN model trains itself for learning by connecting with different nodes (Negnevitsky 2005). The ANFIS incorporates the self-learning ability of neural networks with the linguistic expression function of fuzzy inference. ANFIS is a multilayer feed-forward network. Each node performs a particular function on receiving signals and has a set of parameters about this node. Like ANN, ANFIS can convert unseen inputs to their respective outputs by learning the rules from previously observed data (Bouharati et al., 2008). An ANFIS model can adjust the parameters better in any series and takes into consideration all the edge cases in a rule-viewer interface. ANN model may not take the probabilistic values, but using ANFIS, we can make a set of rules for the same. While ANFIS integrates with fuzzy inference systems and ANNs, it helps to solve non-linear and complex problems within a frame (Okwu et al., 2018). Hybrid-based methods like ANN and fuzzy or ANFIS-GA (Genetic Algorithm) can prove to be extremely useful in dealing with missing data (Keshavarz et al., 2018). Hybrid models significantly increase the accuracy of estimation, especially in non-linear problems (Calp 2019).

Bouharati et al. (2008) proposed a method to detect micro bacterial pollution in freshwater using an ANFIS. The model produced instantaneous results by the measurement of the physical and chemical properties of the sensors. ANFIS methods are based on the concept of Fuzzy set theory, which states that a variable can partially belong to a set and can have a membership value between 0 and 1. In this study, three parameters were selected as an input, i.e., pH, temperature, electrical potential, and the output is considered as the number of bacteria. The author revealed the use of an artificial neural network model of three layers trained and tested on the collected water samples. ANFIS model was used because it combines the advantages of fuzzy systems with transparent knowledge representation and those neural networks which deal with the implicit knowledge that can be acquired using learning. Azeez et al. (2008) compared the performance of ANN and ANFIS in the triage of emergency patients using various vital signs of patients as input parameters.

Chandaran et al. (2012) explained the detection of sulphate reducing bacteria (SRB) using ANFIS, which can be crucial in curbing the corrosion of iron material in the system. The author used three parameters: Voltage, Temperature, and humidity, for training the model. The membership functions were taken to be trapezoidal and bell-shaped. The ANFIS model used three inputs, which finally give the output as either 1 or 0. The predicted results were obtained by the input parameters and the number of epochs was taken as 20. Lastly, the model was tested with testing data up to 250 epochs. The author compared the results and the best membership function was given by trapezoidal shape. Calp (2019) proposed a hybrid model for the estimation of the regional rainfall amount. The proposed model focused on providing efficient water resources management by estimating the amount of rainfall that can occur in the region. While creating the model, the MATLAB package program was used and regression values (R^2) or mean squared error (MSE) were taken into account. The error rate was obtained as 0.9920, 0.9840 and 0.0011, respectively, for the model. The author concludes by stating that this hybrid model is an important support tool for estimating annual rainfall and ensuring the effective management of water resources.

The detection of waterborne bacteria is crucial to prevent health risks. This study uses soft computing techniques based on Artificial Neural Networks (ANN) to detect bacterial pollution in water. The limitation of only relying on sensor-based water quality analysis for detection can be prone to human errors. Hence, there is a need to automate real-time bacterial monitoring to minimize the error, as mentioned above. To address this issue, we implement an automated process of water-borne bacterial detection using a hybrid technique called Adaptive Neuro-fuzzy Inference System (ANFIS) that integrates the advantage of learning in an ANN and a set of fuzzy if-then rules with appropriate membership functions.

In the first part of this section, we explain the design of a fuzzy expert system based on membership functions. Subsequently, an ANFIS model will be introduced based on the fuzzy rule base. This study assigned three fuzzy sets to each water quality parameter, namely desirable, undesirable, and highly undesirable. The individual membership function is assigned to each parameter, as described in Table 3.3.

Table 3.3: Groups defined for water quality parameters

Parameters	**Range**		
	Undesirable	Desirable	Highly Undesirable
Temperature	0-10	5–41	35-48
Dissolved oxygen	0-5	2-14	10-20
pH	0-7.5	6.5-8.5	8-14
Electrical conductivity	0-300	200-1000	800-1000

3.5.1 Data pre-processing

A series of experimental data is obtained from the BITS-UVA project. The detection of bacterial presence is based on four basic water quality parameters, i.e., Temperature, Dissolved Oxygen (DO), pH, and Electrical Conductivity (EC). There are four inputs and one output for each set of data. The output, ‘1’ represents the existence of bacteria, and ‘0’ indicates absence. These four inputs are used to model the ANFIS system, which gives an output that gives output in between 1 and 0. The computation of data for ANFIS is conducted using MATLAB 2019a. However, the ANFIS training algorithms are embedded in MATLAB’s fuzzy logic toolbox (MathWorks). There are four steps for computation. The first step is data input and normalization. After normalization, the total experimental data are divided randomly into 70% for training, 15% for testing, and 15% for validation. The next step is to assign membership functions for data, then train the input using the ANFIS training function. Finally, the predicted result can be obtained by inputting the parameters. Figure 3.5 shows the structure of the ANFIS model.

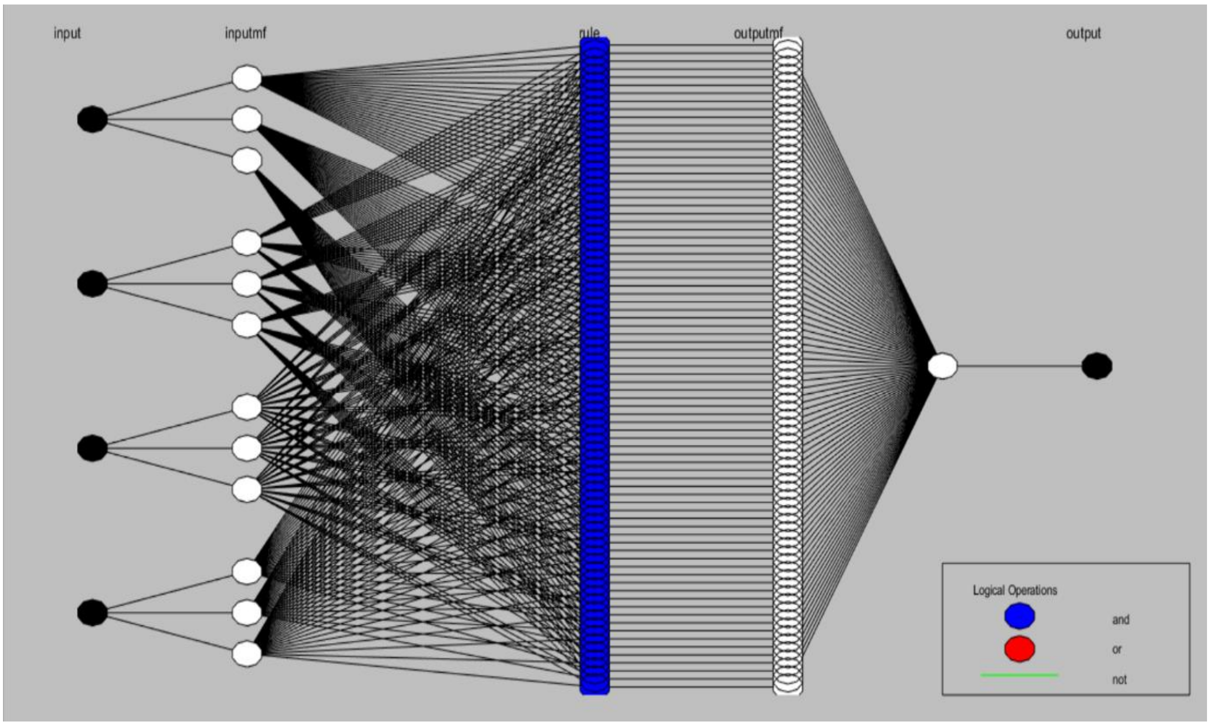
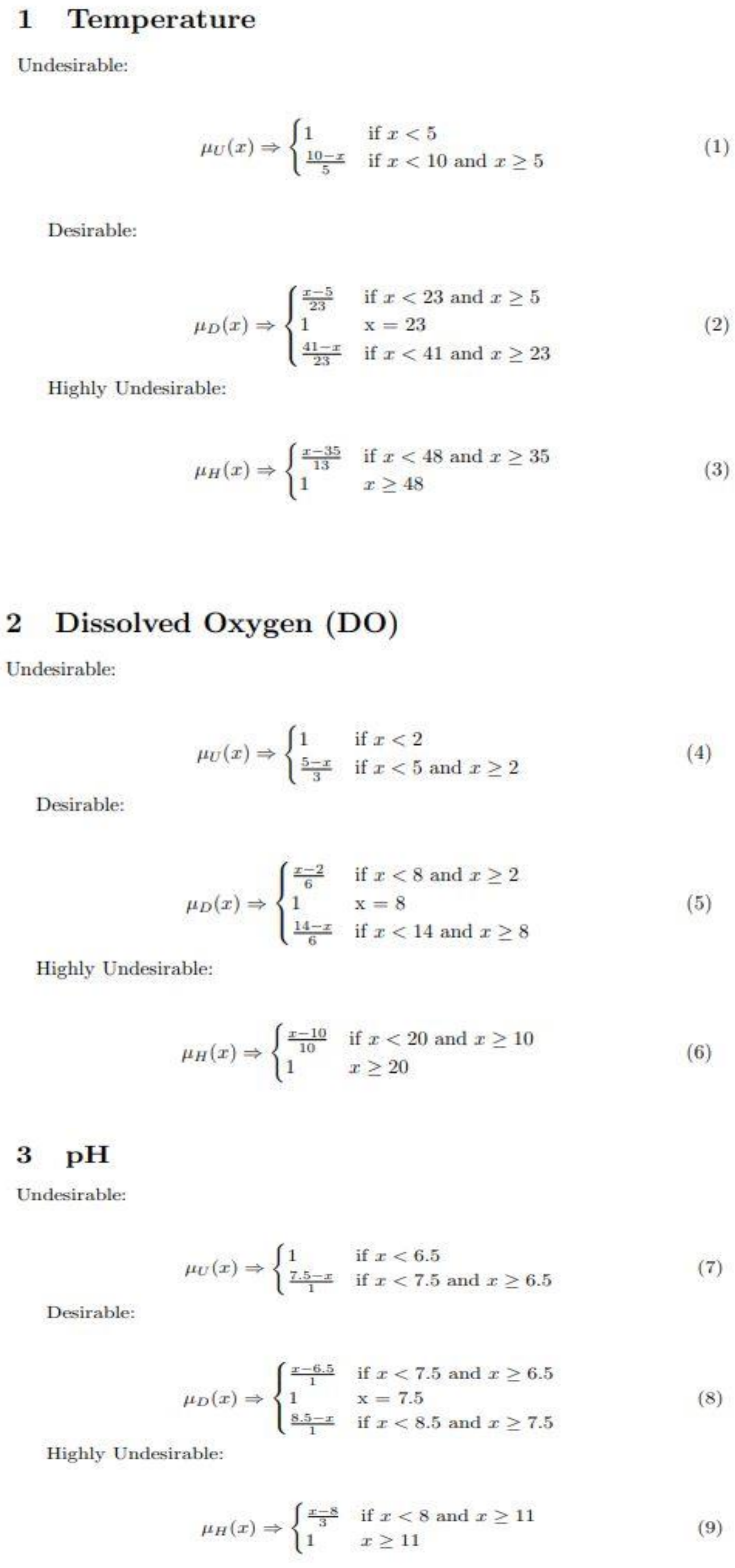


Figure 3.5: Structure of the ANFIS model.

3.5.2 Assigning membership function

In this study, three types of membership functions (Triangular, Trapezoidal, Bell-shaped) are tested to get the best function for the prediction model. The interval contains three fuzzy sets as: “Desirable”, “Undesirable” and “Highly undesirable” Input parameters of the Fuzzy Inference System (FIS) are: Temperature, Dissolved Oxygen, pH, and Conductivity. Temperature accepts values in [0 48] °C. Mathematical equations of membership expressions for temperature are shown from Eq. (1, 2, and 3). Dissolved Oxygen accepts values in [0 20] ppm. Mathematical equations of membership expressions for DO are shown from Eq. (4, 5, and 6). pH accepts values in [0 14]. Membership equations of membership expressions for pH are shown from Eq. (7, 8, and 9). Conductivity accepts values in [0 1000] µS/cm. Membership equations for conductivity are shown from Eq. (10, 11, and 12).



1 Temperature

Undesirable:

$$\mu_U(x) \Rightarrow \begin{cases} 1 & \text{if } x < 5 \\ \frac{10-x}{5} & \text{if } x < 10 \text{ and } x \geq 5 \end{cases} \tag{1}$$

Desirable:

$$\mu_D(x) \Rightarrow \begin{cases} \frac{x-5}{23} & \text{if } x < 23 \text{ and } x \geq 5 \\ 1 & \text{x = 23} \\ \frac{41-x}{23} & \text{if } x < 41 \text{ and } x \geq 23 \end{cases} \tag{2}$$

Highly Undesirable:

$$\mu_H(x) \Rightarrow \begin{cases} \frac{x-35}{13} & \text{if } x < 48 \text{ and } x \geq 35 \\ 1 & x \geq 48 \end{cases} \tag{3}$$

2 Dissolved Oxygen (DO)

Undesirable:

$$\mu_U(x) \Rightarrow \begin{cases} 1 & \text{if } x < 2 \\ \frac{5-x}{3} & \text{if } x < 5 \text{ and } x \geq 2 \end{cases} \tag{4}$$

Desirable:

$$\mu_D(x) \Rightarrow \begin{cases} \frac{x-2}{6} & \text{if } x < 8 \text{ and } x \geq 2 \\ 1 & \text{x = 8} \\ \frac{14-x}{6} & \text{if } x < 14 \text{ and } x \geq 8 \end{cases} \tag{5}$$

Highly Undesirable:

$$\mu_H(x) \Rightarrow \begin{cases} \frac{x-10}{10} & \text{if } x < 20 \text{ and } x \geq 10 \\ 1 & x \geq 20 \end{cases} \tag{6}$$

3 pH

Undesirable:

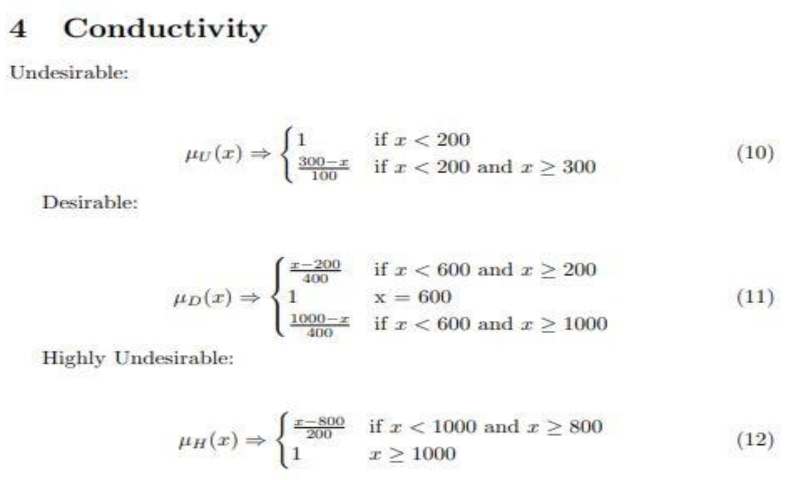
$$\mu_U(x) \Rightarrow \begin{cases} 1 & \text{if } x < 6.5 \\ \frac{7.5-x}{1} & \text{if } x < 7.5 \text{ and } x \geq 6.5 \end{cases} \tag{7}$$

Desirable:

$$\mu_D(x) \Rightarrow \begin{cases} \frac{x-6.5}{1} & \text{if } x < 7.5 \text{ and } x \geq 6.5 \\ 1 & \text{x = 7.5} \\ \frac{8.5-x}{1} & \text{if } x < 8.5 \text{ and } x \geq 7.5 \end{cases} \tag{8}$$

Highly Undesirable:

$$\mu_H(x) \Rightarrow \begin{cases} \frac{x-8}{3} & \text{if } x < 8 \text{ and } x \geq 11 \\ 1 & x \geq 11 \end{cases} \tag{9}$$



4 Conductivity

Undesirable:

$$\mu_U(x) \Rightarrow \begin{cases} 1 & \text{if } x < 200 \\ \frac{300-x}{100} & \text{if } x < 200 \text{ and } x \geq 300 \end{cases} \tag{10}$$

Desirable:

$$\mu_D(x) \Rightarrow \begin{cases} \frac{x-200}{400} & \text{if } x < 600 \text{ and } x \geq 200 \\ 1 & \text{x} = 600 \\ \frac{1000-x}{400} & \text{if } x < 600 \text{ and } x \geq 1000 \end{cases} \tag{11}$$

Highly Undesirable:

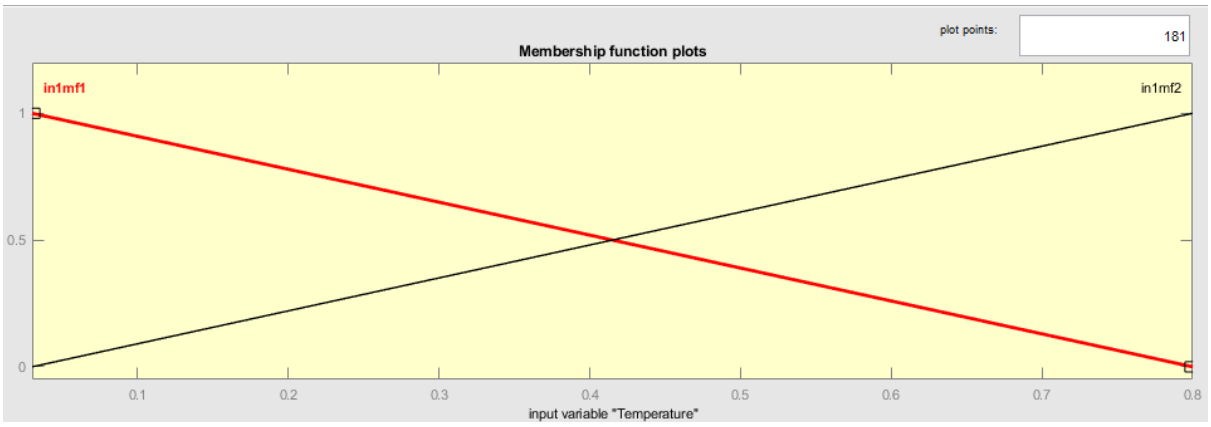
$$\mu_H(x) \Rightarrow \begin{cases} \frac{x-800}{200} & \text{if } x < 1000 \text{ and } x \geq 800 \\ 1 & x \geq 1000 \end{cases} \tag{12}$$

a) Triangular Membership Function

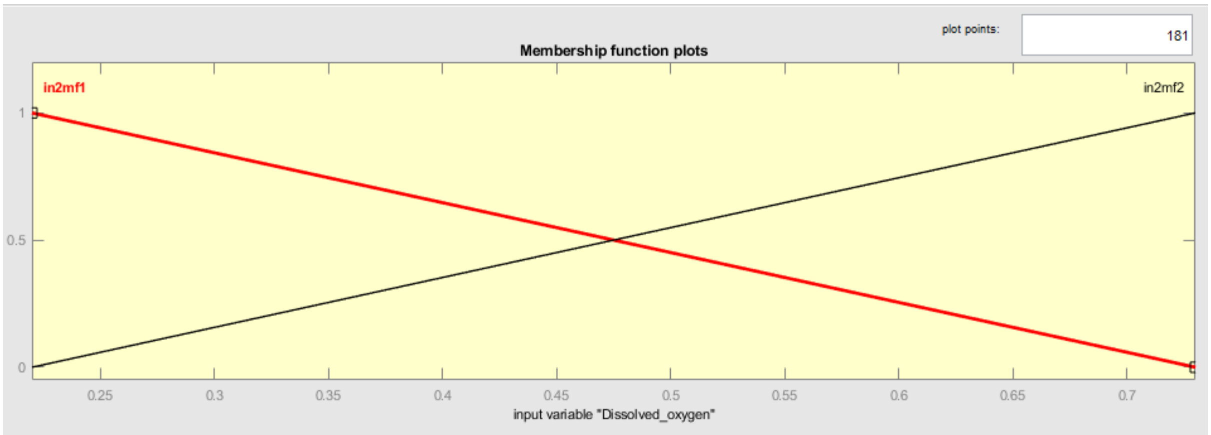
By setting the number of membership functions to three for input data and using a triangular function, the parameters for each input's membership function are tabulated for epochs 1-50. Figure 3.6 and Figure 3.7 show the initial and final membership functions of the input data derived by training via the triangular function. Table 3.4 shows the label of results that are used and their representative of the membership functions.

Table 3.4: Labels of results showing membership function (Where in=input, mf=membership function)

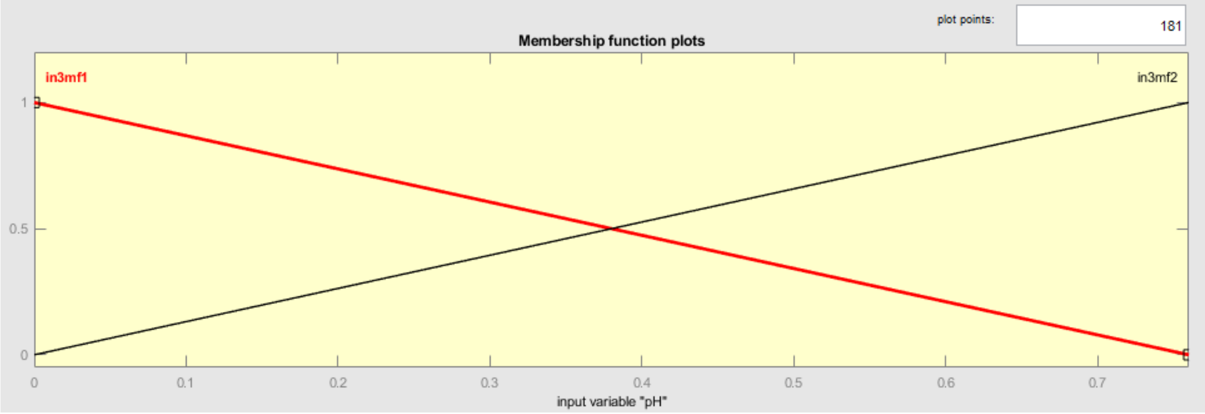
Labels of results	Temperature	Dissolved oxygen	pH	Electrical conductivity
Desirable limits	in1mf1	in2mf1	in3mf1	in4mf1
Undesirable limits	in1mf2	in2mf2	in3mf2	in4mf2
Highly undesirable limits	in1mf3	in2mf3	in3mf3	in4mf3



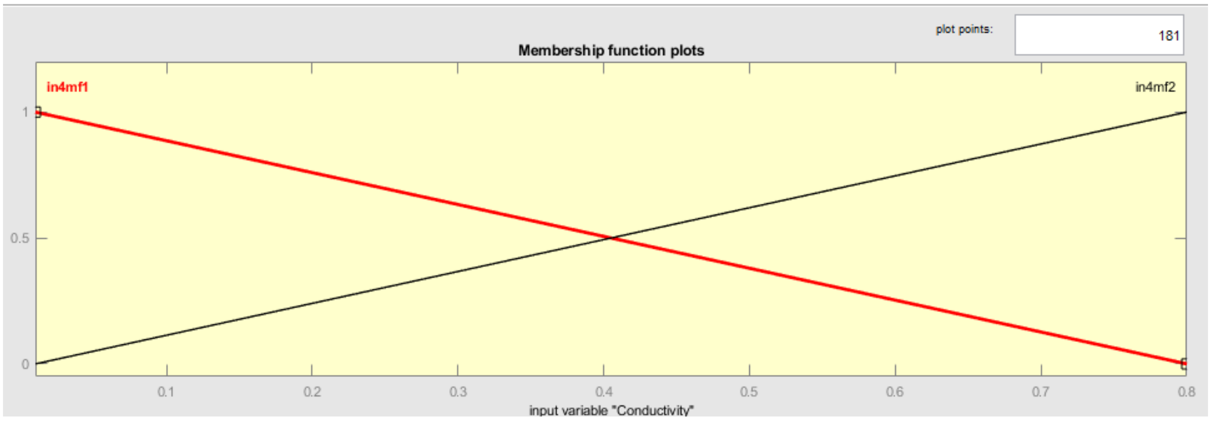
(a) Temperature.



(b) Dissolved oxygen.

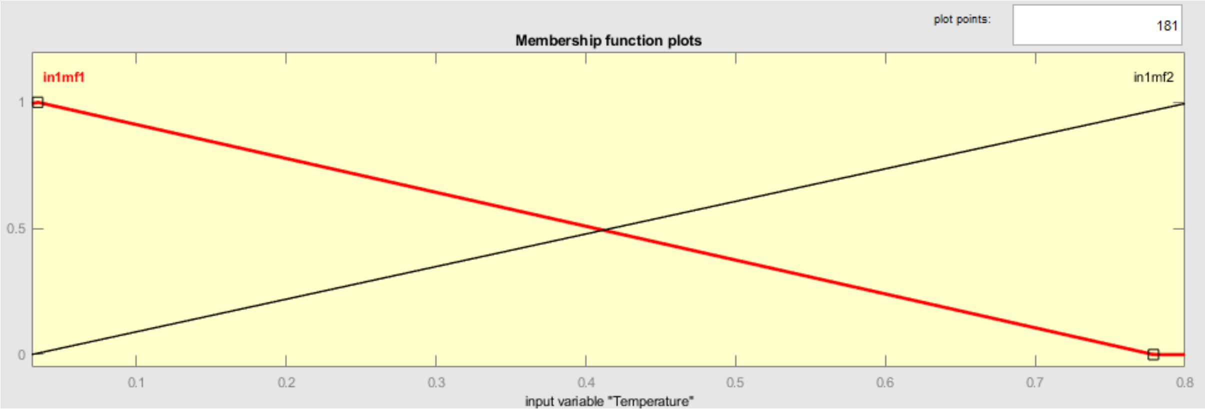


(c) pH.

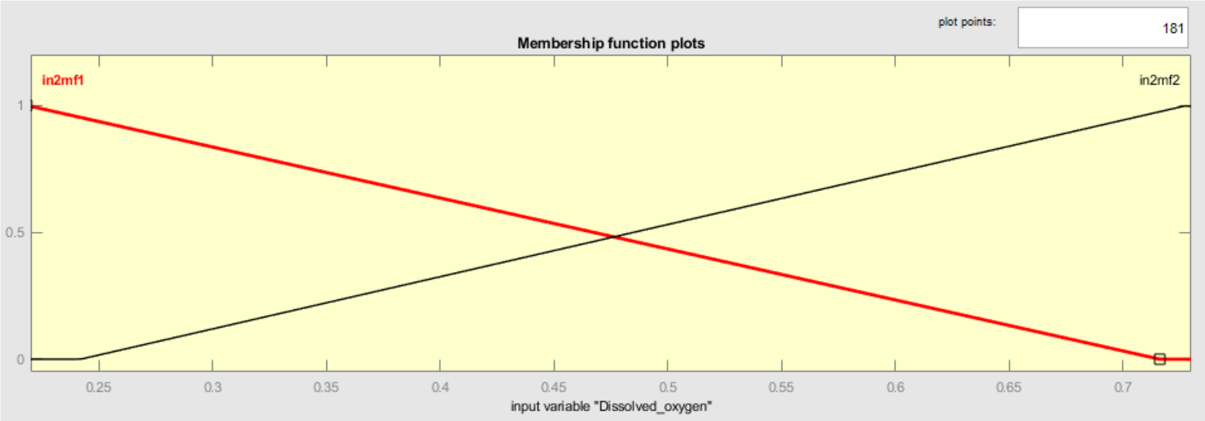


(d) Electrical Conductivity.

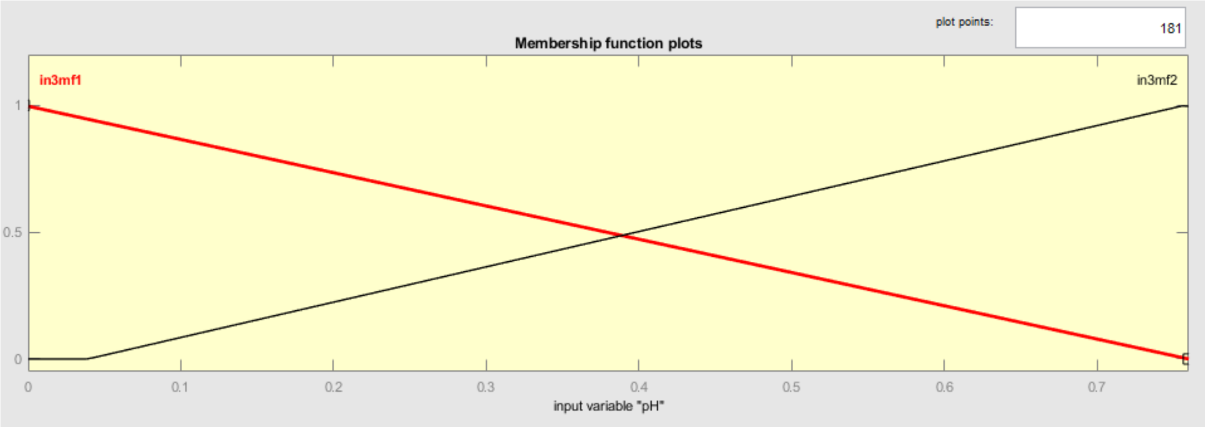
Figure 3.6: Initial membership function (Triangular) of Temperature, Dissolved oxygen, pH, and Electrical conductivity.



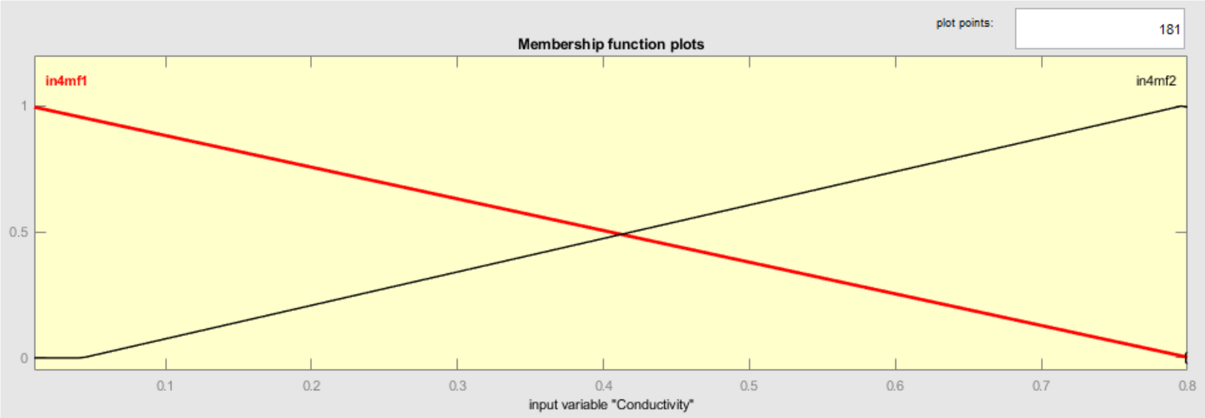
(a) Temperature



(b) Dissolved oxygen



(c) pH

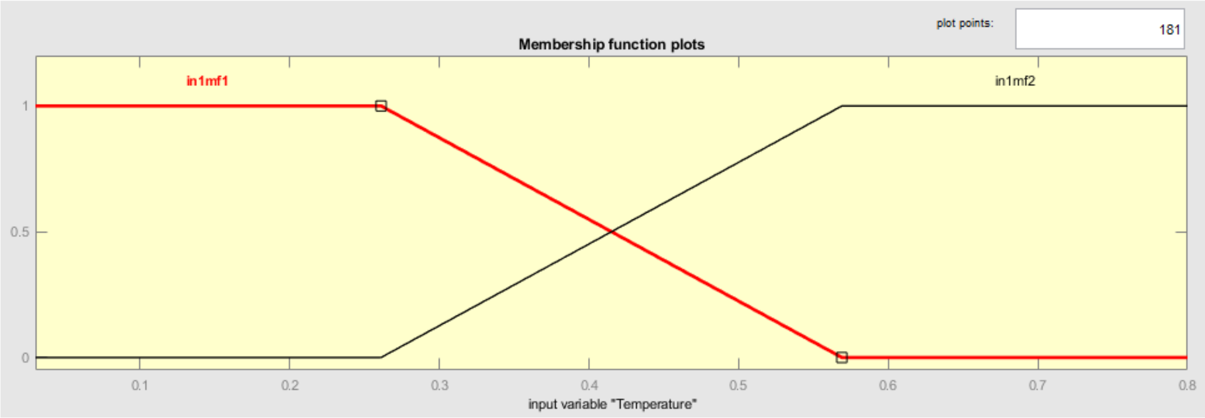


(d) Electrical Conductivity.

Figure 3.7: Final membership function (Triangular) of Temperature, Dissolved oxygen, pH, and Electrical conductivity.

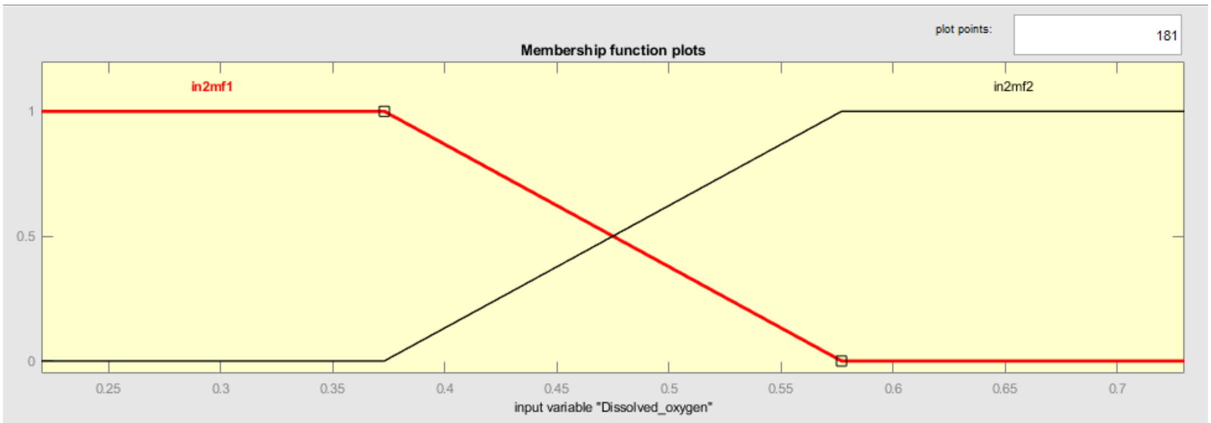
b) Trapezoidal Membership Function

By setting the number of membership functions to three for input data and using a trapezoidal function, the parameters for each input's membership function are recorded and tabulated for epochs 1-20. Figure 3.8 and Figure 3.9 show the initial and final membership functions of the input data derived by training via the trapezoidal function.



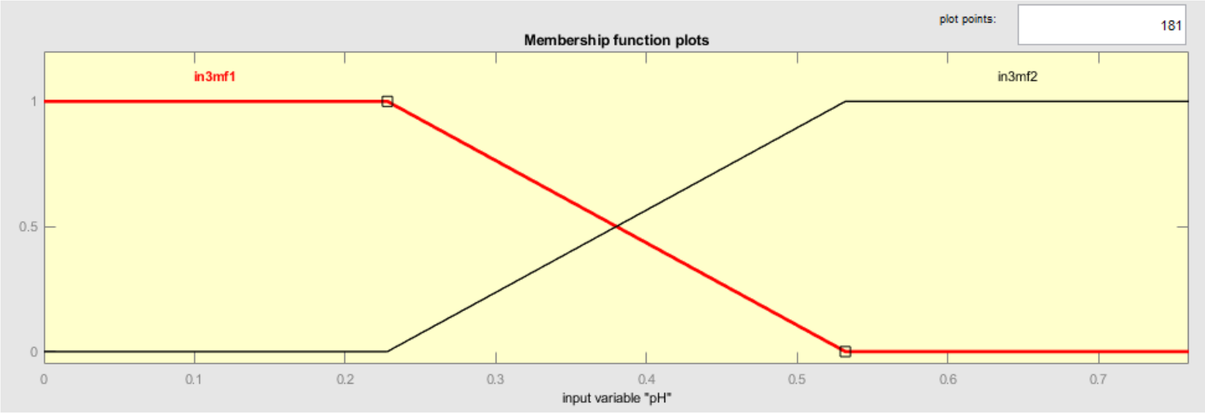
plot points: 181

(a) Temperature



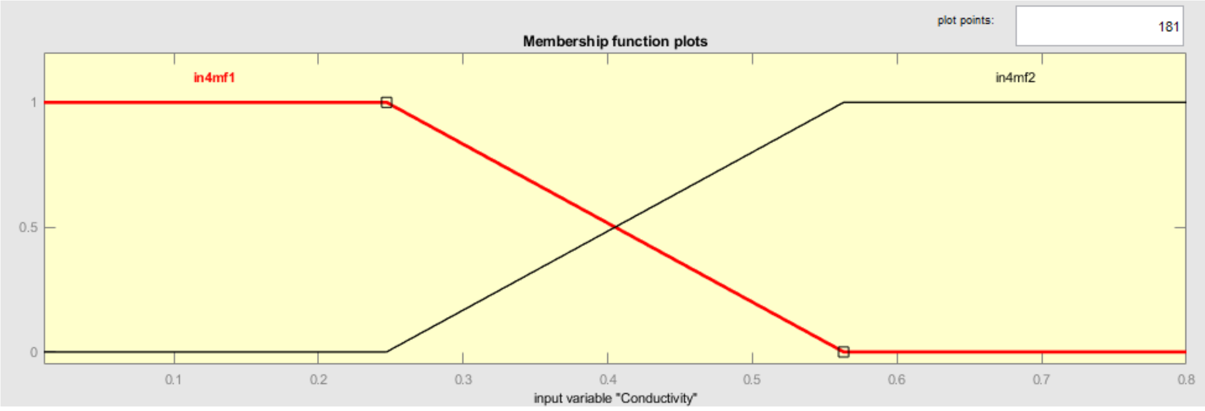
plot points: 181

(b) Dissolved Oxygen



plot points: 181

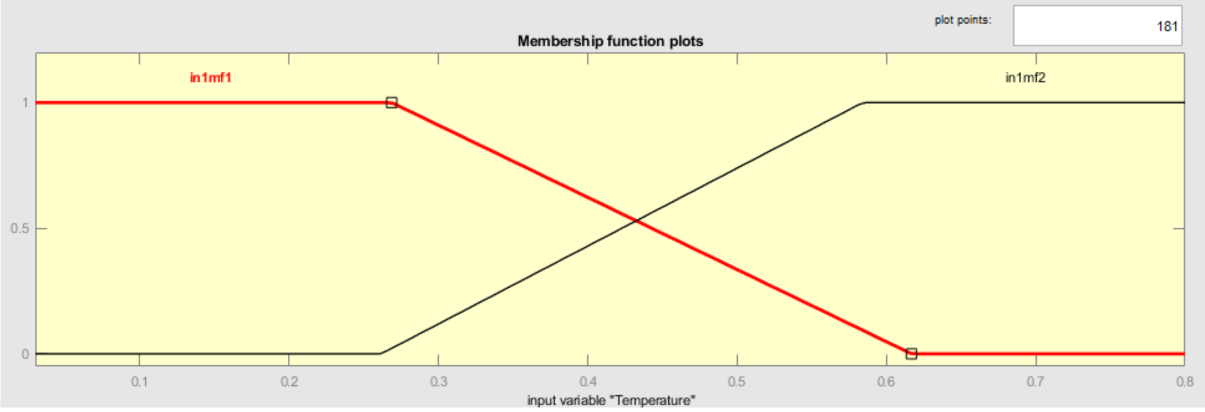
(c) pH



plot points: 181

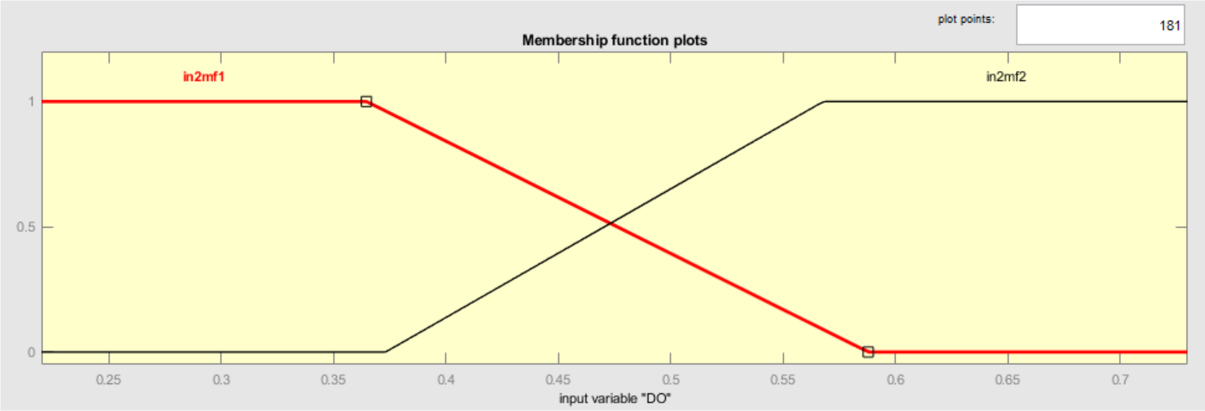
(d) Electrical conductivity.

Figure 3.8: Initial membership function (Trapezoidal) of Temperature, Dissolved oxygen, pH, and Electrical conductivity.



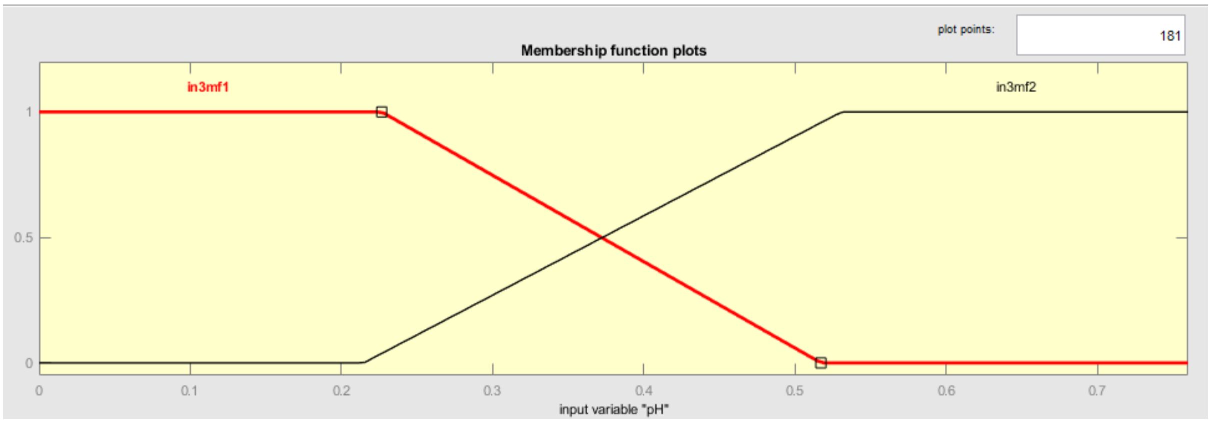
plot points: 181

(a) Temperature

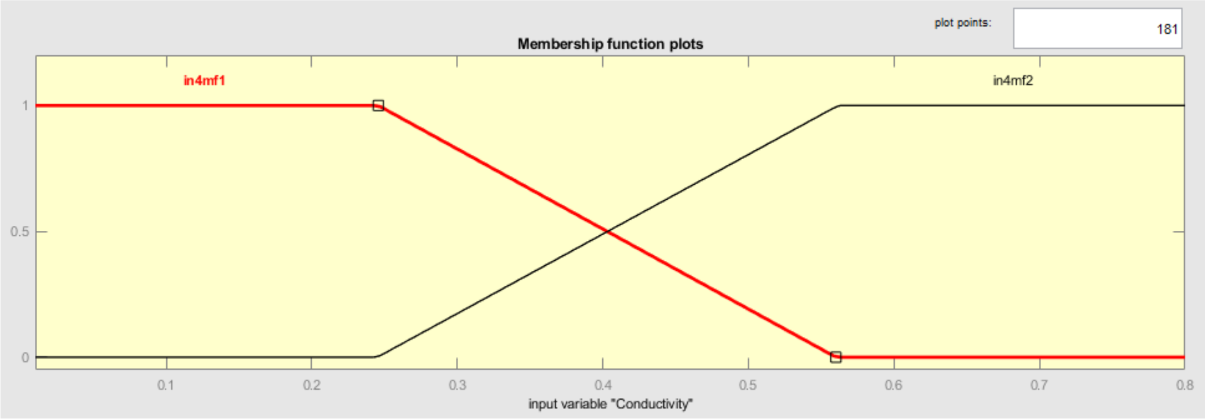


plot points: 181

(b) Dissolved Oxygen



(c) pH

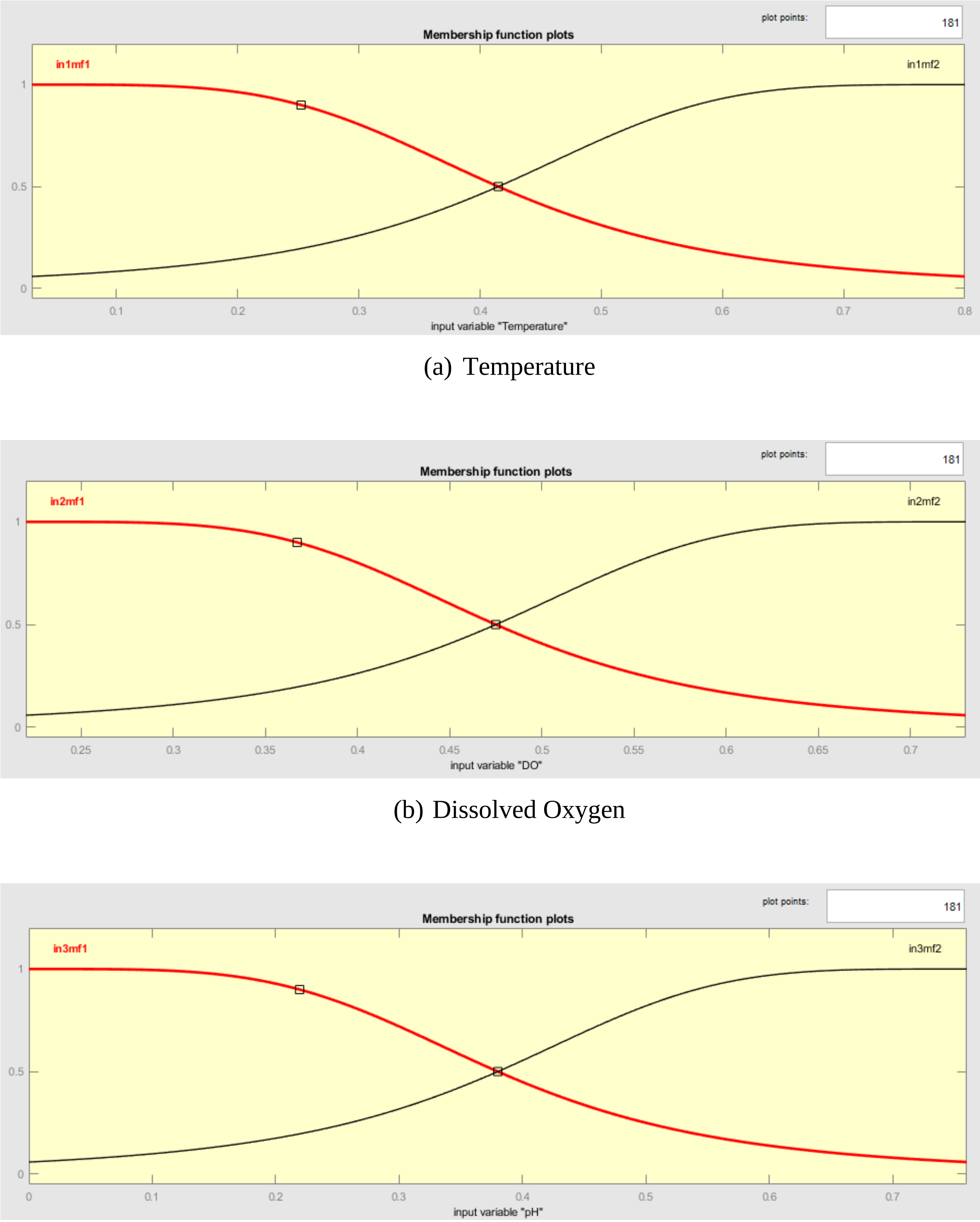


(d) Electrical conductivity.

Figure 3.9: Final membership function (Trapezoidal) of Temperature, Dissolved oxygen, pH, and Electrical conductivity.

c) Generalized bell-shaped Membership Function

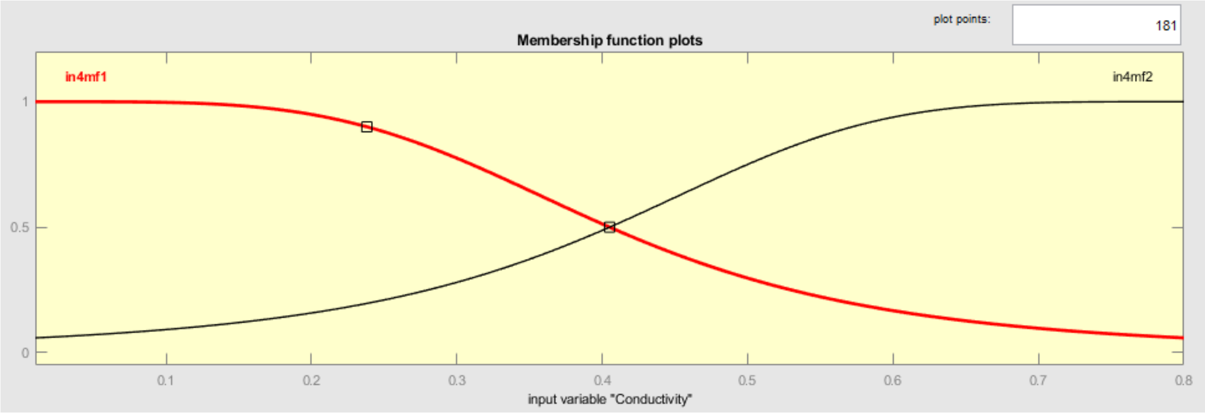
By setting the number of membership functions to three for input data and using a generalized bell function, the parameters for each input's membership function are recorded and tabulated for epochs 1-100. Figure 3.10 and Figure 3.11 show the initial and final membership functions of the input data derived by training via the trapezoidal function.



(a) Temperature

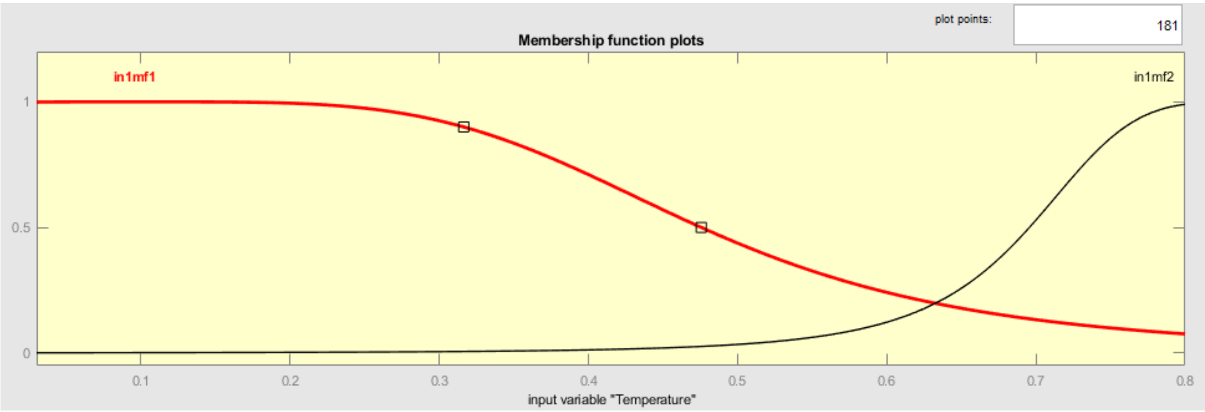
(b) Dissolved Oxygen

(c) pH

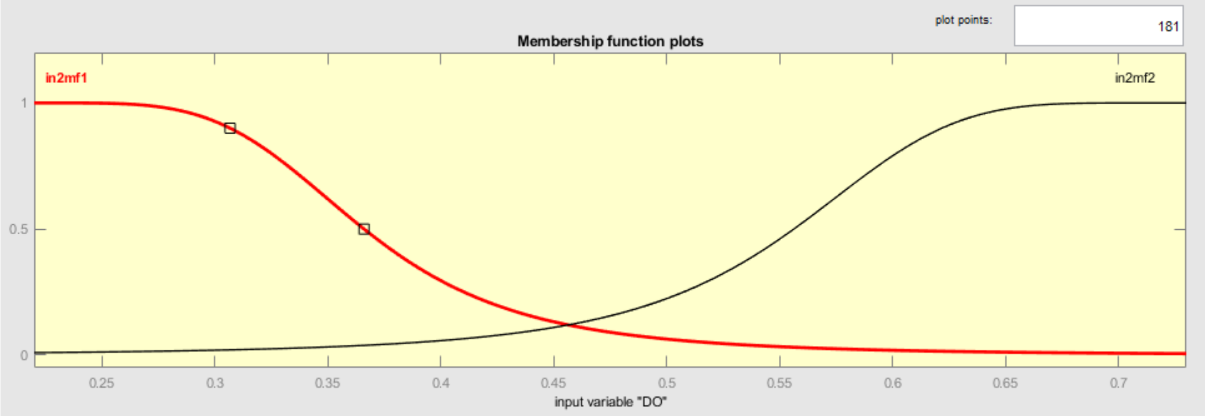


(d) Electrical conductivity.

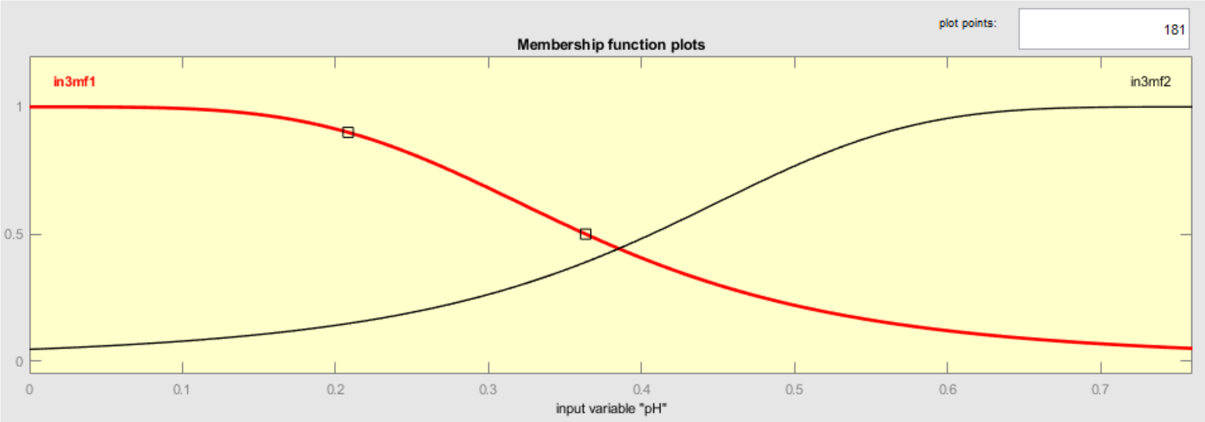
Figure 3.10: Initial membership function (Bell-shaped) of Temperature, Dissolved oxygen, pH, and Electrical conductivity.



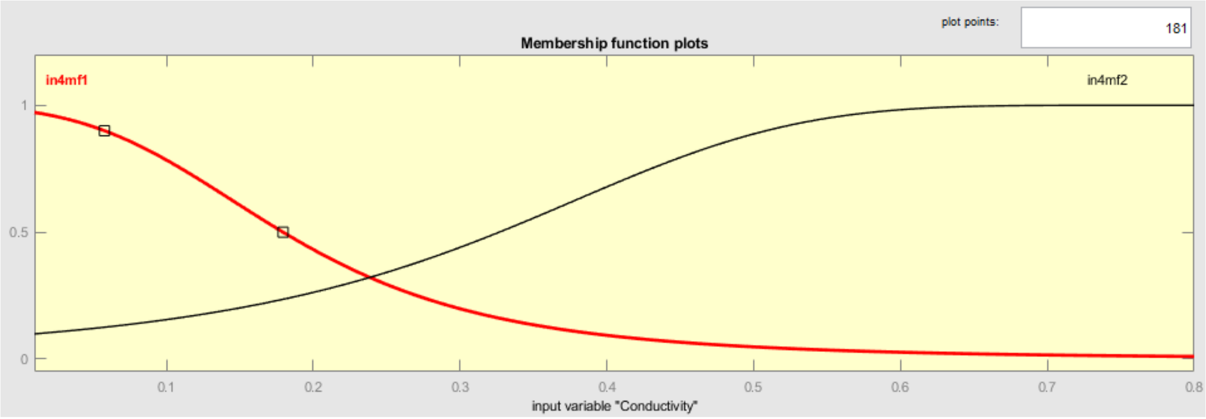
(a) Temperature.



(b) Dissolved Oxygen.



(c) pH

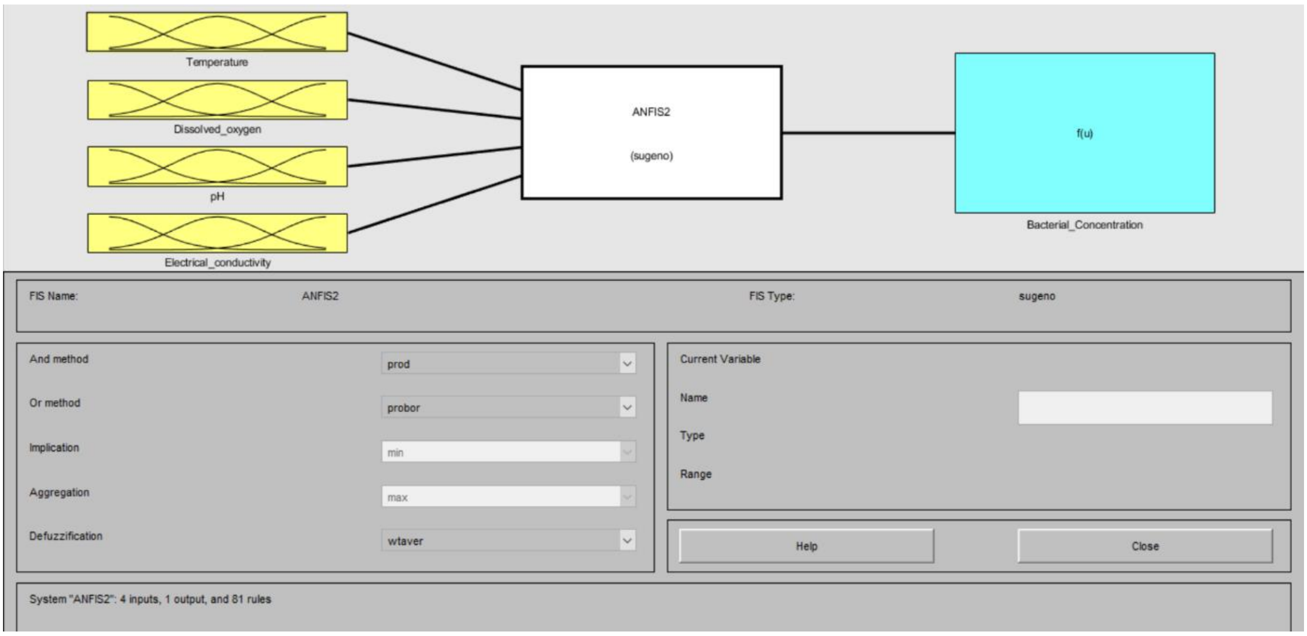


(d) Electrical conductivity.

Figure 3.11: Final membership function (Bell-shaped) of Temperature, Dissolved oxygen, pH, and Electrical conductivity.

3.5.3 Predicted output for ANFIS

The testing data was used to check the capability of the model. We compare the results obtained from three membership functions (triangle, trapezoidal, and bell-shaped) for output prediction. The error rate and parameters for every single membership function are plotted. The resulted FIS was tested using testing data for a hundred epochs. We tabulate the error analysis for each membership function type. Figure 3.12 shows the proposed ANFIS for the detection of bacteria.



FIS Name: ANFIS2

FIS Type: sugeno

And method: prod

Or method: probor

Implication: min

Aggregation: max

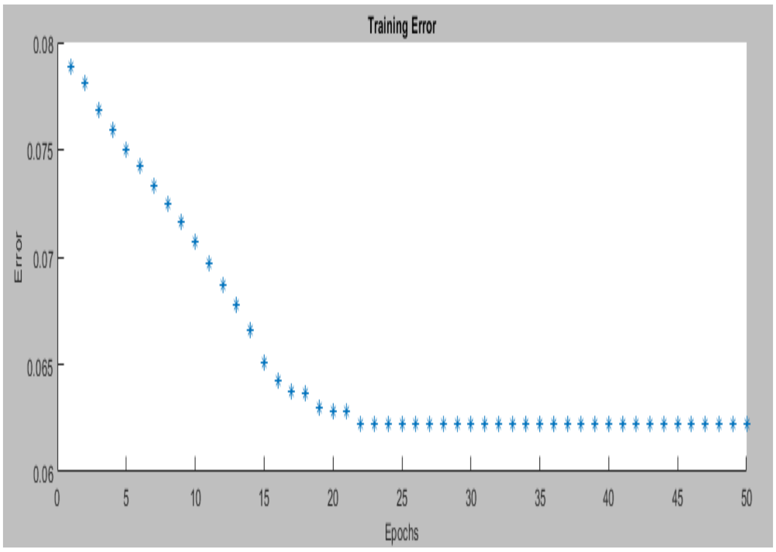
Defuzzification: wtaver

System "ANFIS2": 4 inputs, 1 output, and 81 rules

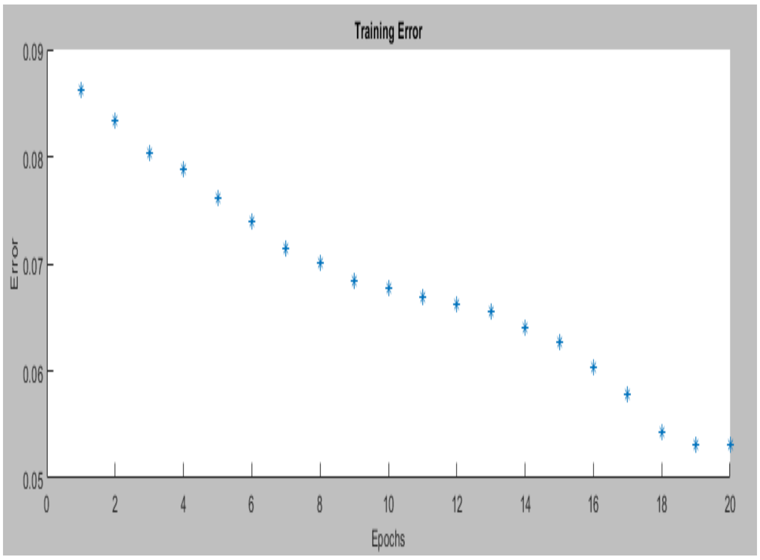
Figure 3.12: Proposed ANFIS for the detection of bacteria in terms of temperature, dissolved oxygen, pH, and electrical conductivity as the model inputs.

3.5.3.1 Error rate

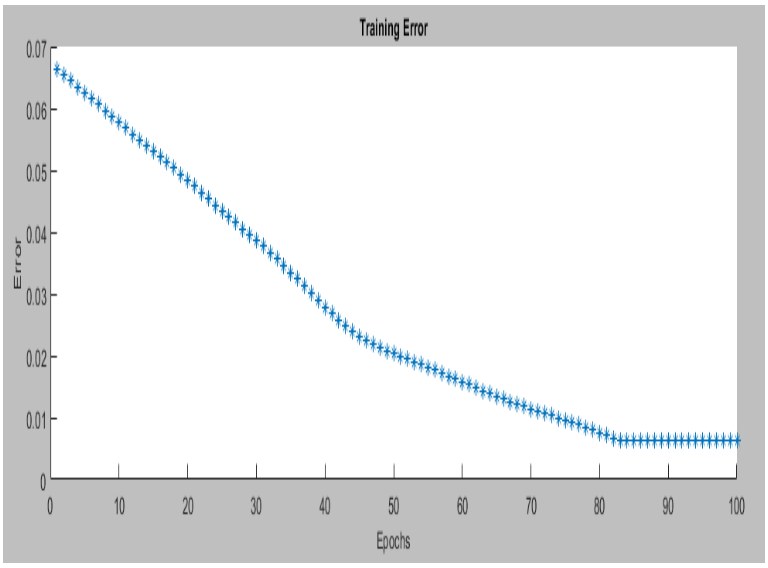
The training error of three membership functions was tested and shown in Figure 3.13. For each type of membership function, the error rate after every single epoch has been recorded. The rate is tabulated in Table 3.5.



(a) Triangle



(b) Trapezoidal.



(c) Bell-shaped.

Figure 3.13: ANFIS training error of triangle, trapezoidal and bell-shaped as membership function.

Table 3.5: Error rate

	Error		
Epoch	Triangle	Trapezoidal	Bell-shaped
1	0.0789	0.07678	0.06645
10	0.0707	0.06321	0.05775
20	0.0628	0.05665	0.04838

50	0.0623	0.05665	0.02036
100	0.0623	0.05665	0.00619

From Table 3.5, it can be seen that the error rate of using bell-shaped membership function is lesser from the starting if compared to triangular and trapezoidal function. The differenced becomes more noticeable when epochs are equal to 100. The error rate is only 0.006 (bell-shaped), which is lesser than 0.06 (triangular) and 0.05 (trapezoidal), and the results show that the bell-shaped function always gives the least error.

3.6 Summary

Due to both biotic and abiotic factors such as starvation, exposure to chlorine, pH, oxygen availability, heavy metals, exposure to white light, temperature changes, salinity, nutrient scarcity, incubation beyond normal growth temperature range, osmotic pressure, copper, harsh environmental conditions, nutrient scarcity and many other factors induced the VBNC state. Specific parameters, such as the method of storage, holding time, and temperature, also influenced the concentration of *E. coli*. It can be concluded that culture-based methods are not accurate for the detection of *E. coli* bacteria in water. Further research is needed to detect the VBNC cells of bacteria in water. *E. coli* entering the VBNC condition could have a detrimental effect on public health. The number of viable cells could be underestimated, and at any time, the VBNC cells could still produce toxins or be resuscitated to become virulent again and again. Various studies have found that resuscitation of *E. coli* post-VBNC may be possible. Some pathogenic *E. coli* strains can produce toxins in VBNC conditions, while others are non-toxic but can regain virulence after regeneration. The results showed that the units forming the colony grew over time. The cell wall of *E. coli* remained intact after one month of laboratory incubation.

The detection of waterborne bacteria is crucial to prevent health risks. The current study uses soft computing techniques based on Artificial Neural Networks (ANN) for the detection of bacterial pollution in water. The limitation of only relying on sensor-based water quality analysis for detection can be prone to human errors. Hence, there is a need to automate real-time bacterial monitoring to minimize the error, as mentioned above. To address this issue, we implement an automated process of water-borne bacterial detection using a hybrid technique called Adaptive Neuro-fuzzy Inference System (ANFIS) that integrates the advantage of learning in an ANN and a set of fuzzy if-then rules with appropriate membership functions. An ANFIS model for the detection of bacteria in drinking water sources has been developed with 81 fuzzy set rules, and the predictive ability of the model is compared with three membership functions. The membership function changes automatically with every iteration during the model training.

The results show that ANFIS with a generalized bell-shaped membership function is the most suitable membership function to model bacterial detection. The least error obtained at epoch 100 is 0.00619 by applying a bell-shaped function through the testing data verification. ANFIS with a bell-shaped membership function gives precisely the same output as experimental output. It can be concluded that culture-based methods are not accurate for detecting *E. coli* bacteria in water. Further research is needed to detect the VBNC cells of water-borne bacteria using sensitive, reliable, and cost-effective methods. The study recommended that *E. coli* bacteria not be used as an indicator organism when the cells are viable but non-culturable.

References

- Alam, M., Farzana, T., Ahsan, C. R., Yasmin, M., & Nessa, J. (2011). Distribution of coliphages against four E. coli virotypes in hospital originated sewage sample and a sewage treatment plant in Bangladesh. *Indian journal of microbiology*, *51*(2), 188-193.

- Arana, I. N. E. S., Muela, A. L. I. C. I. A., Iriberri, J. U. A. N., Egea, L., & Barcina, I. (1992). Role of hydrogen peroxide in loss of culturability mediated by visible light in Escherichia coli in a freshwater ecosystem. *Applied and environmental microbiology*, *58*(12), 3903-3907.

- Arana, I., Seco, C., Epelde, K., Muela, A., Fernández-Astorga, A., & Barcina, I. (2004). Relationships between Escherichia coli cells and the surrounding medium during survival processes. *Antonie van Leeuwenhoek*, *86*(2), 189-199.

- Artz, Rebekka RE, Lisa M. Avery, Davey L. Jones, and Ken Killham. "Potential pitfalls in the quantitative molecular detection of Escherichia coli O157: H7 in environmental matrices." *Canadian journal of microbiology* 52, no. 5 (2006): 482-488.

- Asakura, H., Kawamoto, K., Haishima, Y., Igimi, S., Yamamoto, S., & Makino, S. I. (2008). Differential expression of the outer membrane protein W (OmpW) stress response in enterohemorrhagic Escherichia coli O157: H7 corresponds to the viable but non-culturable state. *Research in Microbiology*, *159*(9-10), 709-717.

- Asakura, H., Panutdaporn, N., Kawamoto, K., Igimi, S., Yamamoto, S., & Makino, S. I. (2007). Proteomic characterization of enterohemorrhagic Escherichia coli O157: H7 in the oxidation-induced viable but non-culturable state. *Microbiology and immunology*, *51*(9), 875-881.

- Atlas, R. M. (1998). *Microbial ecology: fundamentals and applications*. Pearson Education India.

- Aulenbach, B. T. (2010). Bacteria holding times for fecal coliform by mFC agar method and total coliform and Escherichia coli by Colilert®-18 Quanti-Tray® method. *Environmental monitoring and assessment*, *161*(1), 147-159.

- Azeez, D., Ali, M. A. M., Gan, K. B., & Saiboon, I. (2013). Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. *SpringerPlus*, *2*(1), 1-10.

- Barcina, I., & Arana, I. (2009). The viable but nonculturable phenotype: a crossroads in the life-cycle of non-differentiating bacteria?. *Reviews in Environmental Science and Bio/Technology*, *8*(3), 245-255.

- Barer, M. R., Gribbon, L. T., Harwood, C. R., & Nwoguh, C. E. (1993). The viable but non-culturable hypothesis and medical bacteriology. *Reviews in Medical Microbiology*, *4*(4), 183-191.

- Bartram, J., & Ballance, R. (Eds.). (1996). *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes*. CRC Press.

- Besnard, V., Federighi, M., & Cappelier, J. M. (2000). Evidence of viable but non-culturable state in Listeria monocytogenes by direct viable count and CTC-DAPI double staining. *Food Microbiology*, *17*(6), 697-704.

- Besnard, V., Federighi, M., Declerq, E., Jugiau, F., & Cappelier, J. M. (2002). Environmental and physico-chemical factors induce VBNC state in Listeria monocytogenes. *Veterinary research*, *33*(4), 359-370.

- BIS, I. (2012). 10500 Indian standard drinking water–specification, second revision. *Bureau of Indian Standards, New Delhi*.

- Bouharati, S., Benmahammed, K., Harzallah, D., & El-Assaf, Y. M. (2008). Application of artificial neuro-fuzzy logic inference system for predicting the microbiological pollution in fresh water. *Journal of Applied Sciences*, *8*(2), 309-315.

- Calp MH 2019 A Hybrid ANFIS-GA Approach for Estimation of Regional Rainfall Amount. *Gazi University Journal of Science* **32(1)** 145.

- Chandaran, U. D., Halim, Z. A., & Sian, L. K. (2012, October). Study on sulfate reducing bacteria detection using Adaptive Neuro-fuzzy Inference System. In *2012 IEEE International Conference on Circuits and Systems (ICCAS)* (pp. 59-64). IEEE.

- Chowdhury, M. A. R., Xu, B., Montilla, R., Hasan, J. A. K., Huq, A., & Colwell, R. R. (1995). A simplified immunofluorescence technique for detection of viable cells of Vibrio cholerae O1 and O139. *Journal of microbiological methods, 24*(2), 165-170.

- Cook, K. L., & Bolster, C. H. (2007). Survival of Campylobacter jejuni and Escherichia coli in groundwater during prolonged starvation at low temperatures. *Journal of applied microbiology*, *103*(3), 573-583.

- Co-ordination Action Food (CAF) (2007). Methods for detection and molecular characterisation of pathogenic Escherichia coli. In: O'Sullivan J, Bolton DJ, Duffy G, Baylis C, Tozzoli R, Wasteson Y, Lofdahl S (eds.)

- Cunningham, E., O'Byrne, C., & Oliver, J. D. (2009). Effect of weak acids on Listeria monocytogenes survival: evidence for a viable but nonculturable state in response to low pH. *Food control*, *20*(12), 1141-1144.

- Cuny, C., Dukan, L., Fraysse, L., Ballesteros, M., & Dukan, S. (2005). Investigation of the first events leading to loss of culturability during Escherichia coli starvation: future nonculturable bacteria form a subpopulation. *Journal of bacteriology*, *187*(7), 2244-2248.

- Davey, J. W., Hohenlohe, P. A., Etter, P. D., Boone, J. Q., Catchen, J. M., & Blaxter, M. L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nature Reviews Genetics*, *12*(7), 499-510.

- Dolezalova, E., & Lukes, P. (2015). Membrane damage and active but nonculturable state in liquid cultures of Escherichia coli treated with an atmospheric pressure plasma jet. *Bioelectrochemistry, 103*, 7-14.

- Dutka, B. J., & El-Shaarawi, A. (1980). Microbiological water and effluent sample preservation. *Canadian journal of microbiology*, *26*(8), 921-929.

- Evans Jr, D. J., & Evans, D. G. (1996). Escherichia coli in diarrheal disease. *Medical Microbiology. 4th edition*.

- Fakruddin, M., Mannan, K. S. B., & Andrews, S. (2013). Viable but nonculturable bacteria: food safety and public health perspective. *International Scholarly Research Notices*, *2013*.

- Ghezzi, J. I., & Steck, T. R. (1999). Induction of the viable but non-culturable condition in Xanthomonas campestris pv. campestris in liquid microcosms and sterile soil. *FEMS Microbiology Ecology*, *30*(3), 203-208.

- Gourmelon, M., Cillard, J., & Pommepuy, M. (1994). Visible light damage to Escherichia coli in seawater: oxidative stress hypothesis. *Journal of Applied Bacteriology*, *77*(1), 105-112.

- Grey, B., & Steck, T. R. (2001). Concentrations of copper thought to be toxic to Escherichia coli can induce the viable but nonculturable condition. *Applied and Environmental Microbiology*, *67*(11), 5325-5327.

- Harmel, D., Wagner, K., Martin, E., Smith, D., Wanjugi, P., Gentry, T., ... & Hendon, T. (2016). Effects of field storage method on E. coli concentrations measured in storm water runoff. *Environmental monitoring and assessment*, *188*(3), 170.

- Huq, A., & Colwell, R. R. (1996). A microbiological paradox: viable but nonculturable bacteria with special reference to Vibrio cholerae. *Journal of food protection*, *59*(1), 96-101.

- IS10500, B. I. S. (2012). Indian standard drinking water–specification (second revision). *Bureau of Indian Standards (BIS), New Delhi*.

- Jagals, P., Grabow, W. O. K., Griesel, M., & Jagals, C. (2000). Evaluation of selected membrane filtration and most probable number methods for the enumeration of faecal coliforms, Escherichia coli and Enterococci in environmental waters. *Quantitative Microbiology*, *2*(2), 129-140.

- Jones, T., Gill, C. O., & McMullen, L. M. (2004). The behaviour of log phase Escherichia coli at temperatures that fluctuate about the minimum for growth. *Letters in applied microbiology*, *39*(3), 296-300.

- Kana, B. D., Gordhan, B. G., Downing, K. J., Sung, N., Vostroktunova, G., Machowski, E. E., ... & Mizrahi, V. (2008). The resuscitation-promoting factors of Mycobacterium tuberculosis are required for virulence and resuscitation from dormancy but are collectively dispensable for growth in vitro. *Molecular microbiology*, *67*(3), 672-684.

- Keep, N. H., Ward, J. M., Robertson, G., Cohen-Gonsaud, M., & Henderson, B. (2006). Bacterial resuscitation factors: revival of viable but non-culturable bacteria. *Cellular and molecular life sciences*, *63*(22), 2555.

- Keshavarz, Z., & Torkian, H. (2018). Application of ANN and ANFIS models in determining compressive strength of concrete. *Journal of Soft Computing in Civil Engineering*, *2*(1), 62-70.

- Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455-5516.

- Kinjo, Y., & Ueno, K. (2011). iNKT cells in microbial immunity: recognition of microbial glycolipids. *Microbiology and immunology*, *55*(7), 472-482.

- Kolling, G. L., & Matthews, K. R. (2001). Examination of recovery in vitro and in vivo of nonculturable Escherichia coli O157: H7. *Applied and environmental microbiology*, *67*(9), 3928-3933.

- Leclair, R. M., McLean, S. K., Dunn, L. A., Meyer, D., & Palombo, E. A. (2019). Investigating the effects of time and temperature on the growth of Escherichia coli O157: H7 and Listeria monocytogenes in raw cow's milk based on simulated consumer food handling practices. *International journal of environmental research and public health*, *16*(15), 2691.

- Lee, D. W., Gwack, J., & Youn, S. K. (2012). Enteropathogenic Escherichia coli Outbreak and its Incubation Period: Is it Short or Long?. *Osong public health and research perspectives*, *3*(1), 43-47.

- Leininger, D. J., Roberson, J. R., & Elvinger, F. (2001). Use of eosin methylene blue agar to differentiate Escherichia coli from other gram-negative mastitis pathogens. *Journal of veterinary diagnostic investigation*, *13*(3), 273-275.

- Li, L., Mendis, N., Trigui, H., Oliver, J. D., & Faucher, S. P. (2014). The importance of the viable but non-culturable state in human bacterial pathogens. *Frontiers in microbiology*, *5*, 258.

- Linder, K. A. T. H. E. R. I. N. E., & Oliver, J. D. (1989). Membrane fatty acid and virulence changes in the viable but nonculturable state of Vibrio vulnificus. *Applied and Environmental Microbiology*, *55*(11), 2837-2842.

- Liu, Y., Wang, C., Tyrrell, G., & Li, X. F. (2010). Production of Shiga-like toxins in viable but nonculturable Escherichia coli O157: H7. *Water Research*, *44*(3), 711-718.

- Lonsane, B. K., Parhad, N. M., & Rao, N. U. (1967). Effect of storage temperature and time on the coliforms in water samples. *Water Research*, *1*(4), 309-316.

- Maalej, S., Gdoura, R., Dukan, S., Hammami, A., & Bouain, A. (2004). Maintenance of pathogenicity during entry into and resuscitation from viable but nonculturable state in Aeromonas hydrophila exposed to natural seawater at low temperature. *Journal of Applied Microbiology*, *97*(3), 557-565.

- Maas, M. B., Perold, W. J., & Dicks, L. M. T. (2017). Biosensors for the detection of Escherichia coli. *Water Sa*, *43*(4), 707-721.

- Maier, A., Krolik, J., Fan, S., Quintin, P., McGolrick, D., Joyce, A., & Majury, A. (2015). Evaluating appropriate maximum holding times for private well water samples. *Environmental Health Review*, *58*(2), 35-40.

- Makino, S. I., Kii, T., Asakura, H., Shirahata, T., Ikeda, T., Takeshi, K., & Itoh, K. (2000). Does enterohemorrhagic Escherichia coli O157: H7 enter the viable but nonculturable state in salted salmon roe?. *Applied and Environmental Microbiology*, *66*(12), 5536-5539.

- Mary, P., Chihib, N. E., Charafeddine, O., Defives, C., & Hornez, J. P. (2002). Starvation survival and viable but nonculturable states in Aeromonas hydrophila. *Microbial ecology*, 250-258.

- MathWorks: Fuzzy Inference System Modeling – MATLAB & Simulink – MathWorks India, available at, last access: 14 October 2019.

- Negnevitsky, M. (2005). *Artificial intelligence: a guide to intelligent systems*. Pearson education.

- Okwu, M. O., & Adetunji, O. (2018). A comparative study of artificial neural network (ANN) and adaptive neuro-fuzzy inference system (ANFIS) models in distribution system with nondeterministic inputs. *International Journal of Engineering Business Management, 10*, 1847979018768421.

- Olstadt, J., Schauer, J. J., Standridge, J., & Kluender, S. (2007). A comparison of ten USEPA approved total coliform/E. coli tests. *Journal of water and health*, *5*(2), 267-282.

- Oshiro, R. (2002). Method 1604: Total Coliforms and Escherichia coli in water by membrane filtration using a simultaneous detection technique (MI Medium). *Washington, DC: US Environmental Protection Agency*.

- Saxena, T., Kaushik, P., & Mohan, M. K. (2015). Prevalence of E. coli O157: H7 in water sources: an overview on associated diseases, outbreaks and detection methods. *Diagnostic microbiology and infectious disease, 82*(3), 249-264.

- Tamerat N, Muktar Y, Shiferaw D (2016) Application of molecular diagnostic techniques for the detection of E. coli O157: H7: a review. J Vet Sci Technol 7(362):1–9.

- US Environmental Protection Agency. (2002). Method 1103.1: Escherichia coli (E. coli) in water by membrane filtration using membrane-thermotolerant Escherichia coli agar (mTEC). *EPA 821-R-02-020*.

- US EPA. (2009). Drinking water standards and health advisories table.

- WHO. Guidelines for Drinking-water Quality. 4th ed.; WHO (2011) Geneva, Switzerland.

4. Superposition learning-based model for prediction of *E. coli* in groundwater using physico-chemical water quality parameters

The prediction of waterborne bacteria is crucial to prevent health risks. Therefore, there is a need to study groundwater quality by predicting the presence of E. coli. The experimental data for prediction was obtained from BITS-UVA (University of Virginia) groundwater contamination project, having 1301 experimental laboratory results synthesized to test the physical, chemical, and microbiological parameters of water. Sensitivity analysis is performed to study the importance of physico-chemical water quality parameters on E. coli concentration. The superposition-based learning algorithm (SLA) is proposed to study the importance of water quality parameters to predict E. coli in groundwater. The predictive models were developed using MATLAB (R2019b) software.

4.1 Introduction

Water is crucial for human sustenance. An adequate, accessible, and safe supply is needed to be available to the consumers. By improving access to clean drinking water, it will result in substantial health benefits. Efforts should be made to attain groundwater quality as clean as possible for drinking (WHO, 2017). Water helps to maintain the moisture in the internal organs of the body (Gerald, 2011). It also protects the volume and uniformity of blood and lymph fluids (Dooge, 2001), controls body temperature, and eliminates toxins from the body through urine, sweat, and respiration (Molden, 2013), which is essential for maintaining skin functions (Burton et al., 1987). Water pollution can lead to kidney failure and can cause death (David et al., 2011). In the present situation, people are struggling to obtain access to clean water. Generally, infectious diseases are caused due to the presence of human or animal waste in groundwater. Some primary health diseases are caused by micro-organisms, including bacteria, pathogens, viruses, etc., because they can survive, reproduce, and spread in water (Payus et al., 2018). About 37.7 million people in India are affected by waterborne diseases annually, and 1.5 million children have died from diarrhea (WHO, 2017). India had recorded 10,738 deaths from 2012-2017. Uttar Pradesh had recorded the highest deaths due to diarrhea, followed by Assam, West Bengal, Delhi, and Madhya Pradesh (CBHI, 2018; WHO, 2017; India Water Portal, 2019).

4.2 Artificial Intelligence (AI)

Applications of artificial intelligence (AI) in water, ecology and environmental engineering were reported at the beginning of the 1990s. However, in recent years, Artificial Neural Networks (ANNs) have been used intensively for prediction and forecasting in several engineering and water-related areas, including water resource studies by Liong and Sivapragasam, 2002; Muttil and Chau, 2006; El-Shafie et al. 2008; El-Shafie et al. 2011; Noureldin et al. 2011; Najah et al. 2009; oceanography by Makarynska et al. 2008 and environmental engineering by Grubert, 2003. The initial research in coliform bacteria prediction dates back to 2003, which includes analyzing and developing a model to predict bathing waters compliance along the coastline of the Firth of Clyde, situated in the southwest of Scotland, UK. This study used rainfall, river discharge, sunlight, and tidal conditions as inputs of these networks, and fecal coliforms were used as an output. River discharge was found to be the most significant input variable to the bacterial concentration, and the height of the high tide was found to be relatively significant (Lin et al., 2003).

A few studies have evaluated the relation between *E. coli* bacteria and physic-chemical parameters in water samples. The effects of temperature, pH, and water activity were studied, out of which temperature was found to be the most crucial parameter for the thermal inactivation of *E. coli*. The limitation of the model was that it could not provide a prediction equation for the inactivation rate of bacteria (Lou and Nakai, 2001). Results show that the growth of *E. coli* was correlated with pH, dissolved oxygen (DO), specific conductivity (SC), and water temperature (T). An AI-based model on multiple regression analysis was developed to predict coliform bacteria concentrations at the selected sites based on available USGS NWIS data (David and Haggard, 2011). Francy et al. (2013) proposed a model to evaluate the correlation between *E. coli* and other parameters based on water samples collected from eight inland recreational lakes in Ohio. The parameters used were rainfall, wind direction, speed, turbidity, and water temperature, but the model was not developed at sites where the *E. coli* concentration was exceeded. The model was used to predict environmental quality along the Penchala River urban catchment area in Kuala Lumpur, Malaysia, directly affected by human activities (Zamani and Saybani, 2014). The model provided the best training performance, with 70 neurons in the hidden layer.

Mouna et al. (2014) observed factors that have an impact on the growth of *E. coli* with a negative correlation of higher salinity (R=-0.97) and pH (R=-0.98) against a positive correlation of higher turbidity (R=0.93). Islam et al. (2017) developed a linear regression model to evaluate the impact of the atmospheric conditions on bacterial counts. Precipitation and temperature of the water showed a positive correlation with the growth of bacteria. Cheng et al. (2013) studied the correlation between water quality parameters and *E. coli* growth by Pearson's correlation analysis. They observed that the density of bacteria varied negatively with pH and the removal of total suspended solids. The growth of *E. coli* in aeration pond was negatively correlated with the increased dissolved oxygen. Shamsudin et al. (2016) observed *E. coli* growth using Pearson correlation. The results showed a linear relationship with pH (R=0.971), time (R=0.958), turbidity (R=0.885), dissolved oxygen (R=-0.861), and temperature (R=0.763). Rao et al. (2015) examined the relationship between water quality and turbidity. Increased *E. coli* cell counts were significantly associated with increased turbidity ($\beta = 0.003$; $p < 0.0001$) and decreased dissolved oxygen concentrations ($\beta = -0.310$; $p < 0.0001$). The water quality prediction model was developed with water samples from 4 different sampling stations on the Panchaganga River for modeling river quality with BOD and DO parameters (Mulla et al., 2016).

Bisi-Johnson et al. (2017) studied the physico-chemical and microbial properties of water. The relationship between total suspended solids (TSS) and turbidity with *E. coli* was studied by Huey et al. (2010). In all four of the watersheds, a significant correlation was observed between turbidity and *E. coli*. In the three perennial watersheds, a correlation was observed between TSS and *E. coli*. Pachepsky et al. (2018) studied the relationship of temperature, pH, dissolved oxygen (DO), turbidity, nitrate, ammonium with *E. coli* concentration, using Spearman's rank correlation coefficient of -0.247, -0.267, -0.246, -0.293, 0.015, -0.220 respectively. In a few studies, statistical models were used for prediction (Katip, 2018). Mohammed et al. (2018) study was based on observed counts of bacteria and measured water quality parameters, including pH, temperature, conductivity, turbidity, colour, and alkalinity. Considerable improvement in the efficiency of the model was achieved when the input data was normalized before training. Raw water turbidity, colour, and alkalinity were found to have a significant influence on the growth of *E. coli* (Mohammed et al., 2018). Recent studies examined the effects of physico-chemical water quality parameters on the growth of *E. coli* in surface water sources, rivers, beaches, etc. However, no study has been done to examine the effects of physico-chemical water quality parameters on on the growth of *E. coli* in groundwater sources. There is a need to study the impact on *E. coli* growth in groundwater sources.

4.3 Methodology

In this study, we create our dataset using water samples collected from eight districts of Rajasthan, India, under the BITS-UVA groundwater contamination project. The objective of this study is to identify the significant water quality parameters that affect the *E. coli* concentration in groundwater. The dataset was fed into an ANN model using MATLAB (R2019b) software. The model performance was assessed by using two evaluation measures, such as MSE and R^2. We have proposed a superposition-based learning algorithm (SLA) to observe the patterns of ANN models and improve the performance of ANN models to predict *E. coli* in groundwater.

The following methodology has been adopted for the prediction of *E. coli*:

- Water sampling and laboratory testing.
- ANN modeling was performed to obtain a correlation between physico-chemical water quality parameters and *E. coli*.
- Sensitivity analysis was performed to study the importance of different water quality parameters on *E. coli* concentration.
- The randomness of the ANN model was reduced with the application of the SLA algorithm by optimization of weights and bias between the input layer and the output layer of the model.
- ANN modeling was performed using a Superposition-based learning algorithm for the prediction of *E. coli* in groundwater.

4.4 Artificial Neural Networks (ANN)

Artificial intelligence (AI) is a computer science branch mainly concerned with using computational models to understand how humans think and behave (Tanimoto 1987). AI techniques play a leading role in data modeling and providing high-speed computational tools and methods (Rykiel 1989). Many AI-based techniques are used for relationship, estimation, classification, prediction, and segmentation. Each AI technique has its advantages in particular applications. Various factors affect the quality of water and have nonlinear relationships with each other. However, it can be challenging to analyze water without extensive, accurate, and detailed data (Moustakis et al., 1996). Many models have been developed and widely used to resolve water quality analysis problems. Although the models have been adopted for modeling water quality, they require large amounts of data, multiple parameters, and detailed information on the source environment to obtain reliable results (Adriaans et al., 1996).

A feed-forward neural network is the most commonly used supervised learning method in which signals are allowed to travel in the forward direction from input nodes to output nodes. A feed-forward neural network is generally governed by equation (4.1):

$$y_j = f_j\left(\sum_{i=1}^{m} w_{j,i} \cdot y_i + b_i\right) \qquad (4.1)$$

Where,

y_j = output

f_j = Transfer function of the j^{th} neuron in a layer

$w_{j,i}$ = Weight that connects the output y_i of the i^{th}neuron from one layer to the input of the j^{th} neuron in the next layer

b_i = Bias weight on the j^{th}neuron of each layer.

4.4.1 Data Normalization

It is a critical step in the preprocessing level as the machine learning algorithms majorly show improved performance when they deal with features that lie on the same scales. One way of doing this is by normalization. This method works by rescaling the features to a range of (0, 1), which is a particular case of min-max scaling. To normalize the data, we applied the min-max scaling method to each parameter. We checked if the data collected has some missing values which may occur due to improper collection of data, blank values, or NaN or measurements not applicable. Min-max normalization is given by equation (4.2):

$$Xn=\frac{Xr-Xmin}{Xmax-Xmin} \qquad (4.2)$$

Where,

Xn = normalized value

Xr = raw value

$Xmin$ = minimum observed values of X

$Xmax$ = maximum observed values of X

4.4.2 Splitting the datasets into subsets

Initially, a total of 1301 data was divided into training (70%), testing (15%), and validation (15%). For an ANN to generate an output vector that is as close as possible to the target vector, a training procedure was employed, the objective of which was to minimize the mean square. Training is a mechanism by which ANN connection weights are adjusted using a continuous stimulation process. The training process includes the involved modification of the synaptic weights and the bias terms to achieve the primary function of decreasing the error rate. The training algorithms are techniques of optimization which help to fulfill the objective function. Finally, the goal is to make the model generalized to unseen data.

4.4.3 Training functions

Backpropagation applies a gradient descent search through a space of potential network weight, reducing the error (MSE) between training example and target value and network output iteratively. It allowed merely converging to some local minima. Neurons are structured in one layer, with inputs connecting to each neuron and weights. Training in such a network starts by changing the weights connected with the inputs so that the network can identify the input arrangements. There are numerous backpropagation training algorithms in MATLAB (Beale et al., 1992) with this input-output relationship pattern and neural network architecture. In this analysis, two training functions, Levenberg-Marquardt backpropagation (Hayati et al., 2007) and Bayesian regularization backpropagation (MacKay, 1992), were used for training and validation steps.

Levenberg-Marquardt is a network training function that updates the weights and bias values according to Levenberg-Marquardt optimization (trainlm) (Hayati et al., 2007; Foresee et al., 1997; Hagan et al., 1994). If the output function is a subset of the squares, equation (4.3) can be approximate to the Hessian matrix:

$$H = J^T J \qquad (4.3)$$

The gradient can be computed by equation (4.4);

$$g = J^T e \qquad (4.4)$$

Equation (4.5) uses this approximation to the Hessian matrix, used by the Levenberg-Marquardt algorithm:

$$X_{k+1} = X_k - [J^T J + \mu l]^{-1} J^T e \qquad (4.5)$$

Where,

J = Jacobian matrix

e = vector of network errors.

Bayesian regularization backpropagation (trainbr) is a network training function that updates the Levenberg-Marquardt optimization weight and bias values (MacKay, 1992; Torrecilla et al., 2008; Singh et al., 2011). The performance index for the Bayesian regularization method is given by equation (4.6):

$$F = \beta . E_d + \alpha . E_w \qquad (4.6)$$

Where,

α, β = parameters to be optimized.

E_d = mean sum of the squared network errors.

E_w = sum of the squares of the network weights.

4.4.4 Adaption learning functions

The optimized weight values function across artificial neural networks. The method by which we obtain the optimized weight values is called learning. When the equivalent input is presented, the learning process teaches the network to produce the output. Learning is achieved if the trained neural network produces the output within the desired accuracy equivalent to an input pattern with the updated weights. The whole learning method is composed of the following three computations; input layer, hidden layer, and output layer computation.

Gradient descent with momentum weight and bias learning function and gradient weight and bias learning function was used to study the neural network input-output relationship and architecture pattern. The goal of gradient descent (learngd) is to find the values of weight to minimize error. A neural network with one hidden layer, the weight update was determined by the partial derivatives chain rule. Individually, the adjustment of the weight Δw can be calculated using equation (4.7):

$$\Delta w = -\eta \frac{\partial E}{\partial w} \tag{4.7}$$

Where,

E = classification error on iteration

η = learning rate

w = weights

Gradient descent with momentum weight and bias learning function (learngdm) is an advanced gradient descent method that adds an impulse element to the GD algorithm's weight adjustment formula. The equation for weight adjustment of the GDM algorithm on iteration r is given by equation (4.8):

$$\Delta w^{r} = -\eta \frac{\partial E}{\partial w} + \alpha \Delta w^{r-1} \tag{4.8}$$

4.4.5 Activation functions

Activation functions are used in neural networks to calculate the weighted sum of inputs and biases and decide whether a neuron may or may not be fired. It usually manipulates the data presented through gradient descent and then produces an output for the neural network containing the data parameters. The sigmoid activation function is called the logistic function or squashing function (Turian et al., 2009). The Sigmoid is a non-linear activation function with positive derivate and some degree of smoothness (Han and Moraga, 1995), primarily used in feedforward neural networks. The equation (4.9) gives the sigmoid function:

$$f_x = \left(\frac{1}{(1+ exp^{-x})}\right) \tag{4.9}$$

4.4.6 Performance functions

The model performance was evaluated using the value of the coefficient of determination (R^2) and mean square error (MSE). The mean squared error (MSE) quantifies the difference between the predicted and actual values of the measured quantity. The value of R^2 reflects the proportion of variance in the dependent parameter described by the independent parameter. The higher value of R^2 indicates that the model explains variation in the dependent parameter. The value of R^2 and MSE can be calculated using equation (4.10) and equation (4.11), respectively:

$$R^2 = \frac{\sum_{j=1}^{m}(t_{pi} - t_{pi})\left(y_{pi} - y_{pi'}\right)}{\left[\sqrt{\sum_{j=1}^{n}(t_{pi} - t_{pi'})^2 \sum_{i=1}^{n}(y_{pi} - y_{pi})^2}\right]^{½}} \tag{4.10}$$

$$\text{MSE} = \frac{1}{2}\sum_{j=1}^{m}(t_{pi} - y_{pi})^2 \tag{4.11}$$

Where,

t_{pj} = Target or real value p.

y_{pj} = i^{th} output of the final layer or predicted value.

$t_{pj\prime}$ = mean of targeted or real value.

$y_{pj\prime}$ = mean of the predicted value.

n = number of datasets.

4.4.7 Model Architecture

A trial and error procedure was adopted to obtain the optimum structure of the network. A rigorous analysis was carried out with one, two, and three hidden layers. The architecture of the feed-forward neural network (Figure 4.1). The back-propagation of error was carried out by the Bayesian regularization (BR), a standard second-order nonlinear least-squares technique using the back-propagation process to increase the speed and efficiency of the training.

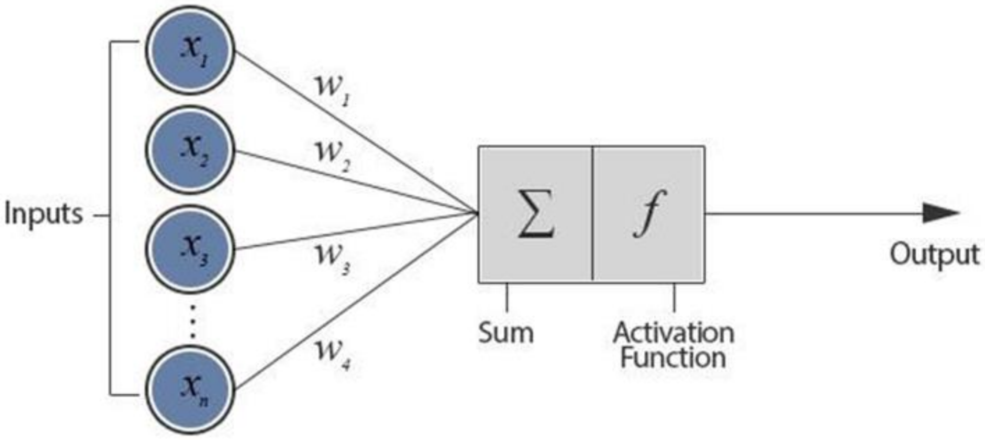


Figure 4.1: The architecture of a feed-forward neural network.

4.4.8 Prediction of *E. coli*

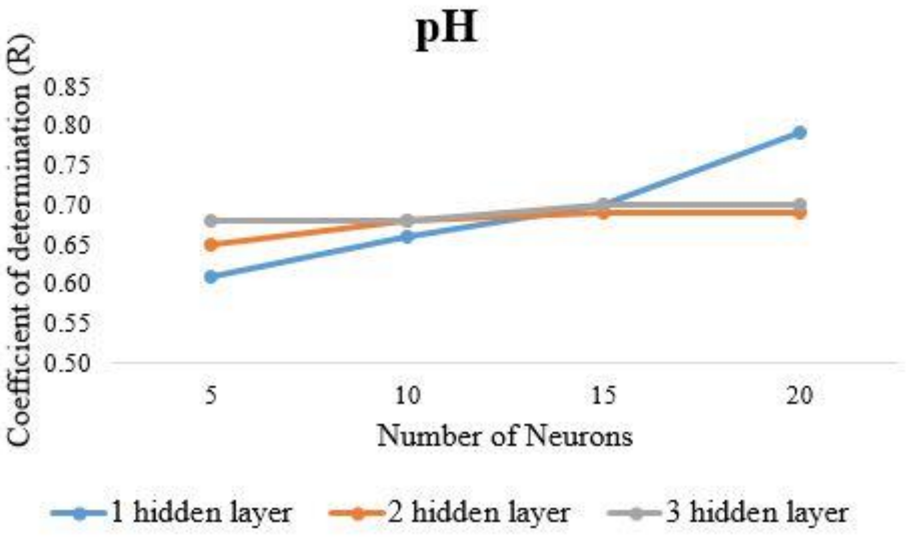
Linear square fitting was used to study the relationship between physico-chemical water quality parameters and *E. coli*. The coefficient of correlation (R^2) was calculated between input parameters and *E. coli*. Let A and B be the two parameters, the values of the constants A and B can be calculated by the equation (4.12), according to the method of least squares:

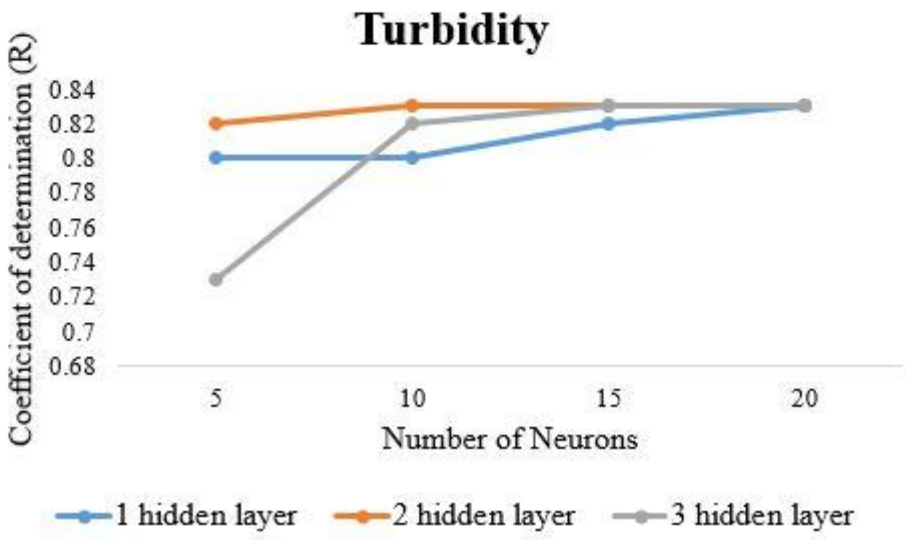
$$B = X.A + Y \qquad (4.12)$$

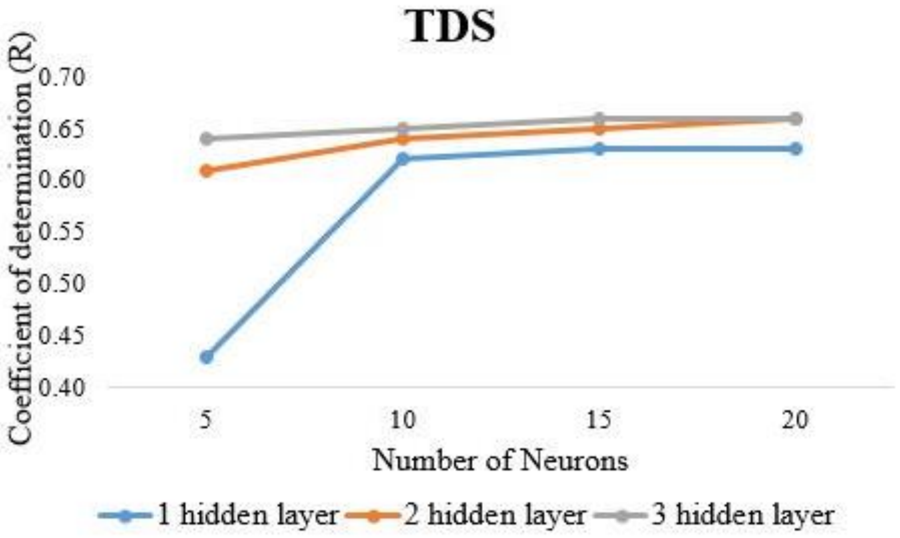
A feed-forward neural network with multiple hidden layers and physico-chemical water quality parameters as input was optimized using the Levenberg Marquardt Training Algorithm (LM), Bayesian regularization backpropagation (BR), and sigmoid activation function had outperformed all other combinations. Figure 4.2 shows the correlation of *E. coli* with physico-chemical water quality parameters. ANN model was optimized by the Levenberg-Marquardt (LM) training function using the sigmoid activation function had outperformed all other combinations. Highest overall correlation was observed between *E. coli* and pH ($R^2 = 0.84$, MSE= 0.0204) with 15 neurons in one hidden layer, Turbidity ($R^2 = 0.83$, MSE= 0.0081) with 15 neurons in one hidden layer, TDS ($R^2 = 0.70$, MSE= 0.0275) with 10 neurons in one hidden layer, Electrical Conductivity ($R^2 = 0.37$, MSE= 0.0503) with 20 neurons in one hidden layer, Fluoride ($R^2 = 0.37$, MSE= 0.0462) with 20 neurons in one hidden layer, ORP ($R^2 = 0.28$, MSE= 0.0282) with 15 neurons in one hidden layer, and Nitrate ($R^2 = 0.26$, MSE= 0.0437) with 15 neurons in one hidden layer. The lowest overall correlation was observed between *E. coli* and Dissolved Oxygen ($R^2 = -0.05$, MSE= 0.0333) with 15 neurons in one hidden layer.

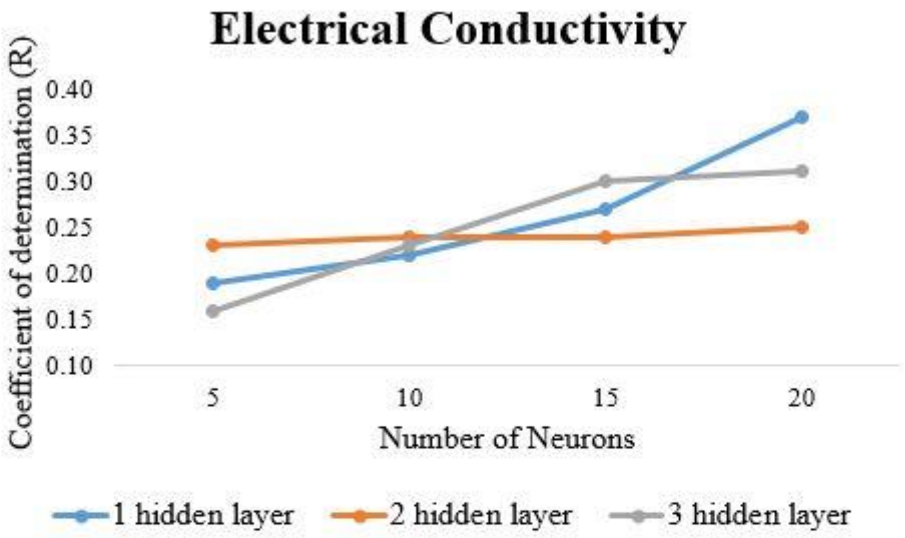
Figure 4.3 shows the performance of the models using the Levenberg-Marquardt (LM) training function having multiple hidden layers and the number of neurons. ANN model using pH as an input parameter displayed the best performance ($R^2 = 0.70$) with three hidden layers and 15 neurons in each layer. ANN model using Turbidity as an input parameter displayed the best performance ($R^2 = 0.83$) with three hidden layers and 15 neurons in each layer. ANN model using TDS as an input parameter displayed the best performance ($R^2 = 0.66$) with three hidden layers and 15 neurons in each layer. ANN model using Electrical Conductivity as an input parameter displayed the best performance ($R^2 = 0.37$) with one hidden layer and 20 neurons. ANN model using ORP as input parameter displayed the best performance ($R^2 = 0.25$) with one hidden layer and 20 neurons. ANN model using Fluoride as an input parameter displayed the best performance ($R^2 = 0.37$) with two hidden layers and 20 neurons in each layer. ANN model using Nitrate as an input parameter displayed the best performance ($R^2 = 0.26$) with three hidden layers and 20 neurons in each layer. ANN model using DO as an input parameter the best performance ($R^2 = -0.03$) with two hidden layers and 20 neurons in each layer.

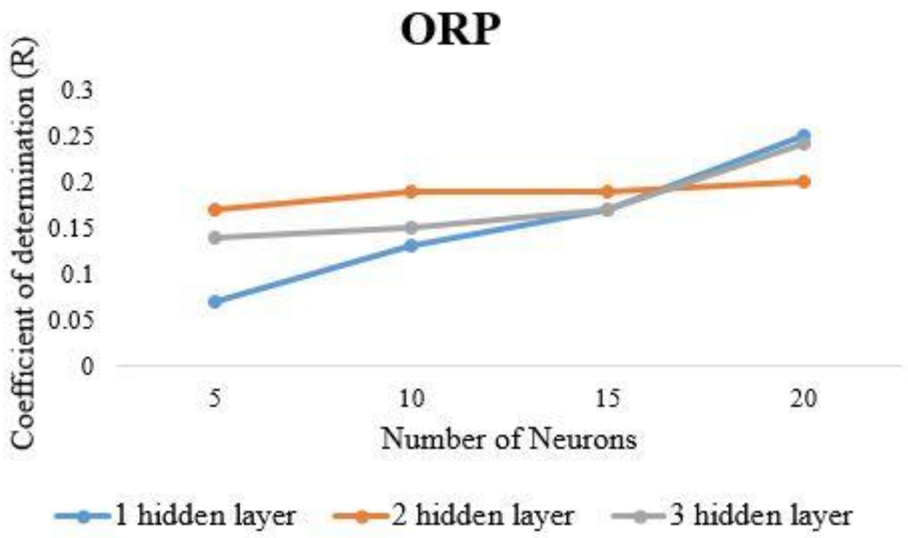
Figure 4.4 shows the performance of the models using the Bayesian regularization (BR) training function having multiple hidden layers and the number of neurons. ANN model using pH as input parameter displayed the best performance ($R^2 = 0.87$) with three hidden layers and 20 neurons in each layer. ANN modeling using Turbidity as an input parameter showed the best performance ($R^2 = 0.83$) with three hidden layers and 20 neurons in each layer. ANN model using TDS as input parameter showed the best performance ($R^2 = 0.70$) with three hidden layers and 20 neurons in each layer. ANN model using Electrical Conductivity as an input parameter displayed the best performance ($R^2 = 0.31$) with two hidden layers and 20 neurons in each layer. ANN model using ORP as an input parameter showed the best performance ($R^2 = 0.28$) with two hidden layers and 20 neurons in each layer. ANN model using Fluoride as an input parameter displayed the best performance ($R^2 = 0.29$) with three hidden layers and 20 neurons in each layer. ANN model using Nitrate as an input parameter showed the best performance ($R^2 = 0.26$) with two hidden layers and 20 neurons in each layer. ANN model using DO as an input parameter displayed the best performance ($R^2 = -0.05$) with three hidden layers and 20 neurons in each layer.

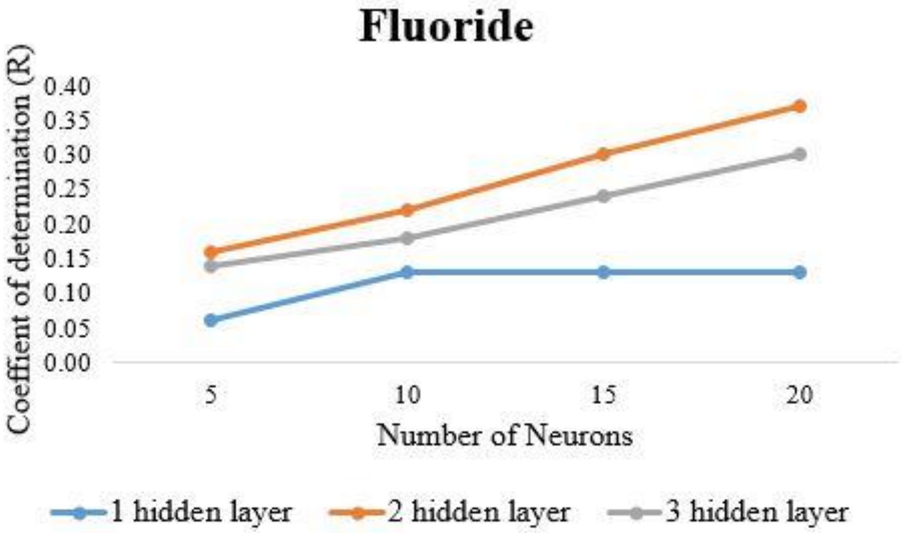


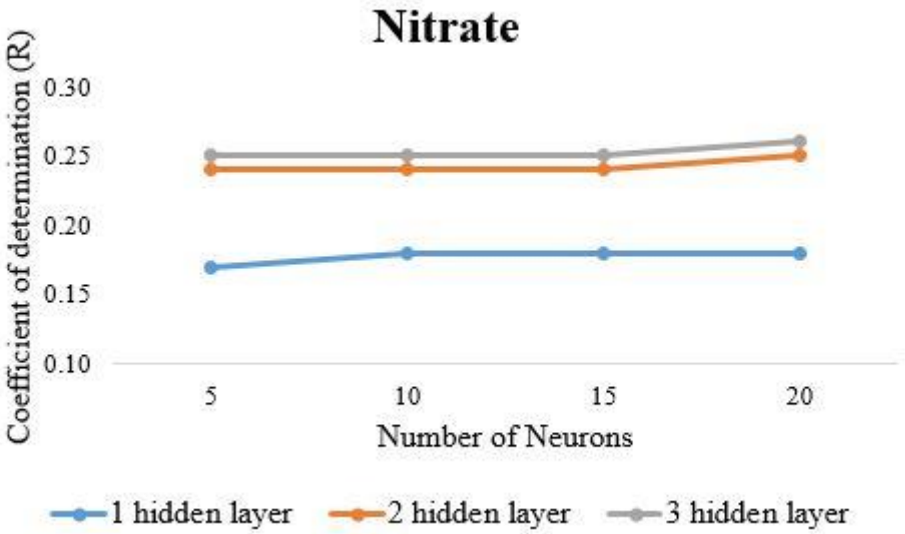

Turbidity
Coefficient of determination (R)
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


TDS
Coefficient of determination (R)
0.70
0.65
0.60
0.55
0.50
0.45
0.40
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


Electrical Conductivity
Coefficient of determination (R)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ORP
Coefficient of determination (R)
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


Fluoride
Coeffient of determination (R)
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


Nitrate
Coefficient of determination (R)
0.30
0.25
0.20
0.15
0.10
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer

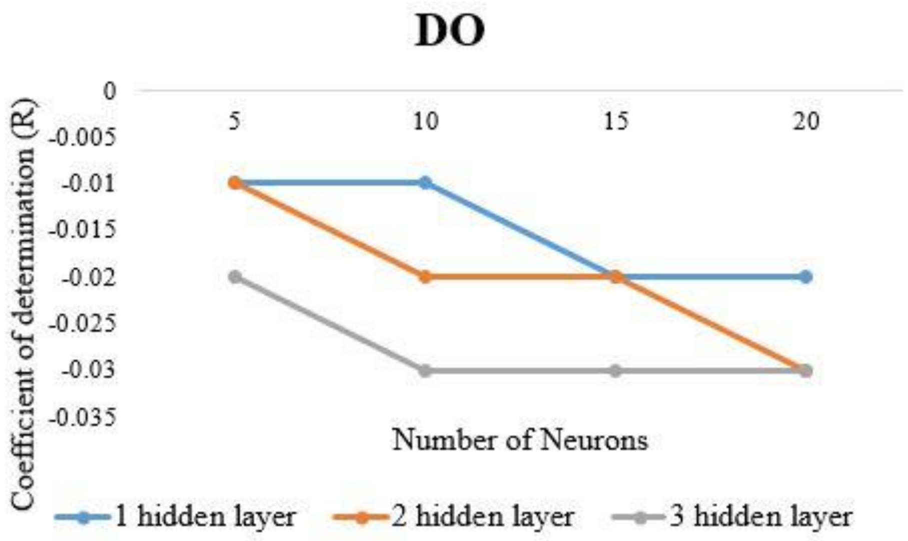
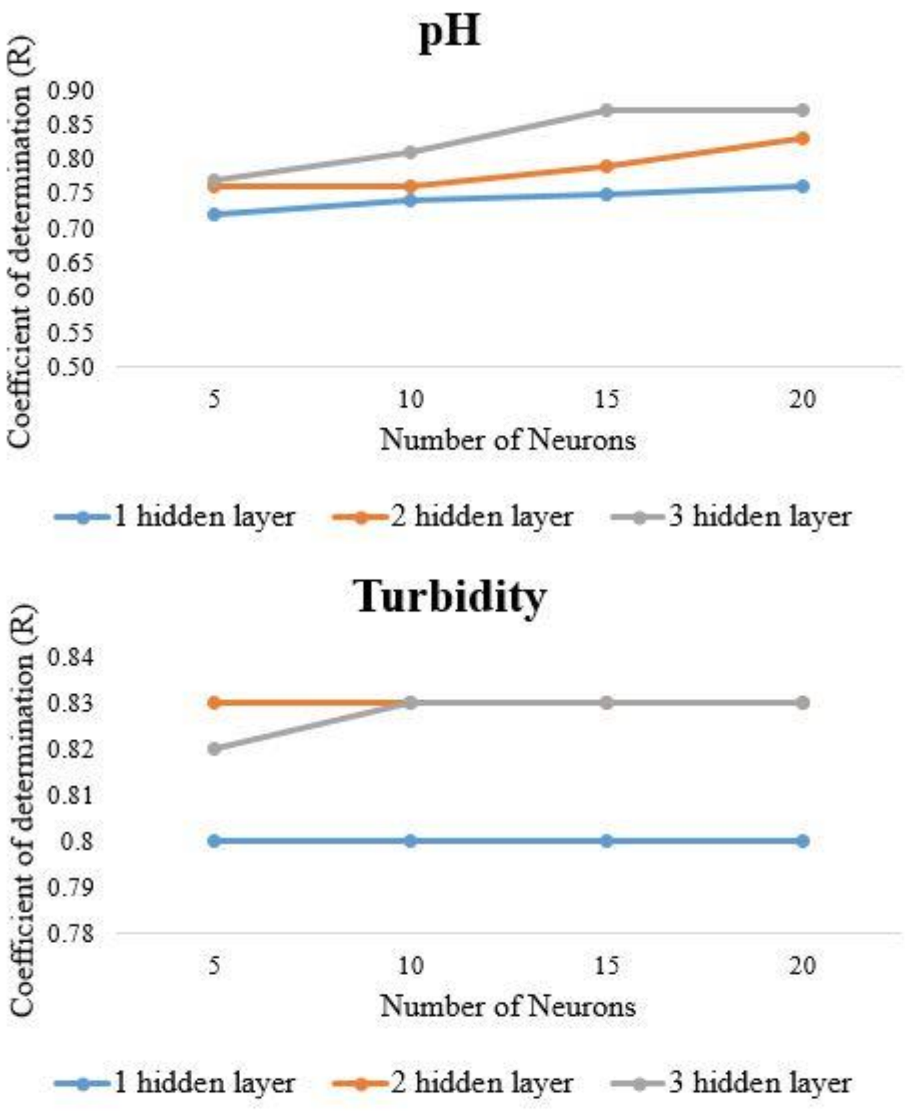
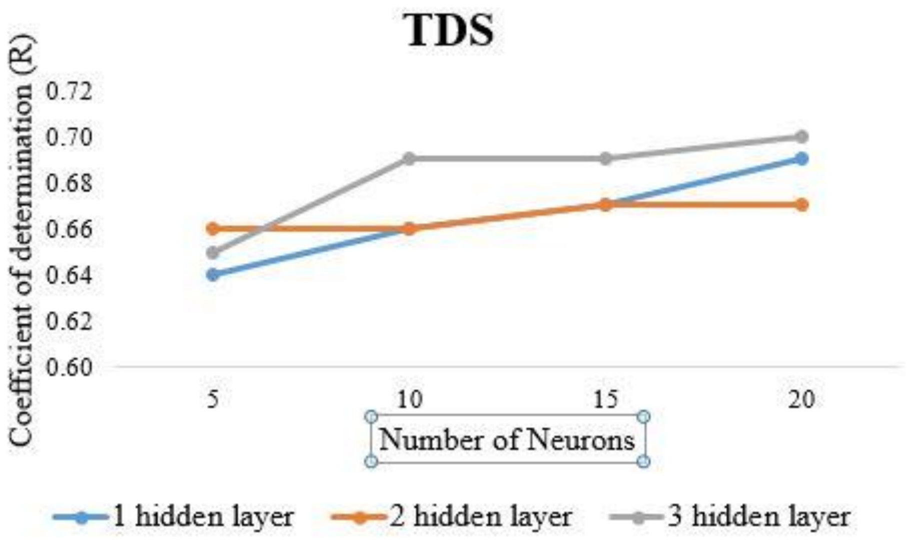
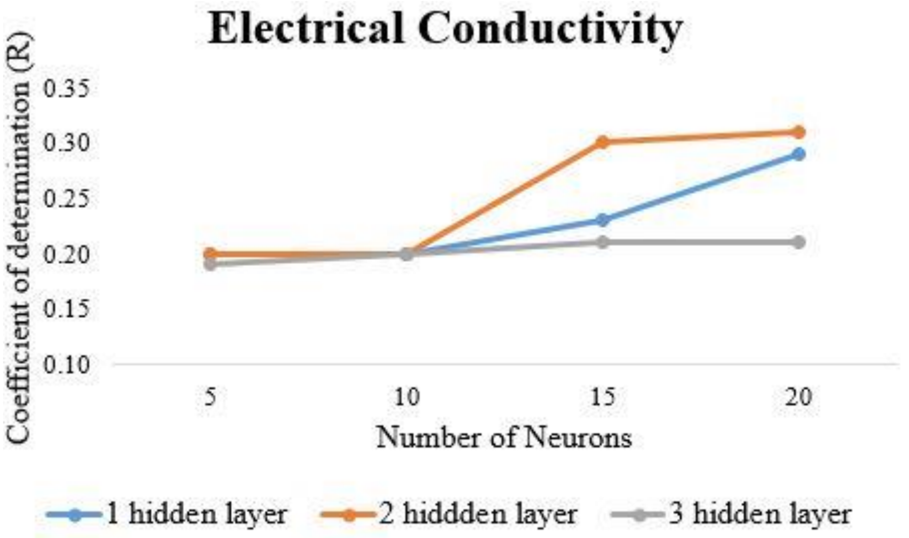
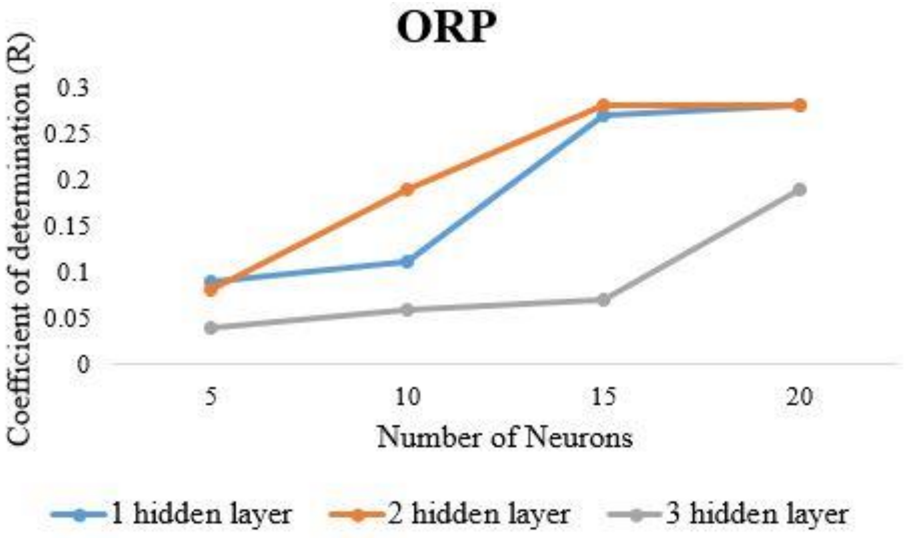


Figure 4.2: The performance of models using Levenberg-Marquardt training function.




TDS
Coefficient of determination (R)
0.72
0.70
0.68
0.66
0.64
0.62
0.60
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


Electrical Conductivity
Coefficient of determination (R)
0.35
0.30
0.25
0.20
0.15
0.10
5
10
15
20
Number of Neurons
1 hidden layer
2 hiddden layer
3 hidden layer


ORP
Coefficient of determination (R)
0.3
0.25
0.2
0.15
0.1
0.05
0
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer

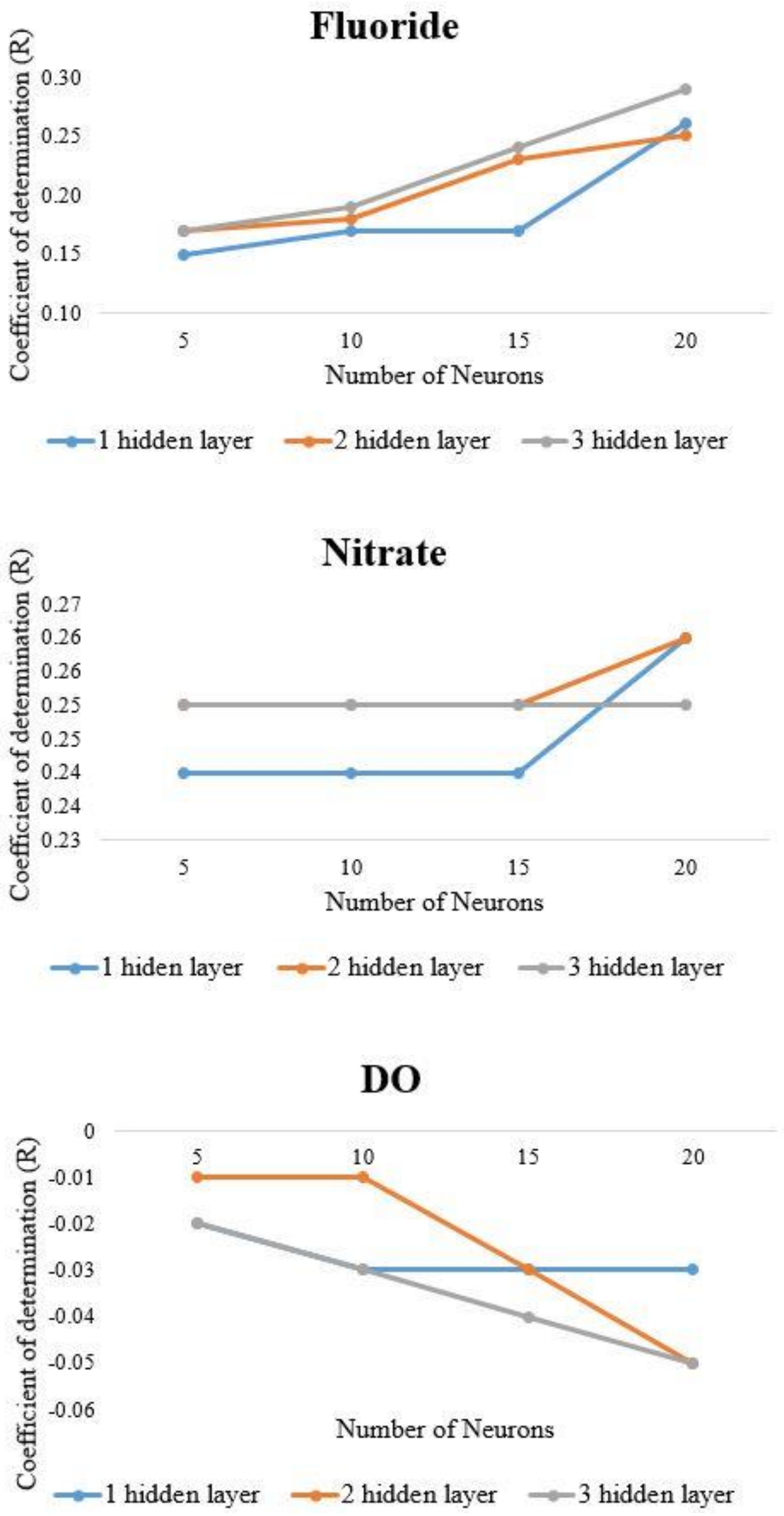


Figure 4.3: The performance of models using Bayesian regularization training function.

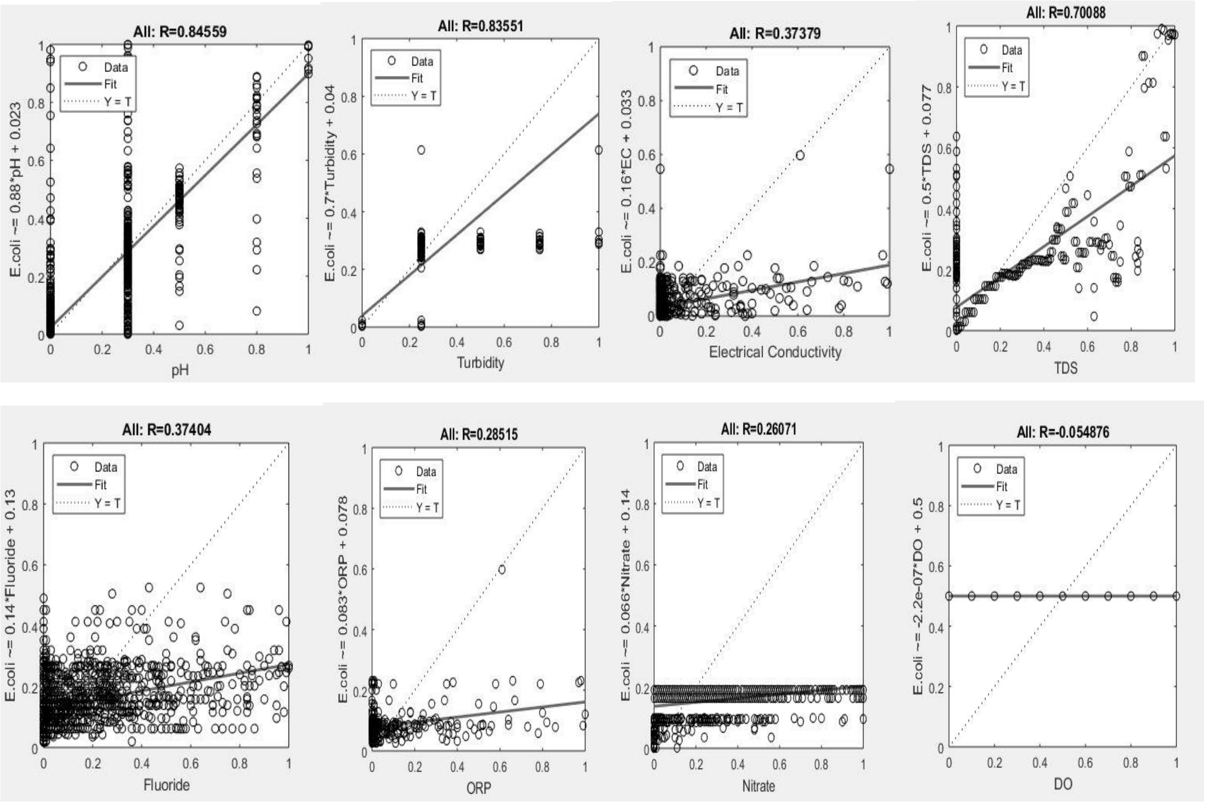


Figure 4.4: Correlation between physico-chemical water quality parameters and *E. coli*.

Using the method of least square, prediction of *E. coli* using equation (4.12):

- *E. coli* = 0.7*Turbidity + 0.04 (4.13)
- *E. coli* = 0.16*Electrical Conductivity + 0.033 (4.14)
- *E. coli* = $-2.2e^{-07}$*Dissolved Oxygen + 0.5 (4.15)
- *E. coli* = 0.14*Fluoride + 0.13 (4.16)
- *E. coli* = 0.066*Nitrate + 0.14 (4.17)
- *E. coli* = 0.083*ORP + 0.078 (4.18)
- *E. coli* = 0.88*pH + 0.023 (4.19)
- *E. coli* = 0.50*TDS + 0.077 (4.20)

No study has been done to predict *E. coli* in groundwater using pH, total dissolved solids, oxidation-reduction potential, dissolved oxygen, electrical conductivity, turbidity, fluoride and nitrate. The majority of the existing techniques are limited to most of the substantial water features to restrict the pH, temperature, turbidity, conductivity, and colour of the water. However, few significant water quality parameters are not considered, which have direct effects on the growth of *E. coli* bacteria. Our study examined the impact of various physico-chemical parameters of water quality on the growth of *E. coli*, and correlation was observed between *E. coli* and water quality parameters.

4.5 Sensitivity analysis

A sensitivity analysis was performed to study the effect on the output parameter when the input parameters are taken as average values. The input parameters were subjected to variability in a range of -10% to +10% of the average measured values. Each of the model input parameters was tested one at a time by keeping the others at their average values. Furthermore, the relative significance of these input parameters was ranked based on a sensitivity index. The first model was developed using all parameters as input parameters and named artificial neural network-E. coli-all parameters (ANN-E. coli-AP), which serve as a reference model. In order to evaluate the significance of all physico-chemical water quality parameters as input parameters for the ANN-E. coli-AP model, a sensitivity analysis was performed by excluding one parameter from eight parameters, and the performance of the ANN output model was evaluated using correlation coefficient (R^2) and mean squared error (MSE). Moreover, sensitivity analysis is instrumental and reliable when sufficient data is available and to assesses the relative importance of the parameter. The second model was developed, referred to as artificial neural network-E. coli-leave DO (ANN-E. coli-LD), which means DO was excluded in predicting the *E. coli* counts. The third model was developed, referred to as artificial neural network-E. coli-leave DO, fluoride (ANN-E. coli-LDF), which means DO and fluoride, were excluded for the prediction of *E. coli*.

The fourth model was developed, referred to as artificial neural network-E. coli-leave DO, fluoride, nitrate (ANN-E. coli-LDFN), which means DO, fluoride, and nitrate were excluded for the prediction. The fifth model was developed, referred to as artificial neural network-E. coli-leave DO, fluoride, nitrate, ORP (ANN-E. coli-LDFNO) means DO, fluoride, nitrate and ORP, were excluded for the prediction. The sixth model was developed, referred to as artificial neural network-E. coli-leave DO, fluoride, nitrate, ORP, conductivity (ANN-E. coli-LDFNOC) means DO, fluoride, nitrate, ORP and conductivity were excluded for the prediction. The seventh model was developed, referred to as artificial neural network-E. coli-leave DO, fluoride, nitrate, ORP, conductivity, and TDS (ANN-E. coli-LDFNOCT) which means DO, fluoride, nitrate, ORP, conductivity, and TDS were excluded for the prediction. The ANN-E.

coli-AP model with 1301 experimental values from 2016 to 2019 was taken as an input dataset, and models were developed using MATLAB R2019b software. The sensitivity index was calculated by equation (4.21):

$$SI = \left(\frac{\acute{Y}_i}{Y} - 1\right) \times 100 \qquad (4.21)$$

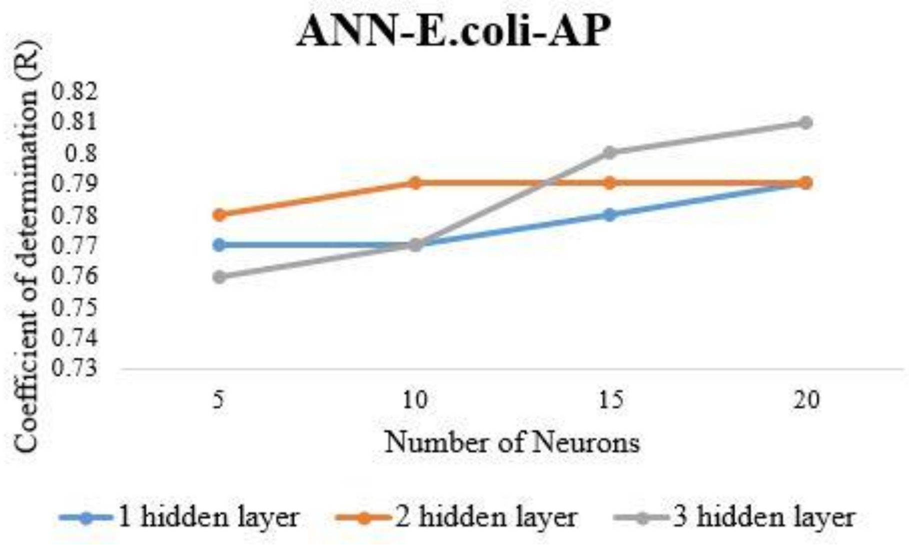
Where,

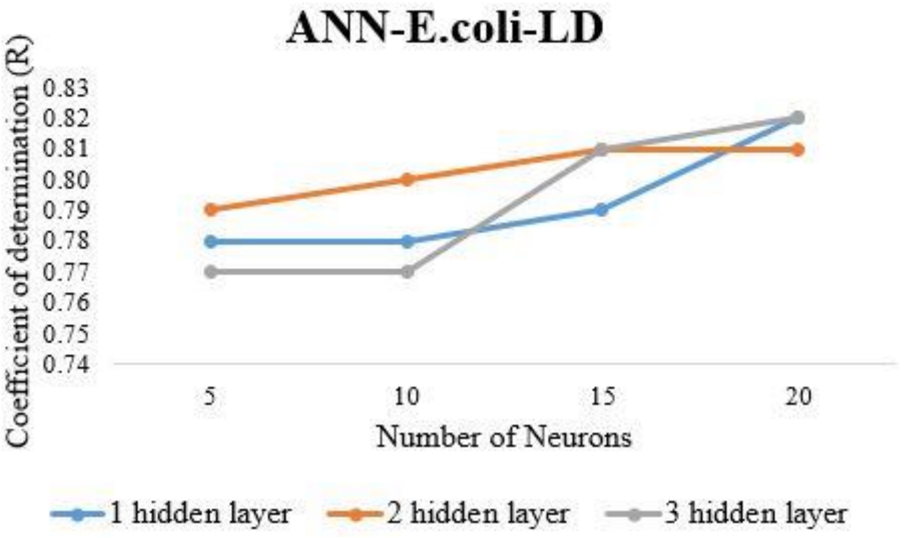
SI = sensitivity index.

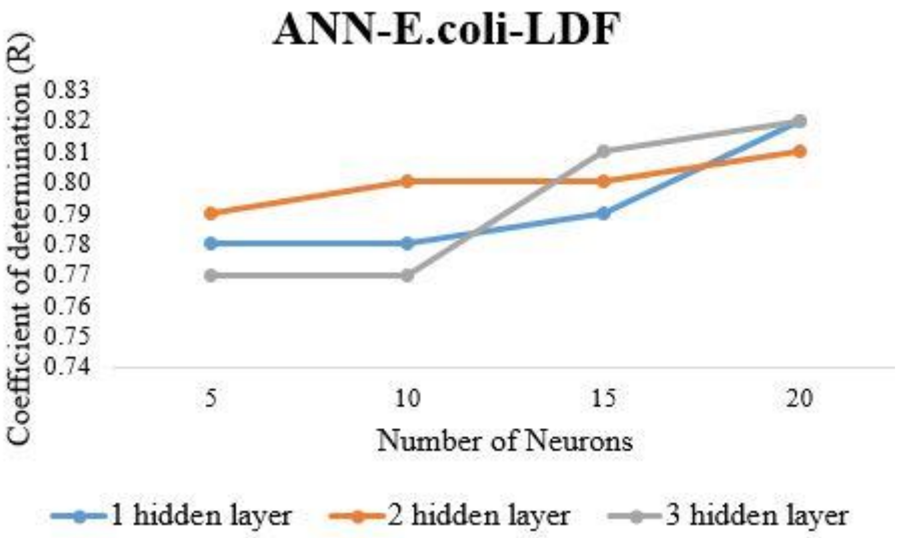
$\acute{Y}_i$ = predicted output value when input value varied.

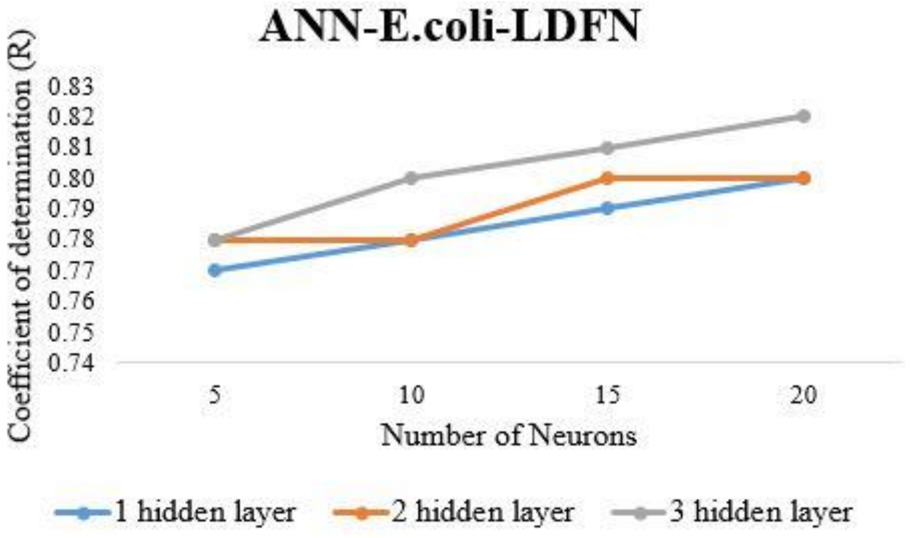
Y = average output value.

In order to evaluate the significance of all physico-chemical water quality parameters as input parameters for the ANN-E. coli-AP model, a sensitivity analysis was performed for seven models by excluding one parameter from eight parameters. The performance of the output model was evaluated using coefficient correlation (R^2) and mean squared error (MSE). Results from the sensitivity analysis showed that ANN-E. coli-AP, ANN-E. coli-LD, ANN-E. coli-LDF, ANN-E. coli-LDFN, ANN-E. coli-LDFNO, ANN-E. coli-LDFNOC, and ANN-E. coli-LDFNOCT have overall highest R^2 values of 0.87, 0.87, 0.86, 0.87, 0.90, 0.86 and 0.86 respectively; minimum lowest MSE values of 0.0315, 0.0215, 0.0587, 0.0298, 0.0892, 0.0179 and 0.1069 respectively. The performance of the sensitivity analysis-based ANN model with multiple hidden layers and the number of neurons using the Levenberg-Marquardt (LM) training functions (Figure 4.5) and Bayesian regularization (BR) training functions (Figure 4.6), respectively.




ANN-E.coli-LD
Coefficient of determination (R)
0.83
0.82
0.81
0.80
0.79
0.78
0.77
0.76
0.75
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ANN-E.coli-LDF
Coefficient of determination (R)
0.83
0.82
0.81
0.80
0.79
0.78
0.77
0.76
0.75
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ANN-E.coli-LDFN
Coefficient of determination (R)
0.83
0.82
0.81
0.80
0.79
0.78
0.77
0.76
0.75
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer

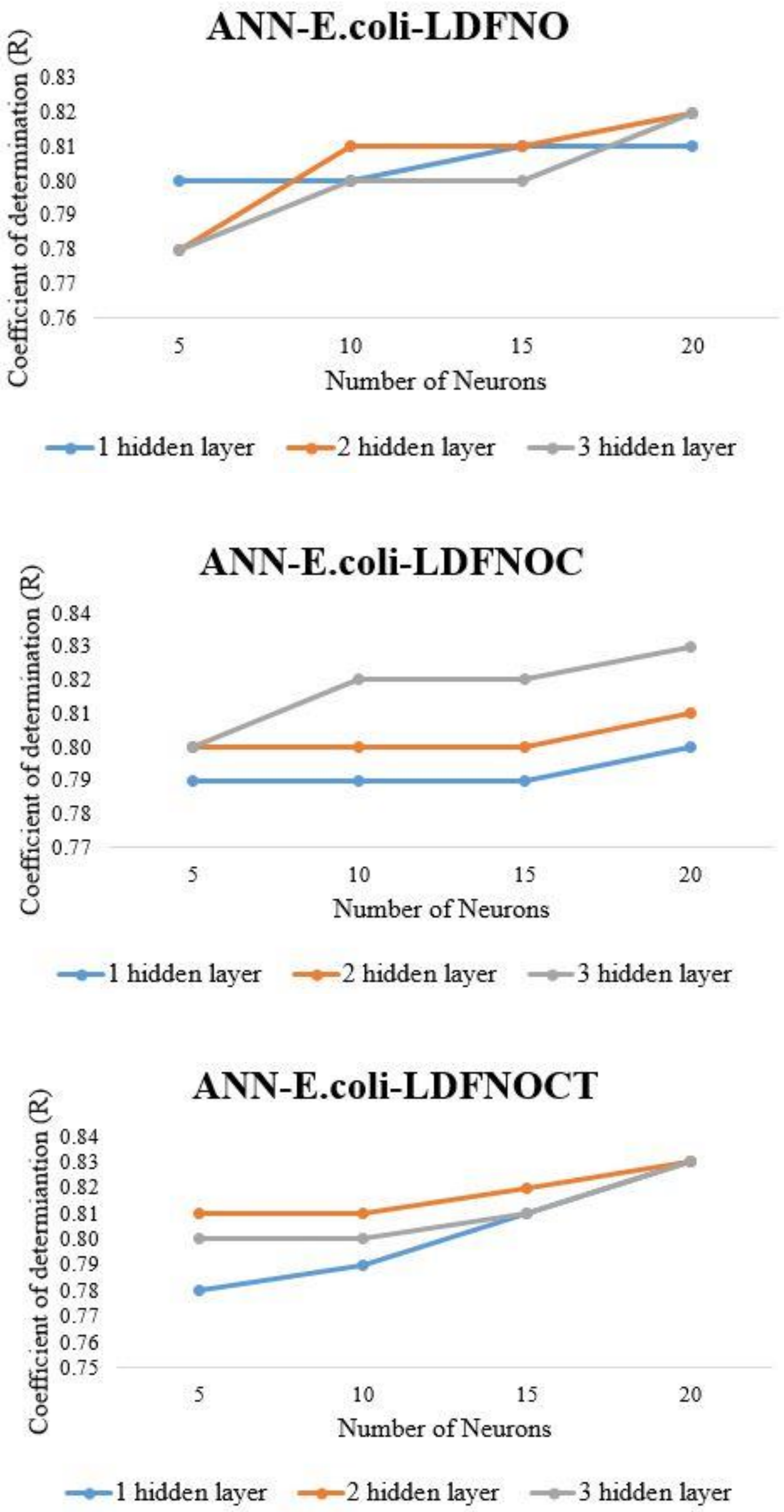
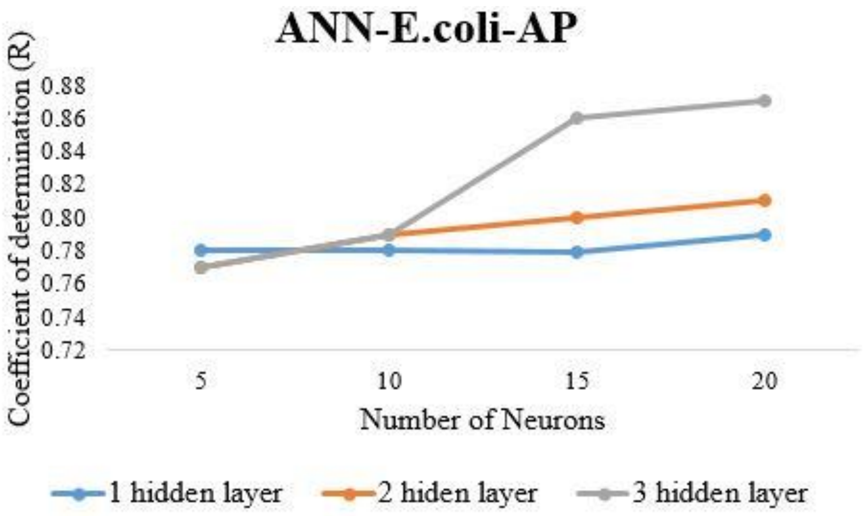
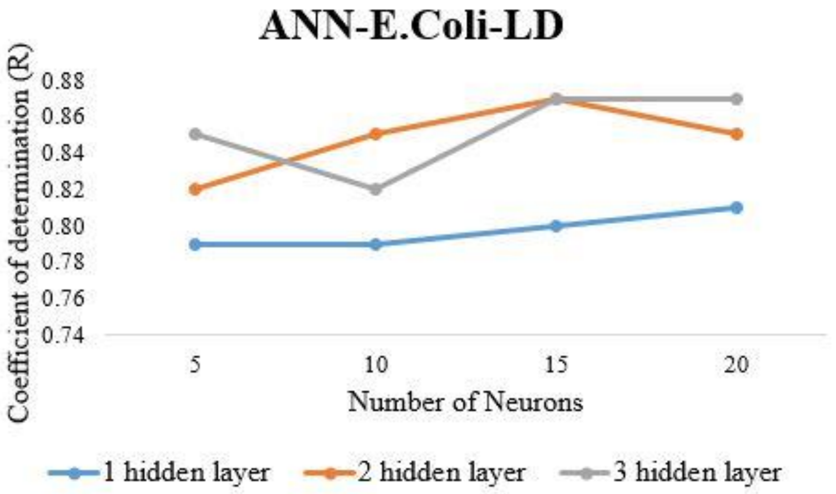
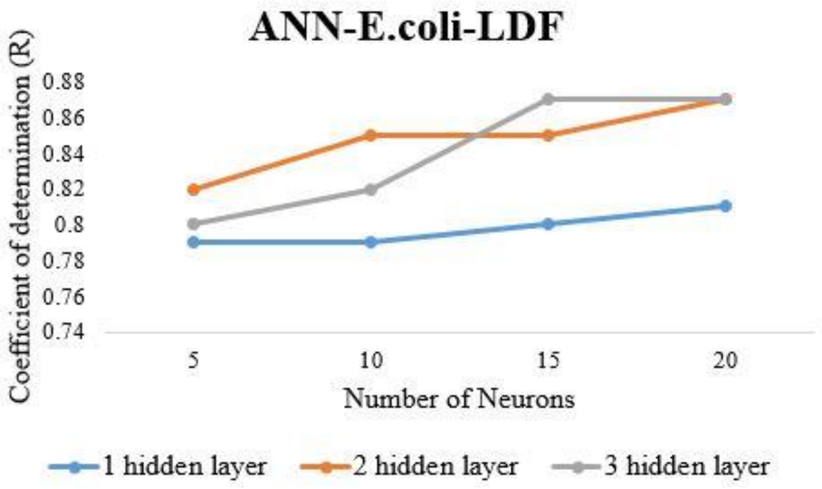
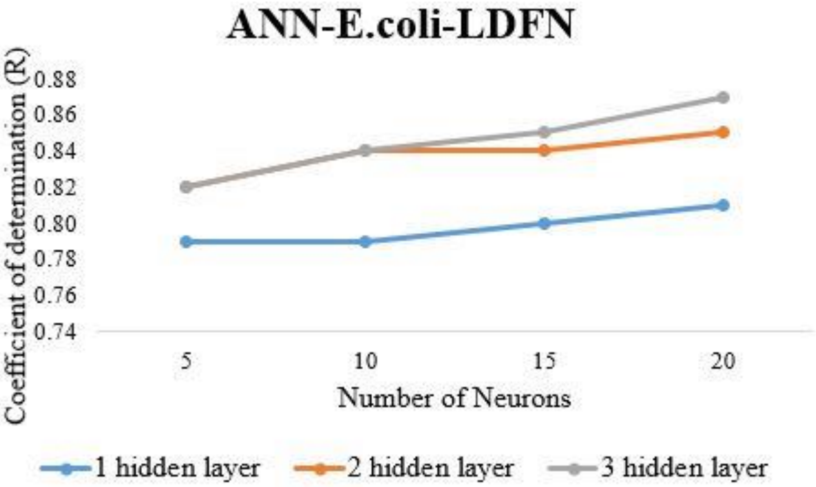


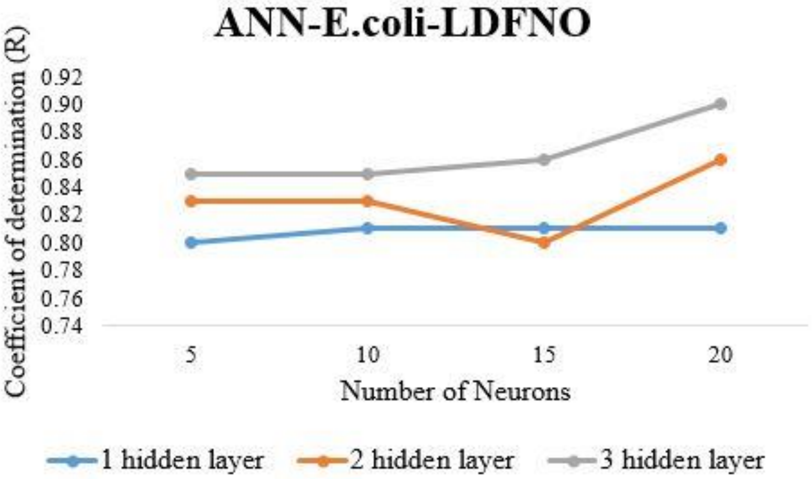
Figure 4.5: The performance of sensitivity analysis-based ANN models using Levenberg-Marquardt (LM) training function.

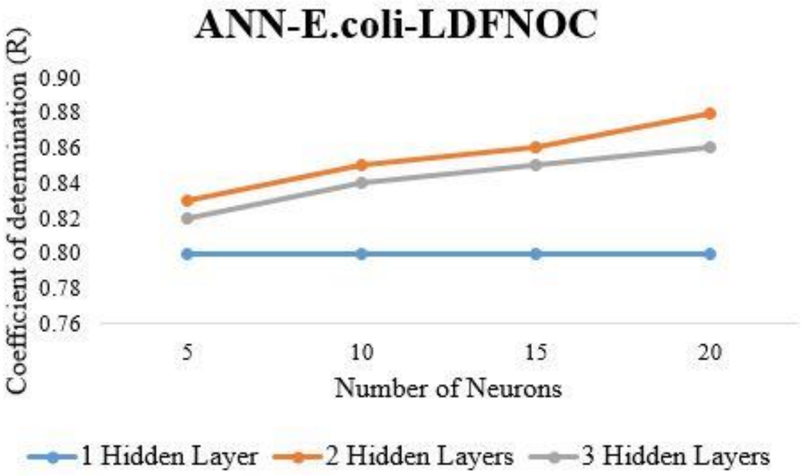

ANN-E.coli-AP
Coefficient of determination (R)
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
0.72
5
10
15
20
Number of Neurons
1 hidden layer
2 hiden layer
3 hidden layer


ANN-E.Coli-LD
Coefficient of determination (R)
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ANN-E.coli-LDF
Coefficient of determination (R)
0.88
0.86
0.84
0.82
0.8
0.78
0.76
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ANN-E.coli-LDFN
Coefficient of determination (R)
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ANN-E.coli-LDFNO
Coefficient of determination (R)
0.92
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
0.74
5
10
15
20
Number of Neurons
1 hidden layer
2 hidden layer
3 hidden layer


ANN-E.coli-LDFNOC
Coefficient of determination (R)
0.90
0.88
0.86
0.84
0.82
0.80
0.78
0.76
5
10
15
20
Number of Neurons
1 Hidden Layer
2 Hidden Layers
3 Hidden Layers

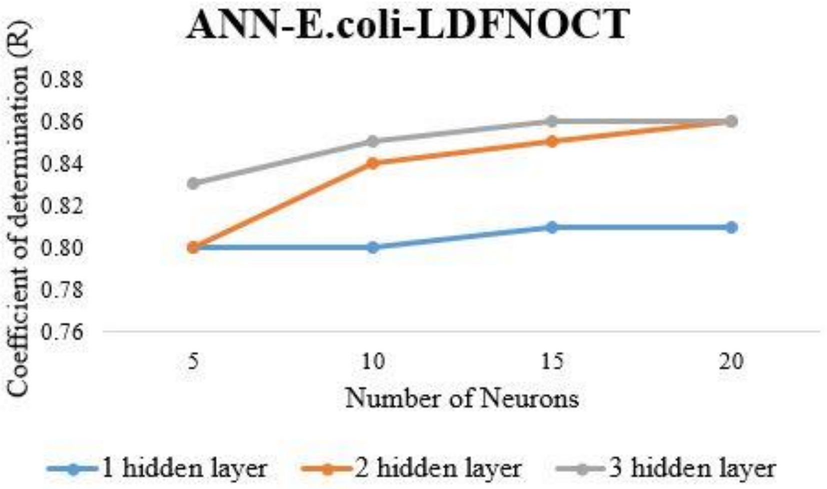


Figure 4.6: The performance of sensitivity analysis based ANN models using Bayesian regularization (BR) training function.

4.6 Analysis of ANN models using Superposition-based learning algorithm

This subsection proposed a learning algorithm that effectively used a superposition algorithm to train an ANN model. The performance of ANN models in superposition was associated with its representation. A non-linear algorithm was used to recover the best neural network configuration. The proposed algorithm was named the superposition-based learning algorithm. In the SLA, the superposition of trained neural networks stored its output for performance evaluation. A comparative analysis was performed using sensitivity analysis for a fixed model architecture with maximum values of coefficient of determining (R^2). The model contains a learning set that was parallel to the ANN-based physico-chemical water quality parameter model. After modeling the architecture of the model, the summation of weights was recorded for an individual network. The connection weights after each iteration for a network were extracted using the MATLAB Neural Network Toolbox. Subsequently, the difference in the weights was calculated for each model after every iteration. The pattern was observed between the difference in weights and iteration in a superposition of a trained neural network. After each iteration for ANN models and sensitivity analysis-based models, the connection weights were extracted and compared with the patterns in a superposition of a trained neural network to validate the proposed SLA. Such superposition-based learning algorithms (SLA) have a polynomial computational cost in the number of features in the training set. The quantum neural networks do not use nonlinear activation functions such as sigmoid or tangent hyperbolic. Hence, this learning phase of the SLA will store the output of the network.

The summation of weights was recorded for an individual network with the overall highest R2 value. The connection weights after each iteration for ANN models were extracted. Subsequently, the difference in the weights was calculated for each model after every iteration. For validation of the model, the patterns observed between the differences in weights after each iteration for ANN were compared with the patterns in a superposition of a trained neural network (Figure 4.7). Results showed that SLA algorithms are non-unitary and non-linear for weighted neural networks. The patterns observed between the differences in weights after each iteration for ANN and sensitivity analysis based models were compared with the patterns in a superposition of a trained neural network for the validation of proposed superposition learning-based algorithms, as shown in Figure 4.7 (4.7a- 4.7g) for Superposition-E.coli-AP, Superposition-E.coli-LD, Superposition-E.coli-LDF, Superposition-E.coli-LDFN, Superposition-E.coli-LDFNO, Superposition-E.coli-LDFNOC, and Superposition-E.coli-LDFNOCT models respectively. Results showed that SLA algorithms are non-unitary and non-linear for weighted neural networks.

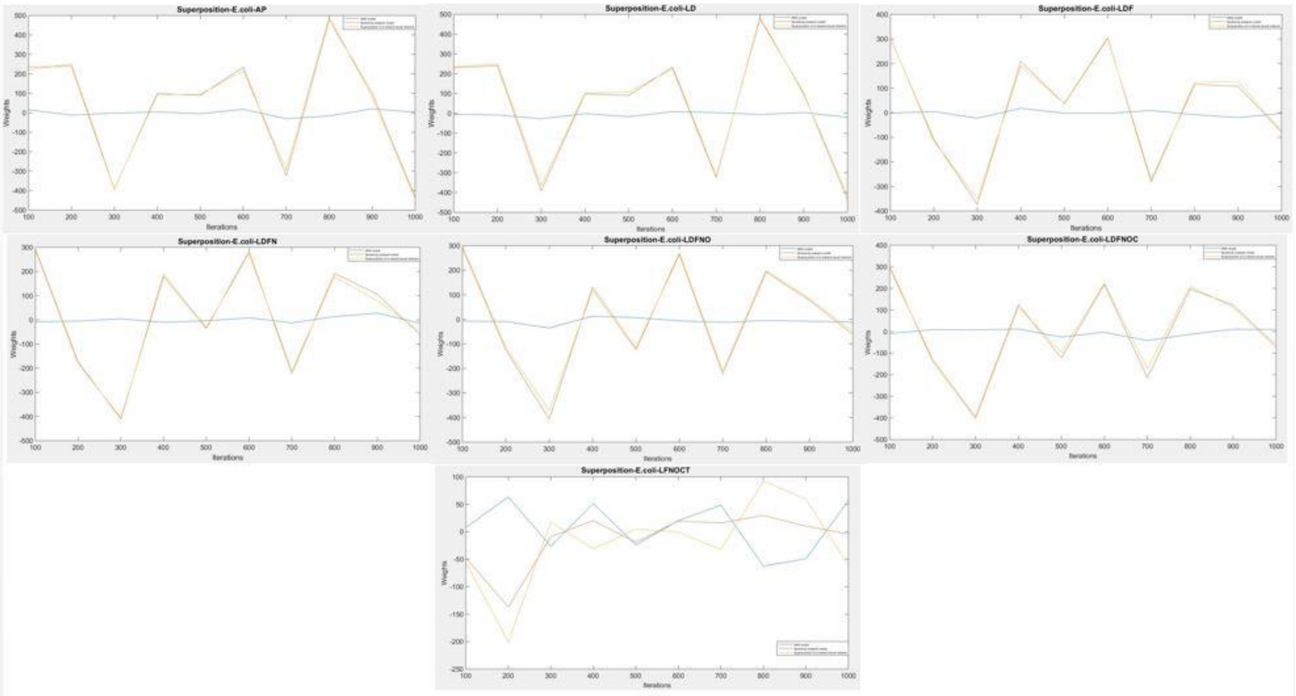


Figure 4.7: Superposition of a trained ANN model.

Prediction of *E. coli* using ANN and SLA-based models is shown in Figure 4.8. The SLA model showed the best performance with the highest R^2 value of 0.92. In contrast, the ANN model showed the best performance with the highest R^2 value of 0.87.

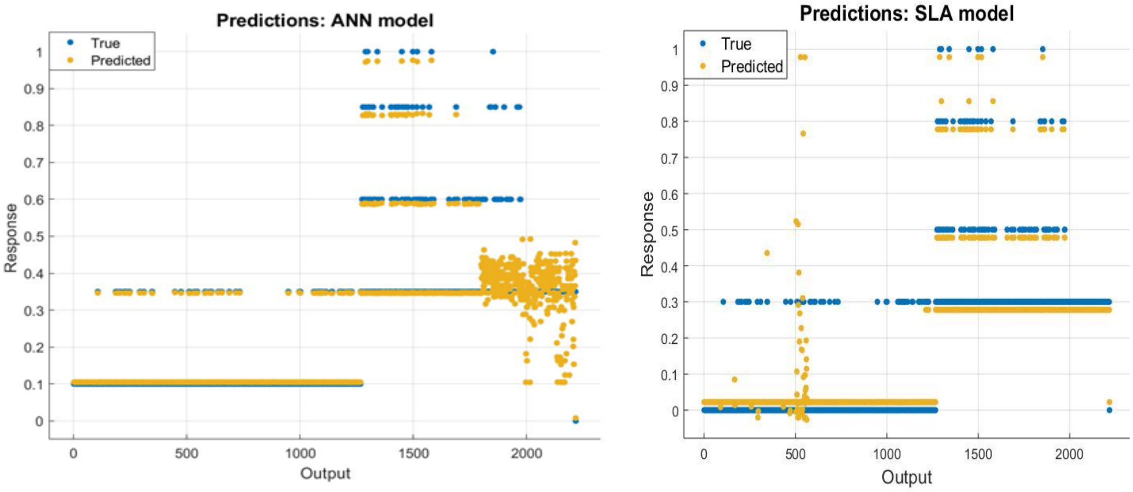


Figure 4.8: Prediction of *E. coli.*

The highlights of the results related to the proposed study are:

- The study generates a novel methodology for predicting *E. coli* concentration in groundwater, which can be used to predict hotspots in terms of continuous exposure of *E. coli*.
- *E. coli* in groundwater samples were highly positively correlated with pH and highly negatively correlated with Dissolved Oxygen.
- The randomness of the ANN model was reduced with the application of the SLA algorithm by optimization of weights and bias between the input layer and the output layer of the model.

4.7 Summary

The limitation of relying on laboratory analysis to detect bacteria can be prone to human errors, which can affect the model's performance and results. The majority of the existing techniques are limited to most of the substantial features of water to limit pH, temperature, turbidity, conductivity, and colour of the water. However, few significant physico-chemical parameters are not considered, which directly affect the growth of *E. coli* bacteria. To overcome the limitations, the artificial intelligence (AI) based technique is used in this study as an alternative to traditional models for predicting *E. coli* to improve accuracy, performance, and cost-effective results. This paper studied the effects of various physico-chemical parameters of water quality on the growth *E. coli*, and a correlation was observed between *E. coli* and water quality parameters. Previous studies show that the predictive models are based on published research reports and testing data, so it is difficult to check their accuracy.

In this study, we create our dataset by using water samples collected from eight districts of the state of Rajasthan under the BITS-UVA (University of Virginia) groundwater contamination project. The field study covers a collection of 1301 groundwater samples. This experimental data set was used to train, test, and validate the results using AI techniques. A superposition-based learning algorithm (SLA) is proposed to observe the patterns of ANN-based sensitivity analysis for automating the prediction process of *E. coli* bacteria in groundwater. The concept of presenting an input pattern for all feasible neural network architectures is unrealistic in classical neural networks. For this concept to be implemented classically, one would need to create multiple neural networks for each configuration and architecture to obtain all the inputs and simulate the related outputs in parallel. After computing the efficiency of each pattern for each configuration of the neural network, one can check the configuration of a trained neural network with the best performance.

The concept of SLA describing a possible configuration of weights in superposition for any architecture. Using SLA, we can obtain all possible neural network configurations in superposition for the specified architectures. The critical property of Grover's algorithms discussed in this paper is the ability to obtain all correlations from the training set in superposition. The result shows that the superposition models based on Grover’s algorithm are more efficient in predicting all patterns in the counts of *E. coli* in groundwater with higher efficiency and low error. We will create a hybrid model of ANN with a fuzzy set theory or genetic algorithms (GA) to get better accuracy for future modifications. We aim to take the AI-enhanced model further to provide a fully automated *E. coli* classification system. Another possible future work is to develop a probabilistic SLA model. We only need to run the neural network twice, one forward and one backward, by modifying Grover’s algorithm.

References:

- Adriaans, P., & Zantinge, D. (1996). Data Mining Addison Wesley Longman Limited. *Edinbourgh Gate, Harlow, CM20 2JE, England*.
- Bisi-Johnson, M. A., Adediran, K. O., Akinola, S. A., Popoola, E. O., & Okoh, A. I. (2017). Comparative physicochemical and microbiological qualities of source and stored household waters in some selected communities in southwestern Nigeria. *Sustainability*, *9*(3), 454.
- Burton Jr, G. A., Gunnison, D., & Lanza, G. R. (1987). Survival of pathogenic bacteria in various freshwater sediments. *Applied and Environmental Microbiology*, *53*(4), 633-638.
- Central Bureau of Health Intelligence, 2018. National Health Profile, 13th issue. https://cdn.downtoearth.org.in/pdf/NHP-2018.pdf (accessed 5 January 2020).
- Cheng, J., Niu, S., & Kim, Y. (2013). Relationship between water quality parameters and the survival of indicator microorganisms–Escherichia coli–in a stormwater wetland. *Water science and technology*, *68*(7), 1650-1656.
- David, M. M., & Haggard, B. E. (2011). Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois River Watershed, Arkansas and Oklahoma, USA. *Water, Air, & Soil Pollution*, *215*(1), 525-547.

- David, M. M., & Haggard, B. E. (2011). Development of regression-based models to predict fecal bacteria numbers at select sites within the Illinois River Watershed, Arkansas and Oklahoma, USA. *Water, Air, & Soil Pollution*, *215*(1), 525-547.

- Dooge, J. C. (2001). Integrated management of water resources. In *Understanding the Earth System* (pp. 115-123). Springer, Berlin, Heidelberg.
- El-Shafie, A., Mukhlisin, M., Najah, A. A., & Taha, M. R. (2011). Performance of artificial neural network and regression techniques for rainfall-runoff prediction. *International Journal of Physical Sciences*, *6*(8), 1997-2003.
- El-Shafie, A., Noureldin, A. E., Taha, M. R., & Basri, H. (2008). Neural network model for Nile river inflow forecasting based on correlation analysis of historical inflow data.
- Foresee, F. D., & Hagan, M. T. (1997, June). Gauss-Newton approximation to Bayesian learning. In *Proceedings of international conference on neural networks (ICNN'97)* (Vol. 3, pp. 1930-1935). IEEE.

- Francy, D. S., Stelzer, E. A., Duris, J. W., Brady, A. M., Harrison, J. H., Johnson, H. E., & Ware, M. W. (2013). Predictive models for Escherichia coli concentrations at inland lake beaches and relationship of model variables to pathogen detection. *Applied and environmental microbiology*, *79*(5), 1676-1688.
- Gerald, P. (2011). Water science. University of Washington. *PMC [serial on the Internet]*.
- Grubert, J. P. (2003). Acid deposition in the eastern United States and neural network predictions for the future. *Journal of Environmental Engineering and Science*, *2*(2), 99-109.
- Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the Marquardt algorithm. *IEEE transactions on Neural Networks*, *5*(6), 989-993.
- Han, J., & Moraga, C. (1995, June). The influence of the sigmoid function parameters on the speed of backpropagation learning. In *International workshop on artificial neural networks* (pp. 195-201). Springer, Berlin, Heidelberg.
- Huey, G. M., & Meyer, M. L. (2010). Turbidity as an indicator of water quality in diverse watersheds of the Upper Pecos River Basin. *Water*, *2*(2), 273-284.
- India Water Portal, 2019. When water kills. https://www.indiawaterportal.org/faqs/waterborne/ (accessed 20 November 2020).
- Islam, M. M., Hofstra, N., & Islam, M. A. (2017). The impact of environmental variables on faecal indicator bacteria in the Betna river basin, Bangladesh. *Environmental Processes*, *4*(2), 319-332.
- Katip, A. (2018). The usage of artificial neural networks in microbial water quality modeling: a case study from the lake Iznik. *Applied Ecology and Environmental Research*, *16*(4), 3897-3917.
- Lin, B., Kashefipour, S. M., & Falconer, R. A. (2003). Predicting near-shore coliform bacteria concentrations using ANNS. *Water Science and technology*, *48*(10), 225-232.
- Liong, S. Y., & Sivapragasam, C. (2002). Flood stage forecasting with support vector machines 1. *JAWRA Journal of the American Water Resources Association, 38*(1), 173-186.

- Lou, W., & Nakai, S. (2001). Application of artificial neural networks for predicting the thermal inactivation of bacteria: a combined effect of temperature, pH and water activity. *Food Research International*, *34*(7), 573-579.
- MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, *4*(3), 415-447.
- Makarynska, D., & Makarynskyy, O. (2008). Predicting sea-level variations at the Cocos (Keeling) Islands with artificial neural networks. *Computers & Geosciences*, *34*(12), 1910-1917.
- Mohammed, H., Longva, A., & Seidu, R. (2018). Predictive analysis of microbial water quality using machine-learning algorithms. *Environmental Research, Engineering and Management*, *74*(1), 7-20.
- Molden, D. (Ed.). (2013). *Water for food water for life: A comprehensive assessment of water management in agriculture*. Routledge.
- Mouna, H., Ahmed, A., & Omar, A. (2014). An evaluation of environmental factors affecting the survival of Escherichia coli in coastal area, Oualidia Lagoon. *International Journal of Current Microbiology and Applied Science*, *3*(10), 710-721.
- Moustakis, V., Lehto, M., & Salvendy, G. (1996). Survey of expert opinion: which machine learning method may be used for which task?. *International Journal of Human-Computer Interaction*, *8*(3), 221-236.
- Muttil, N., & Chau, K. W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*, *28*(3-4), 223-238.
- Najah, A., Elshafie, A., Karim, O. A., & Jaffar, O. (2009). Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of scientific research*, *28*(3), 422-435.
- Noureldin, A., El-Shafie, A., & Bayoumi, M. (2011). GPS/INS integration utilizing dynamic neural networks for vehicular navigation. *Information fusion*, *12*(1), 48-57.
- Pachepsky, Y., Kierzewski, R., Stocker, M., Sellner, K., Mulbry, W., Lee, H., & Kim, M. (2018). Temporal stability of Escherichia coli concentrations in waters of two irrigation ponds in Maryland. *Applied and environmental microbiology*, *84*(3), e01876-17.

- Payus, C., Haziqah, N., Basri, N., & Wan, V. L. (2018). Faecal bacteria contaminations in untreated drinking water (Groundwater well and hill water) from rural community areas.
- Rykiel Jr, E. J. (1989). Artificial intelligence and expert systems in ecology and natural resource management. *Ecological Modelling*, *46*(1-2), 3-8.
- Shamsudin, S. N., Rahman, M. H. F., Taib, M. N., Razak, W. R. W. A., Ahmad, A. H., & Zain, M. M. (2016, August). Analysis between Escherichia Coli growth and physical parameters in water using Pearson correlation. In *2016 7th IEEE Control and System Graduate Research Colloquium (ICSGRC)* (pp. 131-136). IEEE.
- Singh, D. V., Maheshwari, G., Shrivastav, R., & Mishra, D. K. (2011). Neural network–comparing the performances of the training functions for predicting the value of specific heat of refrigerant in vapor absorption refrigeration system. *International Journal of Computer Applications*, *18*(4), 1-5.
- Tanimoto, S. L. (1987). *The elements of artificial intelligence: an introduction using LISP*. Computer Science Press, Inc..
- Torrecilla, J. S., Aragón, J. M., & Palancar, M. C. (2008). Optimization of an artificial neural network by selecting the training function. Application to olive oil mills waste. *Industrial & engineering chemistry research*, *47*(18), 7072-7080.

- Turian, J., Bergstra, J., Bengio, Y., 2009, June. Quadratic features and deep architectures for chunking. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (245-248).

- World Health Organization, 2017. Diarrhoeal disease. http://www.who.int/mediacentre/factsheets/fs330/en/ (accessed 11 January 2020).
- Zamani, M. A. T., & Saybani, M. Artificial Neural Network Model for Prediction of Environmental Status of Urban Catchment of Penchala River, Kuala Lumpur, Malaysia.

5. A Convolutional Neural Network Approach for Detection of *E. coli* Bacteria in Water

The monitoring of bacteriological drinking water quality relies mainly on the analysis of indicator bacteria. E. coli is a more precise indicator of water contamination than other fecal coliform bacteria due to the advancement in testing methods. We implement an automated E. coli bacteria detection process using Convolutional Neural Network (CNN) to address this issue. We have also proposed a mobile application for the rapid detection of E. coli bacteria in water that uses CNN.

5.1 Introduction

Escherichia coli (E. coli) bacteria are gram-negative, non-spore, rod-shaped pathogens that produce gas in the prescribed growth media after fermentation at 35 °C within 48 hours (Greenwood et al., 2003). *E. coli* was first recognized as a pathogen in 1982 (Riley et al. 1983). *E. coli* bacteria should not be present in 100 ml of the water sample, according to WHO (2011), BIS (2012) and USEPA protocols (2007). According to a WHO report (2017), 1.9 billion people worldwide use water that is polluted. About 37.7 million Indians are infected annually by waterborne diseases. Waterborne diseases are still prevalent in India and have resulted in 10,738 deaths in the last five years, as per the Central Bureau of Health Intelligence report (2018). The monitoring of bacteriological drinking water quality relies mainly on the analysis of indicator bacteria. *E. coli* is a more precise indicator of water contamination than other fecal coliform bacteria due to the advancement in testing methods. *E. coli* bacteria can be identified in the laboratory using conventional methods (Co-ordination Action Food 2007), enzymatic methods (Co-ordination Action Food 2007), molecular methods (Tamerat et al. 2016; Saxena et al. 2015), and biosensor-based methods (Maas et al. 2017).

According to the method based on laboratory experiments, it takes 12-48 hours for the concentration of bacterial cells to be recorded. The limitation of relying solely on sensor-based water quality analysis for identification is that it can lead to errors. Therefore, there is a need for real-time monitoring. Enzymatic methods of detection are color-based methods (Rice et al. 1989). The amount of colour appearance can be used to determine the degree of bacterial contamination. The detection method is based on the concept that only *E. coli* bacteria are fed.

No substrate is given for other bacteria. The specified substrate is used as an essential source of nutrients for bacteria. A chromogenic or fluorogenic substance is released from the specified substrate during the substrate utilization period, which indicates the presence of *E. coli*. Manually performing this process is highly time-consuming and difficult. This detection process is analytical. There is always a possibility of human error, which may result in a disastrous decision. The colours of each concentration can be scanned using conventional computer vision methods. It is, however, extremely difficult to determine the colour intensity for each concentration level. This is made simple with deep learning since the algorithm calculates these colour intensities using statistically generated training sets.

Huang L et al. (2018) developed a convolutional neural network (CNN) for bacterial colony classification using digital images. The data from Peking University First Hospital was used for the classification of bacterial colonies. Results show that the network was able to classify 18 bacterial colonies with an accuracy of 73%. Alaslani MG et al. (2018) extracted the learned features from a pre-trained CNN and Support Vector Machine (SVM) algorithm for image classification. The Alex Net pre-trained CNN model was used for feature extraction, and the SVM algorithm was used for classification. The Iris public images were used for the development of an iris recognition system. Results show that the recognition accuracy of the Iris database was 98.3%.

Mohanty SP et al. (2016) used a public dataset containing 54,306 images of healthy and diseased plant leaves. They developed a deep CNN to identify 26 diseases and 14 crop species with an accuracy of 99%. Nehal SA et al. (2019) developed an AI-based lab-on-chip for the detection of bacterial contamination using the Photonic Crystal-based optical biosensor. These biosensors came up with a few limitations of using separate sensors to measure physical, chemical, and bacteriological parameters of water quality which affect sensitivity and accuracy of the results. The method is cost-intensive and requires maintenance. Gunda NSK et al. (2019) developed an AI-based mobile application for water quality monitoring with an accuracy of 99%. The authors have not validated the proposed model with performance functions. Hence, this model is not reliable for bacterial detection.

Previous studies show that the predictive models are based on published research reports, public datasets, and open source testing data, so it is difficult to check the accuracy of the model. Thus, we cannot rely solely on these studies for bacterial detection. However, no study has been conducted to detect *E. coli* bacteria in water using experimental laboratory data with high accuracy and precision. In this paper, we have created our dataset using water samples collected from eight districts of Rajasthan under the BITS-UVA groundwater contamination project. We have also developed an AI-based smartphone application to rapidly detect E. coli in water using laboratory experiment data.

5.2 Artificial Neural Networks

Artificial Neural Networks (ANN) are a model of the Biological Neural Network. Biological Neural Networks assist living organisms to interpret, identify and learn from their environmental patterns for future applications. Humans use these patterns and prior knowledge to process any information and thus come to an output (Fausett 2006). ANNs lend this property to machines. ANNs enable a standard and practical methodology for machines to adapt from instances and modify their operation. ANNs have been shown to be useful, especially in areas where the output of the system is not determined by a specific mapping algorithm between input and output (Gupta et al., 2018). ANNs are most widely used when mapping between inputs and outputs is not linear. McCulloch and Pitts (1943) presented the first-ever model of an artificial neuron, called the perceptron. A layer of perceptrons can perform some tasks. Thus a single layer of perceptrons can form a network. We term such a network as a Single Layer Perceptron. An arrangement of a series of a Single Layer Perceptron is called a Multi-Layer Perceptron (MLP). MLPs are also called feedforward networks.

5.3 Convolutional Neural Networks

A Convolutional Neural Network (CNN) is a subset of the neural network mentioned previously. One or two convolutional layers are present in a CNN, always with a subsampling layer, accompanied by one or more fully connected layers (Khan et al. 2020). The conception of a CNN was sparked by the discovery of a sense system in the brain, the visual cortex. The visual cortex comprises many cells that sense light. Receptive fields are overlapping sub-regions of the visual field. The more complex cells have wider receptive fields, and they serve as local filters over the input space. The convolution layer in a CNN has the same function as the cells in the visual cortex (Hubel et al. 1968). A hand-designed feature extractor collects essential information from the input. It extracts irrelevant variables in the conventional model of pattern recognition (Fukushima et al. 1983). After the extractor, a trainable classifier is used, which is a regular neural network that divides feature vectors into classes. Convolution layers serve as feature extractors in CNNs. They are not, however, handcrafted. The kernel weights for convolution filters are selected during the training phase. Since the receptive fields of the hidden layers are restricted to be local, convolutional layers can extract local features. The weights of the convolutional and fully connected layers are calculated in CNN during the training phase and used for feature extraction (Brownlee 2019) and classification (Huang et al. 2018). The improved network architectures result in reduced memory and computing complexity.

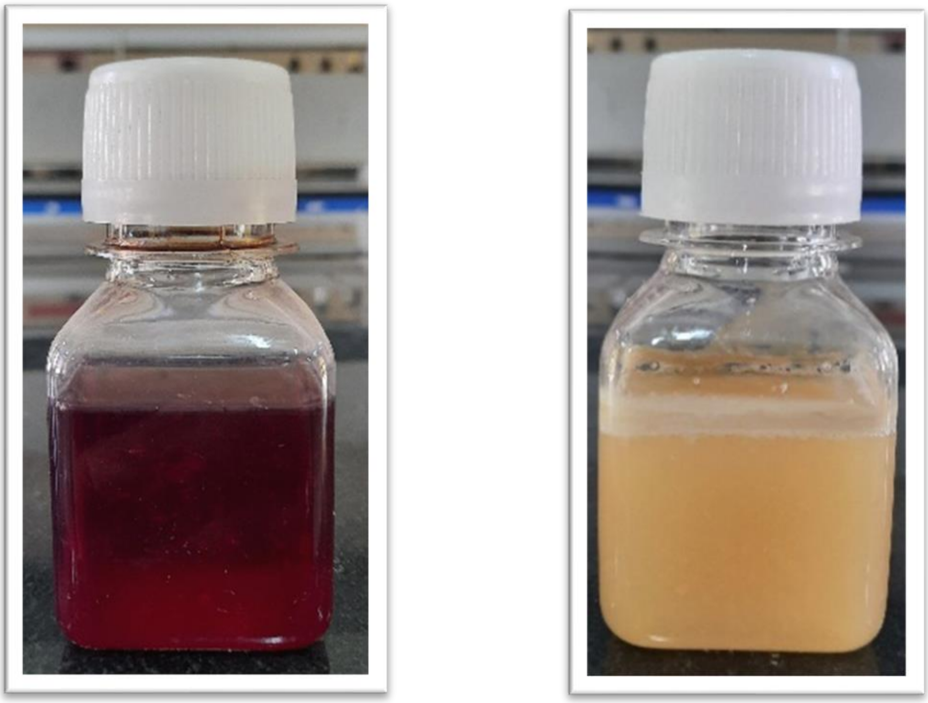
5.4 Methodology

The detection of *E. coli* bacteria is essential to prevent health diseases. According to the laboratory-based methods, 12-48 hours are required to detect bacteria in water. The drawback of depending on laboratory-based methods for the detection of *E. coli* bacteria can be prone to human errors. Hence, the bacterial detection process must be automated to reduce error. The performance of the model was validated using the F-Score, Precision, Sensitivity and Accuracy statistical measures. The following methodology has been adopted for the prediction of bacteria in water.

- Detection of bacteria in water using laboratory experiment method.
- Convolutional Neural Network for prediction of bacteria.
- Mobile application for rapid detection of *E. coli.*

5.5 Present/Absent test (PA test)

The enzymatic method was used to determine the presence or absence of *E. coli* bacteria in groundwater samples (Olstadt et al., 2007). Present/Absent test is a substrate method developed to overcome some constraints of the multiple tube fermentation method (Oshiro 2002) and membrane filter method (Jagals et al. 2000). The detection method is based on the concept that only *E. coli* bacteria are fed. No substrate is given for other bacteria. Firstly 100 ml of water sample was added to the sterile disposable bottle. The powder medium (PA broth) was then swirled into water so that it got dissolved completely. Once dissolved, water samples were incubated for 24-48 hours at 35 °C. After the incubation period, the transition in the colour of the medium from reddish-purple (Figure 5.1a) to yellow (Figure 5.2b) indicated *E. coli*. Figure 5.1 shows the change in colour of the culture medium due to the presence of bacteria.



b)The initial state of culture medium b) Final state of culture medium (if bacteria is present)

Figure 5.1: Colour change in culture medium due to the presence of bacteria.

5.6 Convolutional Neural Network for prediction of bacteria

A total of 1301 images obtained from laboratory testing (Present/Absent test) were used as an input of the CNN model. Image manipulation procedure was used to resize, crop, and rotate the input photos so as to increase the classifier's accuracy in training sets. The pictures were divided into two groups based on the presence or absence of *E. coli*. The experimental data were divided into three sets: training (70%), testing (15%), and validation (15%). We trained the CNN using MATLAB Deep Network Designer 2020b (MathWorks) to distinguish between pictures with *E. coli* present and *E. coli* absent. The CNN model was trained on a machine with 16 gigabytes (GB) of RAM. To detect *E. coli* bacteria on agar plates, the CNN employed image color, red green blue (RGB), and black and white (BW) values. On a scale of 0 to 1, predictions were given for each image. If the predicted value was adjusted to zero, that is, less than 0.5, the final forecast for that picture was *E. coli* absent. If the predicted value was adjusted to 1, that is, a value larger than 0.5, the final forecast for that picture was *E. coli* present. Consequently, the closer the model went to 0, the more certain it was that a picture was an *E. coli* absent image. The nearer it got to 1, the more certain it was that a picture was an *E. coli* absent image. The dataset used in this study contains a total of 1301 images to predict the presence or absence of *E. coli* bacteria in groundwater. All images have a size of 2180x960 pixels. Colour images were used in this study to classify images using colour and texture features. Images from the dataset used for the study are shown in Figure 5.2.

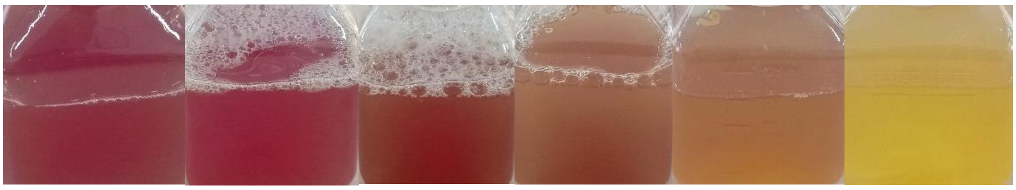


Figure 5.2: A few sample images from *E. coli* dataset.

The CNN model's performance was validated using F-score, precision, sensitivity, and accuracy statistical measures (Dalianis 2018; Prabha et al. 2016). The CNN model used cross-validation to test the networks more thoroughly. This means that all images of the data were used as both training and test data, split into iterations. TP stands for true positive (appropriately recognizing an *E. coli* present image as *E. coli* present). TN stands for true negative (appropriately recognizing an *E. coli* absent image as *E. coli* absent). FP stands for false positive (inappropriately recognizing an *E. coli* present image as *E. coli* absent). FN stands for false negative (inappropriately recognizing an *E. coli* absent image as *E. coli* present). Accuracy was defined as the proportion of properly recognized samples (both present and absent) among all samples (see Equation (5.1)). The ratio of all recognized positive samples to all positive samples is the sensitivity (see Equation (5.2)). If sensitivity is strong, the class was accurately detected. The value of high sensitivity implies that a class has been appropriately recognized. Precision was defined as the ratio of all positively detected positive samples to all positively predicted positive samples (see Equation (5.3)). High precision suggests that a sample classified as positive is, in fact, positive. The weighted average of sensitivity and accuracy was used to get the F-score. In the F-score, the harmonic mean replaces the arithmetic mean. This metric punishes high values much more (see Equation (5.4)). The following equations (Equations (5.1)– (5.4)) can be used to compute F-Score, precision, sensitivity, and accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (5.1)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (5.2)$$

$$Precision = \frac{TP}{TP+FP} \qquad (5.3)$$

$$F - Score = \frac{2*Sensitivity*Precision}{Sensitivity+Precision} \qquad (5.4)$$

Deep Learning was used in the current study to train a CNN model to detect *E. coli*. The CNN model was trained to compare all *E. coli* concentrations to a binary output that stated whether *E. coli* was present or not. The CNN model was developed using MATLAB (R2020b) software. Figure 5.3 shows the architecture of the proposed CNN model for rapid detection of *E. coli* in water.

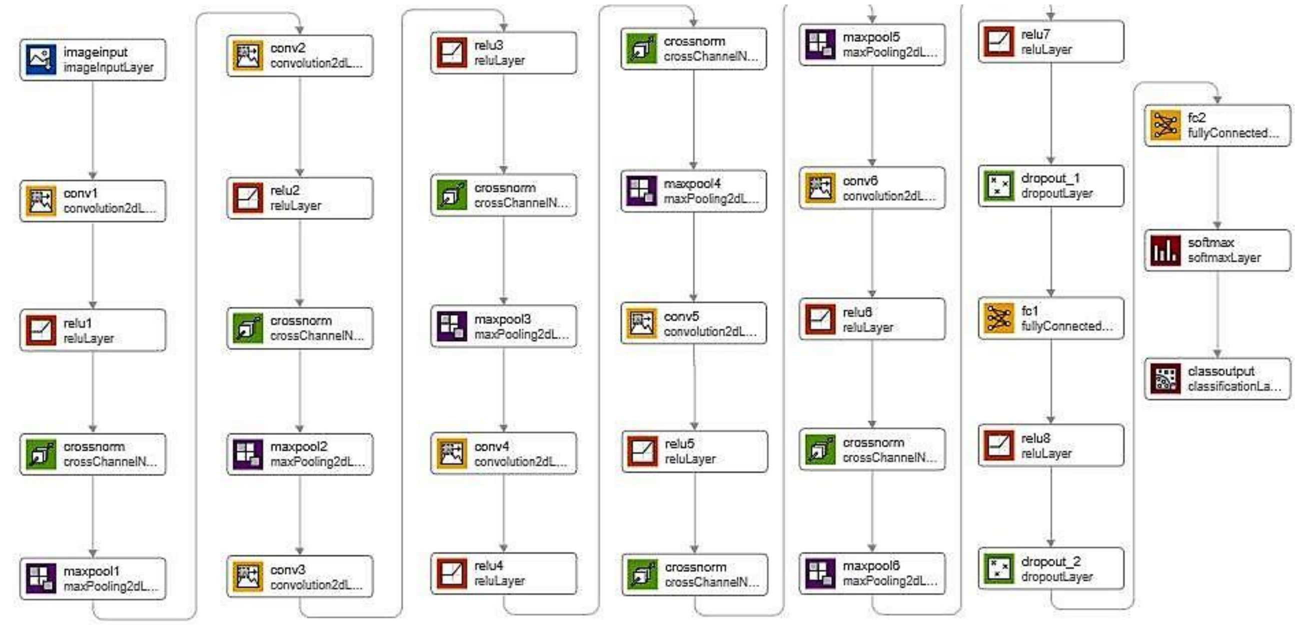
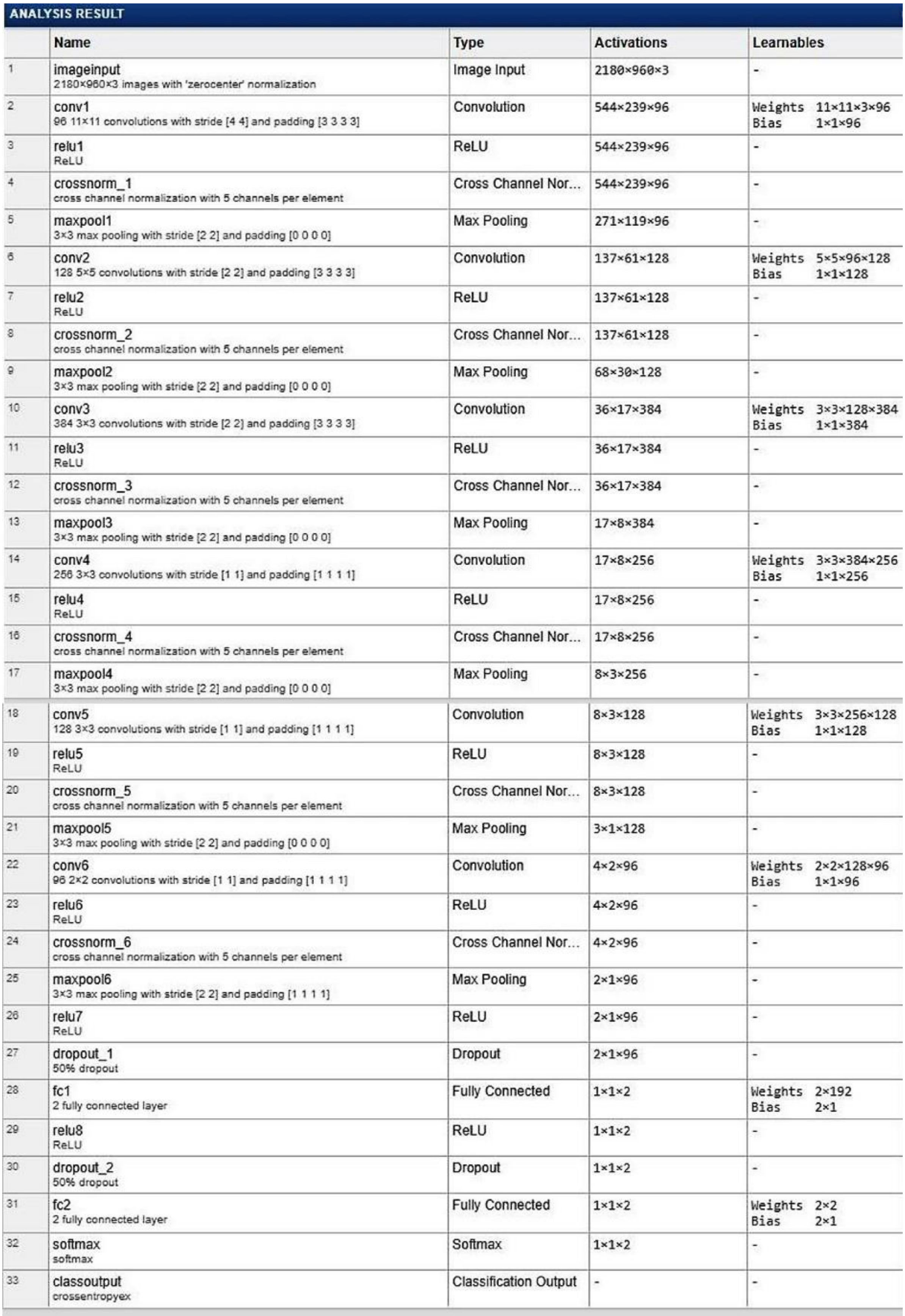


Figure 5.3: Overview of the CNN model.

The developed CNN architecture consisted of six convolution layers and two fully connected layers. Each layer had a different weight, bias, and Rectified Linear Unit (ReLU) linked with it. Weights and bias were assigned based on the number of filters in the first convolutional layer (conv1). The number of filters used in conv1 was 96, and the size of filters were 11x11x3, so the initial weights 11x11x3x96 and bias 1x1x96 were assigned. The first layer was padded with 3x3 kernels with a stride of 4. The next two layers were padded with 3x3 kernels with a stride of 2, while the last three layers were padded with 1x1 kernels with a stride of 1. These convolutional layers were accompanied by two fully connected layers, which were non-strided to prevent overfitting. A detailed summary of the CNN architecture is shown below in Figure 5.4. Before training the data, we optimized the batch size, learning rate, and the maximum number of epochs using grid search, as shown in Table 5.1.



ANALYSIS RESULT

	Name	Type	Activations	Learnables
1	imageinput 2180×960×3 images with 'zerocenter' normalization	Image Input	2180×960×3	-
2	conv1 96 11×11 convolutions with stride [4 4] and padding [3 3 3 3]	Convolution	544×239×96	Weights 11×11×3×96 Bias 1×1×96
3	relu1 ReLU	ReLU	544×239×96	-
4	crossnorm_1 cross channel normalization with 5 channels per element	Cross Channel Nor...	544×239×96	-
5	maxpool1 3×3 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	271×119×96	-
6	conv2 128 5×5 convolutions with stride [2 2] and padding [3 3 3 3]	Convolution	137×61×128	Weights 5×5×96×128 Bias 1×1×128
7	relu2 ReLU	ReLU	137×61×128	-
8	crossnorm_2 cross channel normalization with 5 channels per element	Cross Channel Nor...	137×61×128	-
9	maxpool2 3×3 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	68×30×128	-
10	conv3 384 3×3 convolutions with stride [2 2] and padding [3 3 3 3]	Convolution	36×17×384	Weights 3×3×128×384 Bias 1×1×384
11	relu3 ReLU	ReLU	36×17×384	-
12	crossnorm_3 cross channel normalization with 5 channels per element	Cross Channel Nor...	36×17×384	-
13	maxpool3 3×3 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	17×8×384	-
14	conv4 256 3×3 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	17×8×256	Weights 3×3×384×256 Bias 1×1×256
15	relu4 ReLU	ReLU	17×8×256	-
16	crossnorm_4 cross channel normalization with 5 channels per element	Cross Channel Nor...	17×8×256	-
17	maxpool4 3×3 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	8×3×256	-
18	conv5 128 3×3 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	8×3×128	Weights 3×3×256×128 Bias 1×1×128
19	relu5 ReLU	ReLU	8×3×128	-
20	crossnorm_5 cross channel normalization with 5 channels per element	Cross Channel Nor...	8×3×128	-
21	maxpool5 3×3 max pooling with stride [2 2] and padding [0 0 0 0]	Max Pooling	3×1×128	-
22	conv6 96 2×2 convolutions with stride [1 1] and padding [1 1 1 1]	Convolution	4×2×96	Weights 2×2×128×96 Bias 1×1×96
23	relu6 ReLU	ReLU	4×2×96	-
24	crossnorm_6 cross channel normalization with 5 channels per element	Cross Channel Nor...	4×2×96	-
25	maxpool6 3×3 max pooling with stride [2 2] and padding [1 1 1 1]	Max Pooling	2×1×96	-
26	relu7 ReLU	ReLU	2×1×96	-
27	dropout_1 50% dropout	Dropout	2×1×96	-
28	fc1 2 fully connected layer	Fully Connected	1×1×2	Weights 2×192 Bias 2×1
29	relu8 ReLU	ReLU	1×1×2	-
30	dropout_2 50% dropout	Dropout	1×1×2	-
31	fc2 2 fully connected layer	Fully Connected	1×1×2	Weights 2×2 Bias 2×1
32	softmax softmax	Softmax	1×1×2	-
33	classoutput crossentropyex	Classification Output	-	-

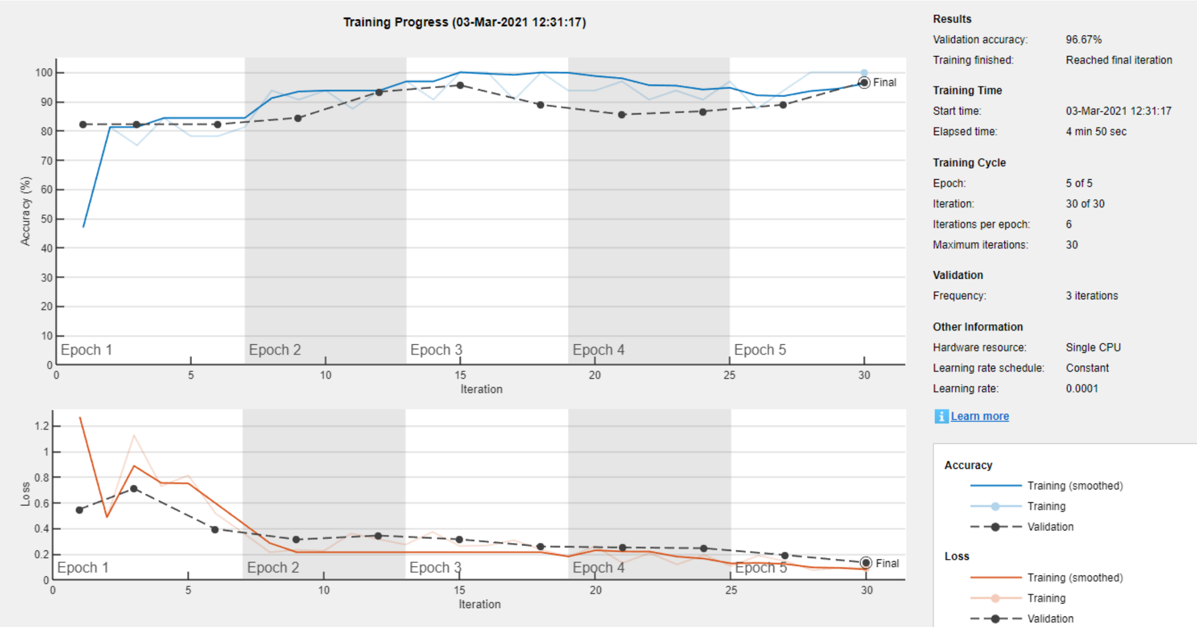
Figure 5.4: Detailed structure of CNN model.

Table 5.1: Grid search of hyperparameters for backend tuning

Hyperparameters	**Values**
Optimizer	SGDM with learning rate (0.001, **0.0001**, 0.00001)
Batch size	(4,16,**32**,64)
Epoch	(10,**30**,50)

The hyperparameters highlighted in bold in Table 5.1 indicate the best performance. The best hyperparameters were used to train the CNN model on training sets. The input images have a size of 2180x960x3 pixels. These numbers correspond to the height, width, and channel size. For a colour image, the channel size is 3, corresponding to the RGB values. Convolution is performed on input images to extract features. The filter size of the convolution layer (conv1) was 11x11, which corresponds to the height and width of the filters used by the training feature when scanning through the images. The number of filters used in the conv1 layer was 96, representing the number of neurons connected to the same input region. The gradients and activations were normalized using a cross-channel normalization layer. After the normalization layer, a nonlinear activation function, i.e., rectified linear unit (ReLU), was applied. After convolution, max pooling was used to downsample each feature map. Since the number of parameters was reduced, it retains only the most essential information from the images and prevents overfitting. The rectangular area was (2, 2) in dimension. Two fully connected layers were added after the convolutional and down-sampling layers.

A fully connected layer has all of its neurons coupled to the neurons of the preceding layer. Such a layer combines all characteristics from the preceding layers to categorize all trends in the picture. The fully connected (FC) layer connected all of the features that were used to identify the pictures. As a consequence, the output parameters in the fully connected layer equaled the number of classes in the predicted data, which in this case was two. The softmax activation function was used to normalize the output of the fully connected layer. The softmax layer generated a series of positive numbers that sum to 1. These values were subsequently used as categorization probabilities by the classification layer. The classification layer was the final layer. Each input was assigned to one of the mutually distinct groups by this layer. It computed the error (loss) based on the probability obtained from the softmax activation function. Our model had an accuracy of 96 percent (1239/1301 images) and an error (loss) of 0.10 after evaluating it on our image dataset, as shown in Figure 5.5. A confusion matrix was used to show the performance of a CNN model in Figure 5.6. The F-Score, Precision, Sensitivity, and Accuracy results calculated using the confusion matrix are shown in Table 5.2.



Results

Validation accuracy: 96.67%

Training finished: Reached final iteration

Training Time

Start time: 03-Mar-2021 12:31:17

Elapsed time: 4 min 50 sec

Training Cycle

Epoch: 5 of 5

Iteration: 30 of 30

Iterations per epoch: 6

Maximum iterations: 30

Validation

Frequency: 3 iterations

Other Information

Hardware resource: Single CPU

Learning rate schedule: Constant

Learning rate: 0.0001

Learn more

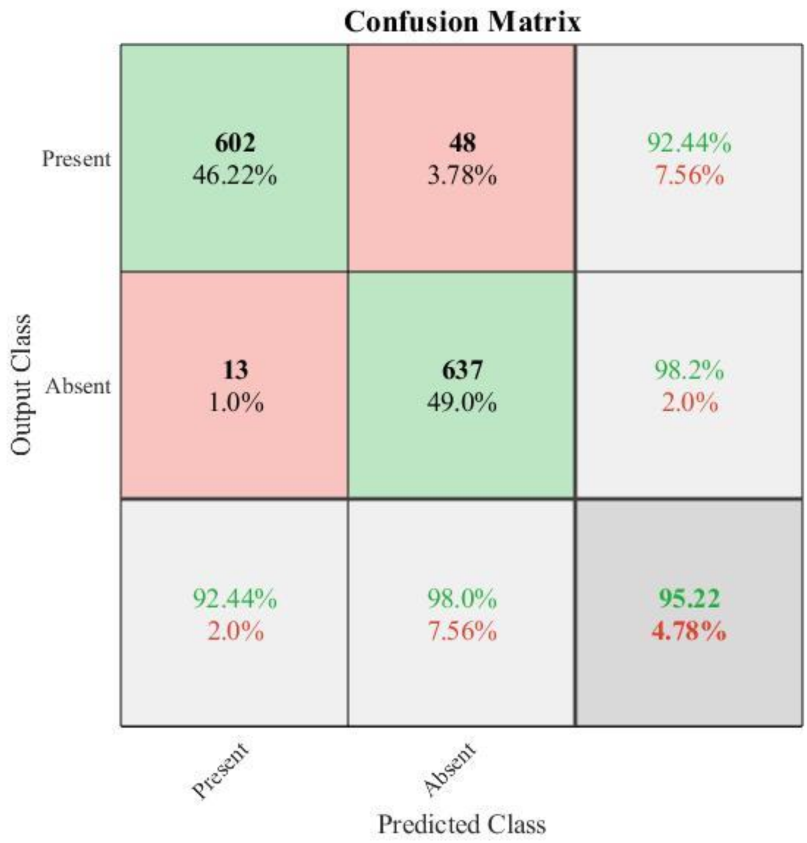
Accuracy

- Training (smoothed)
- Training
- Validation

Loss

- Training (smoothed)
- Training
- Validation

Figure 5.5: The performance of the CNN model.


Confusion Matrix
Output Class
Predicted Class

	Present	Absent	
Present	**602** 46.22%	**48** 3.78%	92.44% 7.56%
Absent	**13** 1.0%	**637** 49.0%	98.2% 2.0%
	92.44% 2.0%	98.0% 7.56%	**95.22** **4.78%**

Figure 5.6: The confusion matrix obtained using *E. coli* datasets.

Table 5.2: The statistical results for performance evaluation.

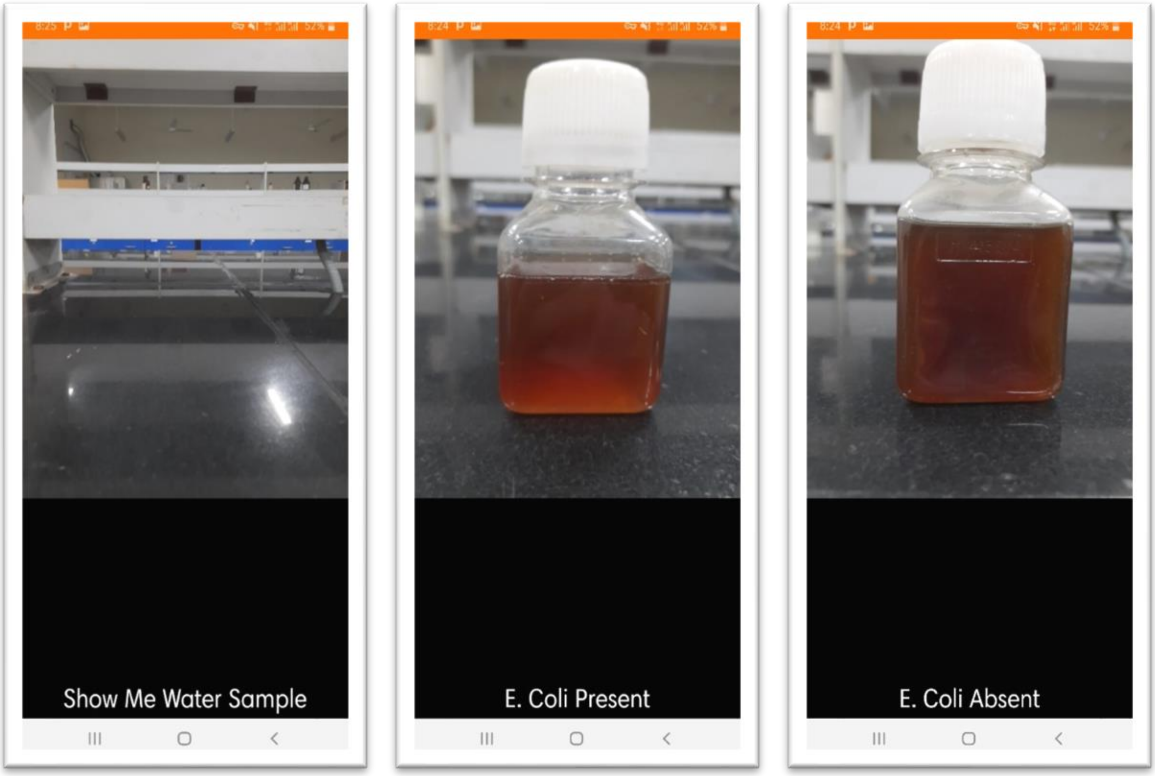
Performance Measure	Results
Accuracy	0.96
Sensitivity	0.92
Precision	0.98
F-Score	0.95

As a result, as seen in Figure 5.6, the CNN had a reasonably high level of confidence in its forecasts. This reliability is reflected in the high accuracy value of 97%. Remarkably, more false negatives were predicted by the CNN than false positives (forty eight *E. coli* present pictures were classified as *E. coli* absent, and thirteen *E. coli* absent pictures were classified as *E. coli* present). Because of the varying levels of blackness in the *E. coli* present photos, the CNN may have misidentified the lesser-brightness *E. coli* present photos as *E. coli* absent. Overall, the pictures were darker, and the water appeared to be more turbid. These findings are crucial to the issue we are considering because 96% accuracy means that people who do not have access to modern technology or complex water testing kits will be able to determine much more efficiently if water is polluted. Increased exposure to accurate testing procedures will assist persons in determining whether water is safe and avoiding the negative repercussions of water contamination. Additionally, with a sensitivity of 0.92 and an accuracy of 0.98, this method has been proven reliable for future applications. Moreover, as the number of sample photos in our data set grows, the machine learning model will develop and results in fewer categorization errors and higher reliability.

5.7 Mobile application for rapid detection of E. coli

TensorFlow 2.5.0 (Abadi et al. 2016) from Google was chosen as the machine learning platform for building an AI-based smartphone application to quickly identify *E. coli* on agar plates. TensorFlow was selected because it enables easy deployment on a smartphone system and offers a user-friendly graphical user interface. The CNN model was created utilizing 1301 images from laboratory experiments and a TensorFlow Lite (.tflite) model. We created the mobile app with Android Studio v. 4.1.1 (Studio 2017; Zapata 2013) and the TensorFlow (. tflite) model (Alsing 2018). The smartphone app allows users to capture photos with their smartphone's built-in camera. The captured pictures are then categorized by interacting with a

CNN. The *E. coli* detection smartphone app that was developed using the model is seen below in Figure 5.7.


Show Me Water Sample

E. Coli Present

E. Coli Absent

Figure 5.7: *E. coli* Detection App.

5.8 Summary

The developed CNN model for rapid detection of *E. coli* in water achieved an accuracy of 96% and an error (loss) of 0.10. The developed model was able to predict *E. coli* bacteria in each water sample within 458ms. The approach was considerably more successful than alternative methods such as polymerase chain reaction (PCR) and traditional techniques. The performance of the model was validated using various statistical measures, which shows that the model is reliable and effective in detecting *E. coli*. We have also developed an AI-based smartphone application using CNN that captures the images using an inbuilt smartphone camera and predicts the bacteria in water based on colour intensity. We demonstrated the effectiveness of our AI-based smartphone application by using it to monitor water quality for bacterial pollution and improve precision over laboratory results. This detection of *E. coli* bacteria in water allows the public health engineering department technicians to use this innovative app without prior knowledge. In the future, we want to expand and expand our study on real-time monitoring of microbiological water quality without complicated testing procedures. Given more time and resources, we could develop a system that can work more effectively upon projects such as Intel Clean Water AI (2018).

References:

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., ... & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)* (pp. 265-283).

- AI-Driven Test System Detects Bacteria in Water. Retrieved from https://software.intel.com/content/www/us/en/develop/articles/ai-driven-test-system-detects-bacteria-in-water.html

- Alaslani, M. G. (2018). Convolutional neural network based feature extraction for iris recognition. *International Journal of Computer Science & Information Technology (IJCSIT) Vol, 10*.

- Alsing, O. (2018). Mobile object detection using tensorflow lite and transfer learning.

- BIS, I. (2012). 10500: 2012. Indian standard drinking water-specification (second revision), Bureau of Indian Standards, New Delhi.

- Bolton, D. J., O'Sullivan, J., Duffy, G., Baylis, C. L., Tozzoli, R., Wasteson, Y., & Lofdahl, S. (2007). Methods for Detection and Molecular Characterisation of Pathogenic Escherichia coli. *Ashtown Food Research Centre, Ashtown, Ireland*.

- Brownlee, J. (2019). A gentle introduction to object recognition with deep learning. *Machine Learning Mastery, 5*.

- Central Bureau of Health Intelligence, National Health Profile (2018). Ministry of Health and Family Welfare, Government of India, New Delhi.

- Dalianis, H. (2018). Evaluation metrics and evaluation. In *Clinical Text Mining* (pp. 45-53). Springer, Cham.

- Diarrhoeal disease. World Health Organization (2017). Available Online at http://www.who.int/mediacentre/factsheets/fs330/en

- Fausett, L. V. (2006). *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.

- Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5), 826-834.

- George, I., Petit, M., & Servais, P. (2000). Use of enzymatic methods for rapid enumeration of coliforms in freshwaters. *Journal of applied Microbiology*, *88*(3), 404-413.

- Greenwood D, Slack R, Peutherer J. Escherichia. In: Medical Microbiology. 16th ed. Edinburgh: Churchill Livingstone, 2003; p. 265–273.

- Gunda, N. S. K., Gautam, S. H., & Mitra, S. K. (2019). Artificial intelligence based mobile application for water quality monitoring. *Journal of The Electrochemical Society*, *166*(9), B3031.

- Gupta, A., Gupta, A., & Gupta, R. (2018). Power and Area Efficient Intelligent Hardware Design for Water Quality Applications. *Sensors & Transducers*, *227*(11), 67-72.

- Huang, L., & Wu, T. (2018). Novel neural network application for bacterial colony classification. *Theoretical Biology and Medical Modelling*, *15*(1), 1-16.

- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215-243.

- Jagals, P., Grabow, W. O. K., Griesel, M., & Jagals, C. (2000). Evaluation of selected membrane filtration and most probable number methods for the enumeration of faecal coliforms, Escherichia coli and Enterococci in environmental waters. *Quantitative Microbiology*, *2*(2), 129-140.

- Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455-5516.

- Maas, M. B., Perold, W. J., & Dicks, L. M. T. (2017). Biosensors for the detection of Escherichia coli. *Water Sa*, *43*(4), 707-721.

- MathWorks: Deep Network Designer Toolbox Release 2020b– MATLAB & Simulink – MathWorks India, available at, last access: 07th March 2021.

- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics, 5*(4), 115-133.

- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in plant science*, *7,* 1419.
- Nehal, S. A., Roy, D., Devi, M., & Srinivas, T. (2019). Highly sensitive lab-on-chip with deep learning AI for detection of bacteria in water. *International Journal of Information Technology*, 1-7.

- Olstadt, J., Schauer, J. J., Standridge, J., & Kluender, S. (2007). A comparison of ten USEPA approved total coliform/E. coli tests. *Journal of water and health, 5*(2), 267-282.

- Oshiro, R. (2002). Method 1604: Total Coliforms and Escherichia coli in water by membrane filtration using a simultaneous detection technique (MI Medium). *Washington, DC: US Environmental Protection Agency*.

- Rice, E. W., Geldreich, E. E., & Read, E. J. (1989). The presence-absence coliform test for monitoring drinking water quality. *Public Health Reports*, *104*(1), 54.

- Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., ... & Blake, P. A. (1983). Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *New England Journal of Medicine*, *308*(12), 681-685.

- Saxena, T., Kaushik, P., & Mohan, M. K. (2015). Prevalence of E. coli O157: H7 in water sources: an overview on associated diseases, outbreaks and detection methods. *Diagnostic microbiology and infectious disease, 82*(3), 249-264.

- Studio, A. (2017). Android studio. *The Official IDE for Android*.

- Tamerat, N., Muktar, Y., & Shiferaw, D. (2016). Application of molecular diagnostic techniques for the detection of E. coli O157: H7: a review. *J Vet Sci Technol, 7*(362), 1-9.

- US EPA. (2007). Drinking water standards and health advisories table. *Edition of the Drinking Water Standards and Health Advisories*.

- WHO. Guidelines for Drinking-water Quality. 4th ed.; WHO (2011) Geneva, Switzerland.

- Zapata, B. C. (2013). *Android studio application development* (p. 110). Packt Publishing.

6. Automated Bacteria Colony Counting on Agar Plates Using Machine Learning

The identification of E. coli bacteria is critical for the prevention of health risks. According to EPA-approved gold standard methods, 24-48 h are required to count viable cells in water. Manual counting of the viable bacterial colony on agar plates is time-consuming and can be prone to human error. The method requires experts to identify and count colonies on agar plates using a microscope. Hence, the bacterial counting procedure must be automated in order to decrease error. The main objective of this study was to develop an automatic system for bacterial colony counting.

6.1 Introduction

In 1982, *E. coli* bacteria was discovered as a human pathogen (Riley et al. 1983). Escherichia coli (*E. coli*) bacteria are gram-negative, non-spore, rod-shaped pathogens that generate gas after 48 h of fermentation at 35°C in the specified growth environment (Greenwood et al., 2003). *E. coli* bacteria should not be present in 100 mL of a water sample according to World Health Organization (WHO) (WHO 2011), IS 10500 (BIS 2012), and USEPA protocols (USEPA 2007). According to a WHO study, 1.9 billion people consume contaminated water throughout the world (Diarrhoeal disease 2017). Waterborne diseases are still widespread in India and have resulted in 10,738 fatalities in the last five years, as per a Central Bureau of Health Intelligence report (CBHI 2018). Waterborne diseases affect around 37.7 million Indians each year. The examination of indicator organisms is mainly used to assess bacteriological drinking water quality. Because of advancements in testing technologies, *E. coli* is a more exact indication of water pollution than other fecal coliform bacteria. *E. coli* bacteria can be identified in the laboratory using conventional methods (Bolton et al. 2007), enzymatic methods (George et al. 2000), molecular methods (Tamerat et al. 2016; Saxena et al. 2015), and biosensor-based methods (Maas et al. 2017). Gram staining (Smith et al. 2005) is a method used to differentiate bacteria on the basis of their cell wall constituents. Staining categorizes bacteria into two classes, that is, gram-positive and gram-negative. Agar is a growth medium used for selective differentiation and detection of *E. coli* bacteria in water samples. The viable cell count method (Jennison et al. 1937) is used to detect and count the number of actively growing bacterial cells in water in terms of colony-forming units (CFUs).

The USEPA-approved gold-standard methods for detecting *E. coli* and counting viable cells are based on culturing water samples on solid agar plates or in liquid media. Viable cell counts can be done by the plate count method (USEPA 2010). In the plate count technique, serial dilutions are made by creating aliquots of a certain volume of liquid culture and plating numerous serial dilutions onto culture plates. A glass spreader is used to spread the volume of culture over the surface of an agar plate, which is then incubated to develop colonies. The bacterial concentration in a water sample can then be calculated, assuming that each viable cell forms a single colony (Harrigan et al., 2014). The number of colonies is counted manually using a bacteria colony counter (Rompré et al. 2002). Manual counting of viable bacterial cells on agar plates is time-consuming and can be prone to human error. The method requires experts to identify and count viable cells. Furthermore, due to bacterial overcrowding, high numbers of colony-forming units on a plate will lead to inaccurate results (Breed et al. 1916).

Eosin methylene blue (EMB) agar is a selective and differential medium used to isolate fecal coliform bacteria. It provides a rapid and accurate method of differentiating *E. coli* from other gram-negative pathogens. *E. coli* bacteria is an indicator of fecal contamination in water. The presence of *E. coli* bacteria indicates the possibility of the presence of pathogenic bacteria and viruses (Khan et al. 2020). Nobody can ferment lactose except *E. coli*. If *E. coli* bacteria are present in water. In this case, a colony will appear on an agar plate with a metallic sheen with a dark center. Gram-positive bacteria growth is typically hindered on EMB agar because of the toxicity of the methylene blue dye. Therefore, only colonies of *E. coli* will appear on agar plates. If no colony appears on the agar plates, it indicates that *E. coli* bacteria are absent in water. Consequently, it can be concluded that only *E. coli* bacteria will grow on agar plates; gram-positive bacteria will not grow on agar plates, so this method is only valid for *E. coli* bacteria.

6.2 Artificial neural networks

Artificial neural networks (ANN) are a kind of biological neural network model. Biological neural networks aid living beings in interpreting, identifying, and learning from their patterns for future applications (Fausett 2006). ANNs are most commonly used when the correlation between inputs and outputs is not linear (Khan et al. 2021a, b). The utility of ANNs has been demonstrated especially in cases in which a system's output is not specified by a unique correlation method between input and output (Gupta et al. 2018). A convolutional neural network (CNN) is a type of neural network. The discovery of a sensory system in the brain, the visual cortex, prompted the idea for CNNs. A CNN has one or two convolutional layers, always with one or more fully connected layers and a subsampling layer (Khan et al. 2020).

The function of a CNN’s convolution layer is the same as that of visual cortex cells (Hubel et al. 1968). A handcrafted feature extractor gathers vital information from input. It removes insignificant features from the typical pattern recognition model (Fukushima et al. 1983). Following the extractor, a trainable classifier, which is a normal neural network that categorizes feature vectors, is employed. CNN's use convolution layers as feature extractors. However, they are not handcrafted. The visual cortex is made up of numerous cells that detect light in tiny amounts. The visual field is divided by overlapping subregions called receptive fields. More complicated cells have larger receptive fields and act as local filters in the input space. During the training phase of a CNN, weights of both fully connected layers and convolutional layers are calculated and utilized for classification (Huang et al. 2018) and feature extraction (Brownlee 2019). Convolution filter kernel weights are selected during the training phase. Convolutional layers can retrieve image patterns because the hidden layer’s receptive fields are constrained to be confined.

6.3 Machine Learning

Machine learning algorithms for object detection are based on autonomous learning and have good detection accuracy. Object detection algorithms based on machine learning have been developed for various applications like face detection (Viola et al. 2004), pedestrian detection (Dollar et al. 2011), medical object detection (Zhu et al. 2016), military object detection (Hua et al. 2018), intelligent transportation systems (Zhang et al. 2011), and intelligent monitoring systems (Chen et al. 2014). An anchor is used in object detection for classification and regression. The algorithm replaces the preceding region proposal network (RPN), feature selection method, and selective search (Kulkarni et al., 2015) with the guided anchor method (Wang et al., 2019). The network module is obtained using function selection (Hu et al. 2018) to elucidate the object compression problem. The skip pooling method (Bell et al. 2016) is used to solve the problem of small object size so as to improve the detection efficiency of faster region-based CNN (R-CNN) in complex scenes.

A public data set comprising 54,306 pictures of healthy and diseased plant leaves was used by Mohanty et al. (2016). They developed a deep CNN that accurately detected images with 99% accuracy. Alaslani et al. (2018) extracted learned features from a support vector machine (SVM) and pre-trained CNN algorithm for image classification. The SVM method was utilized for classification, and an AlexNet pre-trained CNN model was utilized for extracting features. Public pictures of irises were utilized to create an iris recognition system. According to the results, the recognition precision of the database was 98.3%. Huang et al. (2018) used digital pictures to construct a convolutional neural network (CNN) for bacteria colony recognition. The categorization of bacteria colonies was carried out using data from Peking University First Hospital. According to the results, the network was able to classify 18 bacteria colonies with a 73% accuracy.

Gunda et al. (2019) created an AI-based smartphone application for water quality monitoring. The accuracy of the developed model was 99%. However, the developed model was not verified by the authors using performance functions. Using a photonic crystal-based optical biosensor, Nehal et al. (2019) created an AI-based lab-on-chip for detecting bacterial contamination. The limits of employing separate sensors to assess physical, chemical, and bacteriological parameters of water quality, which impact the sensitivity and accuracy of the results, were discovered using these biosensors. The procedure is expensive and time-consuming. Recent studies have used publicly available data sets and platforms for colony counting. Torelli et al. (2018) used the publicly available OpenCV and CellProfiler software platforms for automatic bacterial cell counting. Albaradei et al. (2020) used the CSRNet transfer learning application for cell counting. The training of the model was carried out using Python with the Keras library. Model performance was validated using root-mean-square error (RMSE) values. The average RMSE value of the developed model was 22.38, which is very high. We cannot rely on this model for viable cell counting. The aforementioned studies indicate that prediction models have been based on public data sets, published research reports, and testing data that is freely available on the internet, making it impossible to assess model accuracy. As a result, we cannot depend entirely on these studies to count bacteria colonies. However, no study has been done to identify and count *E. coli* bacterial cells on agar plates with great accuracy using experimental laboratory data.

In order to deal with the problems associated with manual cell counting, this study developed a machine-learning algorithm based on a faster region-based convolutional neural network with higher accuracy. We developed two different networks to detect and count bacterial growth because the developed CNN can always identify an image, not an object. Therefore, it cannot draw bounding boxes around them. To identify objects in an image, we developed a faster R-CNN model to count bacteria colonies using a region proposal network. An RPN generates region proposals by predicting the class and box offsets to collect predetermined bounding box templates known as anchor boxes. The proposed faster R-CNN algorithm combined with the region proposal network, anchor box, region of interest (ROI) pooling layer, convolution layer, and classification output layer. This method can provide improved classification and detection accuracy compared to existing methods.

6.4 Methodology

We built our data set in this study using water samples obtained from eight districts in Rajasthan as part of the Birla Institute of Technology and Science and the University of Virginia (BITS-UVA) groundwater contamination project. We also developed a graphical user interface (GUI) application to rapidly count *E. coli* colony-forming units on agar plates using laboratory experiment data. The main contributions of this paper are as follows:

- A total of 1,301 groundwater samples were collected and analyzed using various water quality tests in the laboratory. This experimental data set was validated using AI techniques.
- A convolutional neural network algorithm was developed to identify and predict *E. coli* bacteria on agar plates.
- A mobile application was developed for the rapid detection of *E. coli* bacteria in water using CNN.
- A faster region-based convolutional neural network algorithm was developed to automate the process of manual cell counting of *E. coli* bacteria on agar plates.
- A graphical user interface application was created to rapidly count *E. coli* colony-forming units on agar plates using faster R-CNN.

6.5 Identification of bacteria and viable cell counting

The most significant bacteriological task is to classify water-borne pathogens. Generally, bacteria display three basic shapes: round, rod-shaped, and spiral. After water samples are collected, bacteria must be grown on culture media to be identified. Gram staining is the first step toward identifying bacteria (Tripathi et al. 2020). Staining is a method used to differentiate bacteria in the cell wall based on their different constituents. By coloring these cells violet or red, the gram staining method categorizes bacteria into two classes: gram-positive and gram-negative. Agar is a growth medium that is used for selective identification and differentiation of *E. coli* in water (Frampton 1993).

Viable cell counts were performed using the plate count method (USEPA 2002). EMB agar (Leininger et al. 2001) was used as a growth media for the identification of *E. coli*. Using 1-mL water samples, serial dilution was performed so that dilution two had a concentration one-tenth that of dilution one and one hundredth that of the water sample. Next, 20 mL of molten cooled agar solution and diluted water samples were mixed well and poured into a sterile petri dish with a diameter of 90 mm. The agar plates were placed in an incubator at 35°C for 24–48 h to distribute the colonies throughout the depth of the medium. Colony-forming units present in the petri dish were counted using a microscope at 10× magnification. The colony-forming units present in a water sample can be determined by multiplying the number of colonies present on the agar plate by the sample's dilution factor (Bartram et al. 1996), as shown in Equation (6.1)

CFU/mL = number of colonies × dilution factor (6.1)

The viable count analysis of the water samples showed *E. coli* bacterial strains with minimum cell counts of 4×10^7 CFU/100 mL and maximum cell counts of 132×10^7 CFU/100 mL, as shown in Figure 6.1. A total of 99 groundwater samples were found positive for *E. coli*.

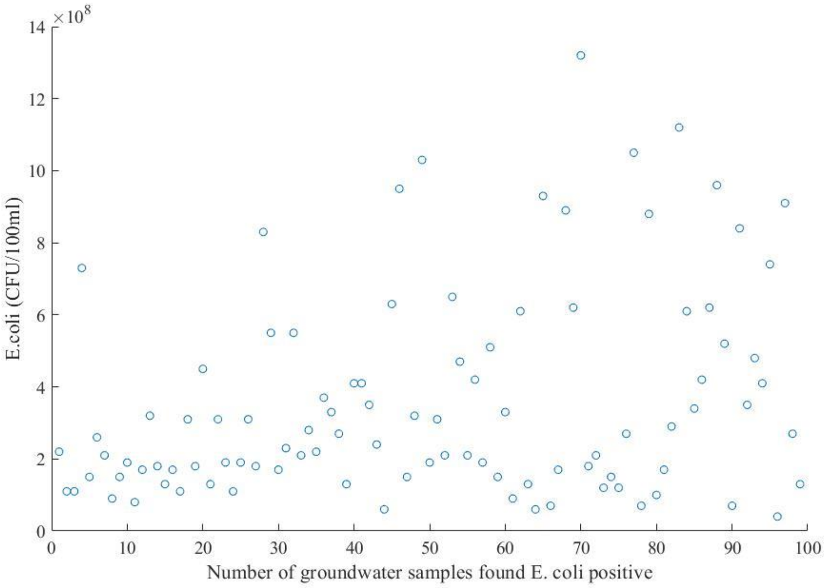


Figure 6.1: Viable cell count of *E. coli* in groundwater samples.

Image segments were obtained from high-resolution images of agar plates using a Canon EOS 3000D (Canon India) digital single-lens reflex (SLR) camera with a resolution of 18 megapixels (MP) and a Sigma 70–300 mm F/4–5.6 DG macro telephoto zoom lens (Sigma). The camera's assembly height was 300 mm above the dish. A universal serial bus (USB) cable was used to connect the camera to the device. A transition in the color of the medium with a metallic sheen with a dark center indicated the presence of *E. coli* bacteria, as shown in Figure 6.2. Image segmentation was performed using thresholding techniques (Zaitoun and Aqel 2015). Each segment was classified into one of five groups based on the number of colonies it included, ranging from 1 to 5, or categorized as an outlier if it contained no colonies but had bubbles, dust, or dirt on the agar surface (Figure 6.3). Segments were labeled manually using a dedicated GUI, and a custom database format was used to store the labeling data.

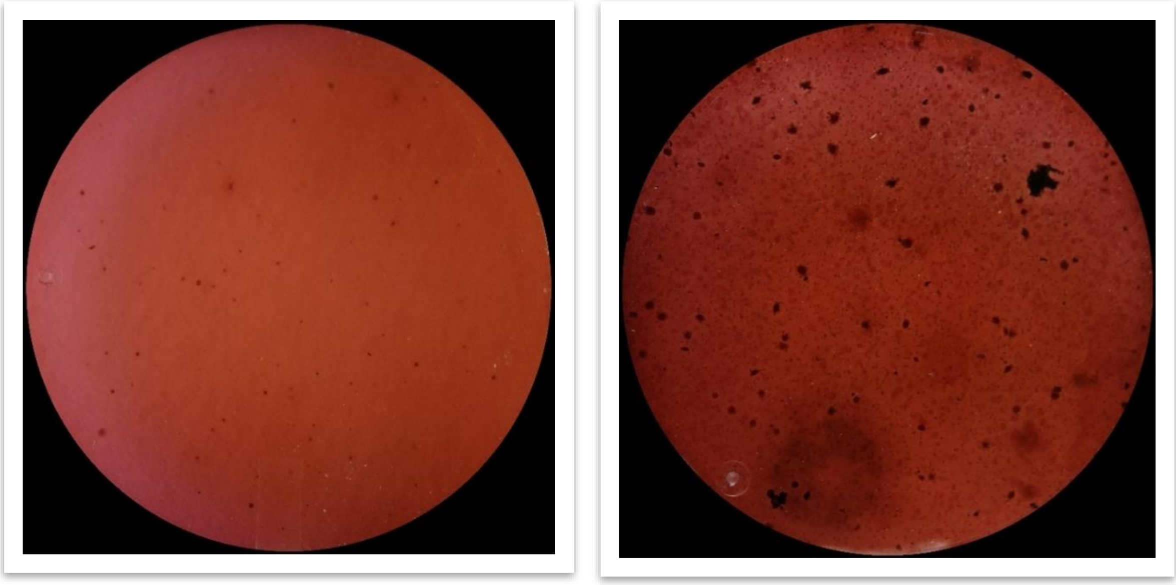


Figure 6.2: Colour change in culture medium with a metallic sheen with a dark center due to the presence of bacteria.

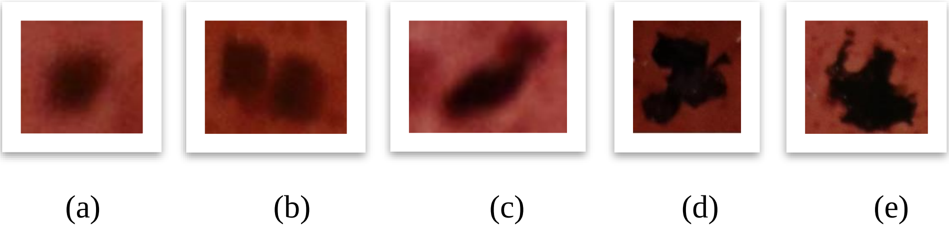


Figure 6.3: A few of *E. coli* dataset images representing a certain number of colonies, from 1 (a) to 5 (e)

6.6 Convolutional Neural Network for identification of E. coli bacteria on Petri dishes

The CNN model was fed 200 pictures collected from laboratory testing (plate count test). Image manipulation procedure was used to resize, crop, and rotate the input photos so as to increase the classifier's accuracy in training sets. The pictures were divided into two groups based on the presence or absence of *E. coli*. The experimental data were divided into three sets: training (70%), testing (15%), and validation (15%). We trained the CNN using MATLAB Deep Network Designer 2020b (MathWorks) to distinguish between pictures with *E. coli* present and *E. coli* absent. The CNN model was trained on a machine with 16 gigabytes (GB) of RAM. To detect *E. coli* bacteria on agar plates, the CNN employed image color, red green blue (RGB), and black and white (BW) values. On a scale of 0 to 1, predictions were given for each image. If the predicted value was adjusted to zero, that is, less than 0.5, the final forecast for that picture was *E. coli* absent. If the predicted value was adjusted to 1, that is, a value larger than 0.5, the final forecast for that picture was *E. coli* present. Consequently, the closer the model went to 0, the more certain it was that a picture was an *E. coli* absent image. The nearer it got to 1, the more certain it was that a picture was an *E. coli* absent image. This study employed a data set of 200 photos to determine whether *E. coli* bacteria were present on agar plates. All pictures were 3,228 × 3,215 pixels in size. In this study, color, RGB, and BW pictures were utilized to categorize images based on color and texture characteristics. Figure 6.4 depicts images from the data set used in the study.

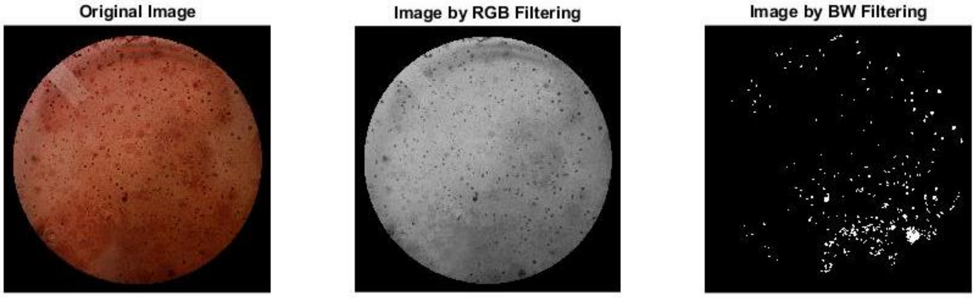


Figure 6.4: A few sample images from the *E. coli* dataset.

The CNN model's performance was validated using F-score, precision, sensitivity, and accuracy statistical measures (Dalianis 2018; Prabha et al. 2016). The CNN model used cross validation to test the networks more thoroughly. This means that all images of the data were used as both training and test data, split into iterations. TP stands for true positive (appropriately recognizing an *E. coli* present image as *E. coli* present). TN stands for true negative (appropriately recognizing an *E. coli* absent image as *E. coli* absent). FP stands for false positive (inappropriately recognizing an *E. coli* present image as *E. coli* absent). FN stands for false negative (inappropriately recognizing an *E. coli* absent image as *E. coli* present). Accuracy was defined as the proportion of properly recognized samples (both present and absent) among all samples [see Equation (6.2)]. The ratio of all recognized positive samples to all positive samples is the sensitivity [see Equation (6.3)]. If sensitivity is strong, the class was accurately detected. The value of high sensitivity implies that a class has been appropriately recognized. Precision was defined as the ratio of all positively detected positive samples to all positively predicted positive samples [see Equation (6.4)]. High precision suggests that a sample classified as positive is, in fact, positive. The weighted average of sensitivity and accuracy was used to get the F-score. In the F-score, the harmonic mean replaces the arithmetic mean. This metric punishes high values much more [see Equation (6.5)]. Intersection-over-union (IoU), also known as the Jaccard index, is a popular metric for measuring how accurate a proposed image segmentation is in comparison to a known/ground-truth segmentation. In segmentation tasks, IoU is recommended over accuracy because it is less impacted by class imbalances that exist in foreground/background segmentation tasks [see Equation (6.6)]. The ratio of genuine negatives to total negatives in the data is defined as specificity [see Equation (6.7)]. The proportion of expected negatives that are true negatives is denoted as the negative predictive value. It expresses the likelihood that a projected negative is a genuine negative [see Equation (6.8)]. The fraction of negative instances wrongly classified as positive cases in the data is referred to as the false positive rate [see Equation (6.9)]. The predicted proportion of type I mistakes is referred to as the false discovery rate (FDR). A type I mistake occurs when we wrongly reject the null hypothesis, resulting in a false positive [see Equation (6.10)]. A false negative is an outcome in which the model forecasts the negative class inaccurately [see Equation (6.11)]. The following equations [Equations (6.2)–(6.11)] can be used to compute F-Score, precision, sensitivity, and accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (6.2)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (6.3)$$

$$Precision = \frac{TP}{TP+FP} \qquad (6.4)$$

$$F - Score = \frac{2*Sensitivity*Precision}{Sensitivity+Precision} \qquad (6.5)$$

$$IoU = \frac{TP}{TP+FN+FP} \qquad (6.6)$$

$$Specificity\ = \frac{TN}{TN+FP} \tag{6.7}$$

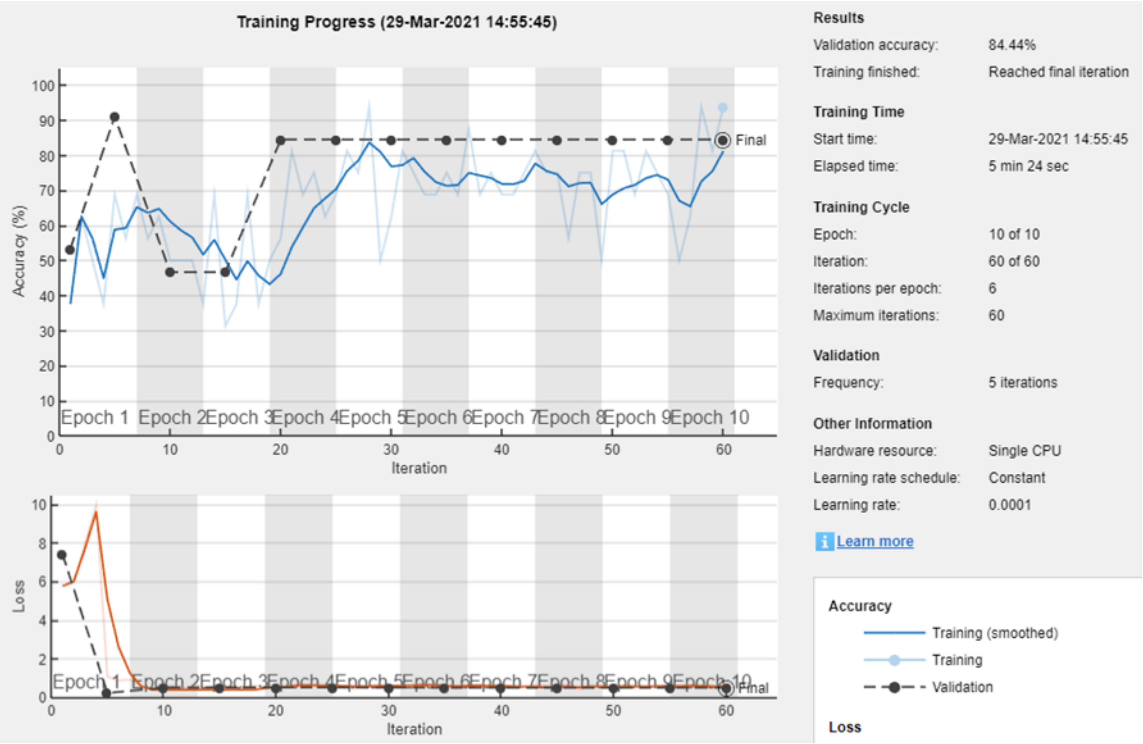
$$Negative\ predictive\ value = \frac{TN}{TN+FN} \tag{6.8}$$

$$False\ positive\ rate = \frac{FP}{FP+TN} \tag{6.9}$$

$$False\ discovery\ rate = \frac{FP}{TP+FP} \tag{6.10}$$

$$False\ negative\ rate = \frac{FN}{TP+FN} \tag{6.11}$$

In the current study, machine learning was utilized to train a CNN model to recognize *E. coli* bacteria on agar plates. The CNN model was trained to match all *E. coli* concentrations to a binary output indicating whether *E. coli* was present. MATLAB R2020b (MathWorks) software was used to create the CNN model. The CNN model's performance was compared using original, RGB, and BW filtering pictures. As demonstrated in Figure 6.5, the CNN model with BW filtered pictures outperformed the other models, with an accuracy of 93.33%, compared to original images (84.44%) and RGB filtered images (86.67%). Figure 6.6 depicts the architecture of the proposed CNN model for detecting *E. coli* on agar plates.



Results
Validation accuracy: 84.44%
Training finished: Reached final iteration

Training Time
Start time: 29-Mar-2021 14:55:45
Elapsed time: 5 min 24 sec

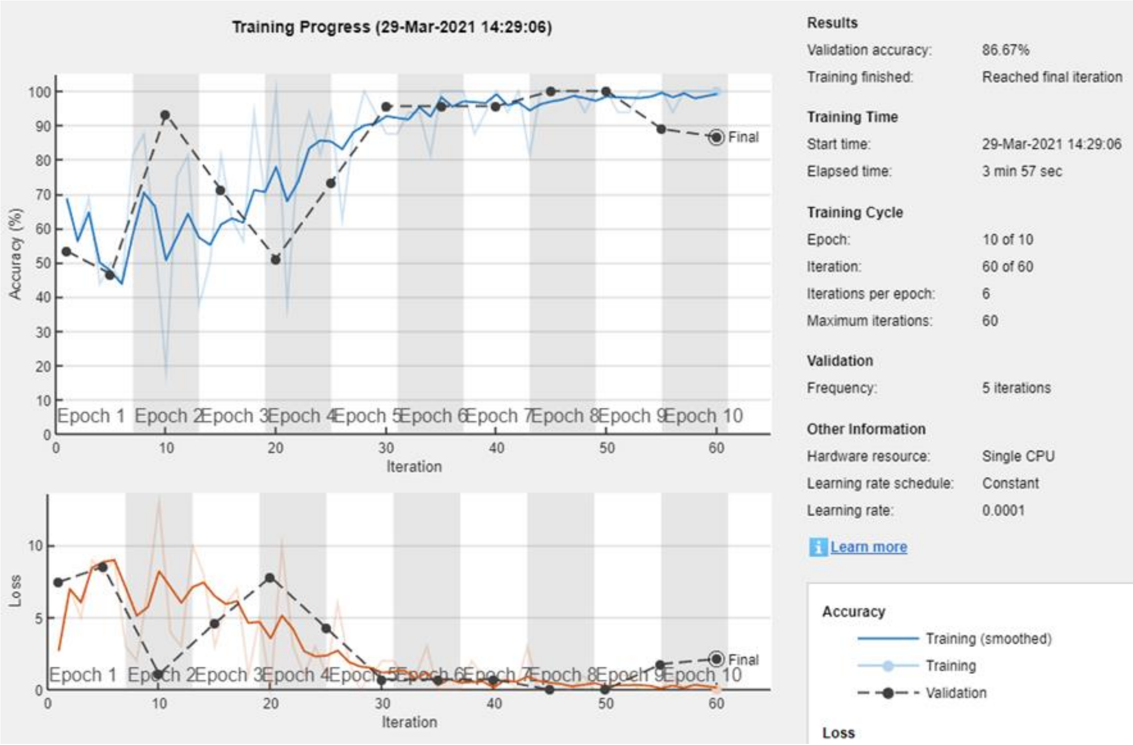
Training Cycle
Epoch: 10 of 10
Iteration: 60 of 60
Iterations per epoch: 6
Maximum iterations: 60

Validation
Frequency: 5 iterations

Other Information
Hardware resource: Single CPU
Learning rate schedule: Constant
Learning rate: 0.0001

Learn more

(a) Original Images


Training Progress (29-Mar-2021 14:29:06)
Accuracy (%)
100
90
80
70
60
50
40
30
20
10
0
Epoch 1 Epoch 2 Epoch 3 Epoch 4 Epoch 5 Epoch 6 Epoch 7 Epoch 8 Epoch 9 Epoch 10
Final
0
10
20
30
40
50
60
Iteration
Loss
10
5
0
Final
Iteration

Results

Validation accuracy: 86.67%

Training finished: Reached final iteration

Training Time

Start time: 29-Mar-2021 14:29:06

Elapsed time: 3 min 57 sec

Training Cycle

Epoch: 10 of 10

Iteration: 60 of 60

Iterations per epoch: 6

Maximum iterations: 60

Validation

Frequency: 5 iterations

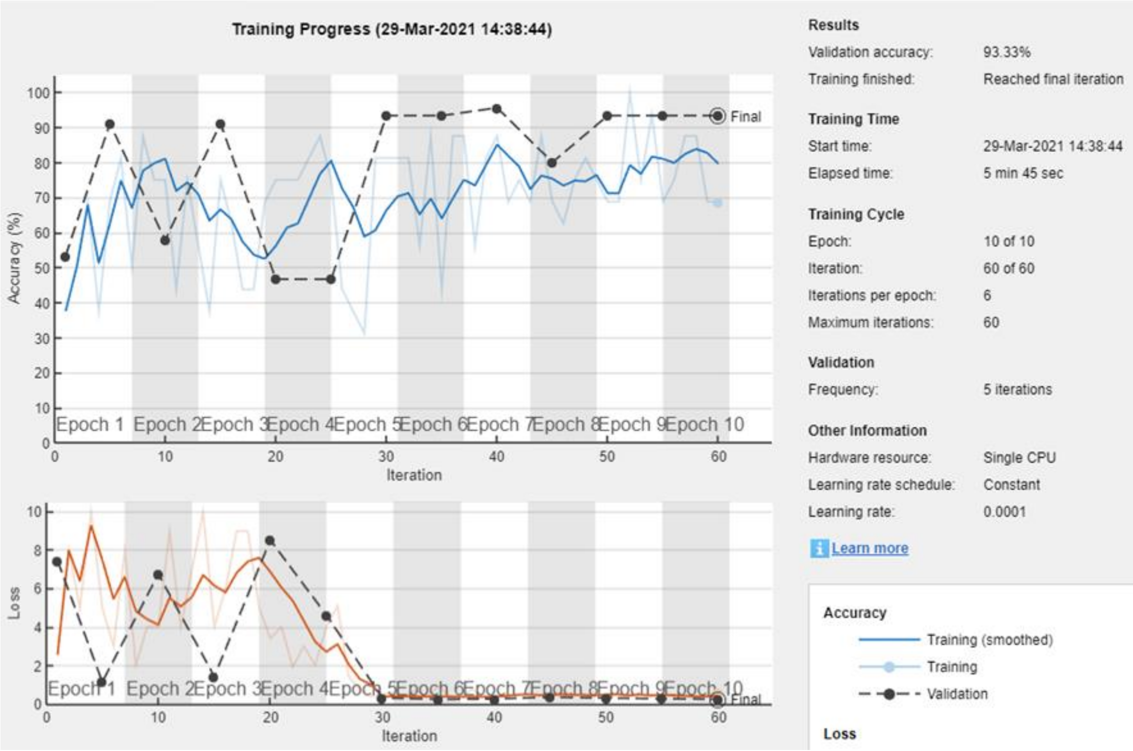
Other Information

Hardware resource: Single CPU

Learning rate schedule: Constant

Learning rate: 0.0001

(b) RGB filtering images


Training Progress (29-Mar-2021 14:38:44)
Accuracy (%)
100
90
80
70
60
50
40
30
20
10
0
Epoch 1 Epoch 2 Epoch 3 Epoch 4 Epoch 5 Epoch 6 Epoch 7 Epoch 8 Epoch 9 Epoch 10
Final
0
10
20
30
40
50
60
Iteration
Loss
10
8
6
4
2
0
Final
Iteration

Results

Validation accuracy: 93.33%

Training finished: Reached final iteration

Training Time

Start time: 29-Mar-2021 14:38:44

Elapsed time: 5 min 45 sec

Training Cycle

Epoch: 10 of 10

Iteration: 60 of 60

Iterations per epoch: 6

Maximum iterations: 60

Validation

Frequency: 5 iterations

Other Information

Hardware resource: Single CPU

Learning rate schedule: Constant

Learning rate: 0.0001

(c) BW filtering images

Figure 6.5: The performance of the CNN model using (a) Original, (b) RGB filter, and (c) BW filter images.

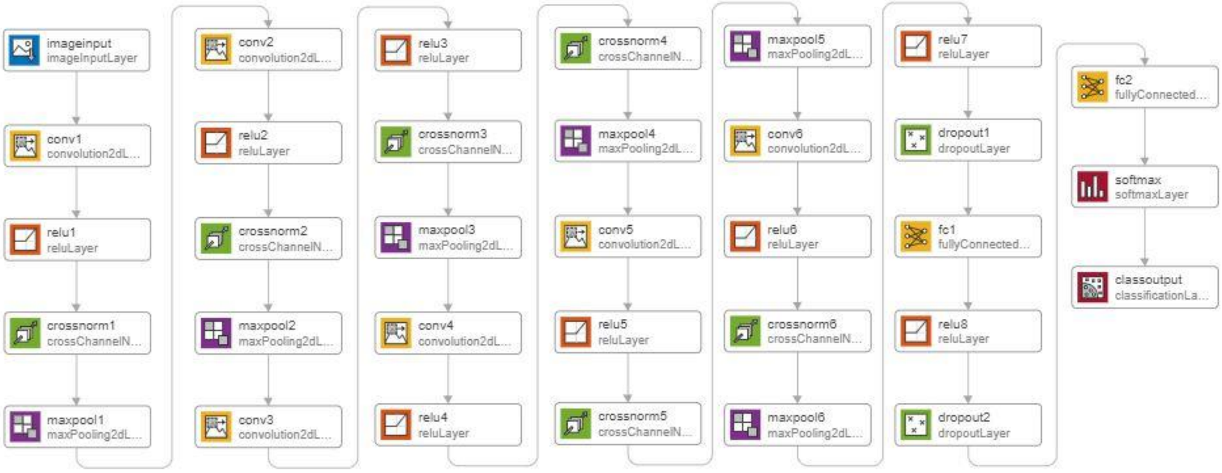


Figure 6.6: Overview of the CNN model.

Convolutional neural networks are a subgroup of artificial neural networks, as discussed previously. Typical CNN architecture consists of one or two convolutional layers with a subsampling layer complemented by one or more fully connected layers. Various regulatory functions like batch normalization and dropout layers are also used to optimize the performance of CNNs (Bouvrie 2006). We trained our CNN model from scratch using MATLAB Deep Network Designer 2020b (MathWorks). The developed model can be modified in case needed as per the requirements. Our developed CNN model consisted of eight layers, and it was more efficient in terms of training time and architecture because it contained fewer layers than other models trained for more than one purpose, like visual geometry group VGG (2014) (19 layers), GoogLeNet (2014) (22 layers) and Inception (2016) (70 layers) (Simonyan et al. 2014; Szegedy et al. 2015, 2017).

The developed CNN model comprised six convolution layers and two fully connected layers. A different weight, bias, and rectified linear unit (ReLU) were applied to each layer. The number of filters in the initial convolutional layer determined the weights and bias (conv1). The first layer was padded using 3 × 3 kernels with a stride of 4. The number of filters utilized in conv1 was 96, and the filter size was 11 × 11 × 3. Therefore, initial weights 11 × 11 × 3 × 96 and bias 1 × 1 × 96 were assigned. The following two layers were padded with 3 × 3 kernels with a stride of 2, and the final three layers with 1 × 1 kernels with a stride of 1. The convolutional layers (conv1) were supplemented by two fully connected layers that were not strided in order to prevent overfitting. Table 6.1 presents a thorough description of the CNN architecture. We used grid search to optimize the batch size, learning rate, and the maximum number of epochs before training the data, as shown in Table 6.2.

Table 6.1: Detailed structure of CNN model.

S. No	Name	Type	Activations	Learnables
1	imageinput 3228x3215x3 images with ‘zerocenter’ normalization	Image Input	3228x3215x3	-
2	conv1 96 11x11 convolutions with stride (4 4) and padding (3 3 3 3)	Convolution	806x803x96	Weights 11x11x33x96 Bias 1x1x96
3	relu1 ReLU	ReLu	806x803x96	-
4	crossnorm1 cross channel normalization with 5 channels per element.	Cross Channel Normalization	806x803x96	-
5	maxpool1 3x3 max pooling with stride (2 2) and padding (0 0 0 0)	Max Pooling	402x401x96	-
6	conv2 128 5x5 convolutions with stride (2 2) and padding (3 3 3 3)	Convolution	202x202x128	Weights 5x5x96x128 Bias 1x1x128
7	relu2 ReLU	ReLu	202x202x128	-
8	crossnorm2 cross channel normalization with 5 channels per element.	Cross Channel Normalization	202x202x128	-
9	maxpool2 3x3 max pooling with stride (2 2) and padding (0 0 0 0)	Max Pooling	100x100x128	-
10	conv3 256 3x3 convolutions with stride (2 2) and padding (3 3 3 3)	Convolution	52x52x256	Weights 3x3x128x256 Bias 1x1x256
11	relu3 ReLU	ReLu	52x52x256	-
12	crossnorm3 cross channel normalization with 5 channels per element.	Cross Channel Normalization	52x52x256	-
13	maxpool3 3x3 max pooling with stride (2 2) and padding (0 0 0 0)	Max Pooling	25x25x256	-
14	conv4 384 3x3 convolutions with stride (1 1) and padding (1 1 1 1)	Convolution	25x25x384	Weights 3x3x256x384 Bias 1x1x384

15	relu4 ReLU	ReLu	25x25x384	-
16	crossnorm4 cross channel normalization with 5 channels per element.	Cross Channel Normalization	25x25x384	-
17	maxpool4 3x3 max pooling with stride (2 2) and padding (0 0 0 0)	Max Pooling	12x12x384	-
18	conv5 128 3x3 convolutions with stride (1 1) and padding (1 1 1 1)	Convolution	12x12x128	Weights 3x3x384x128 Bias 1x1x128
19	Relu5 ReLU	ReLu	12x12x128	-
20	crossnorm5 cross channel normalization with 5 channels per element.	Cross Channel Normalization	12x12x128	-
21	maxpool5 3x3 max pooling with stride (2 2) and padding (0 0 0 0)	Max Pooling	5x5x128	-
22	conv6 96 3x3 convolutions with stride (2 2) and padding (1 1 1 1)	Convolution	3x3x96	Weights 3x3x128x96 Bias 1x1x96
23	relu6 ReLU	ReLu	3x3x96	-
24	crossnorm6 cross channel normalization with 5 channels per element.	Cross Channel Normalization	3x3x96	-
25	maxpool6 3x3 max pooling with stride (2 2) and padding (1 1 1 1)	Max Pooling	2x2x96	-
26	Relu7 ReLU	ReLu	2x2x96	-
27	dropout1 50% dropout	Dropout	2x2x96	-
28	fc1 2 fully connected layer	Fully Connected	1x1x2	Weights 2x384 Bias 2x1
29	relu8 ReLU	ReLu	1x1x2	-
30	dropout2 50% dropout	Dropout	1x1x2	-
31	fc2 2 fully connected layer	Fully Connected	1x1x2	Weights 2x2 Bias 2x1
32	softmax softmax	Softmax	1x1x2	-
33	classoutput crossentropyex	Classification Output	-	-

Table 6.2: Grid search of hyperparameters for backend tuning

Hyperparameters	**Values**
Optimizer	SGDM with learning rate (0.001, **0.0001**, 0.00001)
Batch size	(4,**16**,32,64)
Epoch	(10,30,**60**)

In Table 5.2, the hyperparameters underlined in bold represent the best performance. On training sets, the best hyperparameters were utilized for training the CNN model. The input pictures were 2,180 × 960 × 3 pixels in size. These values represented a channel's height, breadth, and length. The channel size for a color picture was 3, corresponding to RGB values. Convolution was used to extract features from input pictures. The convolution layer (conv1) had a filter size of 11 × 11, corresponding to the height and breadth of the filters employed by the training feature while scanning over the pictures. In the conv1 layer, 96 filters were employed, indicating the number of neurons linked to the same input area. A cross-channel normalization layer was used to normalize the gradients and activations. A nonlinear activation function, that is, a ReLU, was used after the normalization layer. After convolution, each feature map was downsampled using max pooling. Because the number of parameters was decreased, it maintained only the most important information from the pictures and avoided overfitting. The rectangular region had dimensions of (2, 2).

A fully connected layer has all of its neurons coupled to the neurons of the preceding layer. Such a layer combines all characteristics from the preceding layers to categorize all trends in the picture. The fully connected (FC) layer connected all of the features that were used to identify the pictures. As a consequence, the output parameters in the fully connected layer equaled the number of classes in the predicted data, which in this case was two. The softmax activation function was used to normalize the output of the fully connected layer. The softmax layer generated a series of positive numbers that sum to 1. These values were subsequently used as categorization probabilities by the classification layer. The classification layer was the final layer. Each input was assigned to one of the mutually distinct groups by this layer. It computed the error (loss) based on the probability obtained from the softmax activation function. The CNN model was validated using cross-validation with an iteration size of 60. Figure 6.7 depicts the performance of a CNN model using a confusion matrix. Table 6.3 displays the F-score, precision, sensitivity, and accuracy results obtained using the confusion matrix. Assessing it on our image data set, our model had an accuracy of 97% (194/200 photos) and an error (loss) of 0.15, as shown in Figure 6.7.

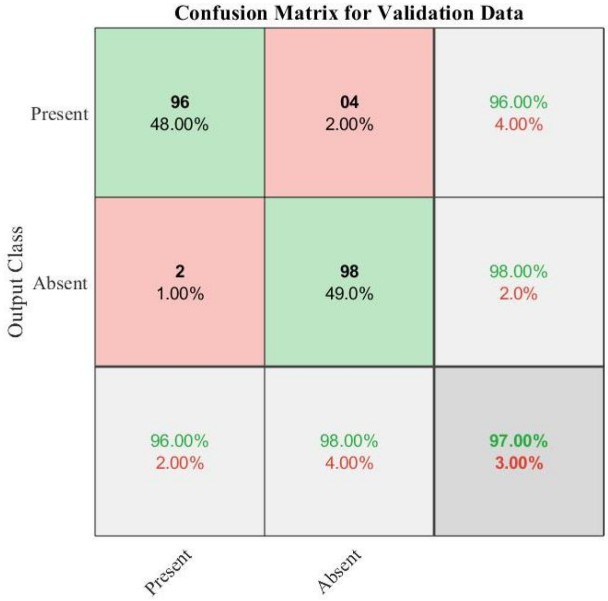


Figure 6.7: The confusion matrix obtained using *E. coli* datasets.

Table 6.3: The statistical results for performance evaluation of CNN.

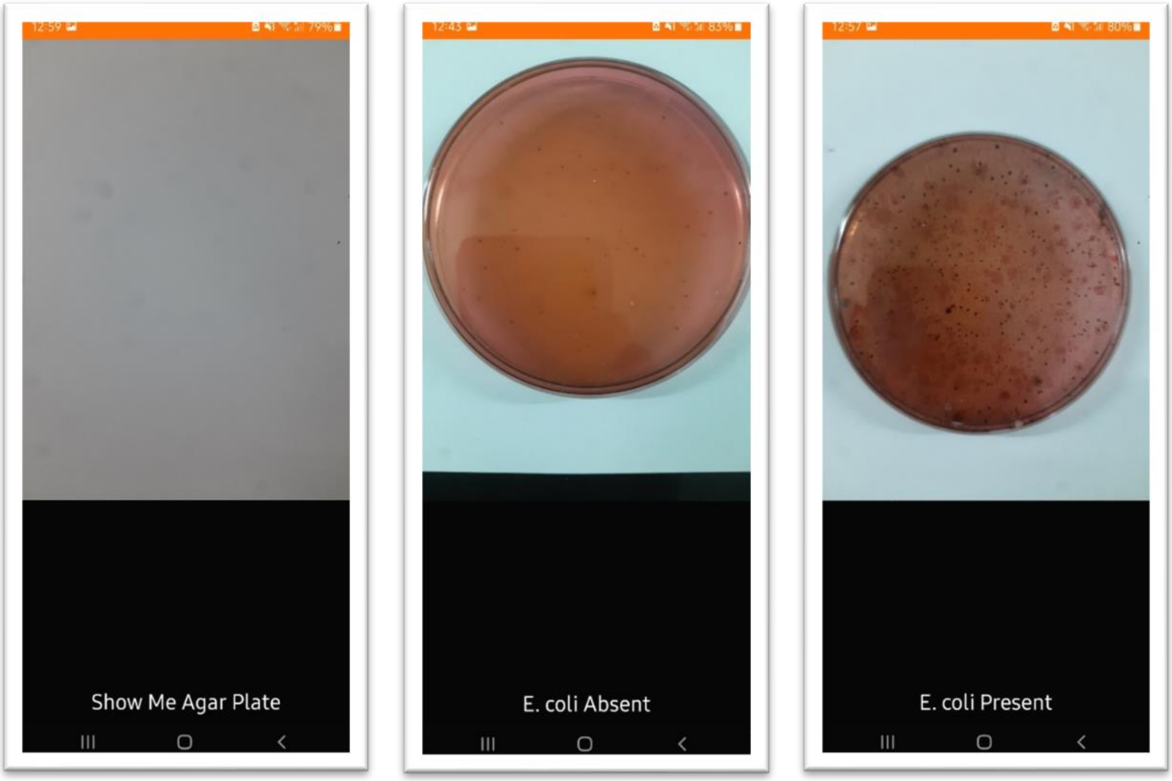
Performance Measure	**Results**
Accuracy	0.97
Sensitivity	0.96
Precision	0.98
F-Score	0.97

As a result, as seen in Figure 6.7, CNN had a reasonably high level of confidence in its forecasts. This reliability is reflected in the high accuracy value of 97%. Surprisingly, more false negatives were predicted by the CNN than false positives (four *E. coli* present pictures were classified as *E. coli* absent, and two *E. coli* absent pictures were classified as *E. coli* present). Because of the varying levels of blackness in the *E. coli* present photos, the CNN may have misidentified the lesser-brightness *E. coli* present photos as *E. coli* absent. Overall, the pictures were darker, and the water appeared to be more turbid.

These findings are crucial to the issue we are considering because 97% accuracy means that people who do not have access to modern technology or complex water testing kits will be able to determine much more efficiently if water is polluted. Increased exposure to accurate testing procedures will assist persons in determining whether water is safe and avoiding the negative repercussions of water contamination. Additionally, with a sensitivity of 0.96 and an accuracy of 0.98, this method has been proven reliable for future applications. Moreover, as the number of sample photos in our data set grows, the machine learning model will develop and results in fewer categorization errors and higher reliability.

6.7 Mobile application for rapid detection of *E. coli*

TensorFlow 2.5.0 (Abadi et al. 2016, Alsing 2018) from Google was chosen as the machine learning platform for building an AI-based smartphone application to quickly identify *E. coli* on agar plates. TensorFlow was selected because it enables easy deployment on a smartphone system and offers a user-friendly graphical user interface. The CNN model was created utilizing 200 images from laboratory experiments and a TensorFlow Lite (.tflite) model. We created the mobile app with Android Studio v. 4.1.1 (Studio 2017; Zapata 2013) and the TensorFlow (. tflite) model (58). The smartphone app allows users to capture photos with their smartphone's built-in camera. The captured pictures are then categorized by interacting with a CNN. Figure 6.8 shows the *E. coli* identification mobile app that was created using the CNN model.



Show Me Agar Plate

E. coli Absent

E. coli Present

Figure 6.8: Smartphone App for Identification of *E. coli* on Agar Plates.

6.8 Faster Region-based Convolutional Neural Network (Faster R-CNN) for bacteria colony counting

Manual counting of viable bacterial cells on agar plates is time-consuming and can be prone to human error. The method requires experts to detect and count viable cells. We developed an algorithm based on a faster R-CNN for automatic bacteria CFU counting with improved detection efficiency to address the aforementioned issues. A total of 200 images obtained from plate count tests were utilized as input for the Faster R-CNN model. We used the ResNet-50 network (Amjoud et al. 2020) for the training of the model. The data set utilized in this study comprised 200 pictures in total. All images had a size of 3,228 × 3,215 pixels. Each segment obtained was assigned to a class based on the number of colonies it included, ranging from 1 to 5, or classified as an outlier if it contained dust, bubbles, or dirt on the agar instead of colonies. A horizontal flip, a BW filter, and image normalization were used to perform data enhancement.

Image manipulation procedure was used to resize, crop, and rotate the input photos so as to increase the classifier's accuracy in training sets. Broad associated areas appeared on the boundary of the agar plate as a result of the binarization procedure. The colonies that came into contact with these boundary sections and areas were excluded. The images were separated by the number of colonies present on the agar plates. The experimental data were divided into three sets: training (70%), testing (15%), and validation (15%). We trained the faster R-CNN to distinguish pictures using MATLAB Deep Network Designer 2020b (MathWorks). The faster R-CNN used BW values of images for the cell counting. The faster R-CNN model used cross-validation to test the networks more thoroughly. This meant that all data images were used as both training and test data, split into iterations. CNN model performance was validated using F-score, precision, sensitivity, and accuracy statistical measures. The methodology presented in Figure 6.9 indicates the application of the faster R-CNN model for bacteria colony counting.

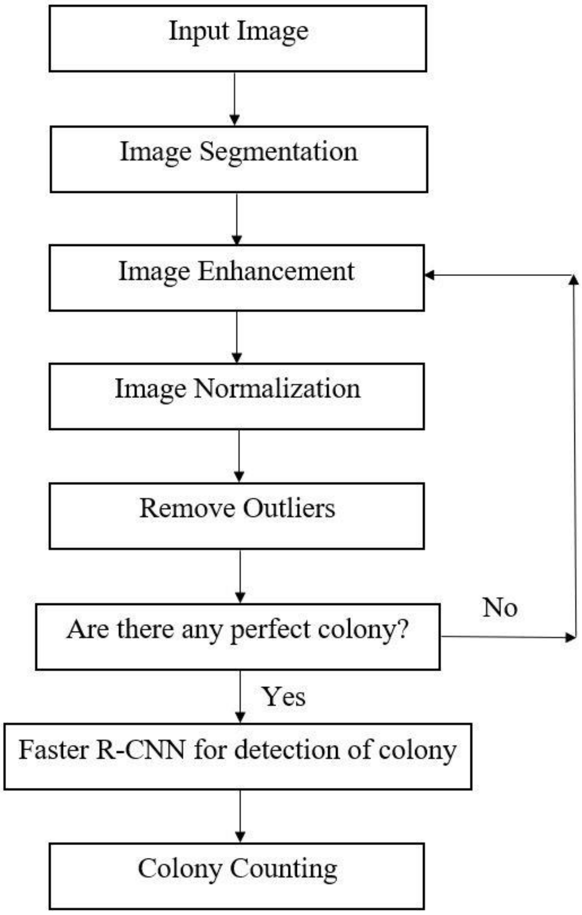


Figure 6.9: Methodology to apply Faster R-CNN algorithm for bacteria colony counting.

The developed CNN-based smartphone app detects bacteria on agar plates. It gives output in the form of present/absent. If *E. coli* bacteria are present on an agar plate, then a second CNN, that is, the faster R-CNN model is proposed for counting the number of actively growing bacterial cells of *E. coli* using an agar plate image. The faster R-CNN model was developed using MATLAB R2020b (MathWorks) software. We used grid search to optimize the learning rate, batch size, and maximum iterations before training the data. The change in the learning rate of the faster R-CNN algorithm was observed iteratively. The learning rate distribution was plotted after every iteration. The learning rate decreased exponentially by three to four times of magnitude. As Figure 6.10 shows, the learning rate continuously decreased exponentially until learning terminated at epoch 80. These trials showed that the learning rate decreased, and it became stationary at epoch 40.

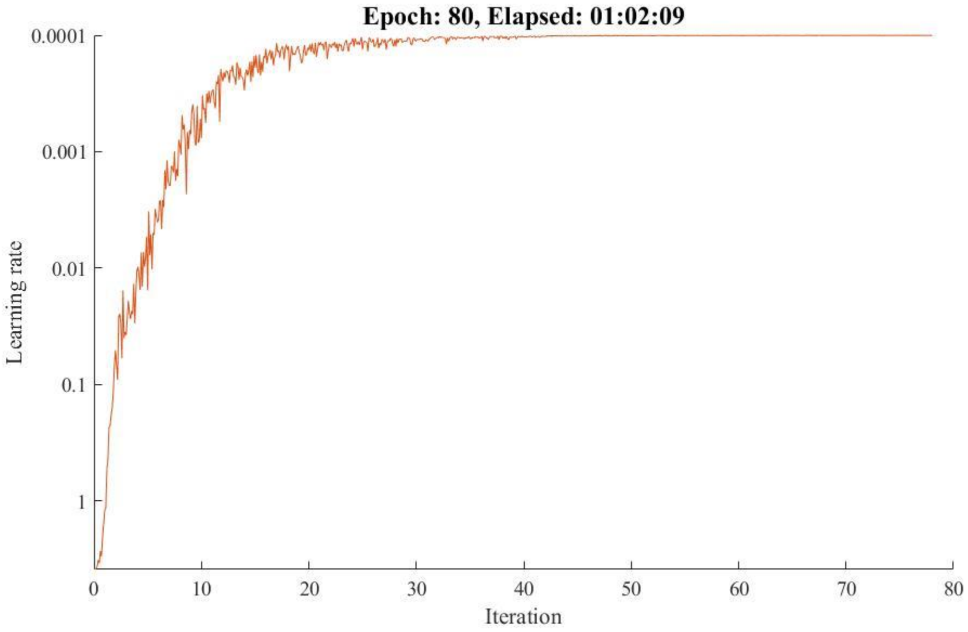
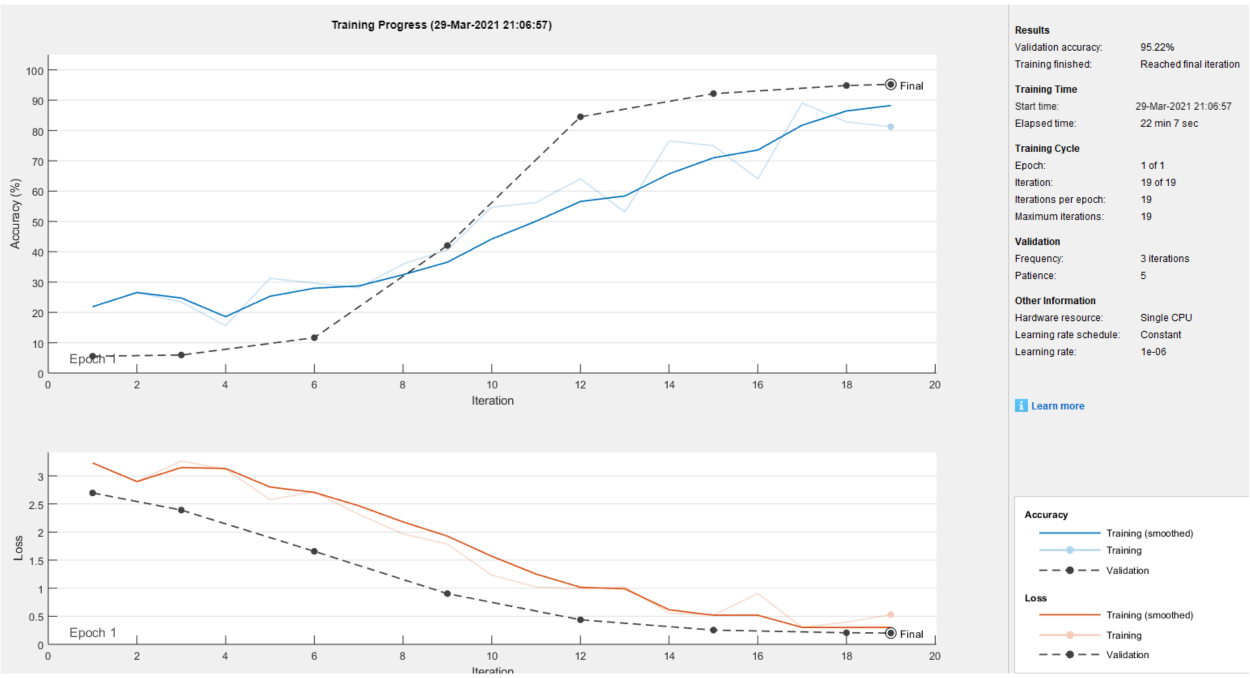


Figure 6.10: Change of learning rate with the Faster R-CNN model.

The pre-trained ResNet-50 network was used to develop the faster R-CNN model for the identification and colony counting of bacteria. The ResNet-50 network was converted into an object detection model using transfer learning. The last three classification layers were replaced with new layers for the nine classes to count bacteria cells. The FC, Softmax, and Classification layers were replaced with rcnnFC, rcnnSoftmax, and rcnnClassification layers, respectively. The faster R-CNN generates region proposals using a region proposal network. An RPN predicts the object or background class as well as the box offsets for a collection of predefined bounding box templates called anchor boxes. The size of anchor boxes is usually calculated based on preliminary information of the scale and aspect ratio of objects in the training data set. The RPN's convolution layers were added to the feature extraction layer, and then the output of the RPN classification layer was added. The classification layer classified each anchor as colony-forming units or CFUs, and then the RPN regression output layers were added. The regression layer predicts the offsets of each anchor box. For each anchor box, the regression layer predicts four box offsets.

Finally, the classification and regression layers were connected to the inputs of the region proposal layer. The ROI pooling layer was connected to the output of the region proposal layer. The input images had a size of 2,180 × 960 × 3 pixels. The faster R-CNN model was optimized using stochastic gradient descent with momentum (SGDM) optimizer with a learning rate of

0.0001, batch size 32, and maximum epochs 30. Our model had a training accuracy of 95.22% and an error (loss) of 0.10 after putting it to the test on a collection of images, as shown in Figure 6.11. The architecture of the developed faster R-CNN model for rapid colony counting of *E. coli* on agar plates is shown in Figure 6.12.



Results	
Validation accuracy:	95.22%
Training finished:	Reached final iteration
Training Time	
Start time:	29-Mar-2021 21:06:57
Elapsed time:	22 min 7 sec
Training Cycle	
Epoch:	1 of 1
Iteration:	19 of 19
Iterations per epoch:	19
Maximum iterations:	19
Validation	
Frequency:	3 iterations
Patience:	5
Other Information	
Hardware resource:	Single CPU
Learning rate schedule:	Constant
Learning rate:	1e-06

Figure 6.11: The performance of the Faster R-CNN model.

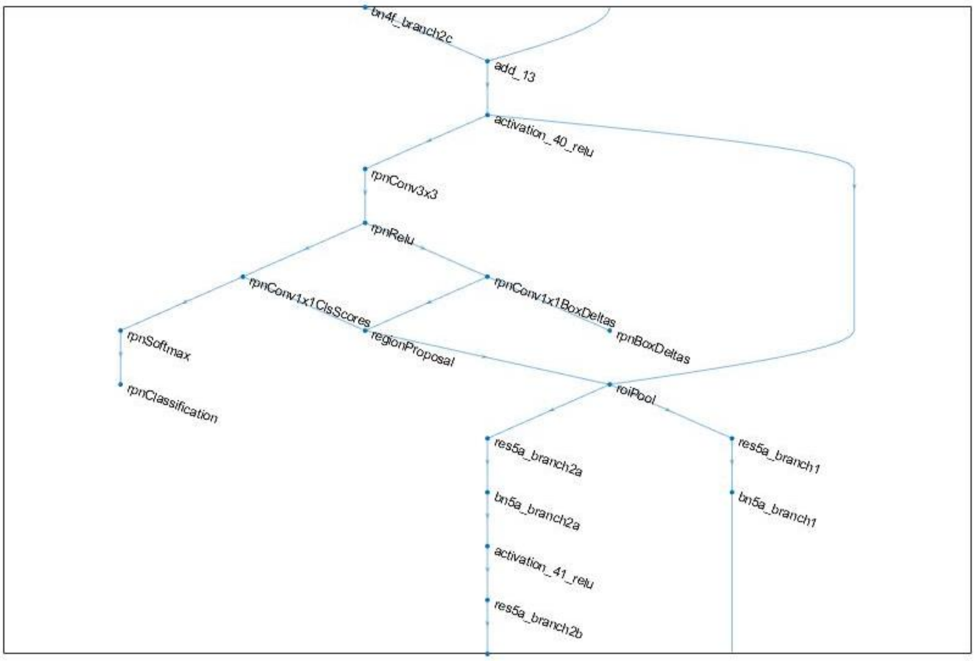


Figure 6.12: Overview of the Faster R-CNN model.

The performance of the model was tested on an image, as shown in Figure 6.13. The identification of *E. coli* bacteria on agar plates was performed using anchor boxes. The faster R-CNN was able to identify almost all colony-forming units on agar plates. Figure 6.14 shows the performance of the faster R-CNN model using a confusion matrix. Our model had an overall accuracy of 97% (193/200 images) and an error (loss) of 0.10 after analyzing it on our picture data set, as shown in Figure 6.14. The validation of the Faster R-CNN model was performed using cross-validation with an iteration size of 30. Table 6.4 shows the F-score, precision, sensitivity, and accuracy results calculated using the confusion matrix. The developed faster R-CNN algorithm exhibited a high degree of certainty in its colony-forming unit predictions, as shown in Figure 6.14. The high precision value of 97% reveals its reliability. Surprisingly, the faster R-CNN predicted more false negatives than false positives due to the varying degrees of darkness in the images. The accuracy of 0.88 and sensitivity of 0.98 show that this method has been proven reliable for future applications. Moreover, as the number of sample pictures in our data set grows, so does the machine learning model, resulting in fewer categorization errors and higher reliability.

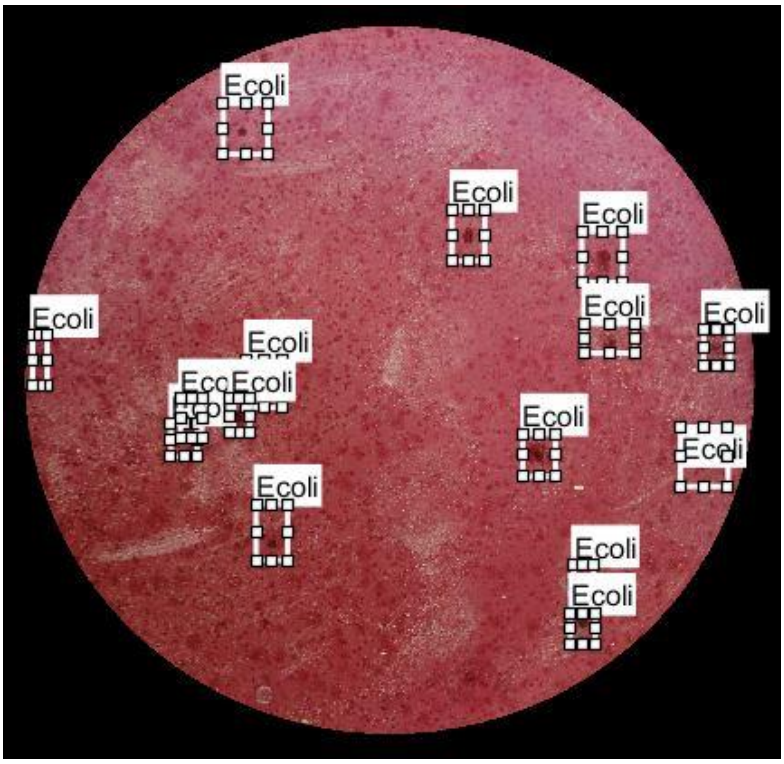
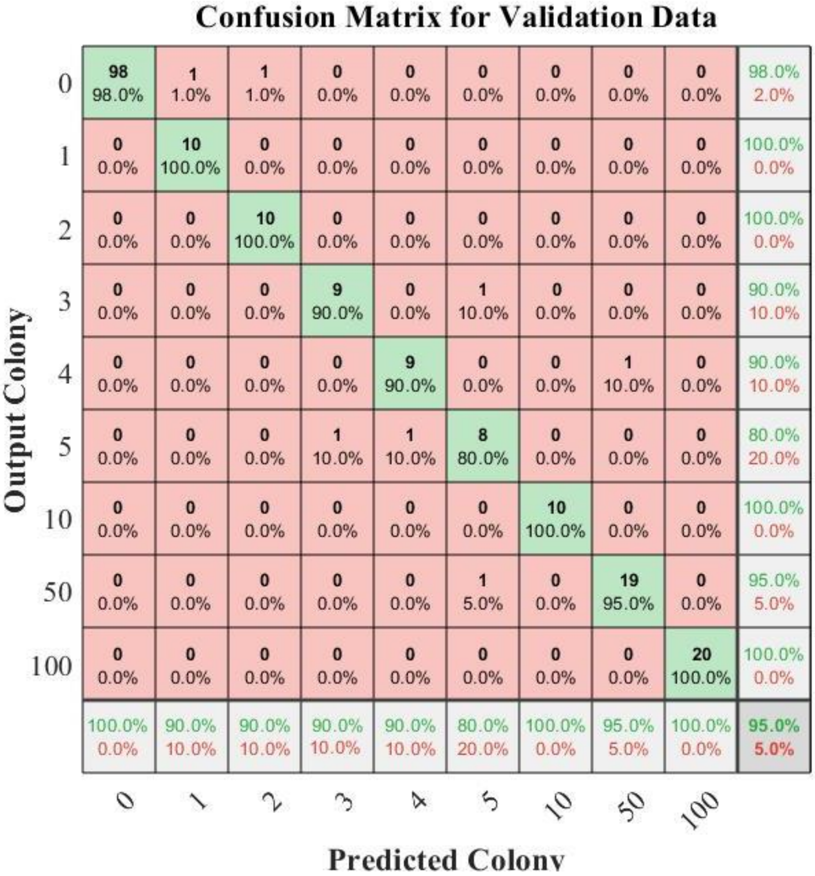


Figure 6.13: Faster R-CNN model prediction on testing data.



Confusion Matrix for Validation Data

Output Colony	0	1	2	3	4	5	10	50	100	
0	98 98.0%	1 1.0%	1 1.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98.0% 2.0%
1	0 0.0%	10 100.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100.0% 0.0%
2	0 0.0%	0 0.0%	10 100.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100.0% 0.0%
3	0 0.0%	0 0.0%	0 0.0%	9 90.0%	0 0.0%	1 10.0%	0 0.0%	0 0.0%	0 0.0%	90.0% 10.0%
4	0 0.0%	0 0.0%	0 0.0%	0 0.0%	9 90.0%	0 0.0%	0 0.0%	1 10.0%	0 0.0%	90.0% 10.0%
5	0 0.0%	0 0.0%	0 0.0%	1 10.0%	1 10.0%	8 80.0%	0 0.0%	0 0.0%	0 0.0%	80.0% 20.0%
10	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	10 100.0%	0 0.0%	0 0.0%	100.0% 0.0%
50	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 5.0%	0 0.0%	19 95.0%	0 0.0%	95.0% 5.0%
100	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	20 100.0%	100.0% 0.0%
	100.0% 0.0%	90.0% 10.0%	90.0% 10.0%	90.0% 10.0%	90.0% 10.0%	80.0% 20.0%	100.0% 0.0%	95.0% 5.0%	100.0% 0.0%	95.0% 5.0%

Predicted Colony

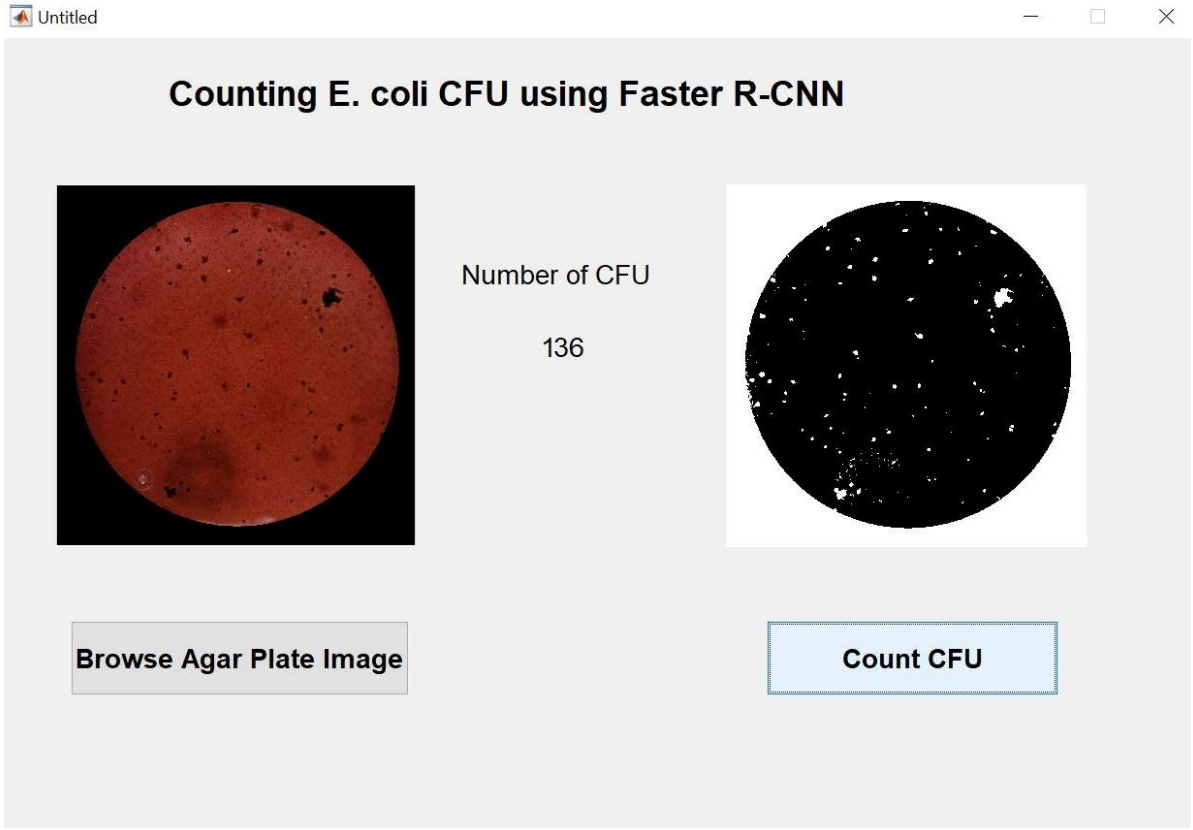
Figure 6.14: The confusion matrix obtained using *E. coli* datasets.

Table 6.4: The statistical results for performance evaluation of Faster R-CNN.

Performance Measure	Results
Accuracy	0.97
Sensitivity	0.98
Precision	0.88
F-Score	0.93
IoU	0.48
Specificity	0.04
Negative predictive value	0.67
False positive rate	0.97
False discovery rate	1
False negative rate	0.02

6.9 Graphical user interface (GUI) for automatic cell counting on agar plates

A graphical user interface (GUI) was developed in MATLAB 2020b (Smith 2006) for the system to automatically count colony-forming units on agar plates using images captured by a built-in smartphone-based sensor. Various components used for developing the GUI included push button, edit text, and axes. When a user presses the push button, axes are used to represent a picture in the space available. Furthermore, when a user clicks on the push button labeled “Browse Agar Plate Image,” all available images are displayed. The user must select a smartphone-captured image of an agar plate from the set of available images. The faster R-CNN algorithm uses the selected input image to count colony-forming units. The output image is shown on the push button labeled “Count CFU,” and the number of colonies in Petri dishes is shown on the push button labeled “Number of CFU.” A GUI application was developed to rapidly count *E. coli* colony-forming units on agar plates using faster R-CNN. The developed MATLAB GUI is shown in Figure 6.15.



Counting E. coli CFU using Faster R-CNN

Number of CFU

136

Figure 6.15: Graphical user interface (GUI) of the automated *E. coli* bacteria colony counter.

6.10 Comparison of models

A comparative analysis of the predictive models (CNN, faster R-CNN) is shown in Figures 6.16 and 6.17. The faster R-CNN method surpassed all the other configurations with the lowest error (loss) value (0.10) and the highest accuracy (95.22%). The results of the comparative analysis show that predictions obtained using the faster R-CNN method had higher decision-making precision and that it can, therefore, be used as a useful method in machine learning.

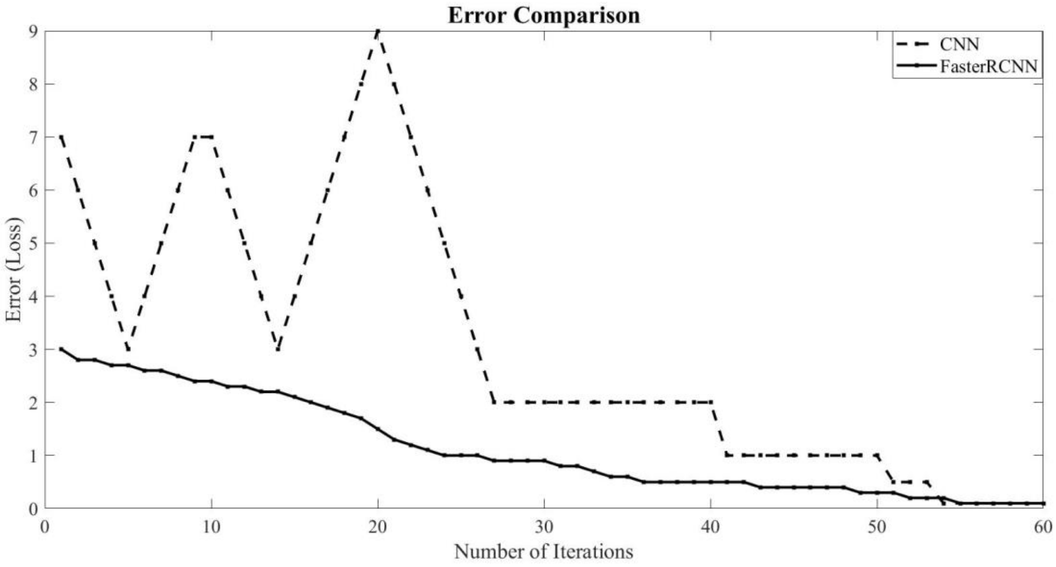


Figure 6.16: Error comparison of CNN and Faster R-CNN models.

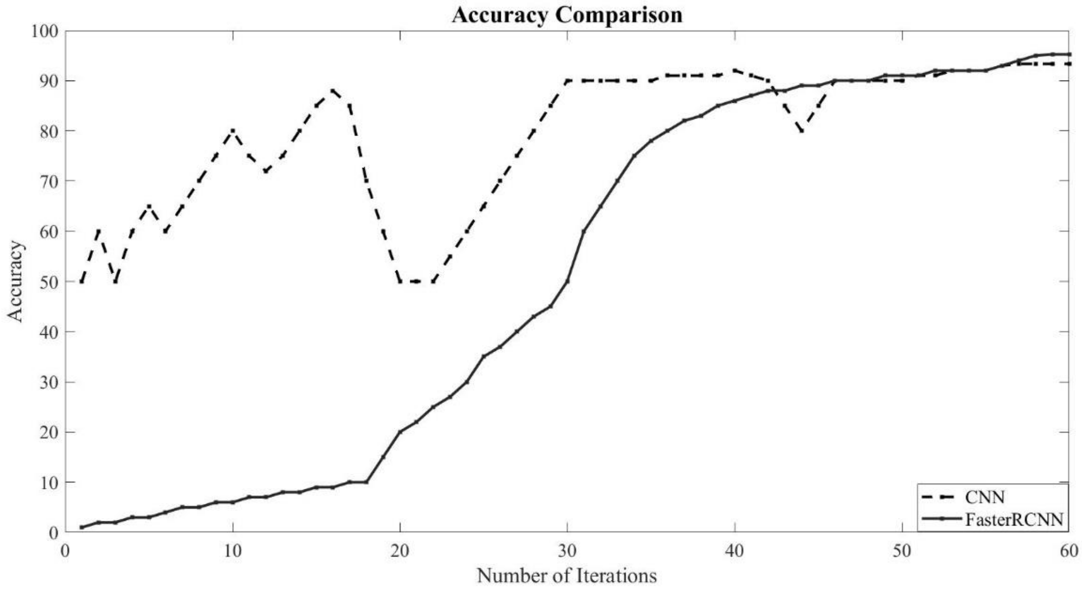


Figure 6.17: Accuracy comparison of CNN and Faster R-CNN models.

Two benchmark techniques were used to perform comparison tests in order to illustrate the performance of the proposed faster R-CNN method. The proposed faster R-CNN method was compared with the CNN methods developed by Huang et al. (2018), Ferrari et al. (2017), and Hay et al. (2018), the DNN method developed by Wang et al. (2020), and the FC-CNN developed by Zieliński et al. (2017). Table 6.5 shows that the proposed faster R-CNN method in this study outperformed the other methods in the detection of *E. coli* with the highest accuracy of 97%.

Table 6.5: Accuracy of detection on *E. coli* datasets

Method	References	Number of Images	Accuracy	Sensitivity	Precision	F-Score
CNN	Huang, L et al. (2018)	404	0.96	0.73	0.78	0.98
AlexNet pre-trained neural network		404	0.96	0.90	0.72	0.97
Unsupervised Autoencoder neural network		404	0.96	0.74	0.84	0.98
CNN	Ferrari, A et al. (2017)	17000	0.92	0.73	0.71	-
CNN	Hay, E. A et al. (2018)	482	0.90	-	-	-
DNN	Wang, H et al. (2020)	71	0.90	0.98	0.99	-
FC-CNN	Zieliński, B et al. (2017)	20	0.82	-	-	-
Proposed Faster R-CNN method	-	200	0.97	0.98	0.88	0.93

6.11 Summary

The bacteria colony counting technique is time-consuming, laborious, and prone to errors. However, experimental biologists often use manual or partially automatic counting methods to count viable cells. As a result, work that automates colony counting or provides labeled data sets for training machine learning models is needed. There is no defined technique for colony counting at the moment, despite the fact that machine learning is popularizing many areas of image processing. This may be due to a lack of annotated picture sets large enough to train artificial neural networks for colony counts on agar plates. We conclude that our proposed method could help solve this issue by using models that have already proficient in other applications and fine-tuning them for this task.

It is feasible to refine a model using a small group of photos in biological studies comprising several colony developments and then count the remaining pictures using our method. While this may accelerate the counting procedure, we feel that the most significant advantage is in making the procedure more precise and reliable. Pretrained models, such as the one we created, may also be used effectively to count CFUs in new pictures without any additional modification. We used a unique data set that we produced via laboratory experimental testing to demonstrate the feasibility of our technique for the first time. The developed faster R-CNN model for rapid counting of *E. coli* colonies on agar plates achieved an overall accuracy of 97% and an error (loss) of 0.10.

We created an AI-based smartphone application utilizing a CNN that takes pictures with the smartphone's built-in camera and forecasts bacteria on agar plates based on color intensity. We demonstrated its applicability for bacterial contamination and increased accuracy over laboratory results by utilizing our AI-powered smartphone app to check water quality. Within 1,032 ms, the developed CNN model predicted the presence of *E. coli* bacteria. The approach was considerably more successful than alternative methods such as polymerase chain reaction (PCR) and traditional techniques. The faster R-CNN model's performance was validated using a variety of statistical metrics, demonstrating that it is accurate and reliable in counting *E. coli* colony-forming units. We also developed a GUI interface to rapidly count *E. coli* colony-forming units on agar plates using faster R-CNN. The automated counting of *E. coli* bacterial cells on agar plates will enable technicians from public health engineering departments to utilize this innovative application without prior expertise.

However, further validation is needed to determine the model's ability to generalize through various experiments. As a result, we want to gather more data and assess the model's capacity for counting in a variety of situations, including higher-quality images with visible cellular components. This would also necessitate a further investigation of network architecture functionality, as well as potentially training additional layers, which will be possible with further input data. Exploring those possibilities will be the primary objective of our future efforts. Faster-RCNN has significant advantages in detection accuracy. However, faster-R-CNN has certain limitations, such as not being capable of real-time detection. The method for obtaining region boxes before classification necessitates a large amount of computation. Because of this limitation, another advanced technique known as mask R-CNN (He et al. 2017) has been developed. Furthermore, the performance of CNNs can be improved by tuning parameters like learning rate, epoch, and the number of layers. All these parameters affect the performance of a CNN. Image augmentation can be used to increase the data sample count using shear, zoom, rotation, and preprocessing functions. CNN model performance is also affected by overfitting and underfitting, which can be solved by training with more data, early stopping, and cross-validation. In the future, we want to expand and expand our study on real-time monitoring of microbiological water quality without the use of laboratory testing procedures. We could create a system that could operate more successfully on initiatives like Intel Clean Water AI if we had more time and resources.

References:

- Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., ... & Zheng, X. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

- Alaslani, M. G. (2018). Convolutional neural network based feature extraction for iris recognition. *International Journal of Computer Science & Information Technology (IJCSIT) Vol, 10*.

- Albaradei, S. A., Napolitano, F., Uludag, M., Thafar, M., Napolitano, S., Essack, M., ... & Gao, X. (2020). Automated counting of colony forming units using deep transfer learning from a model for congested scenes analysis. *IEEE Access*, *8*, 164340-164346.

- Alsing, O. (2018). Mobile object detection using tensorflow lite and transfer learning.

- Amjoud, A. B., & Amrouch, M. (2020, June). Convolutional neural networks backbones for object detection. In *International Conference on Image and Signal Processing* (pp. 282-289). Springer, Cham.

- Arrigoni, S., Turra, G., & Signoroni, A. (2017). Hyperspectral image analysis for rapid and accurate discrimination of bacterial infections: A benchmark study. *Computers in biology and medicine*, *88*, 60-71.

- Bartram, J., & Ballance, R. (Eds.). (1996). *Water quality monitoring: a practical guide to the design and implementation of freshwater quality studies and monitoring programmes*. CRC Press.

- Bell, S., Zitnick, C. L., Bala, K., & Girshick, R. (2016, June). Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2874-2883). IEEE.

- BIS, I. (2012). 10500: 2012. Indian standard drinking water-specification (second revision), Bureau of Indian Standards, New Delhi.

- Bolton, D. J., O'Sullivan, J., Duffy, G., Baylis, C. L., Tozzoli, R., Wasteson, Y., & Lofdahl, S. (2007). Methods for Detection and Molecular Characterisation of Pathogenic Escherichia coli. *Ashtown Food Research Centre, Ashtown, Ireland*.

- Bouvrie, Jake (2006) Notes on Convolutional Neural Networks. (Unpublished)

- Breed, R. S., & Dotterrer, W. D. (1916). The number of colonies allowable on satisfactory agar plates. *Journal of Bacteriology*, *1*(3), 321.

- Brownlee, J. A Gentle Introduction to Object Recognition With Deep Learning, 2019. URL: https://machinelearningmastery. com/object-recognitionwith-deep-learning.

- Caballé-Cervigón, N., Castillo-Sequera, J. L., Gómez-Pulido, J. A., Gómez-Pulido, J. M., & Polo-Luque, M. L. (2020). Machine learning applied to diagnosis of human diseases: A systematic review. *Applied Sciences*, *10*(15), 5135.

- Central Bureau of Health Intelligence, National Health Profile (2018). Ministry of Health and Family Welfare, Government of India, New Delhi.

- Dalianis, H. (2018). Evaluation metrics and evaluation. In *Clinical Text Mining* (pp. 45-53). Springer, Cham.

- Diarrhoeal disease. World Health Organization (2017). Available Online at http://www.who.int/mediacentre/factsheets/fs330/en

- Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2011). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*(4), 743-761.

- Durán, C., Ciucci, S., Palladini, A., Ijaz, U. Z., Zippo, A. G., Sterbini, F. P., ... & Cannistraci, C. V. (2021). Nonlinear machine learning pattern recognition and bacteria-metabolite multilayer network analysis of perturbed gastric microbiome. *Nature communications*, *12*(1), 1-22.

- Fausett, L. V. (2006). *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.

- Ferrari, A., Lombardi, S., & Signoroni, A. (2017). Bacterial colony counting with convolutional neural networks in digital microbiology imaging. *Pattern Recognition*, *61*, 629-640.

- Frampton, E. W., & Restaino, L. (1993). Methods for Escherichia coli identification in food, water and clinical samples based on beta‐glucuronidase detection. Journal of Applied Bacteriology, 74(3), 223-233.

- Fukushima, K., Miyake, S., & Ito, T. (1983). Neocognitron: A neural network model for a mechanism of visual pattern recognition. *IEEE transactions on systems, man, and cybernetics*, (5), 826-834.

- George, I., Petit, M., & Servais, P. (2000). Use of enzymatic methods for rapid enumeration of coliforms in freshwaters. *Journal of applied Microbiology*, *88*(3), 404-413.

- Greenwood D, Slack R, Peutherer J. Escherichia. In: Medical Microbiology. 16th ed. Edinburgh: Churchill Livingstone, 2003; p. 265–273.

- Gunda, N. S. K., Gautam, S. H., & Mitra, S. K. (2019). Artificial intelligence based mobile application for water quality monitoring. *Journal of The Electrochemical Society*, *166*(9), B3031.

- Gupta, A., Gupta, A., & Gupta, R. (2018). Power and Area Efficient Intelligent Hardware Design for Water Quality Applications. *Sensors & Transducers*, *227*(11), 67-72.

- Harrigan, W. F., & McCance, M. E. (2014). *Laboratory methods in microbiology*. Academic press.

- Hay, E. A., & Parthasarathy, R. (2018). Performance of convolutional neural networks for identification of bacteria in 3D microscopy datasets. PLoS computational biology, 14(12), e1006628.

- He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2018). Mask R-CNN. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(2), 386-397.

- Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7132-7141).

- Hua, X., Wang, X., Wang, D., Huang, J., & Hu, X. (2018). Military Object Real-Time Detection Technology Combined with Visual Salience and Psychology. *Electronics*, *7*(10), 216.

- Huang, L., & Wu, T. (2018). Novel neural network application for bacterial colony classification. *Theoretical Biology and Medical Modelling*, *15*(1), 1-16.

- Huang, L., & Wu, T. (2018). Novel neural network application for bacterial colony classification. *Theoretical Biology and Medical Modelling*, *15*(1), 1-16.

- Huang, L., & Wu, T. (2018). Novel neural network application for bacterial colony classification. Theoretical Biology and Medical Modelling, 15(1), 1-16.

- Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215-243.

- Jennison, M. W. (1937). The relations between plate counts and direct microscopic counts of Escherichia coli during the logarithmic growth period. *Journal of bacteriology*, *33*(5), 461.

- Karim, S., Zhang, Y., Yin, S., Bibi, I., & Brohi, A. A. (2020). A brief review and challenges of object detection in optical remote sensing imagery. *Multiagent and Grid Systems*, *16*(3), 227-243.

- Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review*, *53*(8), 5455-5516.

- Kulkarni, A., & Callan, J. (2015). Selective search: Efficient and effective search of large textual collections. *ACM Transactions on Information Systems (TOIS)*, *33*(4), 1-33.

- Leininger, D. J., Roberson, J. R., & Elvinger, F. (2001). Use of eosin methylene blue agar to differentiate Escherichia coli from other gram-negative mastitis pathogens. *Journal of veterinary diagnostic investigation*, *13*(3), 273-275.

- Li, Y., Li, W., Xiong, J., Xia, J., & Xie, Y. (2020). Comparison of Supervised and Unsupervised Deep Learning Methods for Medical Image Synthesis between Computed Tomography and Magnetic Resonance Images. *BioMed Research International*, *2020*.

- Maas, M. B., Perold, W. J., & Dicks, L. M. T. (2017). Biosensors for the detection of Escherichia coli. *Water Sa*, *43*(4), 707-721.

- MathWorks: Deep Network Designer Toolbox Release 2020b– MATLAB & Simulink – MathWorks India, available at, last access: 15th May 2021.

- Mohanty, S. P., Hughes, D. P., & Salathé, M. (2016). Using deep learning for image-based plant disease detection. *Frontiers in plant science*, *7*, 1419.

- Nehal, S. A., Roy, D., Devi, M., & Srinivas, T. (2019). Highly sensitive lab-on-chip with deep learning AI for detection of bacteria in water. *International Journal of Information Technology*, 1-7.

- Prabha, D. S., & Kumar, J. S. (2016). Performance evaluation of image segmentation using objective methods. *Indian J. Sci. Technol*, *9*(8), 1-8.

- Riley, L. W., Remis, R. S., Helgerson, S. D., McGee, H. B., Wells, J. G., Davis, B. R., ... & Blake, P. A. (1983). Hemorrhagic colitis associated with a rare *Escherichia coli* serotype. *New England Journal of Medicine*, *308*(12), 681-685.

- Rompré, A., Servais, P., Baudart, J., De-Roubin, M. R., & Laurent, P. (2002). Detection and enumeration of coliforms in drinking water: current methods and emerging approaches. *Journal of microbiological methods*, *49*(1), 31-54.

- Ryan, F. J. (2019). Application of machine learning techniques for creating urban microbial fingerprints. *Biology direct*, *14*(1), 1-13.

- Sarrafzadeh, O., Dehnavi, A. M., Rabbani, H., & Talebi, A. (2015, October). A simple and accurate method for white blood cells segmentation using K-means algorithm. In *2015 IEEE Workshop on Signal Processing Systems (SiPS)* (pp. 1-6). IEEE.

- Saxena, T., Kaushik, P., & Mohan, M. K. (2015). Prevalence of *E. coli* O157: H7 in water sources: an overview on associated diseases, outbreaks and detection methods. *Diagnostic microbiology and infectious disease, 82*(3), 249-264.

- Simonyan, Karen, and Andrew Zisserman. "Two-Stream Convolutional Networks for Action Recognition in Videos." arXiv preprint arXiv:1406.2199 (2014).

- Smith, A. C., & Hussey, M. A. (2005). Gram stain protocols. *American Society for Microbiology*, *1*, 14.

- Smith, S. T. (2006). *MATLAB: advanced GUI development*. Dog ear publishing.

- Studio, A. (2017). Android studio. *The Official IDE for Android*.

- Szegedy, C., Ioffe, S., Vanhoucke, V., & Alemi, A. A. (2017, February). Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence.

- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015, June). Going deeper with convolutions. In 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1-9). IEEE.

- Tamerat, N., Muktar, Y., & Shiferaw, D. (2016). Application of molecular diagnostic techniques for the detection of *E. coli* O157: H7: a review. *J Vet Sci Technol*, *7*(362), 1-9.

- Torelli, A., Wolf, I., & Gretz, N. (2018). AutoCellSeg: robust automatic colony forming unit (CFU)/cell analysis using adaptive image segmentation and easy-to-use post-editing techniques. *Scientific reports*, *8*(1), 1-10.

- Tripathi, N., & Sapra, A. (2020). Gram Staining. *StatPearls (Internet)*.

- Turra, G., Conti, N., & Signoroni, A. (2015, August). Hyperspectral image acquisition and analysis of cultured bacteria for the discrimination of urinary tract infections. In

2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC) (pp. 759-762). IEEE.

- United States Environmental Protection Agency (US EPA). (2010). EPA Microbiological Alternate Test Procedure (ATP) Protocol for Drinking Water, Ambient Water, Wastewater, and Sewage Sludge Monitoring Methods.

- US Environmental Protection Agency. (2002). Method 1103.1: Escherichia coli (*E. coli*) in water by membrane filtration using membrane-thermotolerant Escherichia coli agar (mTEC). *EPA 821-R-02-020*.

- US EPA. (2007). Drinking water standards and health advisories table. *Edition of the Drinking Water Standards and Health Advisories*.

- Viola, P., & Jones, M. J. (2004). Robust real-time face detection. *International journal of computer vision*, *57*(2), 137-154.

- WHO. Guidelines for Drinking-water Quality. 4th ed.; WHO (2011) Geneva, Switzerland.

- Xu, X., Li, H., Yin, F., Xi, L., Qiao, H., Ma, Z., ... & Ma, X. (2020). Wheat ear counting using K-means clustering segmentation and convolutional neural network. *Plant Methods*, *16*(1), 1-13.

- Zaitoun, N. M., & Aqel, M. J. (2015). Survey on image segmentation techniques. *Procedia Computer Science, 65*, 797-806.

- Zapata, B. C. (2013). *Android studio application development* (p. 110). Packt Publishing.

- Zhang, J., Wang, F. Y., Wang, K., Lin, W. H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1624-1639.

- Zhu, W., Huang, W., Lin, Z., Yang, Y., Huang, S., & Zhou, J. (2016). Data and feature mixed ensemble based extreme learning machine for medical object detection and segmentation. *Multimedia Tools and Applications*, *75*(5), 2815-2837.

- Zieliński, B., Plichta, A., Misztal, K., Spurek, P., Brzychczy-Włoch, M., & Ochońska, D. (2017). Deep learning approach to bacterial colony classification. *PloS one, 12*(9), e0184554.

7. A Novel PCA-FA-ANN based Hybrid Model for Prediction of Fluoride

Fluoride plays an essential role in terms of the health of human beings. Persistent exposure to fluoride, which is present in drinking water mainly, may result in dental, skeletal, and non-skeletal fluorosis. However, being consumed presently, drinking water is not sufficient to indicate the degree of exposure to fluoride. The existing literature indicates that nails can be used as indicative (biomarkers) of not only to exposure of fluoride but also the degree of the same. However, because of differential metabolism rate depending on a number of factors like age, gender, nutritional status, water characteristics, etc., exposure to fluoride is not easily detectable in human beings by just testing the fluoride content in nails. Moreover, due to sensitive chemical analysis and lack of facilities, it is difficult to identify the exact concentration of fluoride in nails. The objective of this study is to identify the significant parameters that affect the fluoride content in nail samples. Apart from laboratories test, the application of different Artificial Intelligence (AI) methods is used to predict fluoride in nails, which will help identify the degree of fluoride exposure to children, females, and males.

7.1 Introduction

Freshwater reserves comprise of fluoride in varying concentrations, from trace amounts to some mg/l and even toxic concentrations (Schmedt et al. 2012, Celinski et al. 2016, O'Donnel 1973, Álvarez et al. 2011, Wenzel et al. 1992). High levels of fluoride are generally found at the foot of high mountains and in geological regions with marine deposits (Koblar et al., 2011). Fluoride is known to have beneficial effects on dental health within permissible limits. On the other hand, extreme fluoride ingestion above the allowable limit can lead to detrimental effects, including the accumulation of dental fluorosis or skeletal fluorosis in both adults and children. The acceptable consumption has been set at 0.05 mg/day/kg weight based on experimental observations. The frequency and intensity of this clinical incidence can differ between persons and communities because of the effects of environmental and physiological influences, the volume of fluoride absorbed, and the duration of exposure (De Carvalho et al. 2011, Buzalaf et al. 2006, Khairnar et al. 2015, Ando et al. 1998).

Fluoride toxicity awareness remains relatively low (Ando et al. 1998). Millions of people around the world are affected by adverse health effects with exposure to a high concentration of naturally occurring fluoride in potable water supplies (Moseley et al. 2003, Yi et al., 2008). Thus, fluoride has been called one of the top ten public health concern chemicals (WHO, 2006). A recent study of the US National Research Council has reported a range of possible health issues linked to elevated exposure to fluoride, including disrupted biochemical and physiological processes, cardiovascular, reproductive, endocrine, gastrointestinal, neurological, and bone fractures (Beir 2005).

7.2 Biomarkers

In order to get relevant results in a large population, a fluoride exposure biomarker should be easily collectable without donor objections, and there should be an accurate, reliable, and legitimate fluoride estimation tool. Samples of the nails and hair can be used as biomarkers to monitor fluoride contamination. Nails have been proposed as appropriate biomarkers for fluoride intake (Pessan et al. 2011, Buzalaf et al., 2006). They can help to detect chronic and sub-chronic exposures to fluoride. The use of nails as fluoride markers is appealing, provided that the samples are easy to obtain (Fukushima et al., 2009), as nails can be collected non-invasively. The user-friendly methodology for assessing nail fluoride and its fast use in an essential laboratory condition exhibits strong ability as a biomarker for epidemiological surveys. The fluoride concentration in nails reflects the total concentration of fluoride absorption and plasma during the processing of nail samples. The concentration of fluoride in the nail samples is thus directly correlated to the average fluoride consumption that happened around three months ago (Whitford et al., 2005).

Recent studies examined the concentration of fluoride in water by the usage of urine (Buzalaf et al. 2012, Antonijevic et al. 2016, Akpata et al. 2014) and nail samples (Buzalaf et al. 2012, Lima-Arsati et al. 2010, Amaral et al. 2014, Linhares et al. 2016, Sousa et al. 2018). Still, no study has been done to examine the correlation between physico-chemical water quality parameters and fluoride concentration in nail samples. Fluoride fingernail analysis has been widely used to determine low-level concentrations in water fluoridation, toothpaste, salt, and milk (Whitford et al. 1999, Buzalaf et al., 2012, Lima-Arsati et al. 2010, De Almeida et al. 2007, Buzalaf et al., 2009, Pessan et al. 2005, Levy et al., 2004). Fukushima et al. 2009 have used nails for investigating the correlation between fluoride exposure biomarkers and total daily intake of fluoride with significant fluoride exposure in drinking water. They studied the impact of age, gender, nail growth rate, and geographic area on the absorption of fluorides in the fingernails and toenails (Elekdag-Turk et al., 2019). They obtained drinking water and nail samples and used an ion-selective electrode to examine fluoride concentration. A comparison mark was created on each nail, and growth levels were calculated. The analysis was done by ANOVA and linear regression. All the factors they considered were directly associated with the fluoride concentration in nail samples. The study recommended that nails should be used as biomarkers of fluoride contamination, with the advantage of being easily obtained. But they do not consider water characteristics. At present, none of the studies on nails as biomarkers of fluoride exposure have examined the impact of age, gender, and factors affecting the bioavailability of fluoride (Clarkson et al., 2000).

There is a need to study the effect of age, weight, gender, water fluoride, nitrate, turbidity, dissolved oxygen, electrical conductivity, and pH levels on fluoride concentration in nail samples since water characteristics might also impact fluoride.

7.3 Methodology

This study focused on costs and remediation of groundwater contamination in India, with particular reference to Rajasthan. 2401 groundwater and fingernail samples were collected from 348 villages and cities in pre and post-monsoon seasons during 2016-2019. 1024 water samples were also collected from the same households from where the nails samples were collected. These water samples were tested for various physical, chemical, and microbiological water quality parameters in laboratories at Birla Institute of Technology and Science, Pilani, India. These parameters are as follows; pH, dissolved oxygen (DO, mg/l), electrical conductivity (EC, s/m), turbidity (NTU), fluoride (mg/l), and nitrate (mg/l) were tested using the titration and spectroscopy method. Nail samples were collected from the various villages, and the data contains the individual weight, height, age, and gender of the family members. The concentration of fluoride in nail samples was measured using an ion-selective electrode.

The objective of this study is to identify the significant water quality parameters and other factors that affect the fluoride content in nail samples. Apart from laboratory tests, different Artificial Intelligence (AI) methods were used to predict fluoride in nails, which will help identify the degree of fluoride exposure to children, females, and males. As a point of reference, we show the relationship between fluoride in drinking water and in nails that we collected and tested in Figure 7.1. We observe no correlation among these two variables though both had elevated levels of fluoride in them.

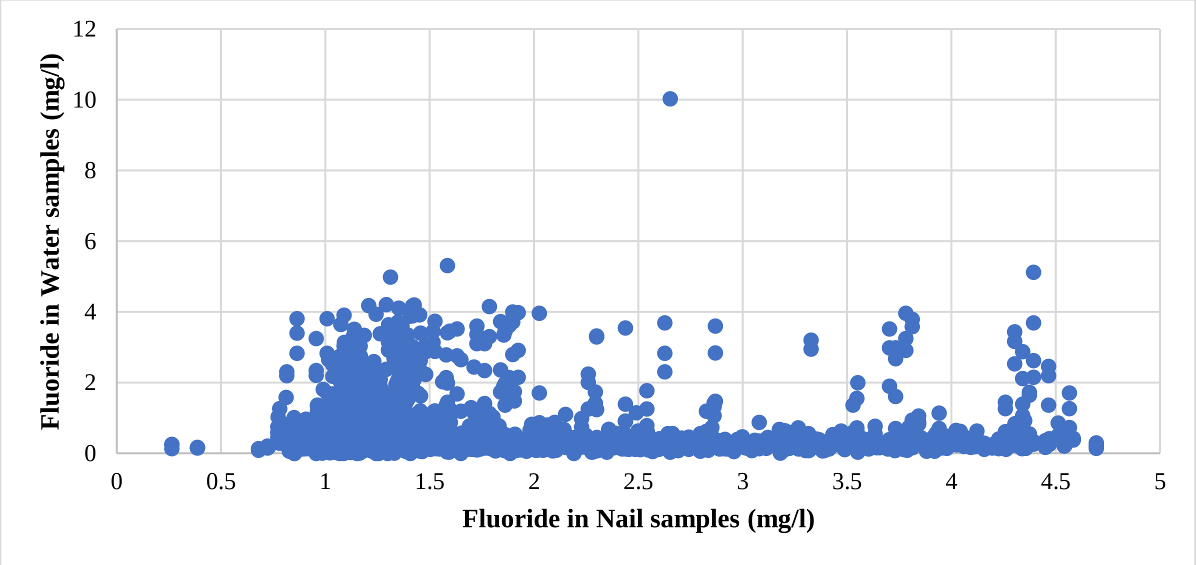


Figure 7.1: Relationship between fluoride in drinking water and nails.

If fluoride monitoring and mitigation policy were based solely on biomarker fluoride concentration such as nails, we would miss the community exposure. Nails fluoride is potentially affected by age, gender, and chemical properties of water. In order to shed light on this relationship, we developed a machine learning model that can help make informed policy choices. This model uses water fluoride levels in conjunction with nails to establish exposure. To this end, it predicts the fluoride in nails using water fluoride, individual, and water characteristics as inputs.

The advantage of the machine learning models is that they provide better prediction accuracy than other mathematical models, including those combined with the PCA algorithm (Reddy et al., 2014). Models do not produce reliable results when both linear and non-linear variables are present in the dataset (Pao 2006). Hence, we rely on hybrid methods to circumvent this issue. We also establish the best architecture of the hybrid model for the prediction of fluoride in nail samples.

In this study, we have proposed a hybrid model (HM) that combines Principal Component Analysis (PCA), Firefly Algorithm (FA), and Artificial Neural Network (ANN) to predict fluoride concentration in nails. In this method, PCA was used for dimensionality reduction of the dataset. FA was applied to optimize the weights and bias between the input and hidden layers of ANN models. The performance of the model was evaluated using five evaluation measures such as MSE, RMSE, MAE, MAPE, and R2. The prediction includes the development of 3 models:

1. Model 1: ANN with all original input parameters.
2. Model 2: ANN with principal components (PC) as input variables, obtained by the PCA algorithm.
3. Model 3: ANN with reduced dataset as input variables, obtained by the PCA-FA algorithm.

7.4 Principal Component Analysis (PCA)

The principal component analysis is a multivariate method used to reduce the dimension of input variables when we have a vast amount of observations and an improved understanding of variables (Lu et al., 2003). The PCA algorithm helps to reduce the dimension of the data into limited numbers of variables for data interpretation and then create basic plots to display essential statistics, including score plot and loading plot, to study the correlation between the broad clustered data set (Stojanovic et al. 2012, Beltran et al. 2006). Such associated variables are known as principal components (Shinde et al., 2009). PCA has its mathematical algorithm in linear algebra, which describes the association between the data containing the variables as columns and the observations or samples as rows. The fundamental purpose is to create a transformed matrix using coefficients of principal components that includes the maximum amount of information and then plot the data using a 2-dimensional plot in MATLAB software (Bell et al. 1997).

After laboratory testing, 2401 experimental observations were used to predict the concentration of fluoride in fingernail samples. The effect of age, weight, height, gender and different water quality parameters like pH, dissolved oxygen, electrical conductivity, turbidity fluoride, and nitrate were taken as input parameters to predict the effects on the concentration of fluorides in the fingernails and toenails. The fundamental purpose of the analysis was to define the interdependence of a large number of variables with a smaller number of simple variables with computing the commonality, then pre-process the factors according to the PCA and extract the principal components to minimize the measurements of the datasets. After normalizing the data, the principal components of the final matrix were calculated when the cumulative variance contribution was more than 85%.

The relationship between Eigenvalues and Eigenvector numbers is plotted in Figure 7.2, which gives the scree plot of PCA.

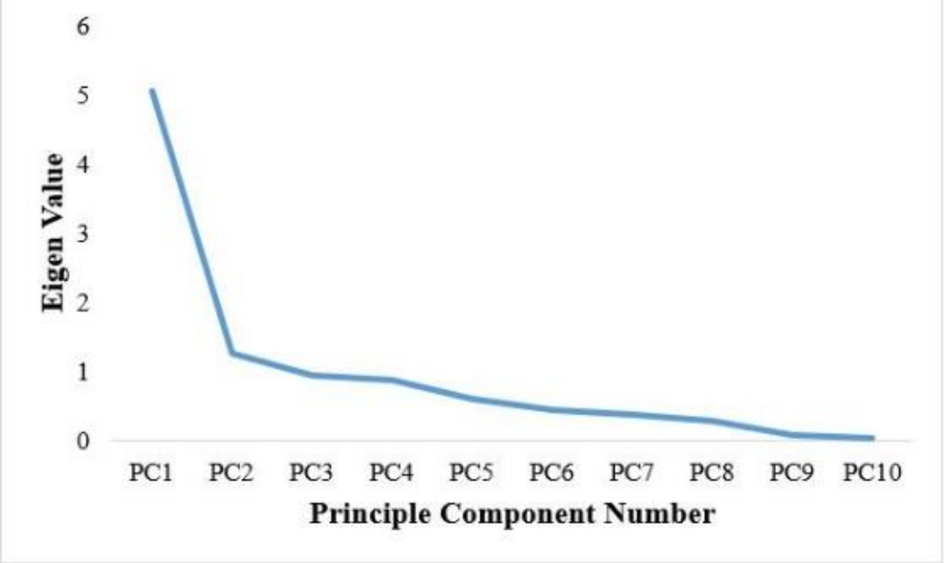


Figure 7.2: Scree plot of the Eigenvalues.

The three coefficients of principal component scores derived from the principal component analysis were used as input for prediction, which reduced the calculation dimension of the model and improved the operation efficiency. When three principal components were taken, the trend began to stabilize. The principal components PC1, PC2, and PC3 were arranged according to their amount of variance in the decreasing order. Also, for the top three principal components, their total cumulative contribution rate had reached 72.64%, as illustrated in Figure 7.3. Hence, we have selected the top three principal components for analysis, as shown in Table 7.1.

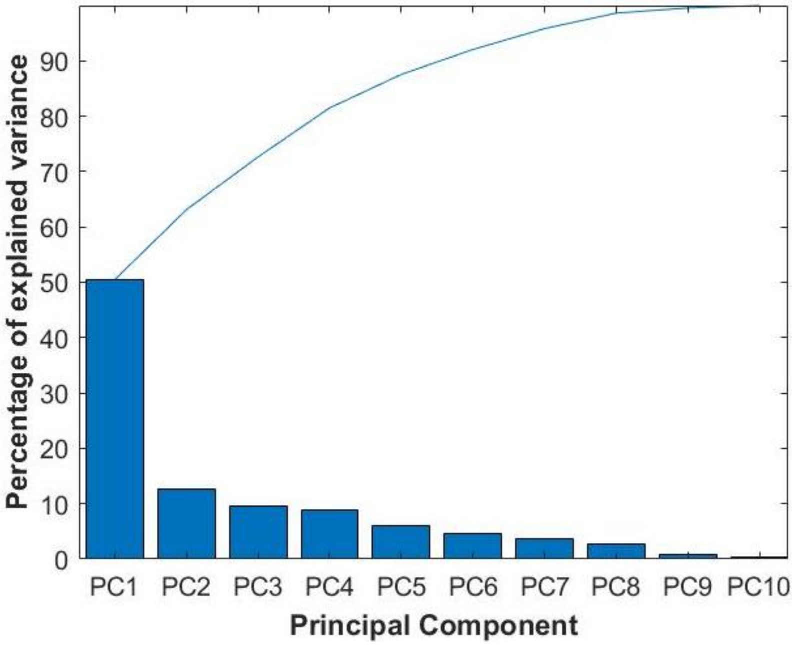


Figure 7.3: The bars display the variance accounted by the first ten principal components.

Table 7.1: The Coefficient of Principal Component Score of Variables

Variable	PC1	PC2	PC3
Weight	0.3660	0.3858	0.2296
Height	0.4230	0.1926	0.0618
Gender	0.1999	-0.2094	-0.3554
Age	0.3494	0.3920	0.2578
Fluoride	0.3218	-0.2819	0.0723
Nitrates	0.2354	-0.5249	0.1909
Turbidity	0.1562	0.1550	-0.8114
Dissolved Oxygen	0.3541	0.1197	-0.2237
Electrical Conductivity	0.2874	-0.4612	0.0012
pH	0.3657	-0.1026	0.0224

Principal components expressions are given in equation’s (7.2-7.4):

$$PC_1 = 0.3660X_1 + 0.4230X_2 + 0.1999X_3 + 0.3494X_3 + 0.3218X_5 + 0.2354X_6 + 0.1562X_7 + 0.3541X_8 + 0.2875X_9 + 0.3657X_{10} \tag{7.2}$$

$$PC_2 = 0.3858X_1 + 0.1926X_2 - 0.2094X_3 + 0.3920X_3 - 0.2819X_5 - 0.5249X_6 + 0.1550X_7 + 0.1197X_8 - 0.4612X_9 - 0.1026X_{10} \tag{7.3}$$

$$PC_3 = 0.2296X_1 + 0.0618X_2 - 0.3554X_3 + 0.2578X_3 - 0.0723 - 0.1909X_6 - 0.8114X_7 - 0.2237X_8 + 0.0012X_9 + 0.0224X_{10} \quad (7.4)$$

The description of the principal components:

- Height (0.423), Weight (0.366), pH (0.3657), Dissolved oxygen (0.3541), Age (0.3494), and Fluoride (0.3218) have significant positive loadings on principal component 1 (PC1).
- Nitrates (-0.5249) and Electrical Conductivity (-0.4612) have significant negative loadings on principal component 2 (PC2).
- Turbidity (-0.8114) and Gender (-0.3554) have significant negative loadings on principal component 3 (PC3).

7.5 Artificial Neural Networks (ANN)

An artificial neural network is a machine-learning algorithm commonly used in multiple problem domains for classification, prediction, and correlation (Bell et al. 1997). ANNs have the potential to estimate any non-linear mathematical function, which is most useful when the correlation between variables is uncertain or complex (Paliwal et al., 2009). ANNs are widely used to extract hidden patterns from complex data (Masters et al. 1995, Haykin et al. 2007). McCulloch et al. 1943 study was inspired by neuronal activity, introduced the idea of considering neural networks as computing machines. Mathematically, Hornik et al. 1989 proved that a multilayer neural network with finite hidden layers and enough hidden neurons is a universal approximator for any Borel measurable function from one finite-dimensional space to another. Several complicated multilayer neural network models have been proposed and used in different fields. Applications of ANN in the groundwater, ecology, and environmental engineering fields were documented in the early 1990s. However, in recent years ANN has been intensively used for prediction and forecasting in a variety of engineering and water-related areas, including water resource analysis by Liong et al. 1999, 2001, Muttil and Chau 2006, El-Shafie et al. 2008, El-Shafie et al. 2011, Najah et al. 2009, oceanography by Makarynskyy, 2004, and environmental engineering by Grubert, 2003.

The principal components extracted by the principal component analysis were used as input datasets of the prediction model. A feed-forward backpropagation neural network was developed, and it used the principal components as inputs to predict the concentration of fluoride in nail samples. The experimental results were compared with the PCA-ANN model results, which use as regressors the original variables.

i. Splitting the datasets into subsets:

Initially, a total of 2401 data was divided into training (70%), testing (15%), and validation (15%).

ii. Training functions

In this analysis, the Bayesian regularization backpropagation (Hayati et al. 2007, MacKay et al. 1992) training function was used to perform training and validation steps.

iii. Adaption learning functions

Gradient descent with momentum weight and bias learning function was used to explain this pattern of neural network input-output relationship and architecture.

iv. Activation functions

A sigmoid activation function was used (Turian et al. 2009).

v. Performance functions

The model performance was evaluated using the value of the coefficient of determination (R2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE).

A feed-forward neural network with multiple hidden layers was optimized using the Bayesian regularization backpropagation (BR) training function, and the sigmoid activation function had outperformed all other combinations. Figure 7.4 illustrated the effect of the number of neurons on the value of R^2 for one, two, and three hidden layers. The model with two hidden layers and 40 neurons shows the best performance with the overall highest R^2 values of 0.85 and lowest errors with MSE values of 0.1, RMSE values of 0.1, MAE values of 0.08, and MAPE values of 0.05, as illustrated in Figure 7.5. Increasing the number of hidden layers led to improved performance; however, this method requires more computational time and does not substantially alter model accuracy. In this study, the maximum number of neurons that could be considered in the ANN model was set at 50 due to the extreme computational time required for the improvement to model accuracy. The response plot was used to visualize the correlation between input and output variables. Figure 7.6(a) shows the response plot of the ANN model, and Figure 7.6(b) shows the plot between predicted output and actual output; a perfect regression model has predicted output equal to actual output with a regression value of 0.85.

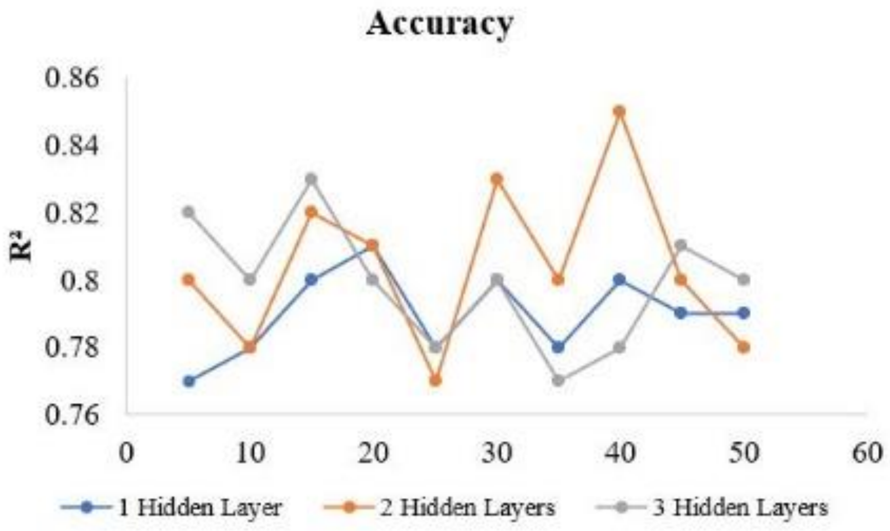


Figure 7.4: Accuracy comparison of ANN models.

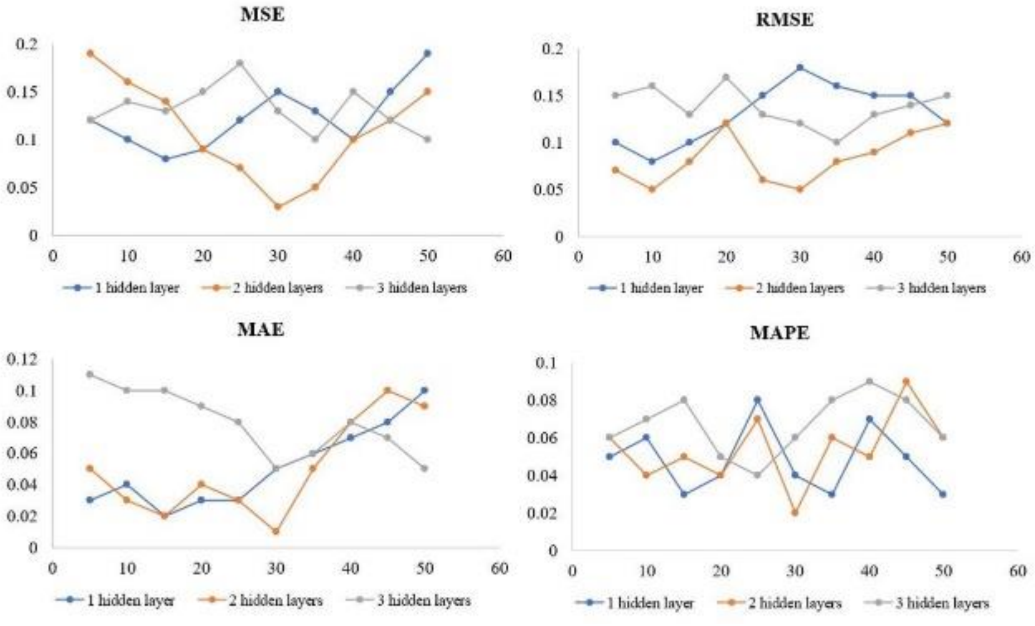
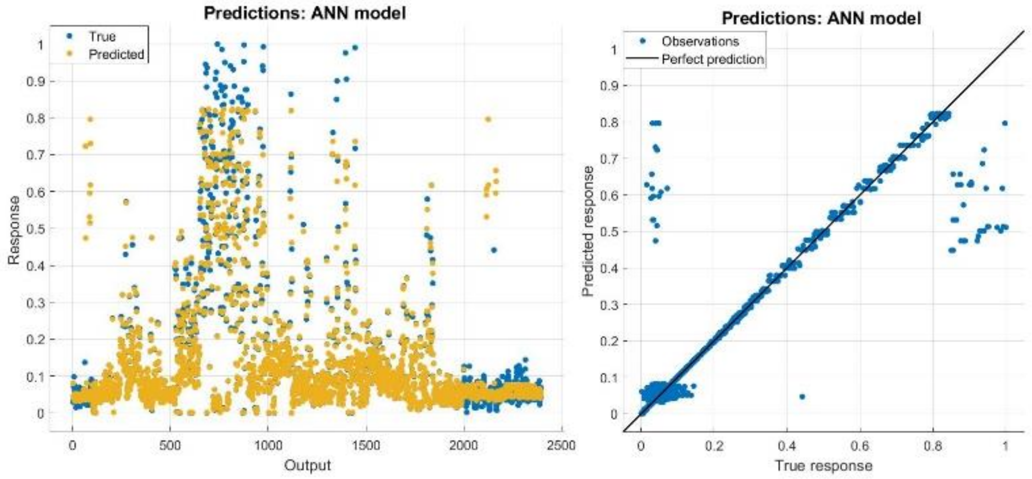


Figure 7.5: Error comparison of ANN models.



a) Response plot b) Actual vs. predicted value

Figure 7.6: a) Response plot of ANN, b) Actual vs. predicted value using ANN.

7.6 PCA-ANN model

The data was analysed using the PCA algorithm, and based upon the principal component scores, feed-forward backpropagation artificial neural network was developed to predict the concentration of fluoride in fingernail samples. The data were divided into training (70%), testing (15%), and validation (15%) and used for fitting and simulation of the model. We have tested the model in the neural network toolbox in MATLAB R2019b. With the selected network type, the input and target data were fixed. With this, we had used fourteen different training functions, two adaptation learning functions, five performance functions, and three transfer functions. Out of all these networks, only one network had completed the process of generating regression plots. The ANN model was optimized by the Bayesian regularization backpropagation (BR) using the sigmoid activation function that has outperformed all other combinations. Figure 7.7 shows the effect of the number of neurons on the value of R^2 for one, two, and three hidden layers. The model with two hidden layers and 30 neurons shows the best performance with the overall highest R^2 values of 0.90 and lowest errors with MSE values of 0.05, RMSE values of 0.08, MAE values of 0.12, and MAPE values of 0.06, as illustrated in Figure 7.8. Figure 7.9(a) shows the response plot of the PCA-ANN model. Figure 7.9(b) shows the plot between predicted output and actual output; a perfect regression model has predicted output equal to accurate output with a regression value of 0.90.

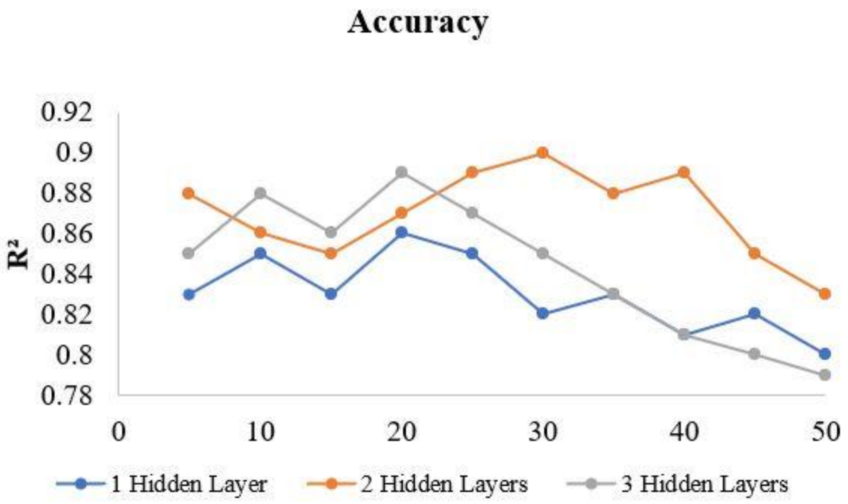


Figure 7.7: Accuracy comparison of PCA-ANN models.

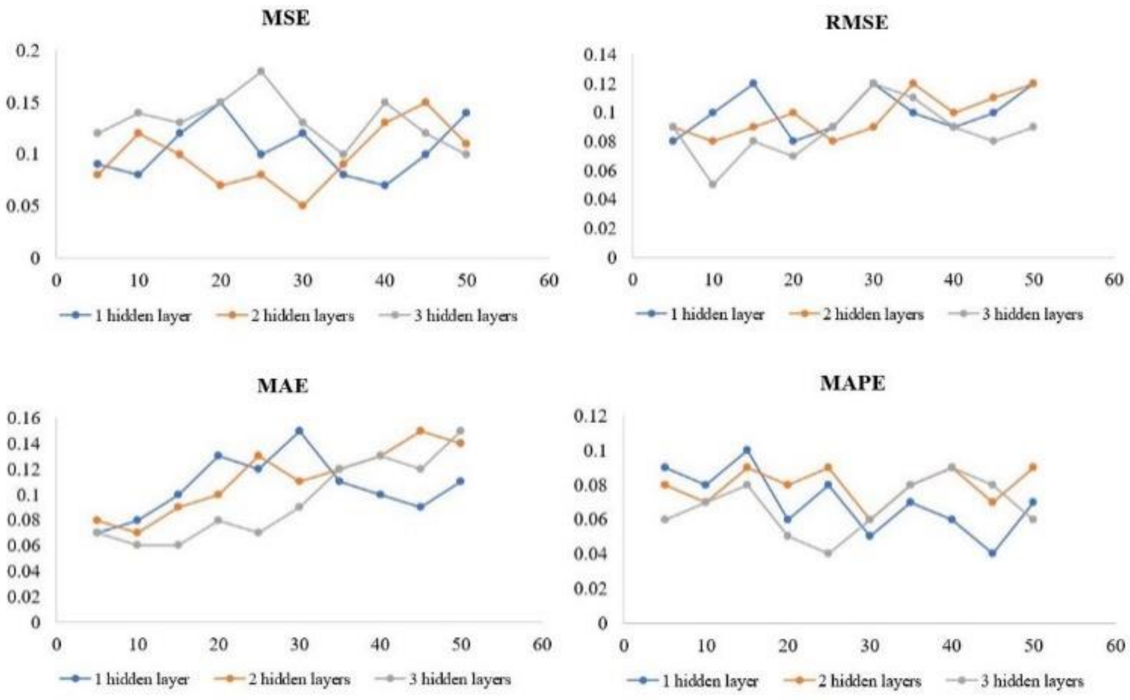
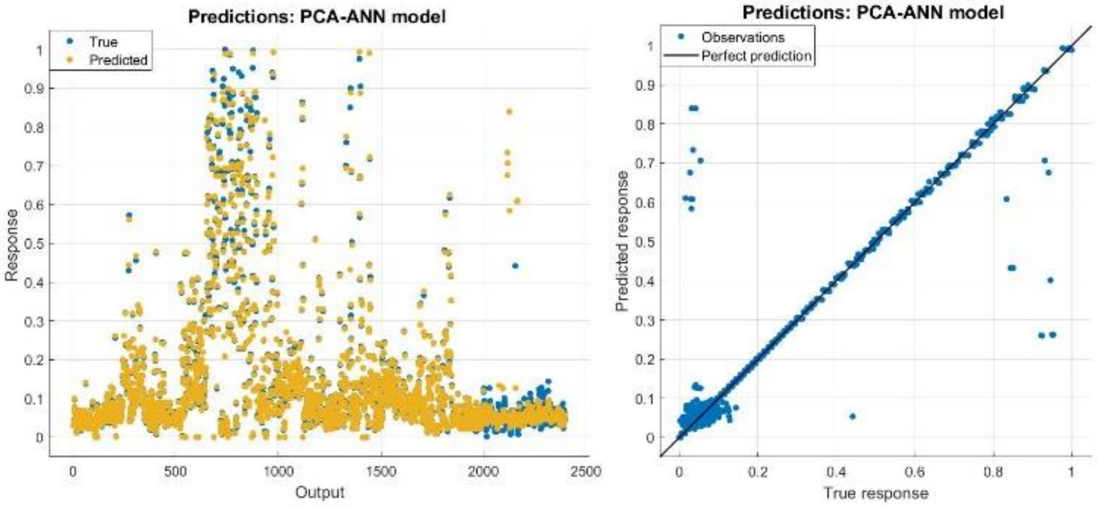


Figure 7.8: Error comparison of PCA-ANN models.



a) Response plot b) Actual vs. predicted value

Figure 7.9: a) Response plot of PCA-ANN, b) Actual vs. predicted value using PCA-ANN.

7.7 Firefly Algorithm (FA)

The firefly algorithm is used in this study to improve the performance of machine learning models by optimizing the weights and bias between the input layer and the hidden layer of the ANN model. Firefly algorithm is one of the swarm intelligent algorithms developed by Yang. It is a metaheuristic algorithm that is inspired by nature and, based on the flashing behavior of fireflies, used to solve complex problems and non-linear optimization problems (Moazenzadeh et al. 2018). The brightness of the fireflies is the main criterion for the optimization of the fitness function (Gandomi et al. 2011, Yang et al., 2011). Yang developed the algorithm based on the following assumptions:

- The attraction of firefly is independent of gender due to unisexuality, and it is directly proportional to the brightness of the emitted light, but it is indirectly proportional to the distance between the fireflies (x_i, x_j). The firefly can move in any direction if the brightness of the neighboring firefly is same.
- The brightness of the light is associated with the optimization of objective function f(x) in the algorithm.

The principal components (PC) extracted by principal component analysis (PCA) were used as input datasets of the hybrid PCA-FA-ANN hybrid model. The firefly algorithm (FA) was applied to select the best attributes from the reduced dataset for optimization of the weights and bias between layers of the ANN model. The same phenomenon was used for dimensionality reduction. The ANN models were developed by training and testing of the dataset. The datasets were divided into subsets using the Firefly algorithm, and each subset was grouped at a single node. The algorithm corrects the errors generated by the ANN model to achieve optimized results for prediction. The model performance was evaluated using the value of the coefficient of determination (R^2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). The methodology presented in Figure 7.10 indicates the application of the hybrid model (HM) to predict fluoride in nail samples.

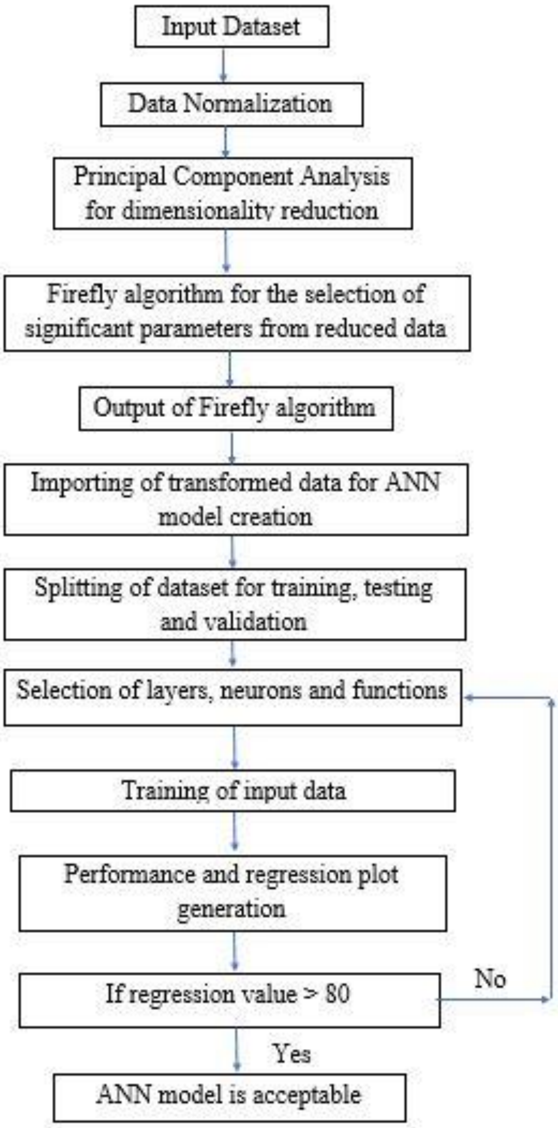

Input Dataset
Data Normalization
Principal Component Analysis for dimensionality reduction
Firefly algorithm for the selection of significant parameters from reduced data
Output of Firefly algorithm
Importing of transformed data for ANN model creation
Splitting of dataset for training, testing and validation
Selection of layers, neurons and functions
Training of input data
Performance and regression plot generation
If regression value > 80
No
Yes
ANN model is acceptable

Figure 7.10: Methodology to apply PCA-FA-ANN algorithm for prediction.

The dataset used for prediction was enormous. Hence, the proposed model was developed in a personal computer with 16 GB RAM using MATLAB 2019b software. The PCA algorithm was applied to the dataset for dimensionality reduction. Further analysis was done on the reduced dataset to reduce the randomness using the proposed firefly algorithm. The optimized dataset was used as input of the ANN model for prediction. The principal components (PC) were fed into the firefly algorithm to generate a random number of solutions (N). The fitness value can be calculated using equation (7.55). The updated solution based on fitness value can be calculated using equation (7.6).

$$F = PC_{PCA} + O_F + C \qquad (7.5)$$

Where,

F = Fitness value used in the proposed algorithm.

PC_{PCA}= Principal components obtained by applying PCA algorithm.

O_F= Objective function used for accuracy evaluation.

C = Constant ranging between 0 and 1.

$$FA_i^{t+1} = FA_i^t - \lambda_o^{xt} e^{-\alpha C_a^2}\left(FA_j^t + FA_i^t\right) + \xi t \psi_i^t \qquad (7.6)$$

Where,

FA_i^t = The real i^{th} solution.

FA_i^{t+1} = The updated i^{th} solution.

FA_j^t = The real j^{th} solution of the brighter firefly.

ξt = Randomized parameter.

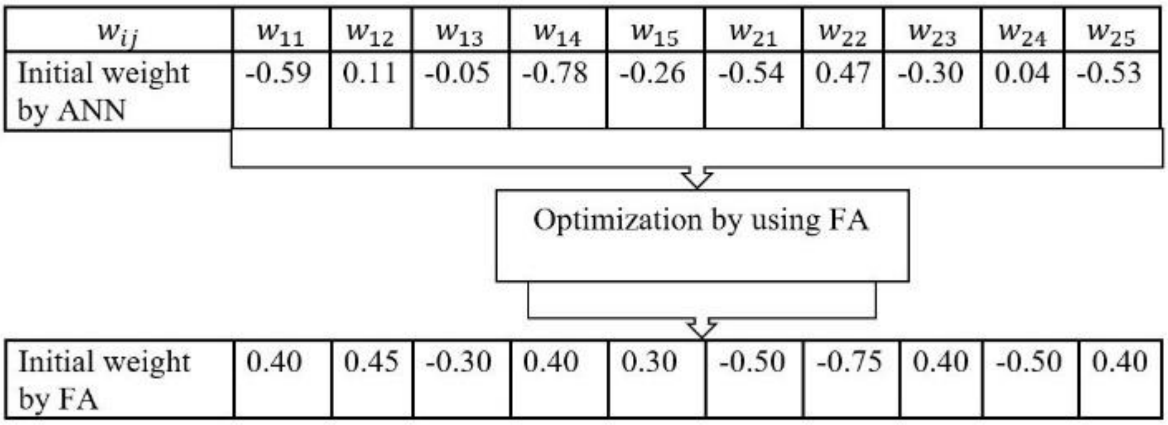
ψ_i^t = Random number generated from Gaussian distribution.

t = Time interval.

λ_o^{xt}= Factor of size scaling.

α = Coefficient of light absorption

In the firefly algorithm, the fitness values were generated using the above equation for each, and the significant parameters were selected using the fitness function. The initial weights were randomly created by the ANN model, and the input data values were multiplied by the suitable weights (w_{ij}) to get output. In the PCA-FA-ANN model, the initial weights were obtained using the firefly algorithm to get optimized weights and bias using the minimal fitness value, as illustrated in Figure 7.11.



w_{ij}	w_{11}	w_{12}	w_{13}	w_{14}	w_{15}	w_{21}	w_{22}	w_{23}	w_{24}	w_{25}
Initial weight by ANN	-0.59	0.11	-0.05	-0.78	-0.26	-0.54	0.47	-0.30	0.04	-0.53

Optimization by using FA

Initial weight by FA	0.40	0.45	-0.30	0.40	0.30	-0.50	-0.75	0.40	-0.50	0.40

Figure 7.11: Description of initial weights.

The population size was 2401, and the maximum number of iterations was 1000. The FA parameters were set as; the factor of size scaling, the randomized parameter, and the coefficient of light absorption was taken as 0.2, 0.9, and 0.9, respectively. The reduced dataset was trained and tested using ANN. A total of 2401 data was divided into training (70%), testing (15%), and validation (15%). The model performance was evaluated using the value of the coefficient of determination (R^2), mean square error (MSE), root mean square error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE). Trial and error procedures were adopted to obtain the optimum structure of the network in which a rigorous analysis is carried out with one, two, and three hidden layers. The ANN model was optimized using the Bayesian regularization backpropagation (BR), and the sigmoid activation function has outperformed all other combinations. Figure 7.12 shows the effect of the number of neurons on the value of R^2 for one, two, and three hidden layers. The model with three hidden layers and 20 neurons shows the best performance with the overall highest R^2 values of 0.94 and lowest errors with MSE values of 0.07, RMSE values of 0.05, MAE values of 0.04, and MAPE values of 0.03, as illustrated in Figure 7.13. Figure 7.14(a) shows the response plot of the PCA-ANN model. Figure 7.14(b) shows the plot between predicted output and actual output; a perfect regression model has predicted output equal to actual output with a regression value of 0.94.

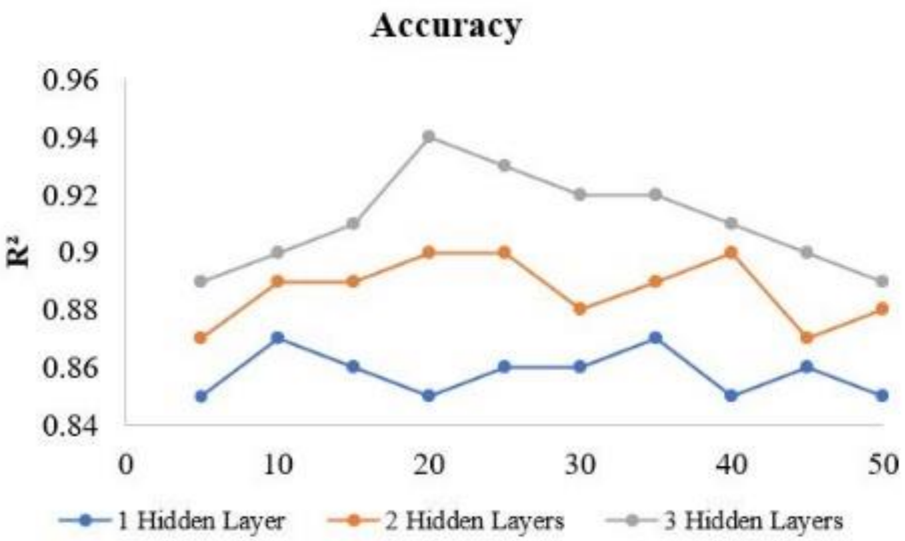


Figure 7.12: Accuracy comparison of PCA-FA-ANN models.

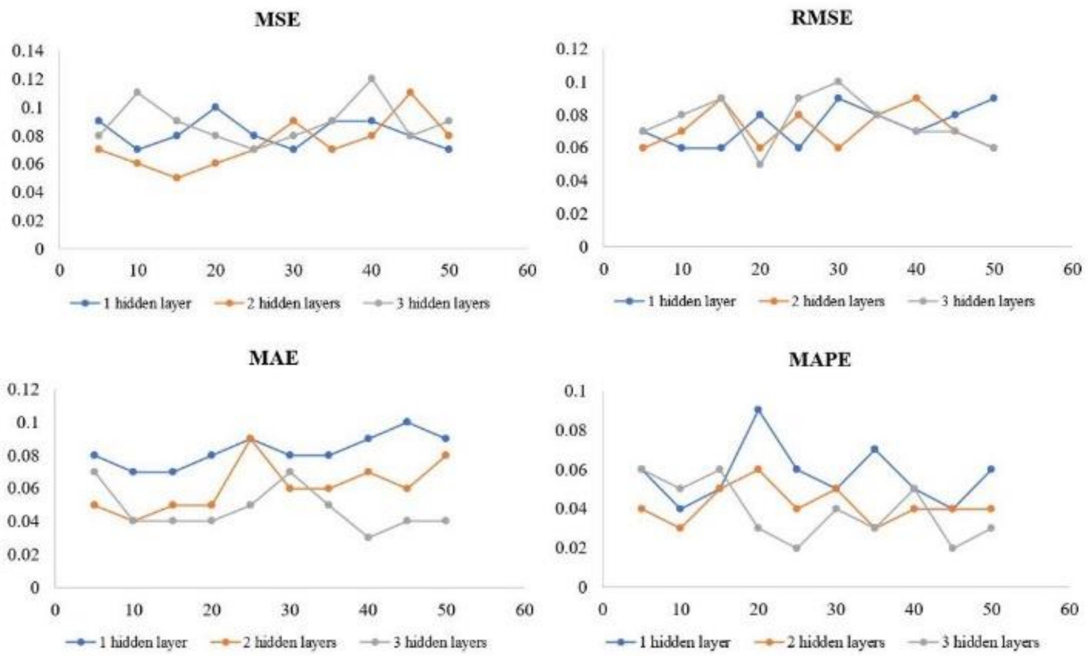
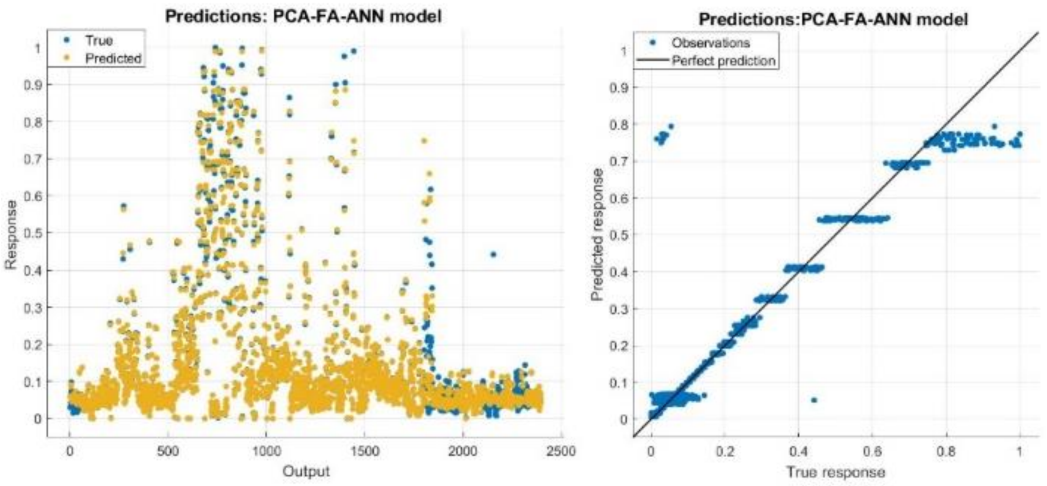


Figure 7.13: Error comparison of PCA-FA-ANN models.



a) Response plot b) Actual vs. predicted value

Figure 7.14: a) Response plot of PCA-FA-ANN, b) Actual vs. predicted value using PCA-FA-ANN.

7.8 Comparison of models

The comparative analysis of predictive models (ANN, PCA-ANN, PCA-FA-ANN) is illustrated in Figure 7.15 and Figure 7.16. The PCA-FA-ANN algorithm outperforms all the other combinations with the lowest MSE values of 0.07, RMSE values of 0.05, MAE values of 0.04, MAPE values of 0.03, and overall highest R^2 of 0.94. The results from the comparative analysis show that the predictions generated from the application of the PCA-FA-ANN algorithm have higher accuracy in decision making and hence can be relied upon as a constructive method in machine learning.

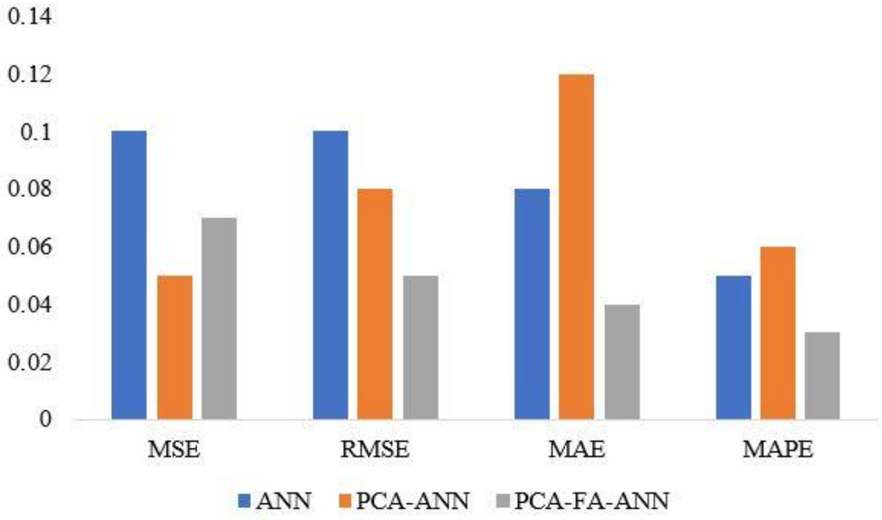


Figure 7.15: Error comparison of ANN, PCA-ANN, and PCA-FA-ANN.

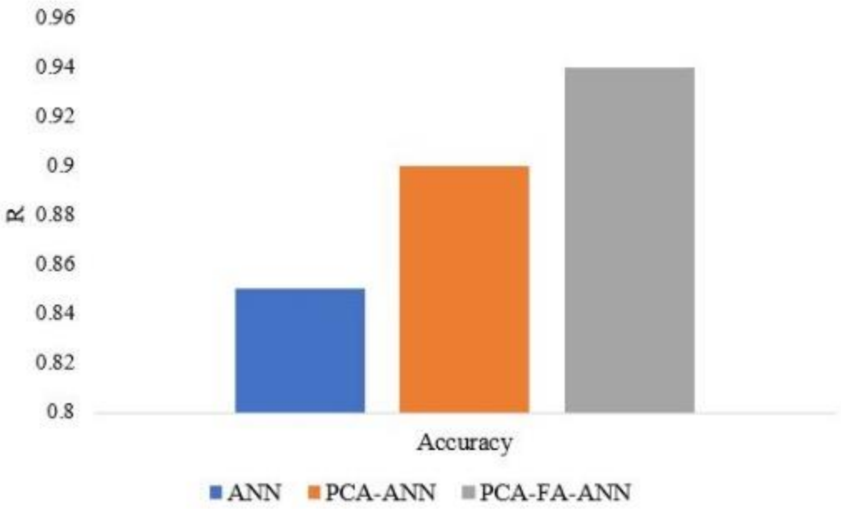


Figure 7.16: Accuracy comparison of ANN, PCA-ANN, and PCA-FA-ANN.

7.9 Sensitivity analysis

A sensitivity analysis was performed to study the effect on the output parameter when the input parameters were taken as average values. The input parameters were subjected to variability in a range of -10% to +10% of the average measured values. Each of the model input parameters was tested one at a time by keeping the others at their average values. Furthermore, the relative significance of these input parameters was ranked on the basis of the sensitivity index. The first model was developed using all ten principal components as input parameters and named AP, which serves as a reference model. In order to evaluate the significance of all input parameters for the AP model, a sensitivity analysis was performed, and the performance of the output model was evaluated using coefficient correlation (R^2) and mean squared error (MSE).

The second model was developed, referred to as AWFpE, used age, weight, water fluoride, pH, and Electrical conductivity as input parameters in predicting the fluoride concentration in fingernail samples. The third model (AFpE) used age, water fluoride, pH, and Electrical conductivity as input parameters. The fourth model AFp, used age, water fluoride, and pH as input parameters. The fifth model, named F, used water fluoride as an input parameter. The sensitivity analysis-based models were compared using ANN, PCA-ANN, and PCA-FA-ANN algorithms, and the model performance was assessed using evaluation measures such as MSE and R^2. The dataset with 2401 experimental values was used for the analysis, and models were developed using MATLAB R2019b software. The sensitivity index was calculated by equation (7.1):

$$SI = \left(\frac{\acute{Y}_i}{Y} - 1\right) \times 100 \qquad (7.1)$$

Where,

SI = sensitivity index.

$\acute{Y}_i$ = predicted output value when input value varied.

Y = average output value.

In order to evaluate the significance of all input parameters (AP), a sensitivity analysis was carried out, and the output of the model was assessed using correlation coefficient (R^2) and Mean Squared Error (MSE) values. Models were developed using ANN, PCA-ANN, and PCA-FA-ANN algorithms and compared for predictive analysis. The PCA-FA-ANN-AWFpE model with five input parameters with 2 hidden layers and 20 neurons in each layer had outperformed all other combinations with the overall highest R^2 of 0.95 and the lowest MSE of 0.002, as illustrated in Figure 7.17 and Figure 7.18, respectively. Figure 7.19 shows the comparative response plot of models using ANN, PCA-ANN, and PCA-FA-ANN algorithm, and Figure 7.20 shows the plot between predicted output and actual output.

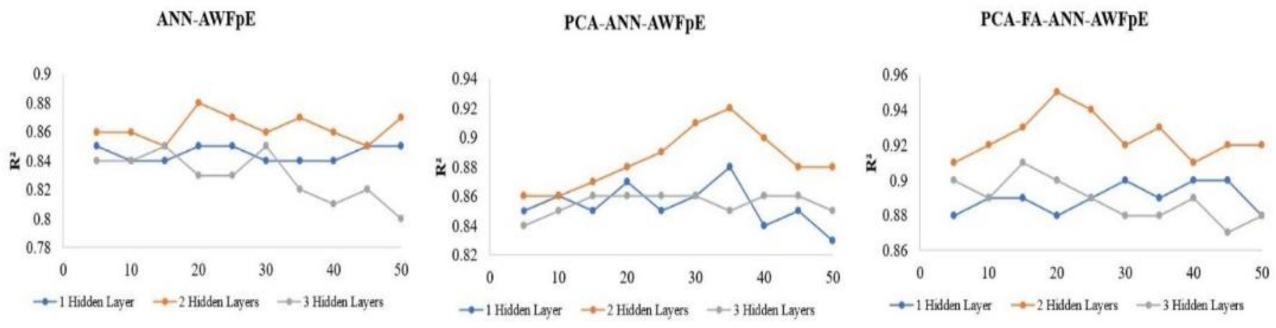


Figure 7.17: Accuracy comparison of AWFpE models.

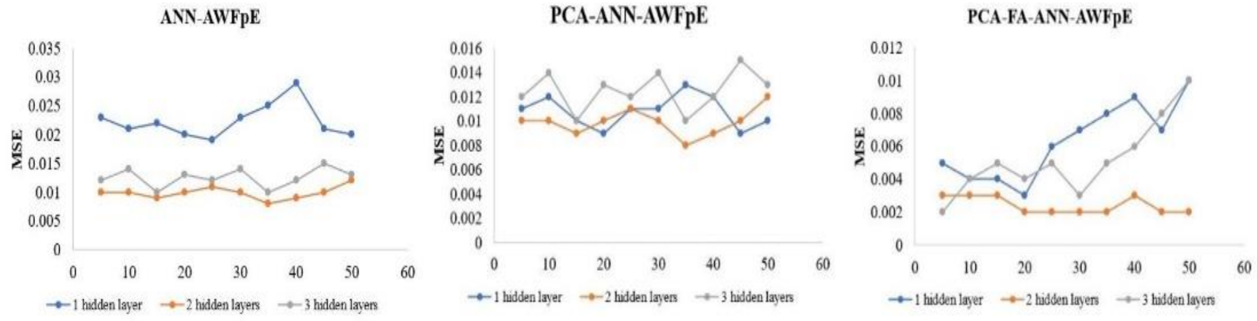


Figure 7.18: Error comparison of AWFpE models.

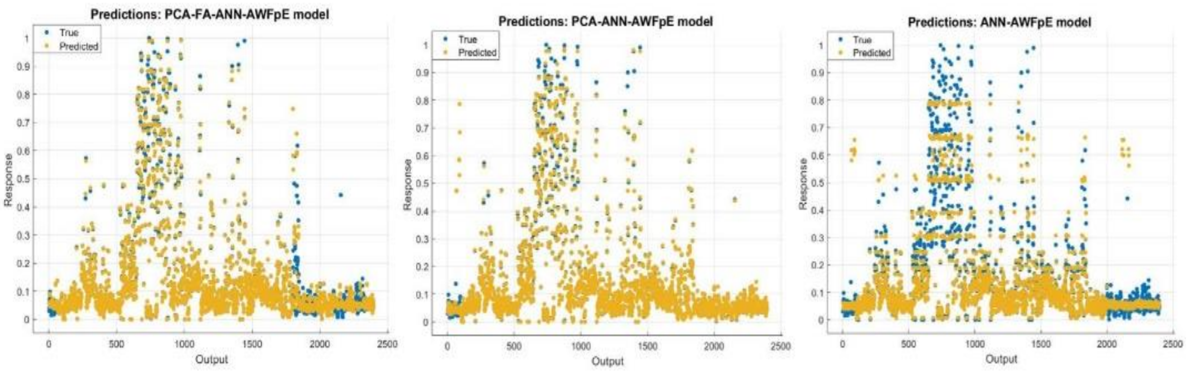


Figure 7.19: Response plot of AWFpE models.

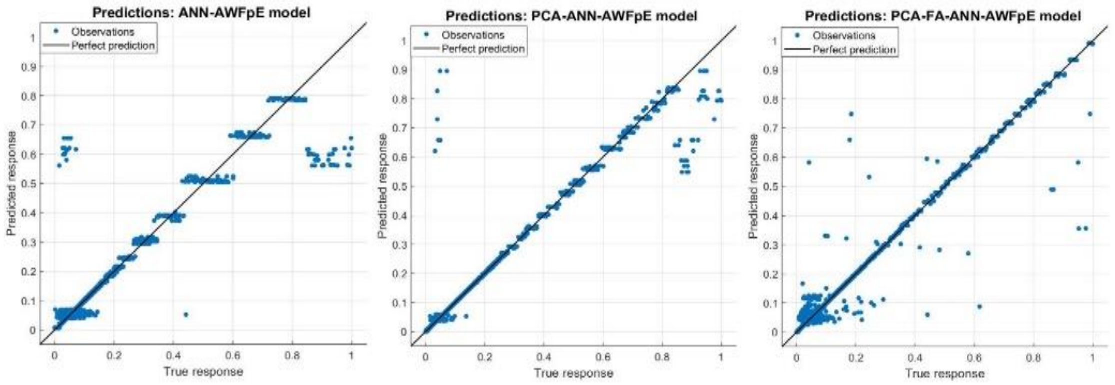


Figure 7.20: Actual vs. predicted value of AWFpE models.

The AFpE model having four input parameters was developed using ANN, PCA-ANN, and PCA-FA-ANN algorithm. The PCA-FA-ANN-AFpE model with 3 hidden layers and 15 neurons in each layer had outperformed all other combinations with the overall highest R^2 of 0.97 and lowest MSE of 0.008, as illustrated in Figure 7.21 and Figure 7.22, respectively. Figure 7.23 shows the comparative response plot of models using ANN, PCA-ANN, and PCA-FA-ANN algorithm, and Figure 7.24 shows the plot between predicted output and actual output.

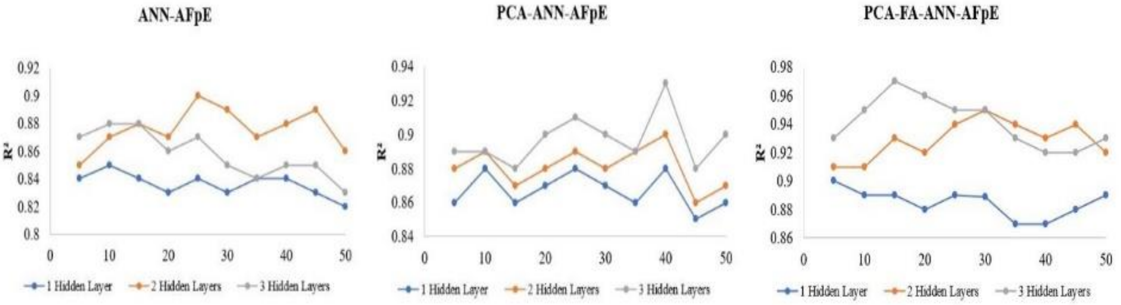


Figure 7.21: Accuracy comparison of AFpE models.

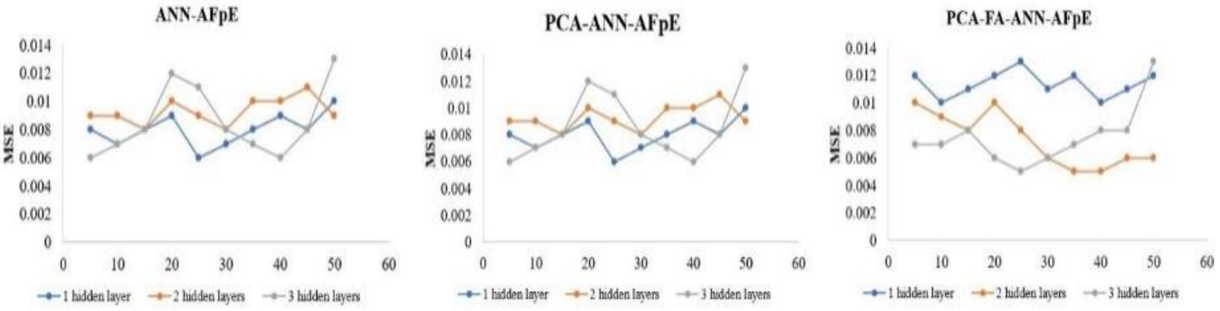


Figure 7.22: Error comparison of AFpE models.

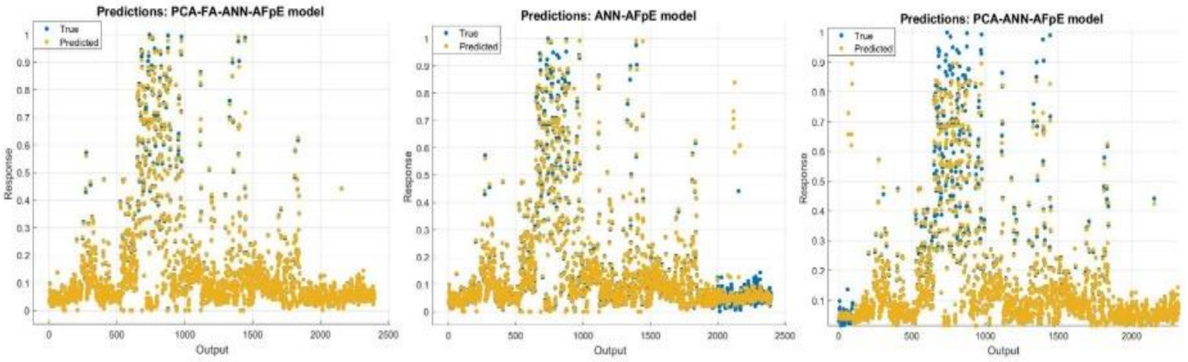


Figure 7.23: Response plot of AFpE models.

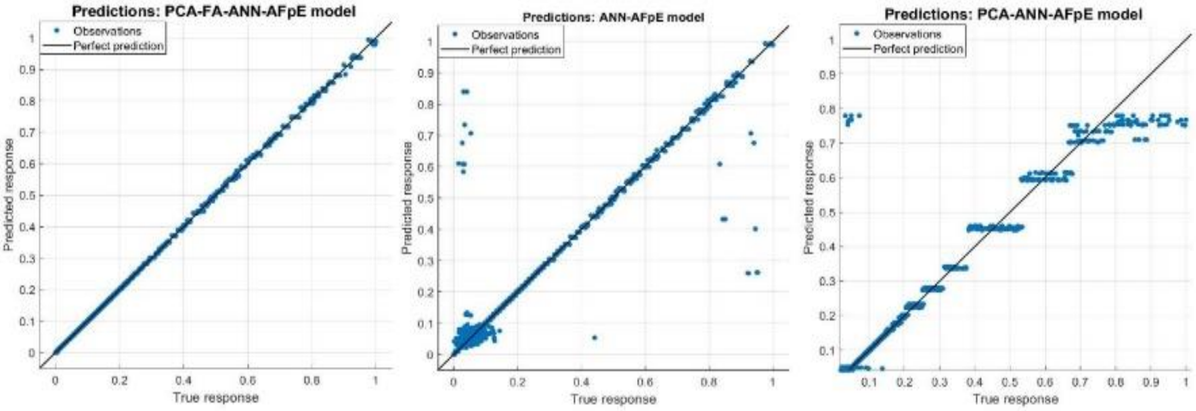


Figure 7.24: Actual vs. predicted value of AFpE models.

The AFp model having three input parameters, was developed using ANN, PCA-ANN, and PCA-FA-ANN algorithm. The PCA-FA-ANN-AFp model with 3 hidden layers and 25 neurons in each layer outperformed all other combinations with the overall highest R2 of 0.95 and lowest MSE of 0.004, as illustrated in Figure 7.25 Figure 7.26, respectively. Figure 7.27 shows the comparative response plot of models using ANN, PCA-ANN, and PCA-FA-ANN algorithm, and Figure 7.28 shows the plot between predicted output and actual output.

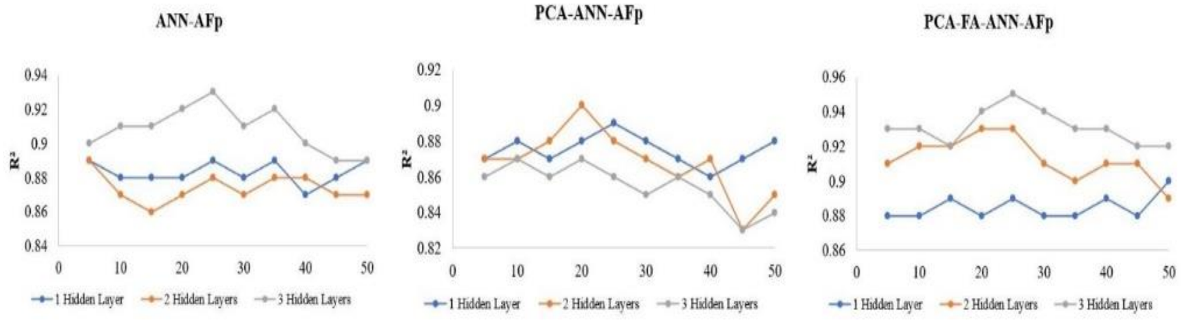


Figure 7.25: Accuracy comparison of AFp models.

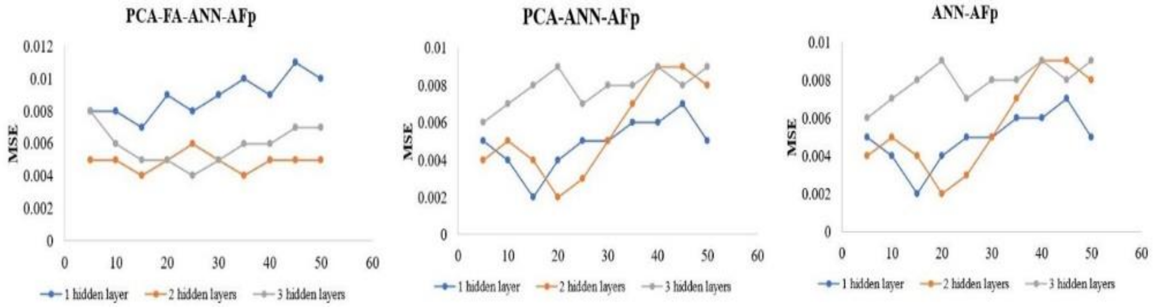


Figure 7.26: Error comparison of AFp models.

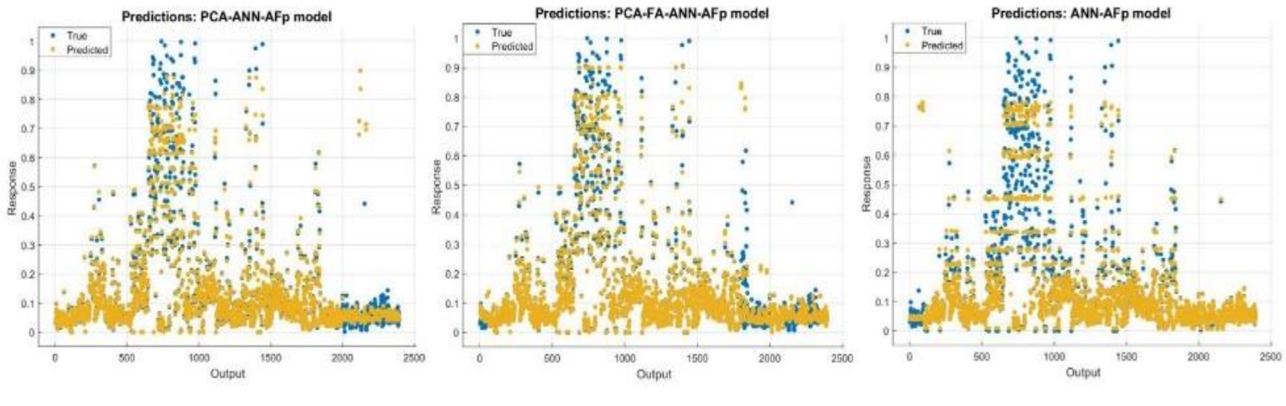


Figure 7.27: Response plot of AFp models.

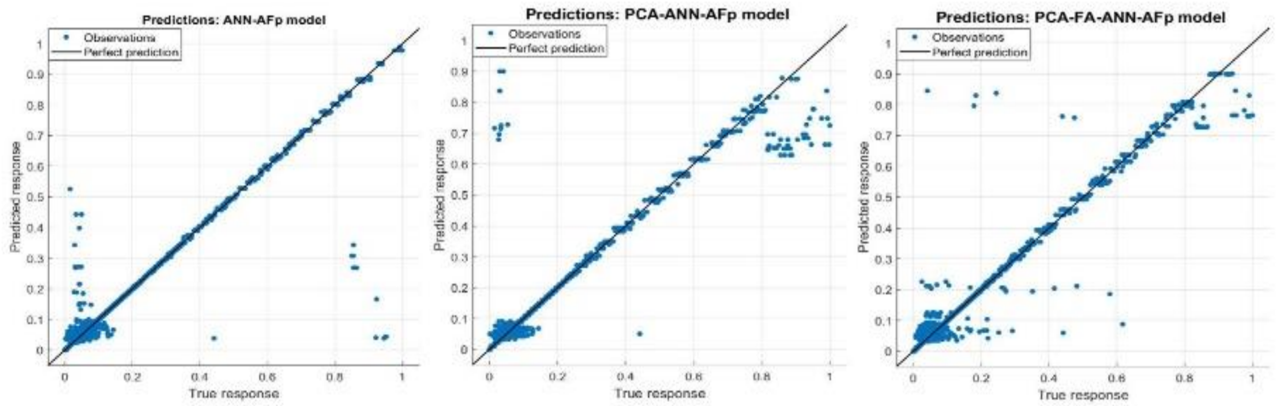


Figure 7.28: Actual vs. predicted value of AFp models.

The F model having one input parameter was developed using ANN, PCA-ANN, and PCA-FA-ANN algorithms. The PCA-FA-ANN-F model with 3 hidden layers and 20 neurons in each layer had outperformed all other combinations with the overall highest R^2 of 0.97 and lowest MSE of 0.007, as illustrated in Figure 7.29 and Figure 7.30, respectively. Figure 7.31 shows the comparative response plot of models using ANN, PCA-ANN, and PCA-FA-ANN algorithm, and Figure 7.32 shows the plot between predicted output and actual output.

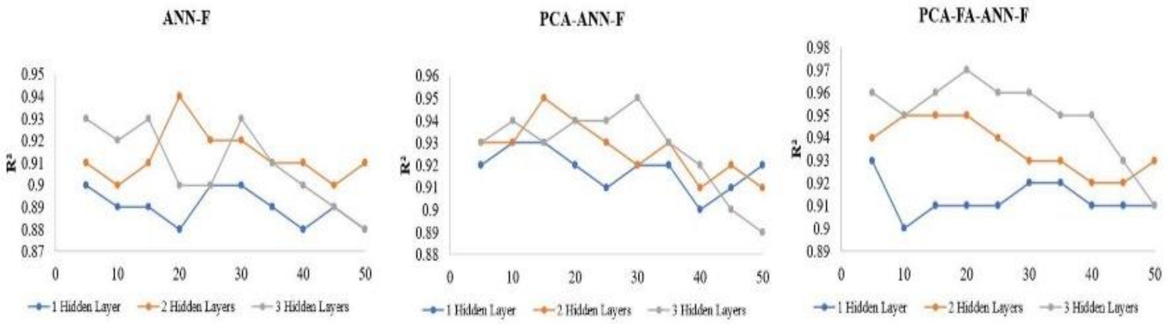


Figure 7.29: Accuracy comparison of F models.

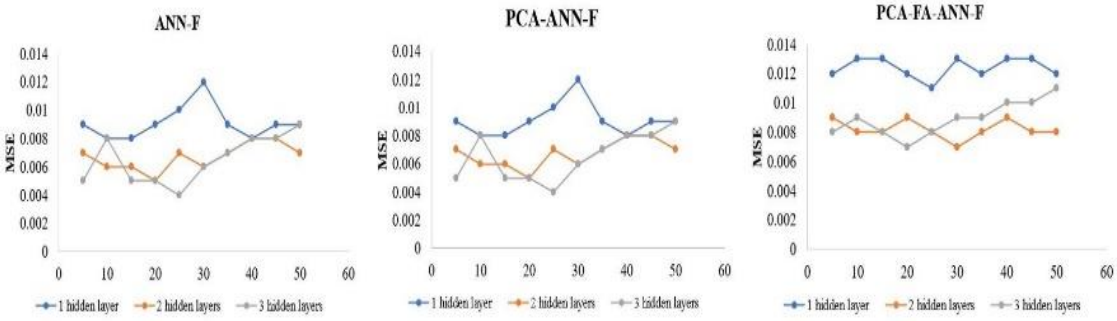


Figure 7.30: Error comparison of F models.

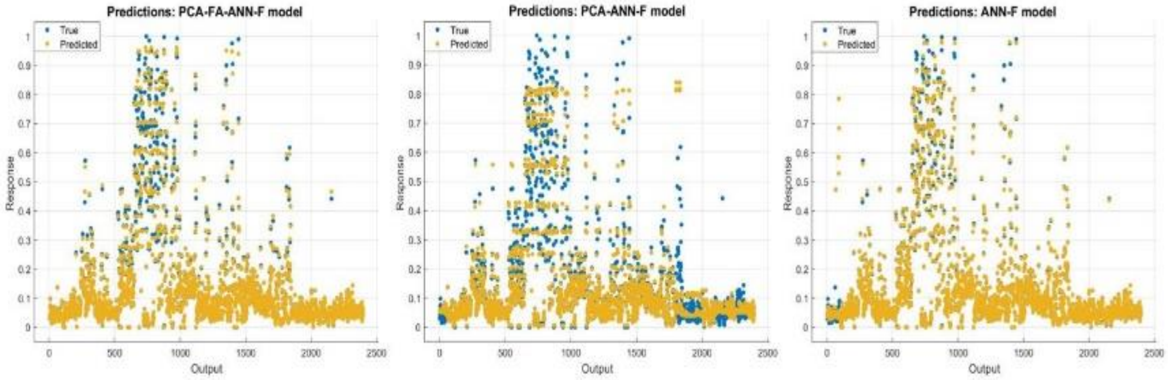


Figure 7.31: Response plot of F models.

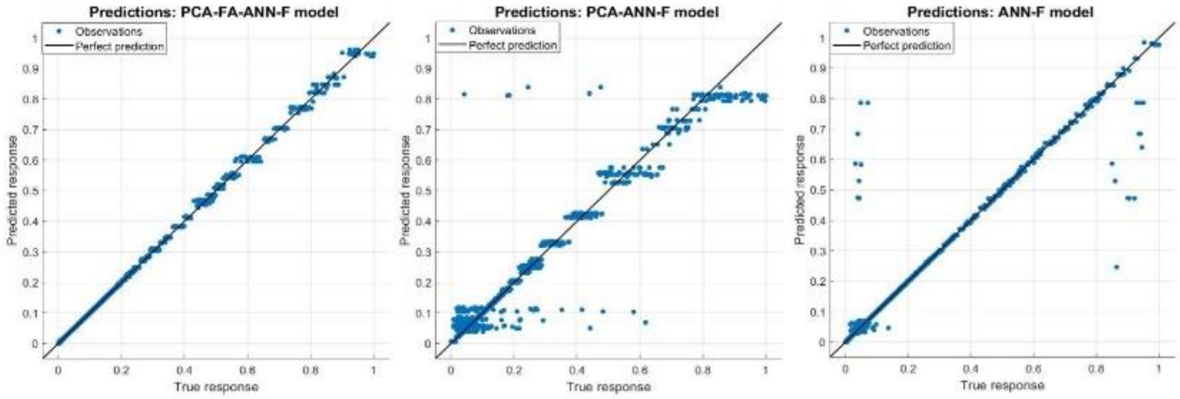


Figure 7.32: Actual vs. predicted value of F models.

The significant fluoride source for human bodies is drinking water, whereas cooking water and food are not so many significant sources. About 90 % of the fluoride content of drinking water is absorbed in the digestive tract. Comparatively, only about 30 – 60% of the fluoride contents of food are absorbed (WHO, 1996). Cerklewski (1997) estimated that 80-95% of fluoride intake in the human body is consumed, 52.6-72.7% of which is excreted by the urine. Assuming a water consumption of 2 L/capita/day for an average body weight of 60 kg, the World Health Organization (WHO) has set the acceptable fluoride value in drinking water at 1.5 mg/L. However, the WHO has proposed that each nation set its acceptable value based on its citizen's water consumption, which depends on the region's environment (WHO, 2017). Since India is situated in a tropical area, its inhabitants prefer to drink more water than those living in temperate or cold regions (Hossain et al., 2013). Thus, for fluorosis prevention, the acceptable value of fluoride in drinking water in India was revised from 1.5 mg/L to no more than 1.0 mg/L (IS 10500; IS 3025 [Part 60]). Also, each person's water consumption within a population varies based on their lifestyle and physical characteristics, such as age, body weight, height, and gender. Therefore, fluoride intake in the body can differ significantly in areas affected by fluoride.

Our study shows that various individual parameters (age, weight, height, and gender) significantly affect fluoride concentration in nail samples. Correlation between nail fluoride concentration with water fluoride, pH, and electrical conductivity as illustrated in Figure 7.33. Height (0.423), Weight (0.366), Age (0.3494) have substantial positive loadings on principal component 1 (PC1), and Gender (-0.3554) have significant negative loadings on principal component 3 (PC3). For the Gender classification, the number of females in the household was taken as the number correlation factor. A negative correlation shows a skew towards the males of the family in fluoride consumption. Therefore, to conduct a risk assessment of fluoride ingestion, people's water consumption in fluoride-affected areas must be correctly measured. So we can conclude that the inclusion of individual water consumption patterns can indeed have a positive correlation to this predictive model. This can be appreciated in the case of Indian standards being set lower than the global standards for fluoride content owing to increased water consumption as a consequence of environmental conditions.

The highlights of the results related to the proposed model are:

- Reducing the dimensions of the dataset using the PCA algorithm had improved the performance of the ANN model.
- The randomness of the ANN model was reduced with the application of the PCA-FA algorithm by optimization of weights and bias between the input layer and the output layer of the model.
- The PCA-FA-ANN model outperforms ANN and PCA-ANN hybrid models in terms of R^2, MSE, RMSE, MAE, and MAPE.
- The concentration of fluoride in nail samples is highly correlated with age, water fluoride, pH, and electrical conductivity of water.

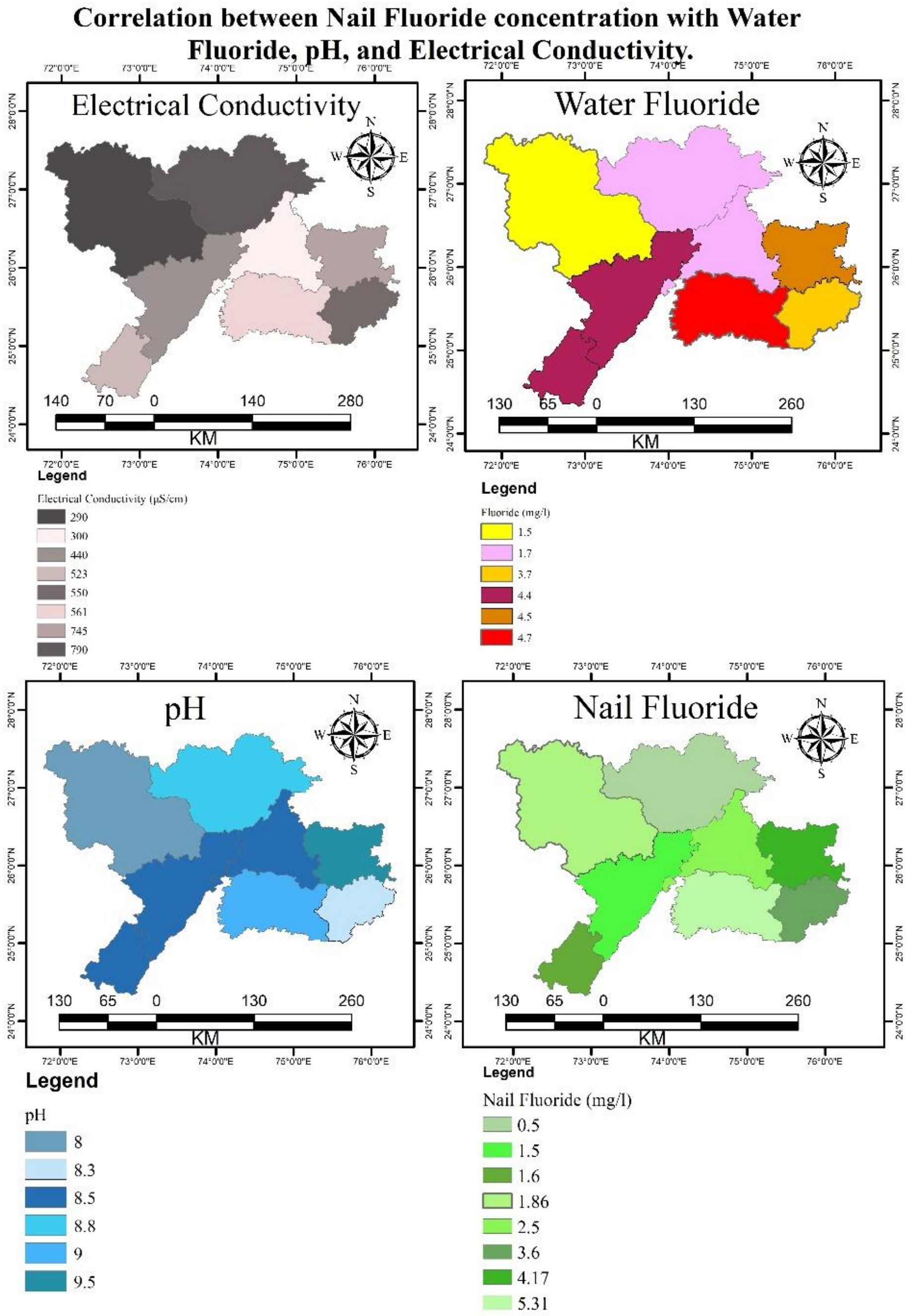


Figure 7.33: Correlation between nail fluoride concentration with water fluoride, pH, and electrical conductivity.

7.10 Summary

Due to environmental, nutritional, dietary, physiological, and cultural factors, fluoride bioavailability may vary among individuals. In this study, we have proposed a hybrid principal component analysis (PCA)- firefly algorithm (FA)- artificial neural network (ANN) machine learning model for establishing the relationship between water fluoride and nails. In this paper, we create our dataset by using water samples collected from eight districts of the state of

Rajasthan under the BITS-UVA groundwater contamination project. We presented a PCA algorithm that would be used to reduce the dimension of the initial data used for data analysis from 2401-dimensions to the 3-dimensional data set. The objective of the algorithm was to restrict the maximum information only in the first three columns named as principal components and ignoring the rest of the columns holding the negligible amount of information. The transformed data were then exposed to a hybrid PCA-FA model with the purpose of reducing the dataset by selecting significant parameters using the fitness value generated by the firefly algorithm. The ANN models were developed on the reduced dataset and compared for performance evaluation. The results obtained from the analysis suggest that the proposed model is more accurate and reliable in comparison to ANN and PCA-ANN models. Furthermore, as these parameters are represented in the 2D plot in MATLAB software, we can view the correlation between the parameters. This study indicates that the concentration of nail fluoride was correlated with age, weight, height, gender, and water quality parameters like pH, dissolved oxygen, electrical conductivity, turbidity, fluoride, and nitrate. In conclusion, our new methodology can be used for the prediction of hotspots of exposure based on a combination of water samples and nail samples testing.

References:

- Akpata, E. S., Behbehani, J., Akbar, J., Thalib, L., & Mojiminiyi, O. (2014). Fluoride intake from fluids and urinary fluoride excretion by young children in Kuwait: a non-fluoridated community. *Community dentistry and oral epidemiology*, *42*(3), 224-233.
- Álvarez-Ayuso, E., Giménez, A., & Ballesteros, J. C. (2011). Fluoride accumulation by plants grown in acid soils amended with flue gas desulphurisation gypsum. *Journal of hazardous materials*, *192*(3), 1659-1666.
- Amaral, J. G., Freire, I. R., Valle-Neto, E. F., Cunha, R. F., Martinhon, C. C., & Delbem, A. C. (2014). Longitudinal evaluation of fluoride levels in nails of 18–30-month-old children that were using toothpastes with 500 and 1100 μg F/g. *Community dentistry and oral epidemiology*, *42*(5), 412-419.
- Ando, M., Tadano, M., Asanuma, S., Tamura, K., Matsushima, S., Watanabe, T., & Cao, S. (1998). Health effects of indoor fluoride pollution from coal burning in China. *Environmental Health Perspectives*, *106*(5), 239-244.
- Antonijevic, E., Mandinic, Z., Curcic, M., Djukic-Cosic, D., Milicevic, N., Ivanovic, M., & Antonijevic, B. (2016). “Borderline” fluorotic region in Serbia: correlations among fluoride in drinking water, biomarkers of exposure and dental fluorosis in schoolchildren. *Environmental geochemistry and health*, *38*(3), 885-896.
- Bell, A. J., & Sejnowski, T. J. (1997). The “independent components” of natural scenes are edge filters. *Vision research*, *37*(23), 3327-3338.
- Beltran, L. (2006). Nonparametric multivariate statistical process control using principal component analysis and simplicial depth.
- BIS, I. (2012). 10500 Indian standard drinking water–specification, second revision. *Bureau of Indian Standards, New Delhi*.
- Buzalaf, M. A. R., Pessan, J. P., & Alves, K. M. R. P. (2006). Influence of growth rate and length on fluoride detection in human nails. *Caries research*, *40*(3), 231-238.
- Buzalaf, M. A. R., & Whitford, G. M. (2011). Fluoride metabolism. In *Fluoride and the oral environment* (Vol. 22, pp. 20-36). Karger Publishers.
- Buzalaf, M. A. R., Granjeiro, J. M., Duarte, J. L., & Taga, M. L. D. L. (2002). Fluoride content of infant foods in Brazil and risk of dental fluorosis. *Journal of Dentistry for Children*, *69*(2), 196-200.

- Buzalaf, M. A. R., Massaro, C. S., Rodrigues, M. H. C., Fukushima, R., Pessan, J. P., Whitford, G. M., & Sampaio, F. C. (2012). Validation of fingernail fluoride concentration as a predictor of risk for dental fluorosis. *Caries research*, *46*(4), 394-400.
- Buzalaf, M. A. R., Vilhena, F. V., Iano, F. G., Grizzo, L., Pessan, J. P., Sampaio, F. C., & Oliveira, R. C. (2009). The effect of different fluoride concentrations and pH of dentifrices on plaque and nail fluoride levels in young children. *Caries research*, *43*(2), 142-146.
- Carvalho, R. B. D., Medeiros, U. V. D., Santos, K. T. D., & Pacheco Filho, A. C. (2011). Influence of different concentrations of fluoride in the water on epidemiologic indicators of oral health/disease. *Ciencia & saude coletiva*, *16*(8), 3509-3518.
- Celinski, V. R., Ditter, M., Kraus, F., Fujara, F., & Schmedt auf der Günne, J. (2016). Trace determination and pressure estimation of fluorine F2 caused by irradiation damage in minerals and synthetic fluorides. *Chemistry–A European Journal, 22*(51), 18388-18393.
- Cerklewski, F. L. (1997). Fluoride bioavailability—nutritional and clinical aspects. *Nutrition research*, *17*(5), 907-929.
- Clarkson, J. J., & McLoughlin, J. (2000). Role of fluoride in oral health promotion. *International dental journal*, *50*(3), 119-128.
- Cotruvo, J. A. (2017). 2017 WHO guidelines for drinking water quality: first addendum to the fourth edition. *Journal-American Water Works Association*, *109*(7), 44-51.
- Davison, A. W., & Weinstein, L. H. (2006). Some problems relating to fluorides in the environment: effects on plants and animals. *Advances in Fluorine Science*, *1,* 251-298.
- De Almeida, B. S., da Silva Cardoso, V. E., & Buzalaf, M. A. R. (2007). Fluoride ingestion from toothpaste and diet in 1-to 3-year-old Brazilian children. *Community dentistry and oral epidemiology*, *35*(1), 53-63.
- Elekdag-Turk, S., Almuzian, M., Turk, T., Buzalaf, M. A. R., Alnuaimi, A., Dalci, O., & Darendeliler, M. A. (2019). Big toenail and hair samples as biomarkers for fluoride exposure–a pilot study. *BMC oral health*, *19*(1), 82.
- El-Shafie, A., Noureldin, A. E., Taha, M. R., & Basri, H. (2008). Neural network model for Nile river inflow forecasting based on correlation analysis of historical inflow data. *J ApplSci*, 8(24), 4487-4499.

- Fukushima, R., Rigolizzo, D. S., Maia, L. P., Sampaio, F. C., Lauris, J. R. P., & Buzalaf, M. A. R. (2009). Environmental and individual factors associated with nail fluoride concentration. *Caries research*, *43*(2), 147-154.

- Fukushima, R., Rigolizzo, D. S., Maia, L. P., Sampaio, F. C., Lauris, J. R. P., & Buzalaf, M. A. R. (2009). Environmental and individual factors associated with nail fluoride concentration. *Caries research*, *43*(2), 147-154.

- Gandomi, A. H., & Yang, X. S. (2011). Benchmark problems in structural optimization. In *Computational optimization, methods and algorithms* (pp. 259-281). Springer, Berlin, Heidelberg.
- Hayati, M., Yousefi, T., Ashjaee, M., Hamidi, A., & Shirvany, Y. (2007). Application of artificial neural networks for prediction of natural convection heat transfer from a confined horizontal elliptic tube. In Proceedings of World Academy of Science, Engineering and Technology (Vol. 22, pp. 269-274).
- Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural networks*, *2*(5), 359-366.

- Hossain, M. A., Rahman, M. M., Murrill, M., Das, B., Roy, B., Dey, S., ... & Chakraborti, D. (2013). Water consumption patterns and factors contributing to water consumption in arsenic affected population of rural West Bengal, India. *Science of the Total Environment*, *463*, 1217-1224.

- Khairnar, M. R., Dodamani, A. S., Jadhav, H. C., Naik, R. G., & Deshmukh, M. A. (2015). Mitigation of fluorosis-a review. *Journal of clinical and diagnostic research: JCDR*, *9*(6), ZE05.
- Koblar, A., Tavčar, G., & Ponikvar-Svet, M. (2011). Effects of airborne fluoride on soil and vegetation. *Journal of Fluorine Chemistry*, *132*(10), 755-759.
- Levy, F. M., Bastos, J. R. D. M., & Buzalaf, M. A. R. (2004). Nails as biomarkers of fluoride in children of fluoridated communities. *Journal of dentistry for children*, *71*(2), 121-125.
- Lima-Arsati, Y. B. O., Martins, C. C., Rocha, L. A., & Cury, J. A. (2010). Fingernail may not be a reliable biomarker of fluoride body burden from dentifrice. *Brazilian dental journal*, *21*(2), 91-97.

- Lima-Arsati, Y. B. O., Martins, C. C., Rocha, L. A., & Cury, J. A. (2010). Fingernail may not be a reliable biomarker of fluoride body burden from dentifrice. *Brazilian dental journal*, *21*(2), 91-97.
- Linhares, D. P. S., Garcia, P. V., Amaral, L., Ferreira, T., Cury, J. A., Vieira, W., & dos Santos Rodrigues, A. (2016). Sensitivity of two biomarkers for biomonitoring exposure to fluoride in children and women: A study in a volcanic area. *Chemosphere*, *155*, 614-620.
- Liong, S. Y., & Sivapragasam, C. (2002). Flood stage forecasting with support vector machines 1. *JAWRA Journal of the American Water Resources Association*, 38(1), 173-186.
- Lu WZ, Wang WJ, Wang XK, Xu ZB, Leung AY. 2003. Using improved neural network model to analyze RSP, NO x and NO 2 levels in urban air in Mong Kok, Hong Kong. Environmental monitoring and assessment. 87(3):235-54.
- MacKay, D. J. (1992). Bayesian interpolation. Neural computation, 4(3), 415-447.
- Makarynska, D., & Makarynskyy, O. (2008). Predicting sea-level variations at the Cocos (Keeling) Islands with artificial neural networks. *Computers & Geosciences*, 34(12), 1910-1917.
- Masters, T. (1995). *Advanced algorithms for neural networks: a C++ sourcebook*. John Wiley & Sons, Inc.
- MathWorks: Neural Network Toolbox Release 2019b– MATLAB & Simulink – MathWorks India, available at, last access: 23rd June 2020.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, *5*(4), 115-133.
- Moazenzadeh, R., Mohammadi, B., Shamshirband, S., & Chau, K. W. (2018). Coupling a firefly algorithm with support vector regression to predict evaporation in northern Iran. *Engineering Applications of Computational Fluid Mechanics*, *12*(1), 584-597.
- Moseley, R., Waddington, R. J., Sloan, A. J., Smith, A. J., Hall, R. C., & Embery, G. (2003). The influence of fluoride exposure on dentin mineralization using an in vitro organ culture model. *Calcified tissue international*, *73*(5), 470-475.
- Muttil, N., & Chau, K. W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution*.
- Najah, A., Elshafie, A., Karim, O. A., & Jaffar, O. (2009). Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of Scientific Research*, 28(3), 422-435.

- Noureldin, A., El-Shafie, A., & Bayoumi, M. (2011). GPS/INS integration utilizing dynamic neural networks for vehicular navigation. *Information fusion*, 12(1), 48-57.
- NRC, U. (2006). United States National Research Council. Committee to assess health risks from exposure to low levels of ionizing radiation. Health risks from exposure to low levels of ionizing radiation: BEIR VII-phase 2. United States National Academy of Sciences.
- O'Donnel, T. A. (1973). The chemistry of fluorine.
- Paliwal, M., & Kumar, U. A. (2009). Neural networks and statistical techniques: A review of applications. *Expert systems with applications, 36*(1), 2-17.
- Pao, H. T. (2006). Comparing linear and nonlinear forecasts for Taiwan's electricity consumption. *Energy, 31*(12), 2129-2141.
- Pessan, J. P., & Buzalaf, M. R. A. (2011). Historical and recent biological markers of exposure to fluoride. *Fluoride and the oral environment*, *22*, 52-65.
- Pessan, J. P., Pin, M. L. G., Martinhon, C. C. R., Silva, S. M. B. D., Granjeiro, J. M., & Buzalaf, M. A. R. (2005). Analysis of fingernails and urine as biomarkers of fluoride exposure from dentifrice and varnish in 4-to 7-year-old children. *Caries research*, *39*(5), 363-370.
- Reddy, S. S., & Momoh, J. A. (2014, September). Short term electrical load forecasting using back propagation neural networks. In *2014 North American Power Symposium (NAPS)* (pp. 1-6). IEEE.
- Schmedt auf der Günne, J., Mangstl, M., & Kraus, F. (2012). Occurrence of difluorine F2 in nature—in situ proof and Quantification by NMR spectroscopy. *Angewandte Chemie International Edition*, *51*(31), 7847-7849.
- Shinde, R. L., & Khadse, K. G. (2009). Multivariate process capability using principal component analysis. *Quality and Reliability Engineering International, 25*(1), 69-77.
- Sousa, E. T. D., Alves, V. F., Maia, F. B. M., Nobre-dos-Santos, M., Forte, F. D. S., & Sampaio, F. C. (2018). Influence of fluoridated groundwater and 1,100 ppm fluoride dentifrice on biomarkers of exposure to fluoride. *Brazilian dental journal*, *29*(5), 475-482.
- Standard, I. (2006). Methods of Sampling and Test (Physical and Chemical) for Water and Wastewater. *Environmental Protection Sectional Committee, CHD, 12*(0), 10.
- Stojanovic, B., & Neskovic, A. (2012, November). Impact of PCA based fingerprint compression on matching performance. In *2012 20th Telecommunications Forum (TELFOR)* (pp. 693-696). IEEE.

- Turian, J., Bergstra, J., & Bengio, Y. (2009, June). Quadratic features and deep architectures for chunking. In Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers (pp. 245-248).
- Wenzel, W. W., & Blum, W. E. H. (1992). Fluoride speciation and mobility in fluoride concentration soil and minerals. *Soil Sci*, *153*, 357-364.
- Whitford, G. M., Sampaio, F. C., Arneberg, P., & Von der Fehr, F. R. (1999). Fingernail fluoride: a method for monitoring fluoride exposure. *Caries research*, *33*(6), 462-467.
- Whitford, G. M. (2005). Monitoring fluoride exposure with fingernail clippings. *Schweizer Monatsschrift fur Zahnmedizin*, *115*(8), 685.
- World Health Organization. (1996). *Trace elements in human nutrition and health*. World Health Organization.
- World Health Organization. (2006). *The world health report 2006: working together for health*. World Health Organization.
- Yang, X. S. (2010). Firefly algorithm, stochastic test functions and design optimisation. *International journal of bio-inspired computation*, *2*(2), 78-84.
- Yang, X. S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver press.
- Yang, X. S. (2010). *Nature-inspired metaheuristic algorithms*. Luniver press.
- Yang, X. S. (2011). Metaheuristic optimization. *Scholarpedia*, *6*(8), 11472.
- Yang, X. S. (2011, May). Metaheuristic optimization: algorithm analysis and open problems. In *International Symposium on Experimental Algorithms* (pp. 21-32). Springer, Berlin, Heidelberg.
- Yi, J., & Cao, J. (2008). Tea and fluorosis. *Journal of Fluorine Chemistry*, *129*(2), 76-81.

8. Portable Hand-Held Smart device for real-time Water Quality Measurement and Water Quality Classification

Conventional water quality measurement techniques include on-site sampling and subsequent laboratory-based tests; both are labor-intensive and cost-intensive processes. The objective of this study is to develop a low-cost system for real-time monitoring of water quality. This ability of an Artificial Neural Network to predict based on some knowledge base is utilized to reduce the number of expensive sensing electrodes and learn the relation between those parameters and the parameters being measured. The device is portable, cost-effective, and usable in all weather conditions. Owing to its portability and frequency of operation, the device enables real-time monitoring of water quality.

8.1 Introduction

Water pollution is one of the most critical challenges for sustainable development. According to a WHO report (World Health Organisation 2017), 1.9 billion people worldwide use water that is polluted. Annually about 37.7 million Indians are affected by waterborne diseases. According to the National Health Profile (Central Bureau of Health Intelligence, 2018) waterborne diseases continue to be prevalent in India and have caused 10,738 deaths over the last five years since 2017. Water quality measurement is an important stepping stone towards finding a solution to this problem. Currently, water quality parameters are measured using methods based on laboratory testing, where the standard laboratory sensors are stationary and water samples are brought in from the field for analysis. In this way, the current water quality monitoring system is a repetitive manual system. It is incredibly tedious with the time-consuming procedure. The test sensor can be mounted in the water sample, and pollution detection can be performed remotely to improve device performance. There are some field usable devices, but those devices are large and cumbersome and way too costly. Basic water quality parameters like pH, Temperature, Turbidity, and TDS are taken as references, as the variations in the value of these parameters indicate the extent of water pollution.

8.1.1 Raspberry Pi

Raspberry Pi is a small single-board computer. It can be used as a fully functional computer by connecting peripherals like a keyboard, a mouse, and a display unit. The performance may not be quite that of a laptop or a desktop, but it is relatively capable a computer. The raspberry pi foundation provides a Debian-based Raspbian OS, which can be loaded onto a micro SD card that can be slotted into the provided slot on the board. The Raspberry Pi board can function as a Linux based computing device (Sangjan et al. 2021). The Raspberry Pi board can be used as a computer to do Image processing, IoT-based applications, Browsing, Python Programming, etc. Raspberry Pi offers more than just computing powers on a small board. The Raspberry Pi provides access to GPIOs, which can be connected to sensors, motors, LEDs, etc., and control them too. Raspberry Pi also provides access to I2C and UART modules, which can communicate with other embedded boards like Arduino Uno, etc (Pi-Teach, 2016).

8.1.2 Arduino Uno

The Arduino Uno is a microcontroller-based embedded board with an onboard Analog-to-Digital Converter (ADC) that can be used to communicate with analog sensors such as the electro-chemical sensors for water quality measurement. Further, the microcontroller can be programmed to convert the analog voltage readings of these sensors into actual parameter readings that can be used to make decisions to control some other signals, or the data can be sent to other devices for further usage. To send the data to other computing devices, the I2C or UART ports available onboard can be used. The Uno board also provides on board power pins for both 5V and 3.3V output, which can be used to drive a variety of sensors (Badamasi et al. 2014).

8.1.3 Artificial Neural Networks

Artificial Neural Networks (ANN) are a model of the Biological Neural Network. Biological Neural Networks help living beings perceive the patterns in their environment, classify them and learn from them for future applications. Humans use these patterns and prior knowledge to process any information and thus come to an output (Fausett, 2006). Applications of ANN in the groundwater, ecology, and environmental engineering fields were documented in the early 1990s. However, in recent years ANN has been intensively used for prediction and forecasting in a variety of engineering and water-related areas, including water resource analysis by Liong and Sivapragasam, 2002; Muttil and Chau, 2006; El-Shafie et al. 2008; El-Shafie et al. 2011; Noureldin et al. 2011; Najah et al. 2009; oceanography by Makarynska et al. 2008 and environmental engineering by Grubert, 2003.

Conventional water quality measurement techniques include on-site sampling and subsequent laboratory-based tests; both are labor-intensive and cost-intensive processes. The measurements are not in real-time. Therefore, there is a need for real-time monitoring of water quality for drinking applications to reduce labor costs and time usage. With the help of Zigbee boards, recorded data is uploaded to the remote data storage in the traditional system. It requires more hardware to set up this technology and is very expensive. There's also no alert indication in that system when parameters are abnormal. In the Solar Powered Water Quality Monitoring System using remote Sensor Network, the advancement of the water sensing network is controlled using sun board. If the sun board is not charged, then the system will not switch on, which is the restriction associated with this method. The system cannot fulfill the objective of real-time monitoring of water quality parameters. This study aims to design and develop a low-cost Raspberry Pi and Arduino Uno-based water quality monitoring system for real-time monitoring using artificial intelligence. A system that is portable, the output is legible for people with limited or no literacy, and it will work in all environmental conditions, unlike a solar-powered water quality network monitoring system.

8.2 Methodology

Groundwater samples were collected from various sources in and around the Birla Institute of Technology and Science, Pilani, Rajasthan, India. The water samples were analyzed in the laboratory using titration and spectroscopy methods. It has been used as the "gold standard" against which the parameter values of the respective electrodes and the predicted parameter values of ANN A1 have been validated. The experimental data were used as the training, testing, and validation data for two Artificial Neural Networks (ANN) A1 and A2.

A Prototype of the device was made using Arduino Uno and the sensor circuits and multi-sensor "Tentacle Shield" (Atlas Scientific, n.d.) from ATLAS Scientific. The parameters measured were pH, DO, ORP, and Electrical Conductivity. The Prototype was tested against the Laboratory experiment results. The prototype was powered by a 10,000mAh Li-ion battery pack as a power supply. This makes the prototype portable.

In order to strictly control the cost of the device, the sensing circuits of DO and EC were done away with. Only the pH and ORP electrodes were used. The respective electrodes were connected to the analog input ports of the Arduino Uno Board. The analog voltage readings from the electrodes measuring pH and ORP were converted into digital readings using the 10-bit ADC present onboard the Arduino Uno microcontroller board (Arduino Inc., n.d.). The other two parameters, DO and EC, are predicted using an ANN.

In order to get pH readings, the voltage readings from the pH Electrode are read by Arduino Uno. These voltages are analog voltage readings. Thus they are connected to the analog input ports. These analog inputs are digitized by the 10-bit ADC on-board the Arduino Uno R3 board before they are displayed on the Serial monitor for Arduino. To convert them back to pH readings, first, the readings are re-quantized to voltage values from 10-bit digital values. Since the voltage swing is between 0V – 5.0V and the quantization is 10-bit. Thus the input voltage values are multiplied by the voltage range and divided by the quantization value: -

$$V_{in} = x \times \frac{5.0}{1023} \tag{8.1}$$

This voltage value is now converted into pH reading by the Nernst equation: -

$$E = E_r + \left(\frac{2.303RT}{nF}\right) \log \left(\frac{unknown\,[H+]}{internal\,[H-]}\right) \tag{8.2}$$

For our electrode equation (2) comes out to be: -

$$pH = \frac{((V_{in} - 512) \times 9.65)}{8.31 \times 2.302 \times 298} + 7 \tag{8.3}$$

7 has been added in the above equation to offset the zero voltage to neutral pH reading of 7.

For measurement of ORP, equation (1) is reused to convert digital readings into voltages (potentials). 2.25 is subtracted from the readings to offset the voltage readings by -225mV to obtain the ORP readings.

$$ORP = x \times \frac{5.0}{1023} - 2.25 \tag{8.4}$$

Measurement of DO and Electrical Conductivity – the sensors for these parameters are prohibitively costly. Hence, to reduce the cost of the final device, these parameters were predicted against pH and ORP values. For the prediction of DO and Electrical Conductivity (EC), an ANN was designed with pH and ORP as input parameters. The electrodes and sensor circuits of two of the four desired parameters are prohibitively expensive, making it impractical to include them in such a low-cost device. Thus, we designed ANN A1, which takes the two parameters, pH and ORP readings from the electrodes via the Arduino Uno Board, and predicts DO and EC values based on the data generated from the laboratory. This aforementioned sensor readings are taken from Arduino Uno into Raspberry Pi (Raspberry Pi Foundation, n.d.) over the Serial connection. The experimental data is stored on the secondary memory (a microSD card) attached to the Raspberry Pi board. The ANN is also coded in Raspberry Pi memory.

The input parameter taken for the ANN A1 were pH and ORP. The outputs of the A1 were DO and EC. The experimental data were divided into training, testing, and validation sets. Randomly selected 70% data points were allocated to the training of the ANN model, 15% each for validation and testing, respectively. The ANN was trained using the experimental data obtained using the Arduino Uno-based prototype.

A Second ANN model, A2, was implemented to determine water quality. The output of ANN A1 along with pH and ORP readings were taken as the input parameters to A2. A2 is designed to classify water quality into three categories, viz. - 1. Potable; 2. Irrigational; 3. Waste Water. The output of A2 has been encoded in a way that it is easily legible by people with limited literacy, which is the target audience of the device. A trial and error-based method were adopted to select the ANN architecture for both the ANN's, A1 and A2, with maximum accuracy using different training functions, activation functions, and performance functions. The ANN results were validated using the coefficient of determination (R^2) values and Root-Mean-Squared Error (RMSE) values.

Both Arduino Uno and Raspberry Pi are very power-efficient devices and can be used to work on a very low power supply. The capabilities of the boards have been put together to use them in tandem to make a handheld device that can deliver real-time water quality readings in the field without having to take the water samples back to the lab for testing of multiple parameters that help to decide the usage of water. The system uses Artificial Neural Networks (ANN) to classify water quality into three categories – Potable, Agricultural usage, and wastewater. To measure the parameters, two primary parameters are selected – pH and Oxidation-Reduction Potential (ORP). Keeping stringent cost control in mind, the raspberry pi is programmed with ANN to predict the values of Dissolved Oxygen (DO) and Electrical Conductivity (EC). The Arduino Uno board has been used as a sensor circuit (Figure 1). The 10-bit ADC on-board the Arduino Uno has been used to convert the analog voltage signals into 10-bit digital signals. These 10-bit strings were thence converted to respective sensor readings as per their voltage-sensor reading mathematical relationships. These readings are then given as output via Serial Output ports. The Serial outputs are then given to the Raspberry Pi board (Figure 2).

The Raspberry Pi has ANN’s programmed in python to utilize the inputs pH (Figure 3) and ORP (Figure 4) to predict the values of Dissolved Oxygen (DO) and Electrical Conductivity (EC) in the water sample. Another ANN takes the values of the pH, DO, ORP and EC and classifies the water quality sampling at that point of time into one of the three categories – potable, agricultural and wastewater. Since the Arduino Uno Board has been set up to take samples every second and send one set of data at the same rate, we can sample data every second. The block diagram of the proposed device is shown in Figure 5. The construction of the device is such that it is packed in a box style with electrodes protruding. The complete setup can be boxed in a handheld device form (Figure 6). The device, thus, can be used in a handheld manner and in any weather condition.

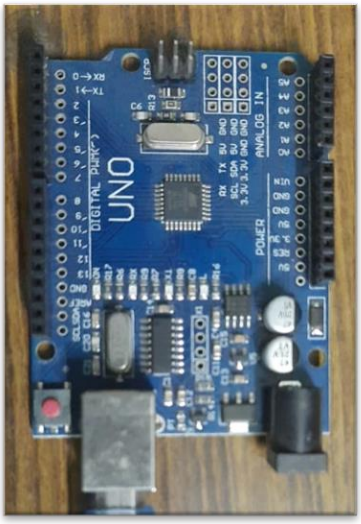


Figure 8.1: Arduino Uno   Figure 8.2: Raspberry Pi 3

Figure 8.3: pH Probe

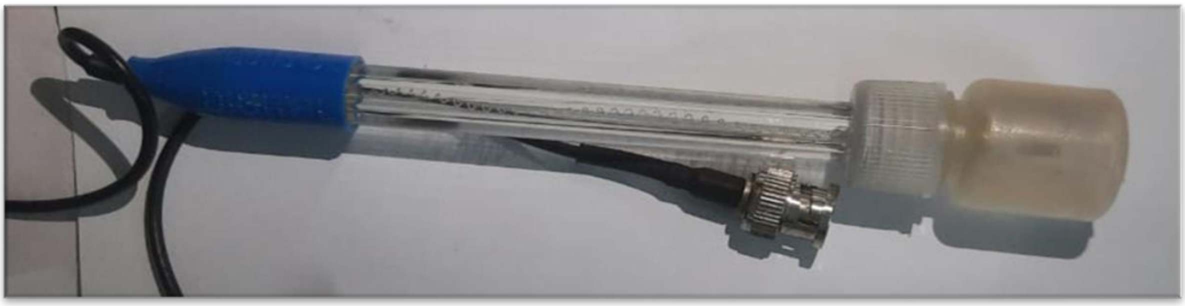


Figure 8.4: ORP Probe

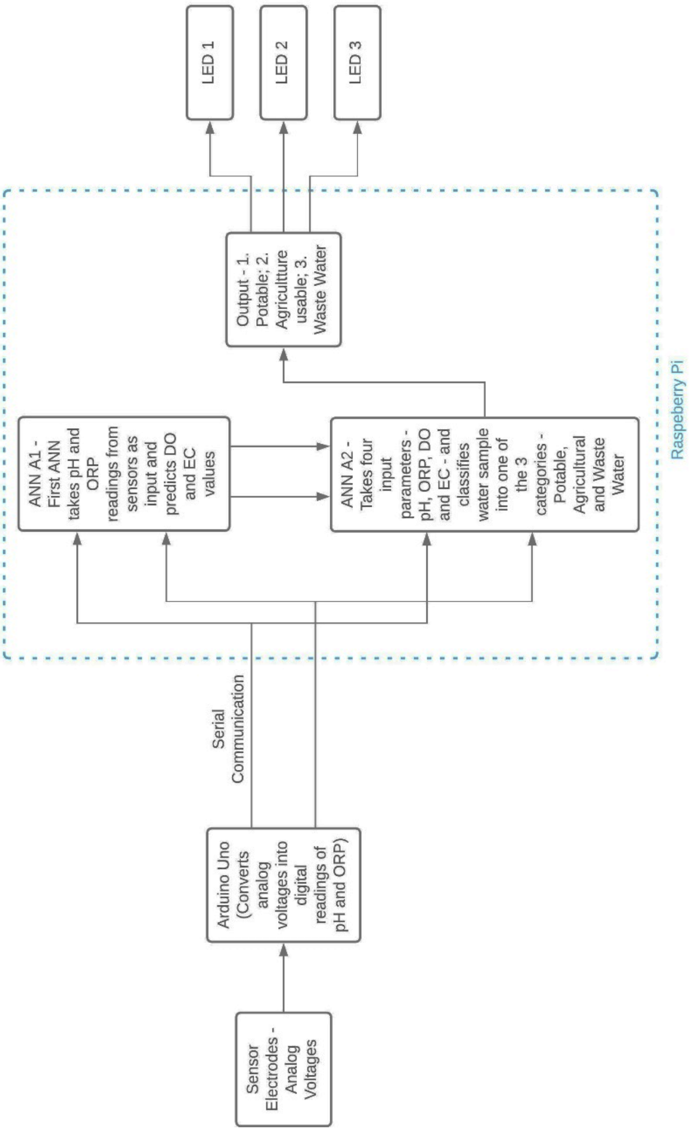

LED 1
LED 2
LED 3
Output - 1. Potable; 2. Agricultture usable; 3. Waste Water
Raspeberry Pi
ANN A1 - First ANN takes pH and ORP readings from sensors as input and predicts DO and EC values
ANN A2 - Takes four input parameters - pH, ORP, DO and EC - and classifies water sample into one of the 3 categories - Potable, Agricultural and Waste Water
Serial Communication
Arduino Uno (Converts analog voltages into digital readings of pH and ORP)
Sensor Electrodes - Analog Voltages

Figure 8.5: Block Diagram of Proposed Device.

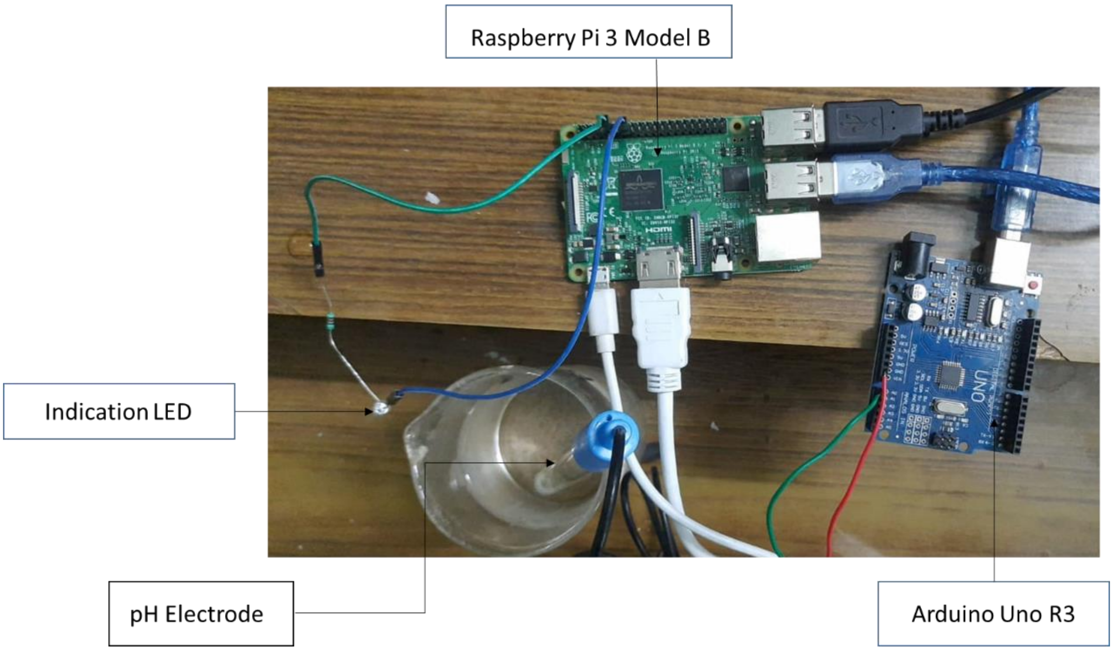


Figure 8.6: Proposed Device Prototype

The study proposes a portable device consisting of 2 electrodes – pH and ORP, an Arduino Uno microcontroller board RaspBerry Pi 3 single-board minicomputer, and a Li-ion battery pack. The RaspBerry Pi 3 board runs Artificial Neural Network and classifies the water sample into three classes, viz, Potable, Agricultural, and Waste Water. The output is generated as illumination of 3 LEDs. The device is portable, cost-effective, and usable in all weather conditions. Owing to its portability and frequency of operation, the device enables real-time monitoring of water quality. The box diagram of the original experimental data sample is illustrated in Figure 7.

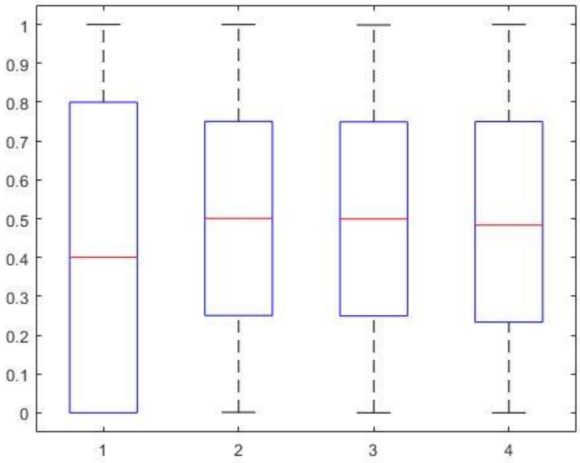


Figure 8.7: Box diagram.

Table 8.1: Cost comparison of the conventional and proposed device

Device Name	**Parameter**	**Cost (Conventional Device)**	**Cost (Proposed Device)**
Atlas Scientific DO Probe	Dissolved Oxygen	INR 21,240 (Atlas Scientific, n.d.)	NA
Atlas Scientific DO Sensor	Dissolved Oxygen	INR 4,299 (Atlas Scientific, n.d.)	NA
Aquasol ORP Electrode	Oxidation-Reduction Potential	INR 1200	INR 1200
Atlas Scientific ORP Sensor	Oxidation-Reduction Potential	INR 3,739 (Atlas Scientific, n.d.)	NA
Aquasol pH Electrode	pH	INR 900	INR 900
Atlas Scientific pH Sensor	pH	INR 3,739 (Atlas Scientific, n.d.)	NA
Atlas Scientific EC Electrode	Electrical Conductivity	INR 11,200 (Atlas Scientific, n.d.)	NA
Atlas Scientific EC Sensor	Electrical Conductivity	INR 5,600 (Atlas Scientific, n.d.)	NA
Battery Pack		INR 1,000	INR 1,000
Memory Card		NA	INR 300
Multiplexer Board		INR 11,869 (Atlas Scientific, n.d.)	NA
Arduino Uno Board		INR 330	INR 330
Raspberry Pi 3 Board		NA	INR 3,000
Total		**INR 75,116/-**	**INR 6,730/-**

The ANN model's performance was validated using the F-Score, Precision, Sensitivity, and Accuracy (Dalianis 2018; Prabha et al. 2016). The ANN model uses cross-validation to test the networks more thoroughly. This means that all experimental data is used as both training, testing and validation data, split into iterations. TP stands for true positive (appropriately recognizing a Potable water sample as Potable). TN stands for true negative (appropriately recognizing an Agricultural or Wastewater sample as Agricultural or Wastewater). FP stands for false positive (inappropriately recognizing a Potable water sample as Agricultural or Wastewater). FN stands for false negative (inappropriately recognizing Agricultural or Wastewater sample as Potable). Accuracy is defined as the proportion of properly recognized samples (Potable, Agricultural or Wastewater) among all samples; see Equation (8.5). The ratio of all recognized positive samples to all positive samples is described as sensitivity; see Equation (8.6). If the sensitivity is strong, the class is accurately detected. The value of high sensitivity implies that the class has been appropriately recognized. Precision is defined as the ratio of all positively detected positive samples to all positively predicted positive samples; see Equation (8.7). A high precision suggests that a sample classified as positive is, in fact, positive. The weighted average of sensitivity and accuracy is used to get the F-score. In F-Score, Harmonic Mean replaces the Arithmetic Mean. It punishes high values much more; see Equation (8.8). The following Equations (8.5-8.8) can be used to compute the F-Score, Precision, Sensitivity, and Accuracy:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (8.5)$$

$$Sensitivity = \frac{TP}{TP+FN} \qquad (8.6)$$

$$Precision = \frac{TP}{TP+FP} \qquad (8.7)$$

$$F - Score = \frac{2 * Sensitivity * Precision}{Sensitivity + Precision} \qquad (8.8)$$

A total of 14 training functions were tested. From these 14 functions, only one training function (trainLM), has completed the task of regression plot and error plots. Hence, a Feedforward-Backpropagation Neural Network (ANN A1) with 2 hidden layers and 16 neurons in each layer was optimized using Lavenberg-Marquardt training function and a sigmoidal activation function (logistic function) had outperformed all other architectures with the highest R^2 value of 0.98 and the lowest RMSE values of 0.00232. Figure 8 shows the effect of the number of neurons on the value of R^2 for one, two, and three hidden layers. Prediction accuracy has been evaluated using the mean square error (MSE) functions, as shown in Figure 9. We can see that there is a significant change in the accuracy of ANN. The model with two hidden layers and 32 neurons shows the best performance with the overall highest R^2 values of 0.99 and the lowest errors with MSE values of 0.04. Figure 10 shows the plot between predicted output and actual output; a perfect regression model has predicted output equal to actual output with a regression value of 0.99. Figure 11(a) and 12(a) shows the response plot of the ANN model for prediction of DO and EC respectively. Figure 11(b) and 12(b) shows the plot between predicted value and actual value DO and EC respectively.

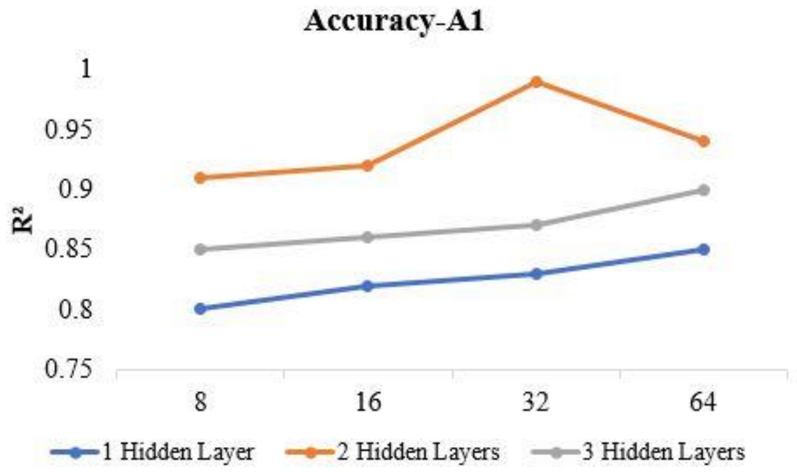


Figure 8.8: Accuracy of A1.

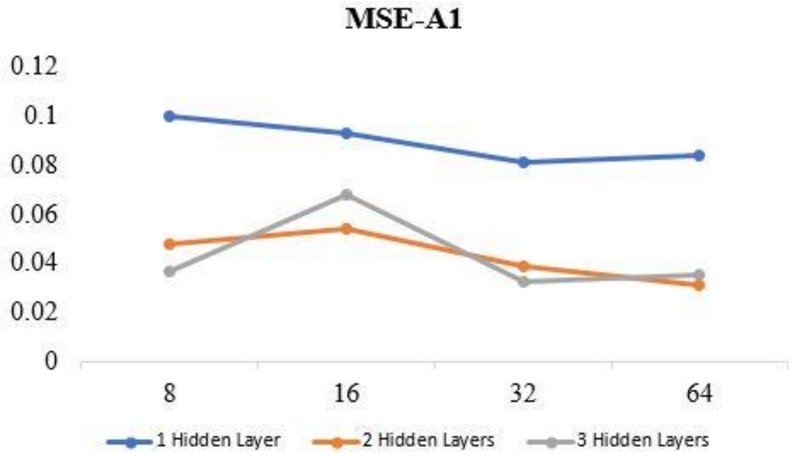


Figure 8.9: Mean Squared Error of ANN A1.

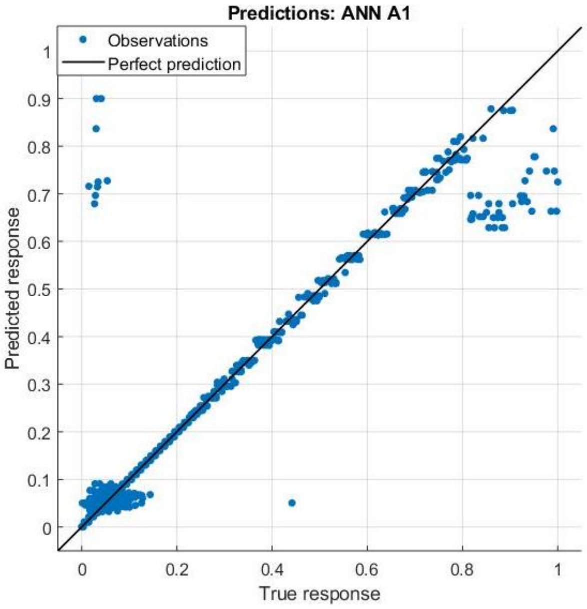
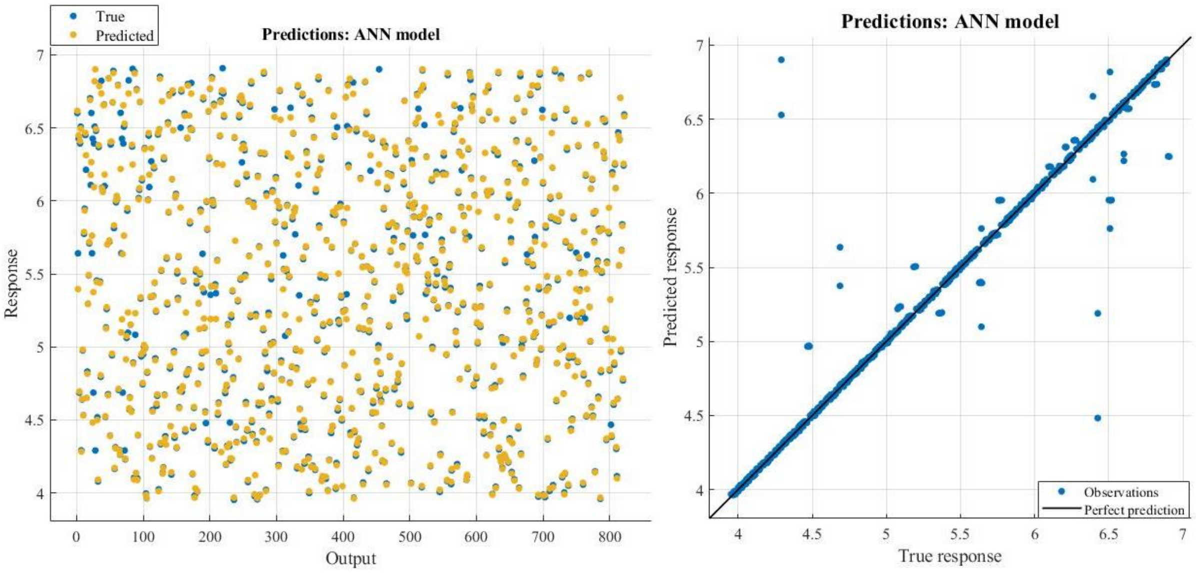
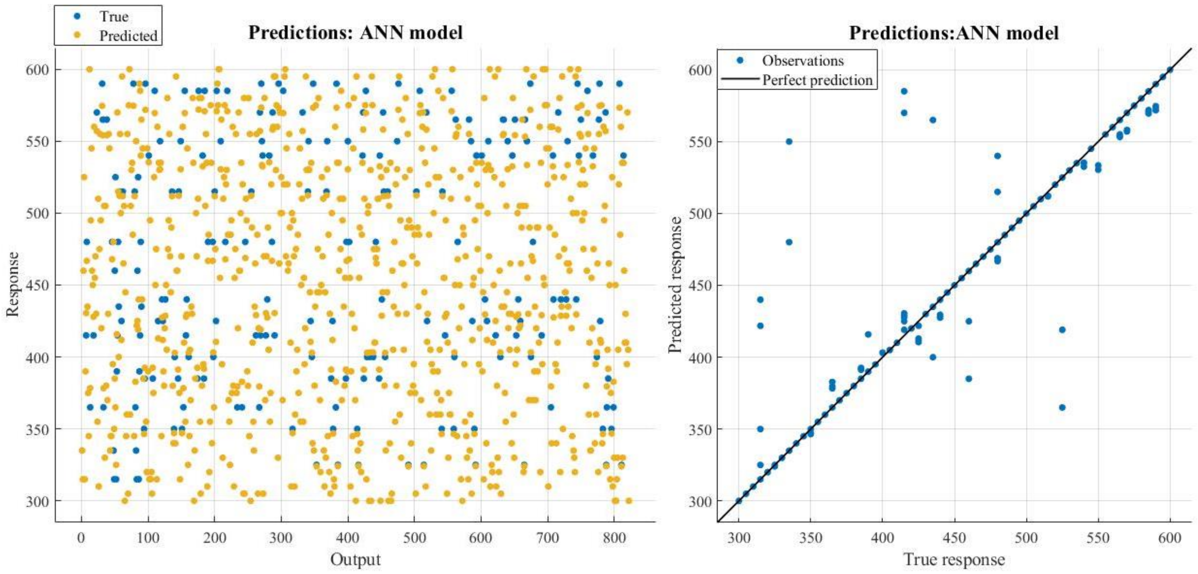


Figure 8.10: Predictions of ANN A1.



b) Response plot b) Actual vs. predicted value

Figure 8.11: a) Response plot of ANN, b) Actual vs. predicted DO value using ANN.



a) Response plot b) Actual vs. predicted value

Figure 8.12: a) Response plot of ANN, b) Actual vs. predicted EC value using ANN.

A Feedforward-Backpropagation Neural Network (ANN A2) with 3 hidden layers and 32 neurons in each layer was optimized using Lavenberg-Marquardt training function and a sigmoidal activation function (logistic function) had outperformed all other architectures with the highest R^2 value of 0.9958 and the lowest RMSE values of 0.00981. Figure 13 shows the architecture of the ANN A2.

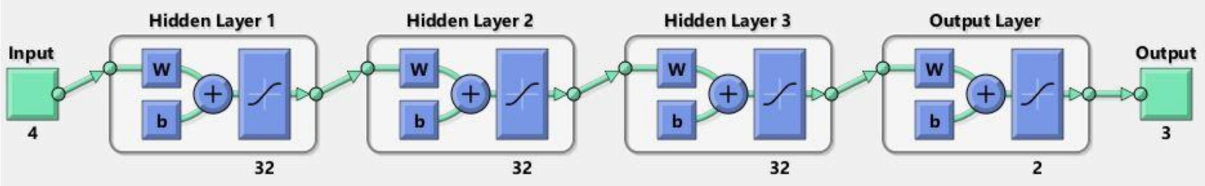


Figure 8.13: ANN A2 Architecture.

Figure 14 shows the effect of the number of neurons on the value of R^2 for one, two, and three hidden layers. Prediction accuracy has been evaluated using the mean square error (MSE) functions as shown in Figure 15. We can see that there is a significant change in the accuracy of ANN. The model with three hidden layers and 64 neurons shows the best performance with the overall highest R^2 values of 0.98 and lowest errors with MSE values of 0.15. Figure 16 shows the plot between predicted output and actual output; a perfect regression model has predicted output equal to actual output with a regression value of 0.98.

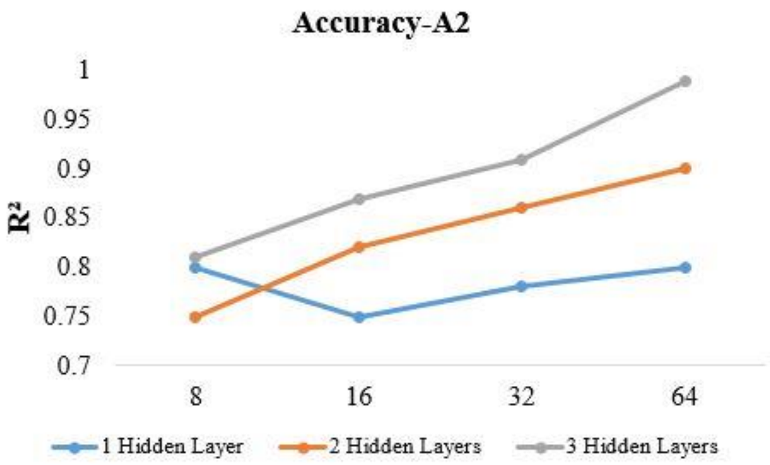


Figure 8.14: Accuracy of ANN A2.

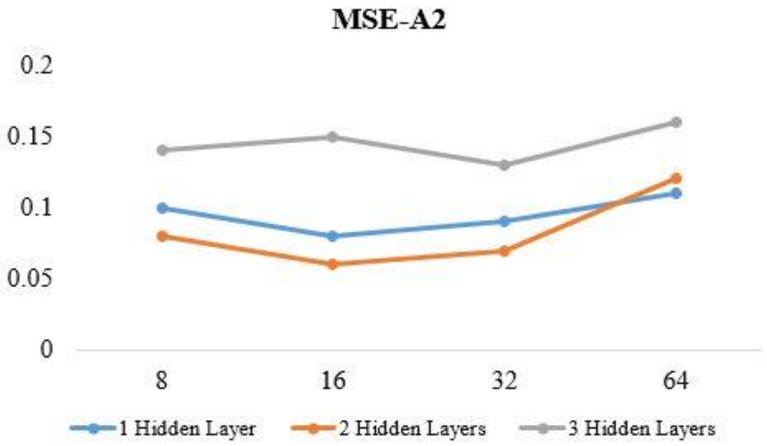


Figure 8.15: Mean Squared Error of ANN A2.

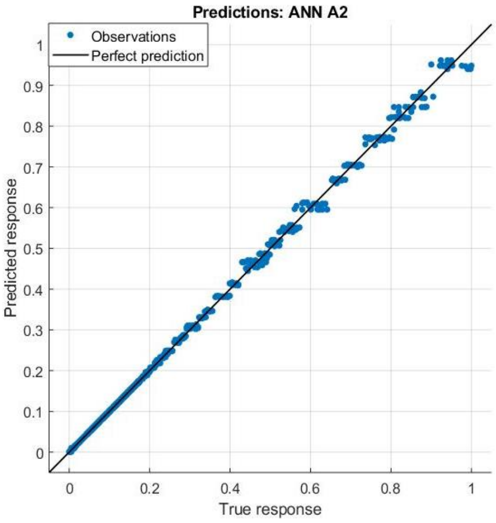


Figure 8.16: Predictions of ANN A2.

A confusion matrix was used to show the performance of an ANN model in Figure 17. The F-Score, Precision, Sensitivity, and Accuracy results calculated using the confusion matrix are shown in Table 2.

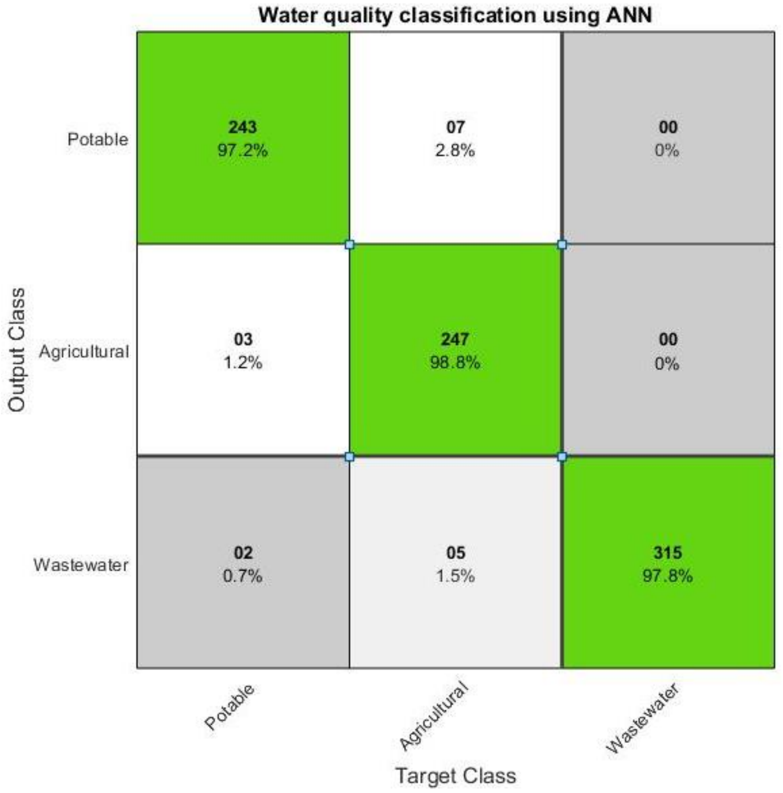


Figure 8.17: The confusion matrix obtained using experimental datasets.

Table 8.2: The statistical results for performance evaluation.

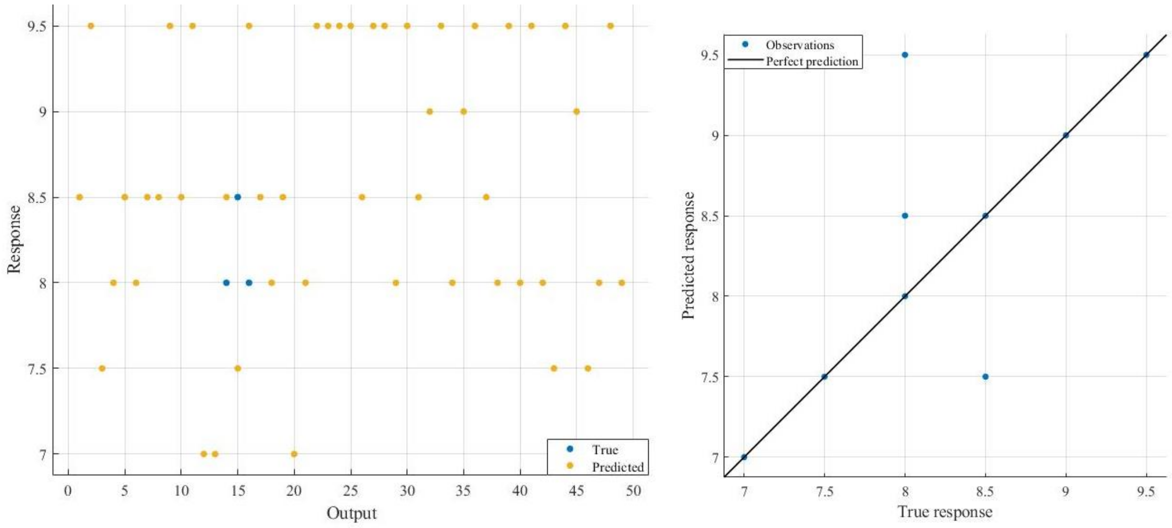
Performance Measure	**Results**
Accuracy	0.98
Sensitivity	0.96
Precision	0.97
F-Score	0.97

As a result, as seen in Figure 17, ANN had a reasonably high level of confidence in its classification. This reliability is reflected in the high accuracy value of 98%. However, more false negatives were predicted by ANN than false positives. These findings are crucial to the issue we're looking at since 98% accuracy means that people who don't have access to modern technology or complex water testing kits will be able to determine whether the water is polluted or not, much more efficiently. Increased exposure to accurate testing procedures will assist persons in determining whether water is safe and avoiding the negative repercussions of water contamination.

8.3 Validation of proposed device

According to the conventional water quality monitoring method, separate sensors are used to measure the various parameters of water quality. It can be prone to human error, and the electrodes such as Dissolved Oxygen and Electrical Conductivity are sensitive to environmental changes. The electrodes in question need careful storage, and their readings can fluctuate depending on the proper usage and storage of the electrodes. Moreover, the electrodes are expensive. These parameters come with a few limitations of using separate sensors to measure water quality which affects the sensitivity and accuracy of the results. Consequently, it becomes difficult to develop a cheap and portable device for real-time monitoring of drinking water quality. After the aforementioned critical literature review, there is a need to automate the process of real-time water quality monitoring needed for minimizing errors caused by these sensors. To overcome the limitations associated with these parameters, we have proposed a device that can measure and classify water quality into three groups using AI. 100 water samples further were collected from different locations in and around BITS Pilani campus. The water samples were analyzed in laboratory and compared using proposed device for validation. The data samples were divided into training (50 %) and validation (50 %) sets. The developed ANN model is used for training of device. The model gains accuracy with each usage as the every new reading is appended to the training data set.

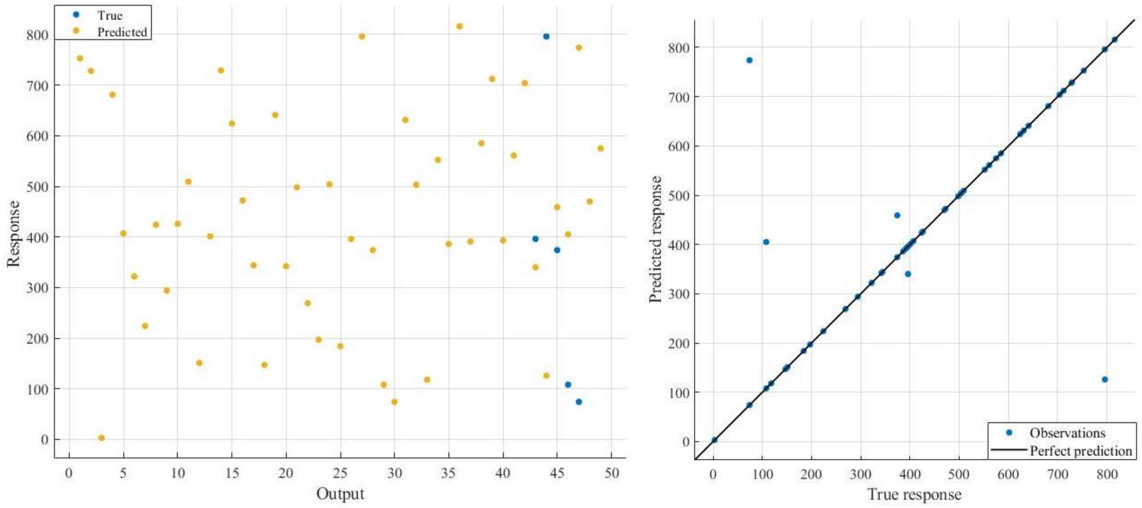
Figure 18(a) and 19(a) shows the response plot of the developed ANN model for prediction of DO and EC respectively, from training data set. Figure 18(b) and 19(b) shows the plot between predicted value and actual value DO and EC respectively. Figure 20 shows the performance graph obtained for the model; the best validation performance of the model was 0.0001 obtained at epoch 291. Error histogram is plotted in Figure 21, error histogram is the histogram of the errors between target values and predicted values after a neural network has been trained. These error values reflect how the expected values vary from the target values, and they may also be negative. The results from the error histogram show that the maximum correlations at 18.31 instances.



a) Response plot

b) Actual vs. predicted value

Figure 8.18: a) Response plot of proposed device, b) Actual vs. predicted DO value using proposed device.



a) Response plot

b) Actual vs. predicted value

Figure 8.19: a) Response plot of proposed device, b) Actual vs. predicted EC value using proposed device.

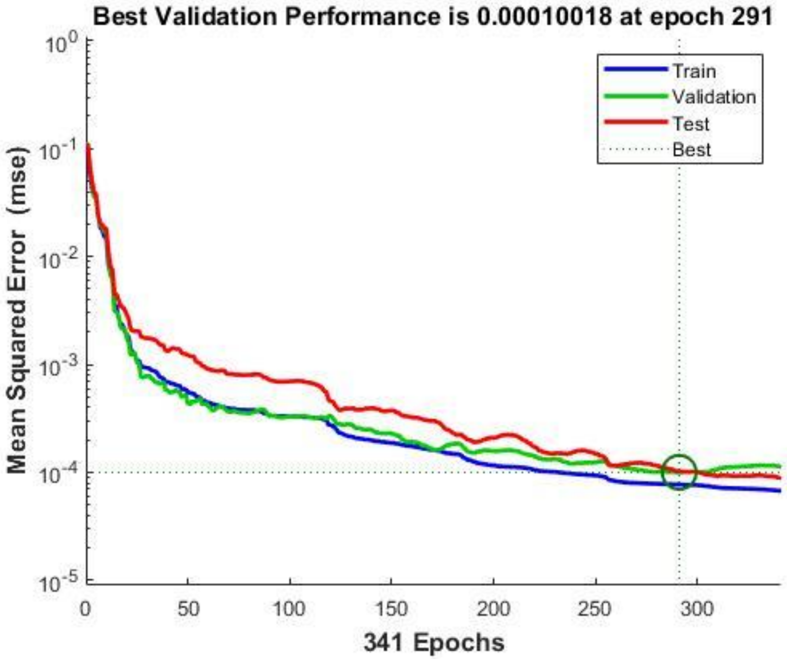


Figure 8.20: Performance of trained neural network.

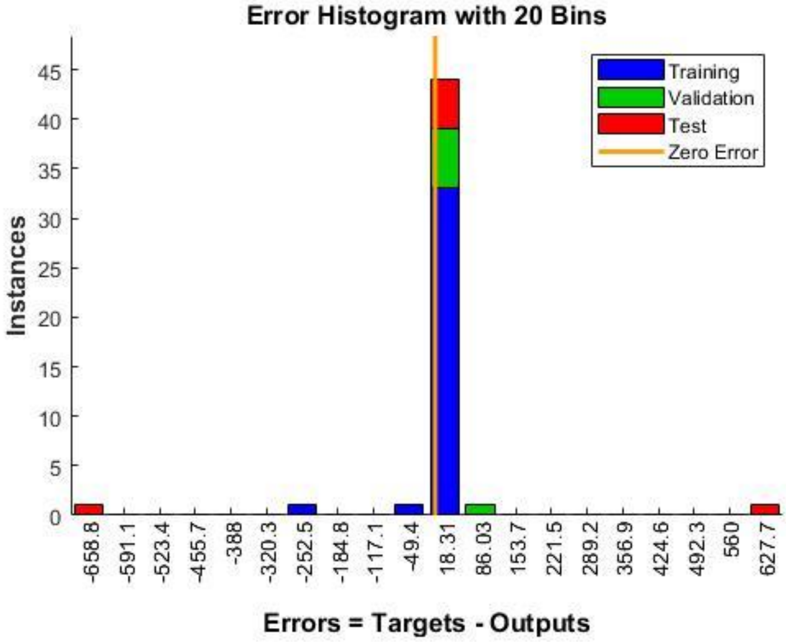


Figure 8.21: Error Histogram of trained neural network.

The remaining 50 water samples were used for validation of device performance and detection accuracy. We have compared the accuracy of our proposed model with lab-based experimental methods, as shown in Table 3.

Table 8.3: Validation of proposed device for real-time water quality measurement

Sample No	DO (Experimental)	DO (Measured using Proposed device)	EC (Experimental)	EC (Measured using Proposed device)
1	9.36	9.3	1777	1745
2	9.32	9.3	1407	1407
3	9.35	9.3	912	912
4	9.36	9.3	1450	1450
5	3.81	3.8	1640	1640
6	7.36	7.3	928	928
7	6.82	6.8	1482	1482
8	7.89	7.8	915	915
9	7.13	7.1	1525	1525
10	5.82	5.8	1225	1225
11	6.34	6.3	1560	1524
12	5.56	5.5	857	857
13	7.31	7.3	1362	1362
14	5.18	5.1	1090	1090
15	8.13	8.1	1402	1402
16	8.37	8.3	1488	1474
17	5.13	5.1	1332	1332
18	7.27	7.2	225	225
19	5.82	5.8	1175	1175
20	5.57	5.5	1036	1036
21	6.15	6.1	1082	1082
22	4.51	4.5	1190	1190
23	8.37	8.3	1180	1180
24	8.54	8.5	577	577
25	5.32	5.3	1322	1322
26	8.72	8.7	1321	1321
27	5.78	5.7	1190	1190
28	7.34	7.3	1126	1126
29	5.76	5.7	1093	1093

30	5.03	5.0	340	340
31	4.41	4.4	550	550
32	6.21	6.2	405	405
33	5.24	5.2	390	390
34	5.46	5.5	305	305
35	4.99	5.0	435	435
36	4.33	4.3	420	420
37	4.01	4.0	555	555
38	6.74	6.7	350	350
39	5.95	6.0	345	345
40	4.26	4.3	360	360
41	5.80	5.8	450	450
42	4.07	4.1	550	550
43	5.06	5.1	490	490
44	5.84	5.8	450	450
45	6.71	6.7	410	410
46	4.97	5.0	560	560
47	6.76	6.7	545	545
48	4.69	4.6	475	475
49	4.91	4.9	505	505
50	6.78	6.7	515	515

It can be seen that the results that the device proposed are accurate, reliable and the readings are repeatable. We have also done cost to performance comparison with the following conventional water testing kits available in the market- YSI Sonde (YSI Incorporated, n.d.), Labtornics LT-59, Atlas Scientific electrodes with sensor IC, Multiplexer unit and Arduino Uno board (Atlas Scientific, n.d.).

8.4 Summary

In this study, we have proposed a low-cost, portable, all weathered ANN-based smart water quality monitoring device for real-time monitoring of water quality in rural arid regions of North-Western India. This technology centers predominantly around the quality checking of water. The study presents a device to test the water quality with the goal that it intent be continuous to protect humankind from polluted water. The proposed device will check the estimation of pH, ORP, DO, and Conductivity of water and decide if the water is reasonable for the usage classes, reducing the entire package cost to 1/10th of the cost of measuring all the parameters. It will work in all environmental conditions. The device can monitor water quality automatically, and it is low in cost and does not require personnel appointments. The proposed device can be maintained and reconfigured both in terms of peripheral hardware and internal software.

The ANN designed can be changed by simple coding in python for Raspberry Pi. The attached sensors can be replaced by any other two sensors, and corresponding changes in training data can reconfigure the device for use on other parameters. The device can be used in many fields like water distribution systems, industries, agricultural fields and can also be used to measure the water quality parameters of lakes & rivers. However, the primary usage is to classify surface water quality. This study lays a foundation for real-time water quality monitoring in rural areas for people with limited or no literacy without the intervention of technical personnel. It can be concluded that AI-based methods can be used for water quality monitoring and also to control hardware costs in the development of Water Quality measuring devices. In future, we hope to elaborate our research for real-time monitoring of water quality without chemical testing methods. Given more time and resources, we could develop a system that can work more effectively upon projects such as Intel Clean Water AI.

References:

- Abyaneh, H. Z. (2014). Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *Journal of Environmental Health Science and Engineering*, *12*(1), 1-8.
- Ahmed, U., Mumtaz, R., Anwar, H., Shah, A. A., Irfan, R., & García-Nieto, J. (2019). Efficient water quality prediction using supervised machine learning. *Water*, *11*(11), 2210.
- Amruta, M. K., & Satish, M. T. (2013, March). Solar powered water quality monitoring system using wireless sensor network. In *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)* (pp. 281-285). IEEE.
- Atlas Scientific. (n.d.). *Atlas Scientific | Environmental Robotics*. (Atlas Scientific) Retrieved March 01, 2021, from https://atlas-scientific.com/#
- Badamasi, Y. A. (2014, September). The working principle of an Arduino. In *2014 11th international conference on electronics, computer and computation (ICECCO)* (pp. 1-4). IEEE.
- Central Bureau of Health Intelligence. (2018). *NationalHealth Profile 2018.* New Delhi: Ministry of Health and Family Welfare, Government of India.
- Chakraborty, U. K. (2018). Reversible and irreversible potentials and an inaccuracy in popular models in the fuel cell literature. *Energies*, *11*(7), 1851.
- Dalianis, H. (2018). Evaluation metrics and evaluation. In *Clinical Text Mining* (pp. 45-53). Springer, Cham.
- Demetillo, A. T., Japitana, M. V., & Taboada, E. B. (2019). A system for monitoring water quality in a large aquatic area using wireless sensor network technology. *Sustainable Environment Research*, *29*(1), 1-9.

- El-Shafie, A., Mukhlisin, M., Najah, A. A., & Taha, M. R. (2011). Performance of artificial neural network and regression techniques for rainfall-runoff prediction. *International Journal of Physical Sciences, 6*(8), 1997-2003.
- El-Shafie, A., Noureldin, A. E., Taha, M. R., & Basri, H. (2008). Neural network model for Nile river inflow forecasting based on correlation analysis of historical inflow data.
- Fausett, L. V. (2006). *Fundamentals of neural networks: architectures, algorithms and applications*. Pearson Education India.
- Faustine, A., Mvuma, A. N., Mongi, H. J., Gabriel, M. C., Tenge, A. J., & Kucel, S. B. (2014). Wireless sensor networks for water quality monitoring and control within lake victoria basin: prototype development. *Wireless Sensor Network, 6*(12), 281.
- Gazzaz, N. M., Yusoff, M. K., Aris, A. Z., Juahir, H., & Ramli, M. F. (2012). Artificial neural network modeling of the water quality index for Kinta River (Malaysia) using water quality variables as predictors. *Marine pollution bulletin, 64*(11), 2409-2420.
- Gopavanitha, K., & Nagaraju, S. (2017, August). A low cost system for real time water quality monitoring and controlling using IoT. In *2017 International conference on energy, communication, data analytics and soft computing (ICECDS)* (pp. 3227-3229). IEEE.
- Grubert, J. P. (2003). Acid deposition in the eastern United States and neural network predictions for the future. *Journal of Environmental Engineering and Science, 2*(2), 99-109.
- Jahan, S., Amareshwar, E., Prasad, S.V.S., & Arulananth, T.S. (2019). Raspberry PI Based Water Quality Monitoring and Flood Alerting System Using Iot. *International Journal of Innovative Technology and Exploring Engineering titled*.

- Kalpana, M. B., & Student, M. T. (2016). Online monitoring of water quality using raspberry Pi3 model B. *International Journal of Innovative Technology And Research, 4*(6), 4790-4795.
- Kumar, M. J. V., & Samalla, K. (2019). Design and development of water quality monitoring system in IoT. *International Journal of Recent Technology and Engineering (IJRTE), 7*, 527-533.
- Liong, S. Y., & Sivapragasam, C. (2002). Flood stage forecasting with support vector machines 1. *JAWRA Journal of the American Water Resources Association, 38*(1), 173-186.
- Makarynska, D., & Makarynskyy, O. (2008). Predicting sea-level variations at the Cocos (Keeling) Islands with artificial neural networks. *Computers & Geosciences, 34*(12), 1910-1917.
- Menon, K. U., Divya, P., & Ramesh, M. V. (2012, July). Wireless sensor network for river water quality monitoring in India. In *2012 Third International Conference on Computing, Communication and Networking Technologies (ICCCNT'12)* (pp. 1-7). IEEE.
- Minu.M.S, P. K. (2019). Wired Sensor Systems for Water Quality Monitoring. *International Journal of Recent Technology and Engineering (IJRTE), 8*(4), 847-852.
- Muttil, N., & Chau, K. W. (2006). Neural network and genetic programming for modelling coastal algal blooms. *International Journal of Environment and Pollution, 28*(3-4), 223-238.
- Najah, A., Elshafie, A., Karim, O. A., & Jaffar, O. (2009). Prediction of Johor River water quality parameters using artificial neural networks. *European Journal of scientific research, 28*(3), 422-435.

- Noureldin, A., El-Shafie, A., & Bayoumi, M. (2011). GPS/INS integration utilizing dynamic neural networks for vehicular navigation. *Information fusion*, *12*(1), 48-57.
- Pi-Teach, R. (2016). learn, and make with Raspberry Pi. *Raspberry Pi*.
- Prabha, D. S., & Kumar, J. S. (2016). Performance evaluation of image segmentation using objective methods. *Indian J. Sci. Technol*, *9*(8), 1-8.
- Puneeth, K. M., Bipin, S., Prasad, C., Kumar, R. J., & Urs, M. K. (2018, May). Real-time Water Quality Monitoring using WSN. In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)* (pp. 1152-1156). IEEE.
- Ranković, V., Radulović, J., Radojević, I., Ostojić, A., & Čomić, L. (2010). Neural network modeling of dissolved oxygen in the Gruža reservoir, Serbia. *Ecological Modelling*, *221*(8), 1239-1244.
- Sangjan, W., Carter, A. H., Pumphrey, M. O., Jitkov, V., & Sankaran, S. (2021). Development of a Raspberry Pi-Based Sensor System for Automated In-Field Monitoring to Support Crop Breeding Programs. *Inventions*, *6*(2), 42.
- The Nernst Equation. (2021, April 14). Retrieved August 9, 2021, from https://chem.libretexts.org/@go/page/262.
- Vijayakumar, N., & Ramya, A. R. (2015, March). The real time monitoring of water quality in IoT environment. In *2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)* (pp. 1-5). IEEE.
- World Health Organisation. (2017). *Diarrhoeal Disease.* World Health Organisation.
- YSI Incorporated. (n.d.). *Continuous, unattendend water quality monitoring sonde | 6920V2*. (YSI Incorporated) Retrieved March 5, 2021, from https://www.ysi.com/6920-V2-2.

- Yue, R., & Ying, T. (2011, March). A water quality monitoring system based on wireless sensor network & solar power supply. In *2011 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems* (pp. 126-129). IEEE.

9. Conclusions and Future scope

This chapter includes the final outcome of the work and provides the scope for future work.

9.1 Conclusions

According to the laboratory experiment based on the conventional analysis method, 24-48 hours are required before the bacteria concentration gets reported (Gautam et al., 2011). As a result of limitations associated with laboratory quantification of microbial water quality, studies have been done to develop real-time or near real-time predictive models to aid in water management decisions. At present, it is not possible to measure bacterial concentrations in water and to obtain an immediate quantitative result to evaluate and prevent human health risks. This study aims to develop a model based on laboratory experiments to predict the count of faecal coliform bacteria for cost-effective water quality management studies. This study evaluated the accuracy of an AI-based modeling approach to predict faecal coliform bacteria concentrations.

The groundwater samples are collected from eight districts of Rajasthan, India, under the BITS-UVA (University of Virginia) groundwater contamination project, containing 1302 water samples used in this study. The viable count analysis of the water samples showed *E. coli* bacterial strains with minimum cell counts of 4×10^7 CFU/100 mL and maximum cell counts of 132×10^7 CFU/100 mL. A total of 99 groundwater samples were found positive for *E. coli*. The limitation of relying on laboratory analysis to detect bacteria can be prone to human errors, affecting the model's performance and results. The majority of the existing techniques are limited to most of the substantial features of water to limit pH, temperature, turbidity, conductivity, and colour of the water. However, few significant physico-chemical parameters are not considered, which directly affect the growth of *E. coli* bacteria. To overcome the limitations, the artificial intelligence (AI) based technique is used in this study as an alternative to traditional models for predicting *E. coli* to improve accuracy, performance, and cost-effective results. This experimental data set was used to train, test, and validate the results using AI techniques.

A superposition-based learning algorithm (SLA) is proposed to observe the patterns of ANN-based sensitivity analysis for automating the prediction process of *E. coli* bacteria in groundwater. The result shows that the superposition models based on Grover's algorithm are more efficient in predicting all patterns in the counts of *E. coli* in groundwater with higher efficiency and low error. The highest correlation is observed between *E. coli* and the pH values, whereas the lowest correlation is observed with Dissolved Oxygen. It can be concluded that culture-based methods are not accurate for detecting *E. coli* bacteria in water. Further research is needed to detect the VBNC cells of bacteria in water. *E. coli* entering the VBNC condition could have a detrimental effect on public health. The number of viable cells could be underestimated. At any time, the VBNC cells could still produce toxins or be resuscitated to become virulent again and again. The study recommended that *E. coli* bacteria should not be used as an indicator organism when the cells are viable but non-culturable.

Enzymatic methods of detection are color-based methods. The amount of colour appearance can be used to determine the degree of bacterial contamination. Manually performing this process is highly time-consuming and challenging. This detection process is analytical. There is always a possibility of human error, which may result in a disastrous decision. The colours of each concentration can be scanned using conventional computer vision methods. It is, however, extremely difficult to determine the colour intensity for each concentration level. We have developed an AI-based smartphone application using CNN to capture images using an inbuilt smartphone camera and predict the bacteria in water based on color intensity. We demonstrated the effectiveness of our AI-based smartphone application by using it to monitor water quality for bacterial pollution and improve precision over laboratory results. The developed CNN model for rapid detection of *E. coli* in water achieved an accuracy of 96% and an error (loss) of 0.10. The developed model was able to predict *E. coli* bacteria in each water sample within 458ms. The approach was considerably more successful than alternative methods such as polymerase chain reaction (PCR) and traditional techniques.

The USEPA-approved gold-standard methods for detecting *E. coli* and counting viable cells are based on culturing water samples on solid agar plates or liquid media. The number of colonies is counted manually using a bacteria colony counter. Manual counting of viable bacterial cells on agar plates is time-consuming and can be prone to human error. The method requires experts to identify and count viable cells. Furthermore, due to bacterial overcrowding, high numbers of colony-forming units on a plate will lead to inaccurate results. In order to deal with the problems associated with manual cell counting, this study developed a machine-learning algorithm based on a faster region-based convolutional neural network with higher accuracy. We automated the process of *E. coli* bacteria identification using a convolutional neural network (CNN). We developed a smartphone application for the rapid detection of *E. coli* bacteria on agar plates using CNN. We also automated the process of bacteria colony counting using a faster region-based convolutional neural network (R-CNN) to overcome manual cell counting process limitations. A graphical user interface (GUI) application was created to rapidly count bacteria colony-forming units on agar plates using faster R-CNN. The developed faster R-CNN model achieved an overall accuracy of 97% and an error (loss) of 0.10. The performance of the CNN and faster R-CNN models were validated using F-score, precision, sensitivity, and accuracy statistical measures. The comparative analysis showed that the faster R-CNN model is reliable and effective in *E. coli* cell counting. The study developed a system for identifying and counting viable cells of *E. coli* bacteria in water that can be used to forecast hotspots of water contamination.

However, further validation is needed to determine the model's ability to generalize through various experiments. As a result, we want to gather more data and assess the model's capacity for counting in various situations, including higher-quality images with visible cellular components. This would also necessitate a further investigation of network architecture functionality and potentially training additional layers, which will be possible with further input data. Exploring those possibilities will be the primary objective of our future efforts. Furthermore, the performance of CNNs can be improved by tuning parameters like learning rate, epoch, and the number of layers. All these parameters affect the performance of a CNN. Image augmentation can be used to increase the data sample count using shear, zoom, rotation, and preprocessing functions. CNN model performance is also affected by overfitting and underfitting, which can be solved by training with more data, early stopping, and cross-validation.

9.2 Further scope of the work:

There is a possibility to undertake the following work in the future as an extension of this study.

- Further studies for the VBNC bacteria, which can be done using molecular technology, such as high-throughput sequencing and qPCR.
- Probabilistic superposition learning algorithm-based modeling can be done by modifying Grover's algorithm. We only need to run the neural network twice, one forward and one backward.

- Development of the low cost optical fiber based water quality monitoring system for rural communities. The system will distinguish between dangerous and harmless bacteria.
- Validation of sensors using simulated and field samples.

Journal

- Khan, F. M., Gupta, R., & Sekhri, S. (2021). Analysis of increase in cell counts of Escherichia coli in groundwater of Rajasthan: Possible presence of VBNC cells. Research Journal of Chemistry and Environment, Vol, 25, 6
- Khan, F. M., Gupta, R., & Sekhri, S. (2021). Superposition learning-based model for prediction of E. coli in groundwater using physico-chemical water quality parameters. *Groundwater for Sustainable Development*, Volume 13, 2021, 100580. https://doi.org/10.1016/j.gsd.2021.100580
- Khan, F. M., Gupta, R., & Sekhri, S. (2021). A convolutional neural network approach for detection of E. coli bacteria in water. *Environmental Science and Pollution Research*, 1-9. https://doi.org/10.1007/s11356-021-14983-3
- Khan, F. M., Gupta, R., & Sekhri, S. (2021). Automated Bacteria Colony Counting on Agar Plates Using Machine Learning. *Journal of Environmental Engineering*, *147*(12), 04021066. https://doi.org/10.1061/(ASCE)EE.1943-7870.0001948
- Khan, F. M., Gupta, R., & Sekhri, S. (2021). A novel PCA-FA-ANN based hybrid model for prediction of fluoride. *Stochastic Environmental Research and Risk Assessment*, 1-28. https://doi.org/10.1007/s00477-021-02001-4
- Portable Hand-Held Smart device for real time Water Quality Measurement and Water Quality Classification, Abheek Gupta, Farhan Mohammad Khan, Anu Gupta, Rajiv Gupta **(Under review)**
- Geospatial Analysis and Hotspots Prediction of Household Iodized Salt Using Artificial Intelligence, Farhan Mohammad Khan, Rajiv Gupta, Sheetal Sekhri **(Under review)**

Book Chapter

- Khan, F. M., & Gupta, R. (2020). Escherichia coli (*FCB*) as an Indicator of Fecal Contamination in Groundwater: A Review. In *International Conference on Sustainable Development of Water and Environment* (pp. 225-235). Springer, Cham, The 3rd International Conference on Sustainable Development of Water and Environment (ICSDWE 2020), 13-14 January, 2020, Inha University, South Korea.

Conference

- Khan, F. M., Sridhar, S., & Gupta, R. (2020). Detection of waterborne bacteria using Adaptive Neuro-Fuzzy Inference System. In *E3S Web of Conferences* (Vol. 158, p. 05002). EDP Sciences, 6th International Conference on Environmental Systems Research (ICESR 2019), 18-20 December, 2019, Melbourne, Australia. ***Best paper award***

Patent

- Portable Hand Held Smart device for real time Water Quality Measurement and Water Quality Classification, Abheek Gupta, Farhan Mohammad Khan, Anu Gupta, Rajiv Gupta, Application No.: 202111017453 ; Filed on 14.4.2021; India; Status: Awaited.

Other Publications

- Khan, F. M., & Gupta, R. (2020). ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. *Journal of Safety Science and Resilience, 1(1)*, 12-18. ***Best Research Award***
- Khan, F. M., Kumar, A., Puppala, H., Kumar, G., & Gupta, R. (2021). Projecting the Criticality of COVID-19 Transmission in India Using GIS and Machine Learning Methods. *Journal of Safety Science and Resilience*.
- Kumar, G., Kumar, A., Khan, F. M., & Gupta, R, Sprawl of the COVID-19 in changing scenario: a methodology based on social interaction, 2021, Library Hi Tech.
- Kumar, A., Khan, F. M., Gupta, R., & Puppala, H. (2020). Preparedness and mitigation by projecting the risk against COVID-19 transmission using machine learning techniques. *medRxiv*.

Brief biography of Candidate

Farhan Mohammad Khan has completed his Bachelor of Engineering (B.E) from RGPV Bhopal in 2015. He has completed Master of Engineering (M.E) from RGPV Bhopal in 2018. At present he is working as a Junior Research Fellow at Department of Civil Engineering, BITS-Pilani, India. Here his work is concentrated on prediction of faecal coliform bacteria in groundwater in Rajasthan. He has attended and delivered an oral presentation in 2019 6th International Conference on Environmental Research (ICESR 2019) in Melbourne, Australia, and it has been selected as one of the best oral presentations. He was awarded for Best Research by Science Father in 2020. He has published research articles in renowned journals and presented papers in conferences.

Prof. Rajiv Gupta is Senior Professor of Civil Engineering, at BITS, Pilani. He has completed his B.E., M.E and Ph.D. from BITS, Pilani. In his last 30 years of teaching and research, he has published more than 150 research papers in peer reviewed journals and presented in conferences in India and abroad and authored a number of books and course development material. He has guided more than 10 Ph.D. scholars apart from being involved in teaching around 30 courses, reviewed more than 125 books, project, and papers of reputed journals. His fields of interest are Water-Energy conservation, GIS and RS, Application of Artificial Intelligence, and Concrete Technology. He is involved in number of research and development projects worth more than Rs. 650 lacs of World Bank, UGC, DST, University of Virginia, and other sponsored organizations. He has also worked in different capacities of administration like Warden, Head of Department and Dean of Engineering Services and Hardware. He was instrumental in developing number of infrastructure facilities at Pilani, Goa, and Hyderabad campuses.