August 2023

Novel Non-Invasive Detection of Thin Film Biofilm and Classification of Deposits Using Machine Learning

Sachin Davis
University of Wisconsin-Milwaukee

NOVEL NON-INVASIVE DETECTION OF THIN FILM BIOFILM AND CLASSIFICATION OF DEPOSITS USING MACHINE LEARNING

by

Sachin Davis

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Engineering

at

The University of Wisconsin-Milwaukee

August 2023

ABSTRACT

NOVEL NON-INVASIVE DETECTION OF THIN FILM BIOFILM
AND CLASSIFICATION OF DEPOSITS USING MACHINE LEARNING

by

Sachin Davis

The University of Wisconsin-Milwaukee, 2023
Under the Supervision of Professor Marcia R. Silva

Clean, safe, readily available water is vital for public health, irrespective of whether it is used for drinking, domestic use, food production, or recreational purposes. Globally, around two billion people use feces-contaminated water sources, which poses a high risk to the safety of drinking water due to the high probability of water contamination. Microbial-influenced corrosion is a significant problem in several industries, including but not limited to wastewater treatment, drinking water distribution systems, food industries, power plants, paper industries, and chemical manufacturing facilities. The presence of microorganisms causes around 70% of the corrosion in gas transmission pipelines, and corrosion accounts for the loss of around 4% of the gross national product. The United States is estimated to spend around $300 billion yearly on corrosion costs. A significant amount of time is spent finding and fixing the problem with a major overhaul or part replacement, saving about 30% of overhead costs. Due to its attachment, sanitization and cleaning methods are ineffective against biofilm in its mature stage. Overall, there is a need for a rapid assessment of pipes or other structures to assist in biofilm monitoring and cleaning procedures.

The study presents and examines a fresh approach that combines non-invasive and non-destructive methods for detecting deposits in near real-time. The detection is accomplished by measuring changes in voltage and time-of-flight of ultrasound sensors and using a random forest

machine learning (ML) algorithm to categorize the deposits into four types: no deposit, biofilm deposit, scaling deposit, and corrosion deposit. This work builds a strong foundation for future novel research for the detection of biofilm using evanescent waves or multiple internal reflections [1]. Additionally, the technique is cost-effective, portable, and requires minimal power. Although random forest learning has been utilized for various classification problems, this study presents a novel application of the ML technique to classify deposits based on voltage and time of flight measurements. Unlike conventional methods like microscopic methods, combining the sensor arrangement with ML techniques allows users to make informed decisions on cleaning strategies, preventing massive biofilm buildup or other deposits in a closed wall piping system.

To

my parents,

and professors.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First and foremost, I sincerely thank my advisor, Dr. Marcia R. Silva, for cultivating a passion for giving back to society and recognizing my skills and potential to be involved in this research. I also thank my dissertation committee, Dr. Jun Zhang, Dr. Nathan Salowitz, Dr. Amit Bhatnagar, and Dr. Lucas Beversdorf, for their guidance and valuable insights at all stages of the research. I would also like to thank Nora Kodis (Sadik) for her help and support during the Howard drinking water treatment plant experiment. Special thanks to Ryan Patrick O'Day for his help and assistance in the machine-learning aspect of this research. I would also like to extend my sincere gratitude to Randolf Metzger from the UWM School of Freshwater Sciences (SFS) machine shop, who was instrumental in helping me design and build the pipe loop. Thanks to Sophia Thompson, Zenab Ali, Emma Alburg, and Eva Oklobdzija for their assistance in various stages of the research.

Last but not least, I would like to thank my parents for encouraging me to pursue my dreams and supporting me in the journey. Their prayers and support have been integral to completing the doctoral program.

Chapter 1

Introduction

1.1 Project Objectives

This research aims to develop a non-invasive and non-destructive technique to detect thin biofilm in real-time and to classify the type of deposit inside the piping system using the random forest machine learning (ML) algorithm. In the early stage of this research, several experiments were conducted with early research aimed at conducting proof-of-concept studies to determine the effectiveness of ultrasound sensors in detecting biofilm presence inside test chambers constructed from different materials mimicking a real-world piping system. The current research stage aims to test the effectiveness of ultrasound sensors in pipe loop setups at the UWM Water Technology Accelerator Laboratory at the Global Water Center and the Howard Ave Water Treatment Plant. A ground truth experiment was also conducted to train and test the dataset using a ML technique. The ML model was used to help classify the type of deposit – no deposit, biofilm deposit, or corrosion deposit in a plastic container mimicking a household water distribution system. The current research work also aims to build a strong foundation for a future novel research, which makes use of evanescent waves or multiple internal reflections for non-invasive detection of biofilm [1]. The future research stage involves the development of a portable, low-cost standalone device that can indicate the presence of any deposits in a closed-wall piping system.

1.2 Background

Clean, safe, and readily available water is vital for public health, irrespective of whether it is used for drinking, domestic use, food production, or recreational purposes. Globally, at least two billion people use feces-contaminated water sources, which poses a high risk to the safety of drinking water since it increases the probability of microbial contamination. It is believed that billions of people live in water-stressed countries, which will exacerbate based on population or climate change [2]. Historically, diseases that spread through water, such as cholera and typhoid fever, were highly concerning. However, the discovery of drinking water treatment plants and separating wastewater discharge has helped mitigate these concerns. These modern concepts have mostly eradicated the presence of *S. typhi, V. cholerae O1, and Shigella spp* and are rarely found in water distribution systems. However, pathogens such as *Escherichia coli (E. coli), Legionella spp., Aeromonas spp., Mycobacterium spp.,* and *Pseudomonas aeruginosa* can grow in water distribution systems. The emergence of these pathogens has been directly linked to a change in water usage habits – the increased use of heated drinking water and the advent of warm water reservoirs, which provides an ideal habitat for biofilm growth [3]. Different microbes can survive in distribution systems, with some capable of growing and producing biofilms. The organisms that cause diseases in healthy individuals are classified as primary pathogens, while those that facilitate infections in individuals with existing health conditions are classified as opportunistic pathogens [4]. Both primary and opportunistic waterborne pathogens have transmission routes other than water and are agents of foodborne outbreaks [5]. Table 1 shows a list of bacterial pathogens known to cause waterborne diseases and, in addition, have the potential to attach to long-term or short-term biofilms. A long-term example is *Heliobacter pylori*, which survived at least 192 hours on stainless steel coupons used to monitor biofilm build-up [6]. Two non-pathogenic *E. coli* injected

into a pilot distribution system with a biofilm (20 °C) grew slightly in the biofilm before eventually dying out [7]. It was also reported in an article by Swerdlow that an *E. coli* outbreak persisted for weeks after the contaminated water meters and main breaks were replaced or repaired. Although biofilm presence was not indicated, biofilms will probably prolong some microbes' survival [8].

Table 1: Bacterial pathogens that are the known causes of waterborne diseases. D.O. represents U.S. disease outbreaks, and CCL represents EPA's Contaminant Candidate List [8].

Organism	**Major Disease**	**Primary Source**	**DO**	**CCL**
Salmonella typhi	Typhoid fever	Human feces	Y	
Salmonella paratyphi	Paratyphoid fever	Human feces	Y	
Salmonella typhimurium	Gastroenteritis	Human/animal feces	Y	
Other *Salmonella sp.*	Gastroenteritis	Human/animal feces	Y	
Shigella	Bacillary dysentery	Human feces	Y	
Vibrio cholerae	Cholera	Human feces, coastal	Y	
E. coli	Gastroenteritis	Human feces	Y	
Yerisinia enterocolitica	Gastroenteritis	Human/animal feces	Y	
Campylobacter jejuni	Gastroenteritis	Human/animal feces	Y	
Legionella pneumophila	Legionnaires disease, Pontiac fever	Warm water	Y	
Helicobacter pylori	Peptic ulcers	Saliva, Human feces		Y

Table 2 shows the various routes of pathogen entry into distribution systems by the potential health consequences, considering the severity of the disease, probability of waterborne

disease outbreak, volume contaminated, and frequency of intrusion, recently ranked by an expert panel [9].

Table 2: Routes of pathogen entry into distribution systems [9].

Risk Level	**Pathway**
High	Treatment breakthrough, intrusion, cross-connections, main repair/break.
Medium	Uncovered water storage facilities.
Low	The central installation covered water storage facilities and purposeful contamination.

Microbial-influenced corrosion (MIC) is another significant problem in several industries, including but not limited to wastewater treatment, drinking water distribution systems, food industries, power plants, paper industries, chemical manufacturing facilities, offshore pipelines, and membrane application facilities. A study conducted by the National Bureau of Standards in 1968 found that light affected the current required to protect iron or steel against microorganisms. It was observed that the corrosion rate in the dark was much lower than the rate in indirect sunlight and required less current for specimen protection [10]. The corrosion is caused due to the removal of hydrogen from metal surfaces, which then combines with electrons reducing sulfate, forming hydrogen sulfide, commonly called the cathodic depolarization process. The presence of anaerobic sulfate-reducing bacteria in iron and steel is the most common cause of corrosion. Electron acceptance at cathodic sites due to metal dissolution from anodic sites is a classic example of an electrochemical corrosion reaction. The chemical reactions therein result in end-product removal, accelerating corrosion [11]. Microorganisms cause around 70% of the corrosion in gas transmission pipelines, and corrosion accounts for the loss of around 4% of the gross national

product (GNP) [12]. The United States is estimated to spend around $300 billion yearly on corrosion costs. A significant amount of time is spent finding and fixing the problem with a major overhaul or part replacement, saving about 30% of overhead costs [13].

The microorganisms associated with corrosion can be divided into three groups – algae, fungi, and bacteria. It was found that the fungus *Cladosporium resinae* caused corrosion in subsonic aircraft fuel tanks, which in turn caused wing perforation and loss of fuel [14]. In addition, it was also found that reclaimed water promoted corrosion in comparison to sterile water. Settled bacteria, extra-cellular polymeric saccharides (EPS), and corrosion in biofilm heavily influenced the corrosion process. Biofilm-influenced corrosion is an ever-present problem, especially in cast iron pipes used in reclaimed water distribution systems [15]. It is estimated that microorganisms were the root cause of about 40% of damages in sewer networks in a study by Kaempfer and Berndt in 1999. Approximately $100 billion was used for the repair and upkeep of private and public sewage systems in Germany, which were around 70 years old. [16]. Several other factors can cause corrosion – galvanic corrosion, pitting, and hydrogen grooving are other forms of corrosion.

Galvanic corrosion occurs when a metal is exposed to an electrolyte with different concentrations or when different metals in a common electrolyte are in contact with each other either physically or electrically. The more noble metal corrodes at a slower rate, while the active metal corrodes at an accelerated rate. This type of corrosion is a common problem in the marine industry or pipe structures in contact with saline water. Some factors that affect galvanic corrosion are temperature, humidity, and salinity [17].

Pitting is the most common and damaging form of corrosion in passivated alloys [18]. Pitting corrosion occurs due to either low oxygen or high species concentrations. In the worst case,

tiny local fluctuations will degrade the film at critical points while the most surface remains protected. The corrosion at these localized points is amplified and causes corrosion pits. This area becomes anodic, while the remaining metal becomes cathodic, resulting in a localized galvanic reaction. In extreme cases, the long and narrow corrosion pits can cause stress concentration that may cause small holes or cause tough alloys to shatter.

Hydrogen grooving is a type of corrosion observed in the chemical industry and is commonly caused due to the interaction of a pipe surface with corrosive agents, corroded pipe constituents, or hydrogen gas bubbles. When a steel pipe comes into contact with sulfuric acid, the iron in the steel reacts with the acid to form a passivation coating of iron sulfate and hydrogen gas. While the iron sulfate coating protects the steel from further corrosion, hydrogen bubbles will remove this coating, and the traveling bubble exposes more steel to the acid, causing a vicious cycle [18].

Biofilms are also a significant threat in the food industry, significantly affecting the quality and safety of dairy products. Several perishables (e.g., cheese and butter) and semi-perishable (e.g., casein and milk powder) foods are manufactured in the dairy industry. The dairy industry adheres to strict microbiological guidelines to maintain the products' stability, flavor, and functionality. The source of contamination can occur at any stage of dairy processing. Below are some of the familiar sources of contamination and some of the possible reasons [19].

- Milking – contamination could occur due to bacteria on the udder or biofilm in the machines used for milking.
- Transportation and Storage – contamination is possible due to bacteria in transfer lines, storage vessels, or improper refrigeration techniques.

- Processing – contamination is possible due to improper pasteurization or improper manufacturing techniques.

Although there is limited data on food poisoning, it is believed that bacteria, fungi, viruses, animals, plants, and chemicals are the major contributing factors to food-induced illnesses, not including allergies. Table 3 shows the food-induced illness data in the USA. It can be observed that *Salmonella* non-typhi, *Campylobacter* spp., and *Staphylococcus aureus* are the most common causes of food poisoning in the United States. These strains of bacteria recorded the highest number of cases and deaths in a study by Snyder in 1995 [20]. Overall, there is a need to provide manufacturers with a rapid assessment of their plants to assist in monitoring the effectiveness of cleaning procedures. The techniques for detecting biofilms involve detecting bacterial molecules, proteins, or polysaccharides on surfaces or in water flushed through the pipes [21].

Table 3: Data showing food-induced illness in the USA due to bacteria [20].

Cause	**Cases**	**Deaths**
Bacteria		
Staphylococcus aureus	8,900,000	7,120
Streptococcus (Group A)	5,000,000	175
Salmonella non-typhi	3,000,000	2,000
Campylobacter spp.	2,100,000	2,100
Clostridium perfringens	650,000	6-7
Shigella spp.	300,000	600
Escherichia coli	200,000	400
Vibrio (non-cholera)	30,000	300-900
Listeria monocytogenes	25,000	1000

Biofilm formation is a phenomenon that occurs in artificial and natural environments. Figure 1.1 shows the attachment of *Staphylococcus aureus* biofilm on an indwelling catheter, proving that bacterial adhesion can occur inside the human body if proper precautions are not taken. The microorganisms' physiological status and surface material type also contribute to bacterial adhesion. If the surface is rough, the colonized areas are protected from the effects of shear stress, turbulent flow, and biocide activity. The EPS strengthens the bacterial adhesion and captures other bacterial species forming a second layer. The strength of the bacterial adhesion is increased a hundred-fold against biocide treatment due to this new layer. Besides the resistance to biocides, the new EPS layer also increases heat treatment resistance [22].

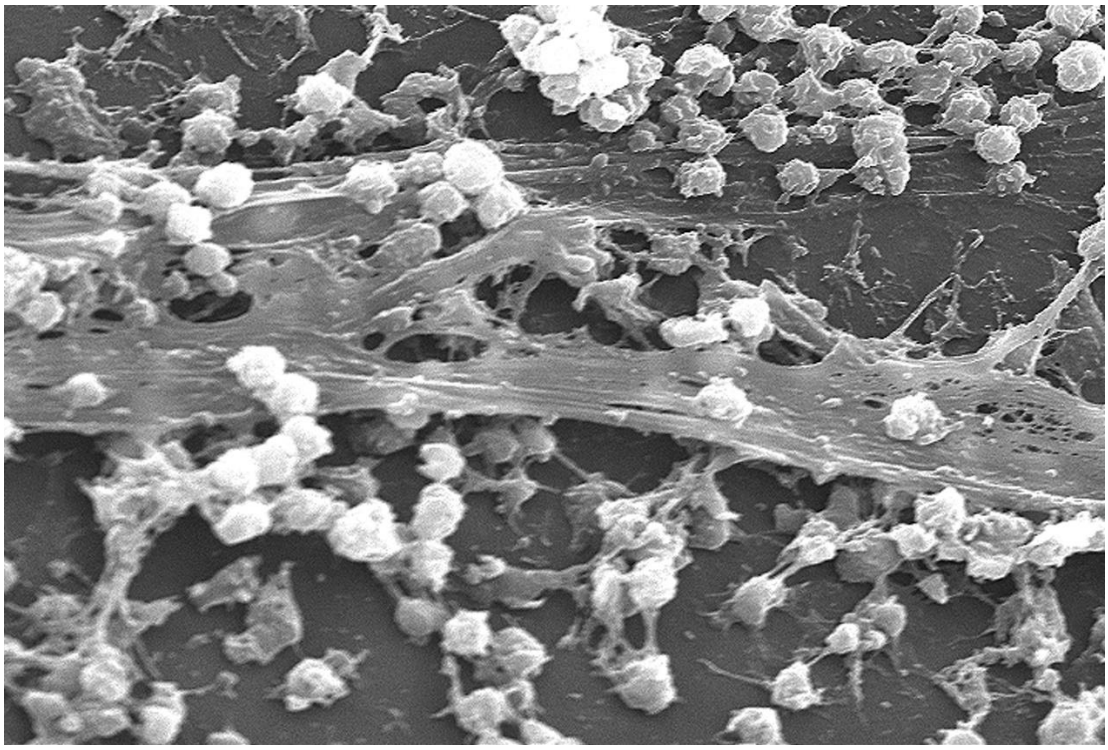

Figure 1.1: *Staphylococcus aureus* biofilm on an indwelling catheter [23], Public Domain Image.

Bacterial attachment is a severe problem in many industries. Within the biofilm are continuous growth, multiplication of bacteria, and active release of bacterial cells into the environment, leading to subsequent product contamination. Once the biofilm is established,

sanitization and cleaning become extremely difficult. Sanitizers and detergents cannot penetrate the EPS matrix to destroy bacterial cells. Prevention is the best method to eliminate biofilm build-up [24]. However, prevention is not always feasible, as the type of biofilm to cause a build-up is difficult to predict. Early detection of biofilm is the best way to ensure proper sanitization and cleaning since detergents and sanitizers are effective at an early stage of biofilm maturation. The gold standard for detecting biofilm is the Tissue Culture Plate (TCP) method, which Christensen et al. first introduced in 1985 [25]. All the other invasive and non-invasive methods used to detect biofilm are described below.

1) **Tissue Culture Plate (TCP) method**

   Britannica describes tissue culture as transferring animal or plant tissue fragments (a single cell, a population of cells, or a whole or part of an organ) to an artificial environment where they survive and function [26]. The TCP method involves screening isolates for their ability to form biofilm. Isolates from agar plates are inoculated in respective media for 18 hours at 37 °C and diluted 1/100 with fresh medium. Figure 1.2 shows a cultured animal cell growing in a growth medium. The optical density (O.D.), considered an index of biofilm-forming capacity and adhesion of bacteria to surfaces of stained adherent bacteria, is analyzed using a micro-Enzyme-Linked Immunosorbent Assay (ELISA) auto reader (Bio-Rad, model 680) at a wavelength of 570 nm. The O.D. value is directly proportional to the strength of bacterial adherence and the capacity of cells to form biofilm. A higher O.D. value means the isolate can form biofilms and adhere solidly to surfaces [27].

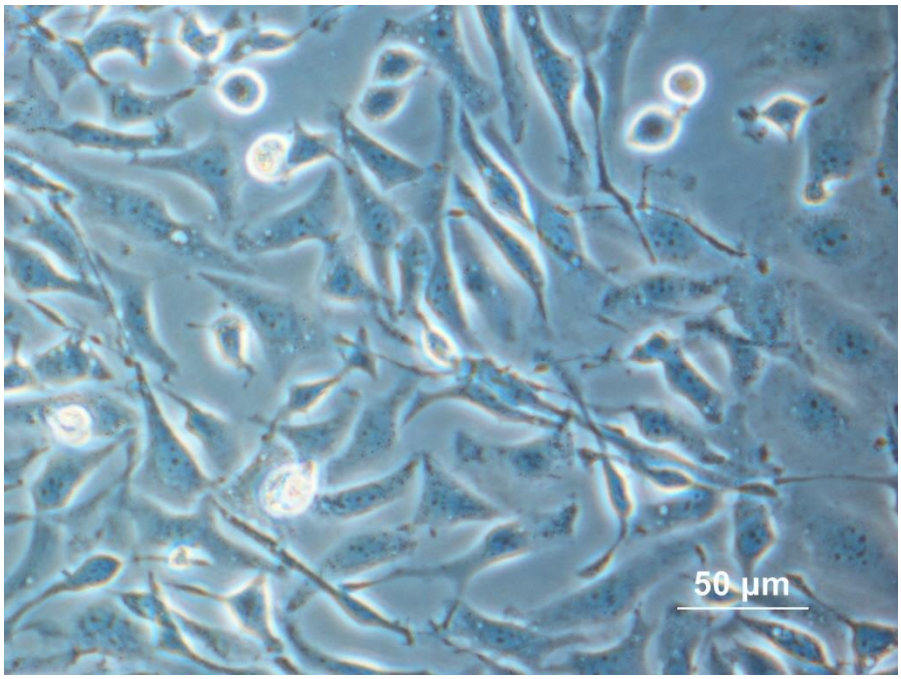

Figure 1.2: Adherend Chinese hamster ovary cells in a cell culture flask observed under the microscope [28], Public Domain Image.

2) Tube method (T.M.)

The Tube method, described by Christensen et al. [29], is a qualitative assessment technique of biofilm formation. Tubes are incubated for 24 hours at 37 °C after inoculating with a loopful of microorganisms from overnight culture plates, washed with phosphate buffer saline, dried, and stained with crystal violet. After removing the excess stain, the tubes are washed with deionized water and dried in an inverted position to observe biofilm formation. A visible film on the walls or the bottom of the tube indicates the presence of biofilm, and the amount of biofilm formation is scored as 3 - strong, 2 - moderate, 1 - weak, and 0 - no biofilm formation [27].

3) Congo red Agar method (CRA)

Freeman et al. [30] described the CRA method as an alternative for screening biofilm formation by *Staphylococcus* isolates requiring specially prepared brain heart infusion broth supplemented with 5% sucrose and Congo red. Congo red is prepared separately and added to the agar plates, then inoculated and incubated aerobically for 24-48 hours at 37 °C. Dry crystalline black colonies indicate a positive result for biofilm-forming capacity, while dark-centered pink colonies indicate weak biofilm-forming capacity [27].

4) DAPI method

The DAPI method was first formulated by Porter and Feig in 1980 and used a particular DNA stain - 4'6-diamidino-2-phenylidole (DAPI) to detect a nucleic acid. The DAPI produces a bright blue, fluorescent glow in direct proportion to the cellular content when excited by light at a wavelength of 365 nm. An example of the blue glow can be seen in Figure 1.3, where specific nuclei of endothelial cells were stained with DAPI. Any material other than the DNA molecules appears pale yellow when stained with DAPI. DAPI-stained cells become visible more than the limit of light microscopy resolution. The staining time required for DAPI is relatively short and can be stored for up to 24 weeks at 4 °C as it does not fade as commonly as other fluorochrome dyes [31].

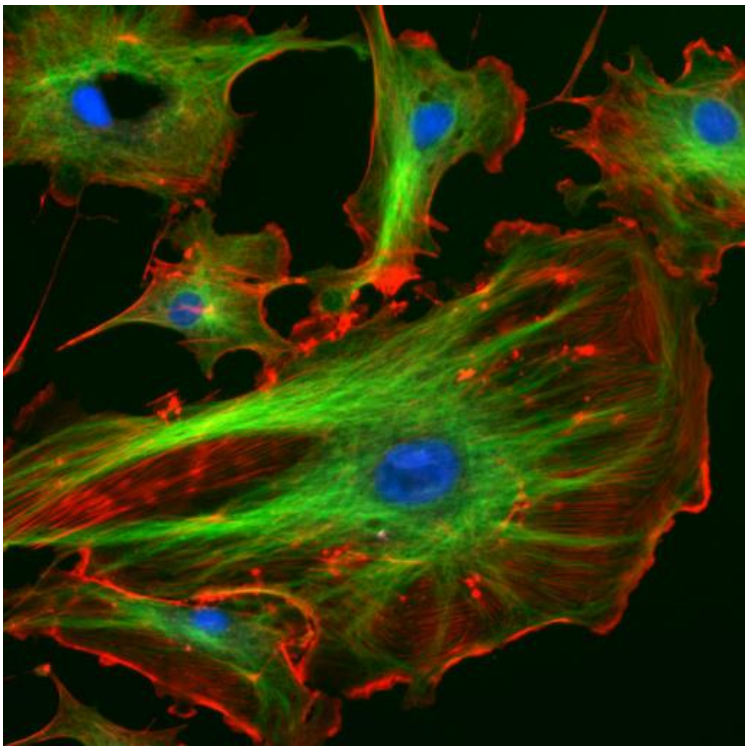

Figure 1.3: Endothelial cells as observed under the microscope. The nuclei stained with DAPI are blue [32], Public Domain Image.

5) Standardized Single-Disc Method

Wilkins et al. developed a single-disc diffusion technique combined with the incorporation of inoculum in the pour method to determine the susceptibility of anaerobic bacteria. The authors presented a modified method of the standard Bauer-Kirby procedure [33] commonly used for aerobic pathogens. This method aims to develop a standardized technique for susceptibility testing of anaerobic bacteria. An 18 – 24-hour bacteria culture is added to a cooled Brain Heart Infusion Supplemented (BHI-S) agar medium and solidified at room temperature. The plates are then placed into an anaerobic jar after placing antibiotic discs. The jars are placed in the incubator and left undisturbed at 37 °C for 18 – 24 hours. Zones are observed against a bright black background using a high-intensity lamp, and the zone diameters are measured using a ruler. An area of inhibition is seen

outside the hazy light growth in an inner area of the antibiotic disc for some strains, and the outer zone of inhibition is measured. The diameter of the area with inhibited growth demonstrates the antibiotic's effect and the resistance of bacteria to antibiotics [34]. Figure 1.4 shows the growth of bacteria isolated from a shark in the presence of an antibiotic disc.

Figure 1.4: Standard Kirby – Bauer testing. White antibiotic discs are placed on agar plates of bacteria. Poor bacterial growth zones indicate antibiotic susceptibility [35], Public Domain Image.

6) Fluorescence "In Situ" Hybridization (FISH)

Biomedical researchers P. R. Langer-Safer, M. Levine, and D. C. Ward developed fluorescence *in situ* hybridization (FISH) in 1982 as a molecular cytogenic technique using fluorescent probes binding specific parts of nucleic acid sequences with a high degree of complementarity. This technique can detect the presence or absence of specific DNA sequences on chromosomes [36]. The technique can also be used for pathogen

identification in medical microbiology. FISH allows identifying a wide range of pathogens, and the diagnosis time is shorter than many biochemical differentiation techniques. This technique is most commonly used when there is a need for immediate identification, especially identification of blood cultures. FISH can be considered an easy and economical technique for rapid preliminary diagnosis. Figure 1.5 shows a simple schematic of the FISH technique, which binds chromosomes or their portions with fluorescent molecules and is helpful for the identification of chromosomal abnormalities and gene mapping. Comparing two biological species to identify evolutionary relationships using FISH is possible. FISH is widely used to identify microorganisms, including biofilm and complex multi-species bacterial organizations. A single DNA probe can be used to visualize the distribution of specific species within the biofilm. Preparing multiple probes for two species allows the visualization of multiple species in the biofilm and the determination of the architecture of the biofilm [37].

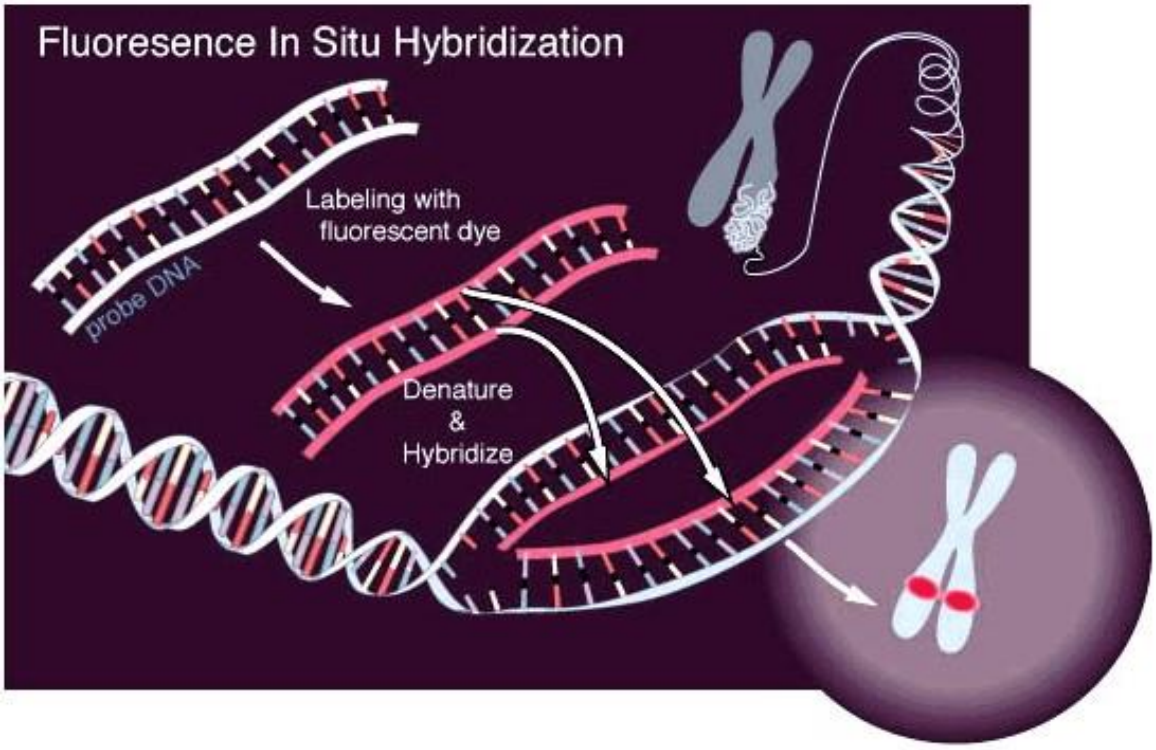

Figure 1.5: A simple schematic of the fluorescence in situ hybridization (FISH) technique [38], Public Domain Image,

7) Surface-enhanced Raman scattering (SERS)

Surface-enhanced Raman scattering (SERS) imaging is a technique for the chemical characterization of biological systems and has high sensitivity, can be applied in aqueous environments, and yields informative spectra. Silver or gold nanoparticles are used for the *in situ* SERS analysis. Silver colloids were used in the SERS imaging technique developed by Ivleva et al. in 2010. These colloids were prepared at room temperature by reducing silver nitrate with hydroxylamine hydrochloride at alkaline pH and can be stored in a dark and cool (4 °C) place for up to three weeks. Figure 1.6 shows a simple schematic for the SERS technique, where the analyte is mixed with gold or silver nanoparticles, a laser is reflected on the microscope slide with specific wavelengths, and the resultant SERS signals

are analyzed to determine the characteristics of the biofilm. The Renishaw 2000 Raman microscope with a He-Ne laser and a wavelength of 633 nm was used in the analysis. Compared to confocal laser scanning microscopy, Raman microscopy does not require staining, provides chemical information about complex biofilm matrixes, and is a non-destructive technique for biofilm analysis [39].

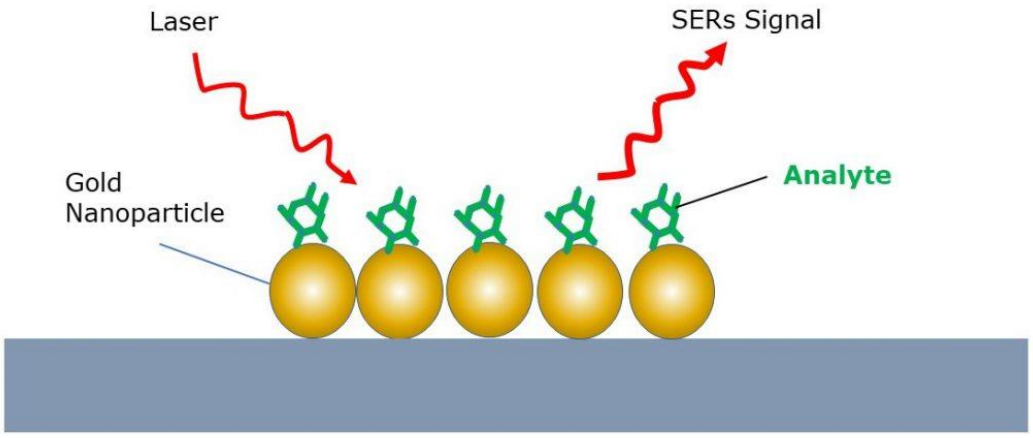

Figure 1.6: An example schematic of the Surface-enhanced Raman scattering technique [40], Public Domain Image.

8) Hyperspectral Microscope Imaging (HMI)

The spatial and spectral information provided by the Hyperspectral microscope imaging (HMI) method is a solid optical detection technique for foodborne pathogens. Figure 1.7 shows a sample hyperspectral data of the Earth captured by the National Aeronautics and Space Administration (NASA). The gold standard for detecting foodborne pathogens is the conventional microbiological method for cell counting. However, this method requires extensive labor and a long time, from days to weeks. Compared to this method, the HMI is very sensitive and is a rapid pathogen detection method. The HMI method proposed by Park et al. in the SPIE Defense conference involved using a Nikon upright microscope with an acousto-optic tunable filter (AOTF), high-performance cooled Electron multiplying

CCD, 16-bit camera, and dark-field illumination lighting sources. The AOTF-based hyperspectral microscope imaging method can be used to characterize the spectral properties of *Salmonella enteritidis* and *E. coli.* Since no standard protocol exists for hyperspectral microscopy, numerous combinations of imaging (reflectance and transmittance) and illumination (brightfield, darkfield, phase contrast, and autofluorescence) are required to reduce data variation from the imaging method. Additionally, multiple image acquisitions at varied wavelengths must be utilized to reduce random noise [41].

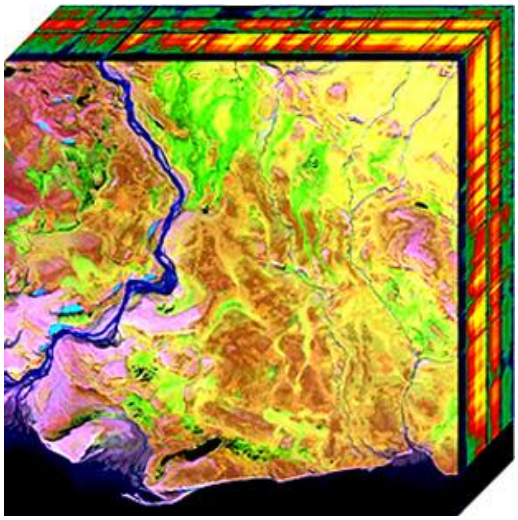

Figure 1.7: A graphic representation of hyperspectral data [42], Public Domain Image.

Table 4: Advantages and limitations of microscopic and spectrometric methods for detecting biofilm or corrosion deposits [43].

Techniques	**Advantages**	**Limitations**
Confocal laser microscopy (CLM)	• *In situ* analysis of biofilm structure. • Visualization of corrosion formation.	• Staining methods of bacteria may alter growth conditions. • Limitations of focal length on the volume of media measured.
Fourier transform infrared spectrometry (FTIR)	• *In situ* analysis of biofilm composition.	• The presence of water absorbs infrared light. • It is unable to distinguish dead cells and living cells.
X-ray photoelectron spectroscopy (XPS)	• Analysis of changes in chemical states of surfaces. • It does not require a considerable amount of products.	• The lack of water or dehydration changes the chemical states of the products. • The chemical states determined by XPS require confirmation by a secondary technique.
Auger electron spectroscopy (AES)	• Analyze smaller areas in comparison with XPS.	• The high energy density of electrons can cause more radiation damage than XPS.
Extended X-ray absorption fine structure (EXAFS)	• It provides information about the molecular structure of biofilms.	• X-ray energy might be too high for *in situ* monitoring.

Atomic force microscopy (AFM)	• Visualization of the topography of biofilm and corroded samples.	• It does not provide compositional information.
Scanning electron microscopy (SEM) and energy dispersive X-ray (EDX)	• Straight-forward technique. • Failure analysis of corroded materials.	• It requires an abundant amount of corrosion. • Samples need cleaning or conductive coating.
Electrochemical impedance spectroscopy (EIS)	• It can detect pitting initiation.	• The presence of biofilm can complicate and convolute the measurement.
Quartz crystal microbalance	• It indicates corrosion formation by detecting weight changes.	• The presence of biofilm can overshadow the dissolution of a metal film.

Table 4 represents the advantages and limitations of microscopic and spectrometric methods for detecting biofilm or corrosion. While some methods described in the table are used for detecting both corrosion and biofilm, some methods are used exclusively in detecting either biofilm or corrosion. Water testing is the most effective and straightforward method for detecting hard water scaling. Water testing usually involves a kit with water test strips and a color chart corresponding to the water's hardness measured in grains per gallon, milligrams per liter, or parts per million. A high grains per gallon value indicates that the water is tough, with a high concentration of calcium and magnesium indicating a solid presence of hard water scaling [44].

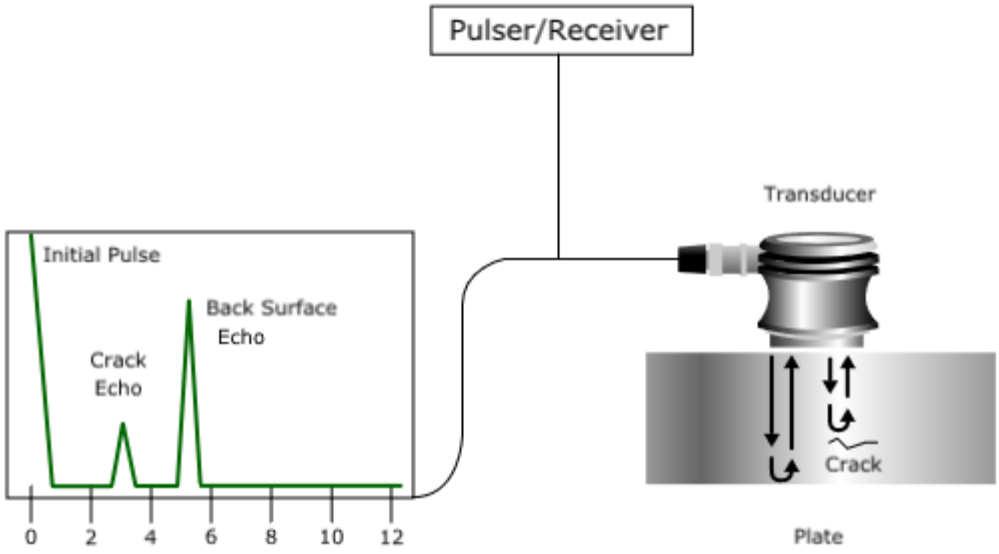

Figure 1.8: Ultrasonic detection technique illustrated by a single ultrasound sensor and a solid surface with an internal crack and an oscilloscope screen portraying signals showing information about the approximate position and size of the defect [45].

The most modern method for non-invasive biofilm or corrosion detection is ultrasound sensors. Lazzaro Spallanzani proved that bats could navigate accurately in the dark through echo reflection from high-frequency inaudible sounds early in 1793. Richardson invented the echo locator in 1912 based on the idea of ultrasound used for navigation and detection of objects in the water. The beginning of Sonar and ultrasound for medical imaging is traced back to the sinking of the Titanic. Within a month of the Titanic tragedy, British scientist L.F Richardson (1913) filed patents to detect icebergs using ultrasound. French scientists Chilowski and Langevin started developing a device to detect submarines using Ultrasound during World War 1 [46]. Ultrasound techniques have been used to detect cracks in solid surfaces for several years. An ultrasound sensor is placed on one side of the solid surface, ultrasound signals are passed through the surface, and the reflected signals are recorded with the help of an oscilloscope. When there is no crack or defect inside the solid surface, the wave reflects the sensor after a delay. The presence of a crack or a

defect will cause a second wave detected at an earlier interval with a smaller amplitude before the echo of the ultrasound waves, as seen in Figure 1.8 [47]. Since rail inspections were initially performed, ultrasound sensors have been used to detect internal defects on railway tracks. Visual inspections were found to be ineffective as they are unable to detect internal defects due to their simple nature. Ultrasound techniques have since been used to detect internal defects with approval from the National Transportation Safety Board [48]. A typical UT inspection system consists of several functional units, such as the pulser/receiver, transducer, and display devices. A pulser/receiver is an electronic device that can produce high-voltage electrical pulses. The transducer generates high-frequency acoustic waves, which propagates through materials and part of the waves is reflected due to flaws in the material. The reflected wave is transformed into an electrical signal, and is displayed on a screen, as shown in Figure 1.8. The reflected signal strength is displayed versus the time from signal generation to when an echo was received, and the signal can sometimes be used to gain information about the features of a defect [49]. Ultrasound has been widely used in industrial applications to detect structural defects and provide biomedical imaging of cells, tissues, and organs. Ultrasound is now a valuable and flexible modality in medical imaging and often provides an additional or unique characterization of tissues. An ultrasound transducer sends an ultrasound pulse into tissue and receives echoes back. The echoes contain spatial and contrast information. The concept is analogous to sonar used in marine applications, but the technique in medical ultrasound is more sophisticated, gathering enough data to form a rapidly moving two-dimensional grayscale image. Some characteristics of returning echoes from tissue can be selected to provide additional information beyond a grayscale image. Doppler ultrasound, for instance, can detect a frequency shift in echoes and determine whether the tissue is moving toward or away from the transducer. This technique is invaluable for evaluating some structures,

such as blood vessels or the heart (echocardiography) [50]. A. Aubry et al. developed an experimental setup that uses an array of sources/receivers placed before the medium. The impulse responses between every couple of transducers were measured and formed into a matrix. Single-scattering contributions exhibit a deterministic coherence along the antidiagonals of the array response matrix, whatever the distribution of inhomogeneities. This property is taken advantage of to discriminate single from multiple-scattered waves. Experimental results were observed with ultrasonic waves in the MHz range on a synthetic sample (agar-gelatin gel) and breast tissues. The authors found that the multiple scattering contributions are negligible in the breast, around 4.3 MHz [51]. The attenuation of sound waves and the dispersion of waves in cancellous bones in humans were studied with the help of ultrasound. The experiments were performed with a bone model miming phantom and human cancellous bones. The experiment focused on analyzing the physical mechanisms of ultrasonic wave propagation in a cancellous bone that governs phase velocity and attenuation coefficient as a function of frequency and porosity [52]. The properties of a liquid, such as viscosity and absorption, are significant for acoustic investigations because these factors affect the proper choice of the measuring method and temperature−pressure conditions. Ionic liquids (ILs) are generally much more viscous than conventional molecular organic liquids, i.e., the viscosity values of most ILs at room temperatures are two to three orders of magnitude larger than almost all molecular organic liquids. The propagation terms in most ILs are rather like those in highly associated viscous polyhydroxy liquids compared to those in low-viscous conventional molecular organic liquids [53]. The basic principle behind an ultrasound sensor is the piezoelectric effect which converts one form of energy into another, especially mechanical energy, into electrical energy and vice-versa. The Curie brothers first discovered this effect in 1880. Piezoelectric is derived from the Greek word 'piezo,' which means pressure. An electric

charge can be applied to piezoelectric crystals creating deformations in the crystal and converting the electric signal into a pressure signal. This effect is commonly seen in piezoelectric speakers. Any mechanical deformations in the crystal can contribute to an electric charge. This effect is commonly seen in microphones [54]. Piezoelectric sensors are susceptible, and the piezoelectric effect is used in many applications involving generating and detecting sounds like sonar, microphones, and electronic frequency generation. The piezoelectric effect of change in polarity during the compression and stretching of the plates is observed in Figure 1.9.

Some of the most common applications of piezoelectric transducers are as follows [55]:

- Diagnostics and ultrasonic imaging in the field of medicine and infertility treatments.
- Electric lighters – the sudden electric signal, due to the pressure applied to the piezoelectric sensor, causes the fire.
- Seatbelt lock in response to rapid deceleration.
- Automatic door opening systems.
- Microphones and speakers.

The advantages of piezoelectric transducers are as follows:

- No external force is needed for the operation.
- Compact, portable, and reasonably easy to use.
- Parameters change rapidly due to the sensitive nature of the piezoelectric sensor.

The limitations of piezoelectric transducers are as follows:

- Measured parameter values can vary with temperature changes.
- Due to the low voltage, external circuitry may be necessary depending on the application.
- Under static conditions, measurements may not be suitable.
- Desired shape and strength cannot be defined for the ultrasound material.

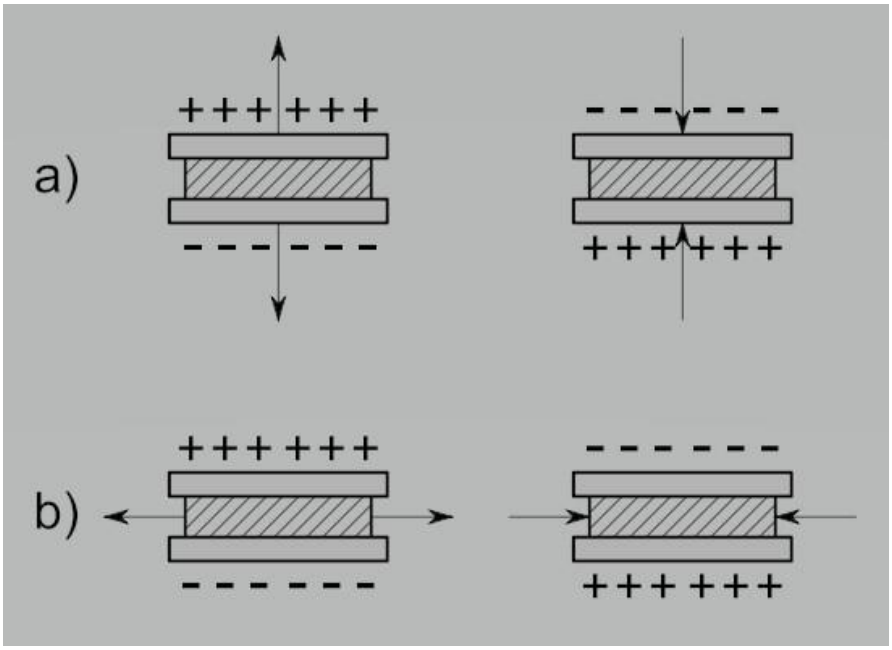

Figure 1.9: A graphic representation of the piezoelectric effect occurring during the compression and stretching of a piezoelectric plate. (a) The effect is observed when the plate is stretched and compressed along the X-axis. (b) The effect is observed when the plate is stretched and compressed along the Y-axis [56], Public Domain Image.

This research aims to use a non-invasive and non-destructive method to detect biofilm with the help of commercially available 1 MHz ultrasound sensors. In the early stage of the research, it was found that ultrasound sensors can detect various daily-use objects like plastic bags, print-quality paper, and household aluminum foils. The sensors were also tested on agar coating to identify the range of measurements for the research. It was found that voltage and time of flight measurements from the ultrasonic sensor arrangement can detect materials with thicknesses more significant than 40 µm. It was also observed that the technique could detect internal deposits in different materials like Polyvinyl Chloride (PVC), Copper and Galvanized Iron [47]. The biofilm detection method involves using two commercially available 1 MHz sensors attached on the same side of a testing chamber or piping utilizing multiple internal reflections of acoustic waves. In the early stage of this research, we explored using the sensor arrangement in detecting everyday

materials, PolyHEMA, and *E. coli*. It was observed that this sensor arrangement would require high-power equipment to observe and analyze data. However, an alternative sensor arrangement with a transmitter on one side and a receiver on the opposite side could be used to detect the presence of deposits. Some everyday materials used were A4-sized paper, household aluminum foil, and Ziploc bags. As the thickness of the test object increased, the voltage measurement observed on the oscilloscope decreased. It was concluded that the sensor arrangement could detect biofilm of thickness greater than or equal to 40 µm in a closed-loop piping system [57]. The sensitivity of the sensors to changes in environmental parameters such as temperature, liquid concentration, turbidity, and conductivity in the piping system was tested to understand the effectiveness of the sensor arrangement. In addition, the response of the sensors to various frequencies was also tested to understand the best operating frequency of the ultrasound sensor for biofilm detection. The experiment concluded that the sensors are sensitive to the change in turbidity and conductivity of liquids in the piping system, which allows the sensors to detect non-attaching foreign deposits in the liquid flow within the piping system. Compared to the other sensors tested, 1 MHz ultrasound sensors provided more voltage range for data analysis, making them a good candidate for biofilm detection [58]. Using two ultrasound sensors allows the signals to attenuate less than the single-sensor arrangement because the signals travel half the distance [47]. In addition, using ML with a non-invasive, non-destructive biofilm detection technique is a newer area that has not been explored much but has great potential. Table 5 illustrates some of the existing non-invasive biofilm detection techniques and their limitations compared to the dual sensor arrangement in this research and our previous research.

Table 5: Strengths and limitations of existing non-invasive techniques for detecting biofilm.

Technology	**Strengths**	**Limitations**
Ultrasonic monitoring of early-stage biofilm on polymeric surfaces [59]	• Fast Results.	• Single sensor using reflection method. • They have a limited detection range. • High frequency can destroy biofilm allowing it to disperse into the liquid.
High-frequency ultrasound imaging of single-species biofilm [60]	• Fast Results.	• Single sensor using reflection method. • They have a limited detection range. • High frequency can destroy biofilm allowing it to disperse into the liquid.
Novel acoustic sensor for early detection of biofouling [61]	• Fast Results. • Non-destructive technique.	• Single sensor using reflection method. • Signals undergo higher attenuation due to the use of single sensors.
Device and method for detecting deposits [62]	• Fast Results. • Non-destructive technique.	• Single sensor using reflection method. • A secondary device (Light Addressable Potentiometric Sensor) is required to confirm biofilm presence. • Expensive equipment is needed.
Device and method for detecting and analyzing deposits [63]	• Fast Results. • Non-destructive technique.	• Single sensor using reflection method. • A secondary device (Temperature Sensor) is required to confirm biofilm presence. • Expensive equipment is needed.

In addition to detecting biofilm, characterization and classification of the deposit helps the user make crucial decisions on the corrective strategy. ML techniques would be the best method for classifying and characterizing foreign deposits. ML focuses on improving the performance of computers in the execution of different tasks by leveraging the data. The applications of ML range from data mining programs used to detect fraudulent credit card transactions to information filtering to understand users' reading behavior and autonomous vehicles. An ML algorithm's formation depends on its ability to answer the following questions [64].

- Which algorithm would have the best performance in solving the problem?
- What is the amount of training data required?
- What is the benefit of prior knowledge in the selection of ML algorithms?
- What is the amount of tasks the algorithm needs to learn?
- What specific functions of each task should the ML algorithm learn?
- Can the process be automated?
- Can the learner improve the ability to represent and learn the target function?

ML is often seen as a broad subfield of artificial intelligence. Arthur Samuel, a pioneer in artificial intelligence and computer gaming and an employee of IBM, coined 'machine learning' in 1959 [65]. The ML approach can be classified into three broad categories depending on the signal or feedback available to the learning system [66].

- Supervised learning: This method aims to map the inputs to the outputs. An example of this method is a teacher presenting data with example inputs and their desired outputs to the computer.

- Unsupervised learning: This method aims to discover the hidden patterns in data or be used to approach the end of the learning process. In this method, the algorithm does not require any labels and should be capable of finding structure in the input provided.
- Reinforcement learning: This method aims to navigate the problem space by constant feedback analogous to its rewards which the algorithm tries to maximize. In this method, the algorithm interacts with a dynamic environment to meet a specific goal. An example of this method is the autonomous driving vehicle or an autonomous opponent in games.

Detecting biofilms and classifying foreign deposits uses a supervised learning approach where data sets containing inputs and the desired outputs are required. This data is known as training data [67]. The training data is often represented as a matrix, and an array of vectors represents each training example, called a feature vector. Supervised learning algorithms can predict new inputs' outputs with iterative objective function optimization [68]. A learned algorithm improves the accuracy of its predictions over time. The optimal algorithm will be able to detect the outputs for in'uts that were not included in the training data [69]. The most predictive approach used in data mining, statistics, and ML is a decision tree that can visualize the decision-making and decisions. The target variables use a discrete set of values called classification trees in which leaves represent class labels and branches represent the features that lead to the class labels [70]. Figure 1.10 shows an example of the decision tree indicating the survival probability of passengers on the Titanic.

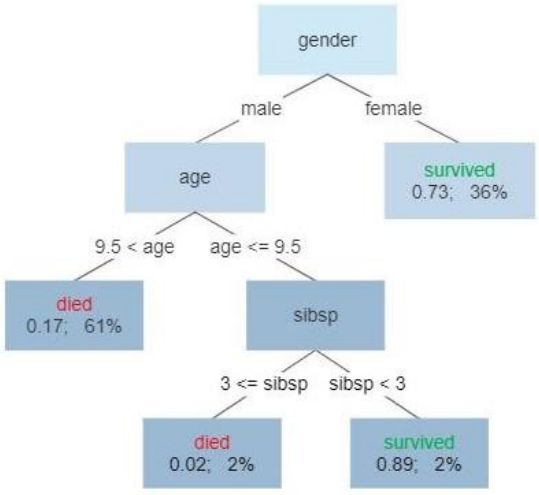

Figure 1.10: A decision tree showing the probability of survival and the percentage of observations of passengers on the Titanic. Sibsp represents the number of siblings or spouses aboard [71], Public Domain Image.

The advantages of the decision tree analysis are as follows [72]:

- Easy to interpret and understand.
- Can handle numerical and categorical data.
- Easy and quick data preparation.
- It can be compared to human decision-making closely.
- Robust against co-linearity.
- It can handle large datasets.
- Built-in feature selection.

The most significant limitation of the decision tree model is the overfitting problem which may fail to fit additional data or predict future observations [73]. Poor generalization of samples and overfitting of the training data is the risk of large trees. A small change in the training data significantly changes the tree and the final predictions. A small tree, on the other hand, does not capture structural information about the sample space. The problem of overfitting can be eliminated using the pruning method (data compression technique used in search algorithms and

ML), which reduces the size of decision trees by removing redundant and non-critical tree sections. Pruning is most effective when the tree nodes contain fewer instances, and removing some nodes does not interfere with the model's accuracy [74]. Tin Kam Ho proposed the general method of random decision forests in 1995. In this method, the splitting with hyperplanes can allow the forests of trees to grow without suffering from overtraining, as the model is sensitive to selected feature dimensions [75]. Leo Breiman properly introduced a method of building a forest of uncorrelated trees using a decision tree method combined with random node optimization and bagging. The method uses the out-of-bag error to estimate the generalization error and measure variable importance through permutation [76]. Decision trees are prone to problems like bias and overfitting, but together, multiple decision trees can predict accurate results when individual trees are not correlated. A simple schematic of the random forest method is shown in Figure 1.11. random forest algorithms utilize three primary hyperparameters – node size, number of trees, and number of features sampled. This algorithm is a collection of decision trees, where each tree consists of data samples drawn from a training set with replacement, called the bootstrap sample. One-third of the training sample is set aside as the test data, also known as an out-of-bag sample. An additional random instance is injected through feature bagging, reducing the 30orrelateon among Trees and adding diversity to the dataset. The final prediction is made after averaging the individual trees for a regression task or a majority vote for a classification task and cross-validation of the out-of-bag sample [77].

The key benefits of the random forest method are as follows:

- Reduced risk of overfitting.
- Flexibility.
- Easy determination of feature importance.

The critical challenges of the random forest method are as follows:

- Time-consuming process.
- Resource-consuming process.
- Complex to interpret predictions.

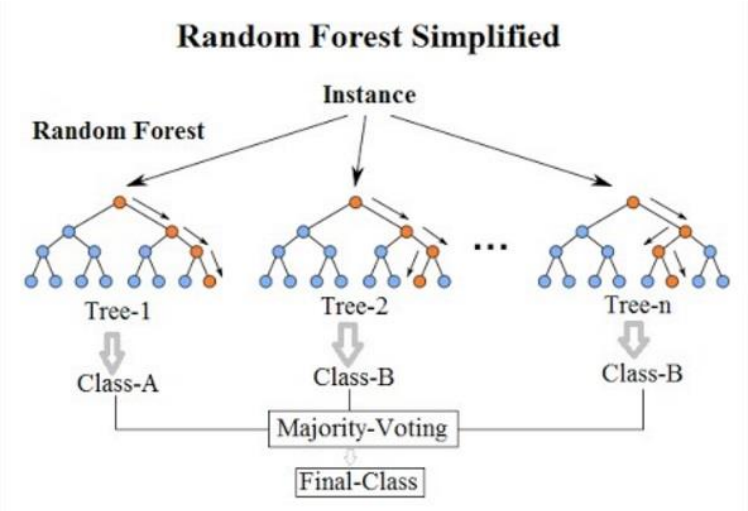

Figure 1.11: A schematic of the Random Forest decision tree [78], Public Domain Image.

Andrade et al. introduced a technique that automatically detects biofilm in tooth surfaces. This technique involves image analysis of intra-oral photographs and uses a neural network algorithm to detect dental biofilm to improve oral hygiene [79]. In similar research by Dimauro et al., biofilm samples were prepared into microscopic slides, images were captured using an optical microscope and analyzed using a convolutional neural network algorithm, producing an accuracy of about 99% [80]. However, both these methods involve the preparation of biofilm strains. Table 6 illustrates additional ML techniques used to classify biofilm using image analysis and biofilm strain preparation. Compared to existing ML techniques, the algorithm described in this research uses data from a non-invasive, non-destructive technique and can classify between scaling, corrosion, and biofilm deposits with a higher accuracy of about 99.8%. The method discussed in

this research allows plant managers or operation engineers to make rapid decisions on effective cleaning strategies, which is critical to several industries.

Table 6: Contributions and accuracy of the machine learning model in classifying or identifying biofilm or corrosion.

Main Contributions	**Target process**	**Model organism**	**Accuracy Score**
Identify chemical components responsible for bacterial biofilm using binary classification [81]	Essential oil chemical components	*Pseudomonas aeruginosa*	69 – 98%
Identify chemical components that modulate biofilm production using binary classification [82]	Essential oil chemical components	*Staphylococcus aureus and Staphylococcus epidermis*	68.7 – 90.6%
Use of lanthanide nanoparticles to detect pathogenic biofilms using random forest [83]	Biofilm infection	*Staphylococcus aureus, Pseudomonas aeruginosa, Acinetobacter baumannii, E. coli, and Stenotrophomonas maltophilia,*	95 – 100%
Semantic segmentation of corrosion using a Fully-Convolutional Network (FCN) [84]	Corrosion detection	Corrosion	55%
Finding local minima using the Ensemble method [85]	Corrosion detection	Corrosion	86 – 93%

1.3 Significance and Novelty

In many industries, it is crucial to obtain rapid results to take corrective actions promptly, especially in the food industry. The study presents and examines a fresh approach that combines non-invasive and non-destructive methods for detecting deposits in near real-time. This is accomplished by measuring changes in voltage and time-of-flight of ultrasound sensors and using a random forest ML algorithm to categorize the deposits into four types: no deposit, biofilm deposit, scaling deposit, and corrosion deposit. This work builds a strong foundation for future research, which makes use of evanescent waves or multiple internal reflections for the non-invasive detection of biofilm, is part of an invention disclosure filed in June 2023 [1]. The sensors utilized in this study are priced at $2 each, and the overall expense of all equipment employed in this investigation falls below $500. The total power needed for all the equipment is below 10 W, resulting in low energy expenses. All the equipment employed in this research can be incorporated into a portable tablet-like interface for gathering data. Consequently, the approach is cost-effective, portable, and demands minimal power. Although random forest learning has been utilized for various classification problems, this study presents a novel application of the ML technique to classify deposits based on voltage and time of flight measurements. The biofilm deposit in this research is defined as increased bacterial activity in the pipe loop or plastic container, scaling is defined as the mineral build-up in the pipe loop or container due to hard water, and corrosion is defined as the presence of metal deposits in the pipe loop or container.

Chapter 2

Materials

2.1 Sensors and Electronic Boards

2.1.1 Sensors

The sensors used in this research include multiple ultrasound sensors of varied frequencies. The details of the sensors are as follows:

(a) 1 MHz Ultrasound sensor (1ME21TR-1, Osenon Technology)

The 1ME21TR-1 is a dual-use ultrasound sensor that can be used for multiple applications, including but not limited to flow calculation, detection of objects, and measuring distance in liquids [86], which can be used as either an ultrasound transmitter or receiver. The primary characteristics of this sensor are described in Table 1 below.

Table 7: Characteristics of the 1 MHz (1ME21TR-1) ultrasound sensor.

Nominal Frequency	1.0 MHz ± 5%
Bandwidth	200.0 kHz
Max. Input Voltage	300 V_{pp}
Directivity and Sensitivity	8° ± 2° (-6dB), -35dB (min.)
Protection Level and Material	IP65, Plastic
Maximum Pressure	1.6 MPa
Operating Temperature	-20 °C ~ +80 °C
Distance of Detection	0.1 ~ 5 m (reflection in liquid)

Figure 2.1 depicts the 1 MHz ultrasound sensor manufactured and sold by Osenon Technology. These sensors can be activated with the help of a 3-volt eight-burst sinusoidal signal. The sensors are

- Compact and portable,
- High sensitivity and can withstand high sound pressure,
- Low power consumption, and
- High reliability.

Figure 2.1: An image of the 1 MHz ultrasound sensor [86].

(b) 400 kHz Ultrasound sensor (400E10TR-1, Osenon Technology)

The 400E10TR-1, like the 1ME21TR-1, is a dual-use ultrasound sensor that can be used as an ultrasound transmitter or receiver. It is generally used for ultrasonic distance measurement, liquid level detection, land leveling, thickness gauging, path edge detection, and other similar applications [87]. The primary characteristics of the sensors are described in Table 2 below.

Table 8: Characteristics of the 400 kHz (400E10TR-1) ultrasound sensor.

Nominal Frequency	400 kHz ± 16 Hz
Bandwidth	30.0 kHz
Max. Input Voltage	300 V_{pp}
Directivity and Sensitivity	7° ± 2° (-6dB), -75dB (min.)
Protection Level and Material	IP65, Aluminium Alloy
Operating Temperature	-40 °C ~ +80 °C
Distance of Detection	0.05 ~ 0.3 m

Figure 2.2 depicts the 400 kHz ultrasound sensor manufactured and sold by Osenon Technology. These sensors can be activated with the help of a 20-volt fifty-burst sinusoidal signal. The sensors are

- Compact and portable,
- High sensitivity and can withstand high sound pressure,
- Low power consumption, and
- High reliability.

Figure 2.2: An image of the 400 kHz ultrasound sensor [87].

(c) 2.5 MHz Ultrasound sensor (2ME20TR-1, Osenon Technology)

The 2ME20TR-1, like the 1ME21TR-1, is a dual-use ultrasound sensor that can be used as an ultrasound transmitter or receiver. It is mainly used for ultrasonic bubble sensors in an infusion pump [88]. The primary characteristics of the sensors are described in Table 3 below.

Table 9: Characteristics of the 2.5 MHz (2ME20TR-1) ultrasound sensor.

Nominal Frequency	2.5 MHz ± 5%
Bandwidth	10%
Max. Input Voltage	< 20 V
Sensitivity and Material	-30 dB (min.), Plastic
Operating Temperature	-20 °C ~ +70 °C
Distance of Detection	0.02 ~ 2.6 m

Figure 2.3 depicts the 2.5 MHz ultrasound sensor manufactured and sold by Osenon Technology. These sensors can be activated with the help of a 3-volt continuous sinusoidal signal. The sensors are

- Compact and portable,
- High sensitivity and can withstand high sound pressure,
- Low power consumption, and
- High reliability.

Figure 2.3: An image of the 2.5 MHz ultrasound sensor [88].

2.1.2 Electronic boards

The various electronic boards used in the research are described in detail below.

(a) Raspberry Pi 4 Model B (8 GB RAM)

The Raspberry Pi used in this research was purchased from CanaKit. The Raspberry Pi 4 is a tiny computer about the size of a credit card that provides desktop performance comparable to entry-level x86 computers. The product's key features include a high-performance 64-bit quad-core processor, dual-display output via two micro-HDMI ports providing upto 4K resolution, dual-band 2.4/5 GHz wireless LAN, Bluetooth 5.0, Gigabit

Ethernet, USB 3.0, and Power over Ethernet (PoE) capability. Figure 2.4 shows the top view of the Raspberry Pi 4 Model B [89].

Figure 2.4: Top view of the Raspberry Pi 4 Model B, 8 GB RAM variant [89].

The Raspberry Pi 4 Starter kit comes with the following modules:

- Raspberry Pi 4
- CanaKit USB-C Power Supply
- Set of 3 Aluminium Heat Sinks
- CanaKit Quick-Start Guide
- SanDisk 32 G.B. MicroSD with NOOBS (New Out of Box Software)
- Premium Black Case
- CanaKit Low Noise Fan
- USB Card Reader
- Micro HDMI Cable

The advantages of the Raspberry Pi 4 in comparison with other models are as follows [90]:

- The fanless, energy-efficient Pi runs silently and uses less power,
- It consists of two USB 3.0 ports in addition to two USB 2.0 ports, and
- It enables fast networking with onboard wireless networking and Bluetooth.

(b) Raspberry Pi 10.1-inch Touchscreen Display with a rear housing

The 10.1-inch touchscreen monitor by EVICIV has a built-in cooling fan to guarantee heat dissipation. It is enclosed in a durable, hard-wearing case to improve its appearance and protect the board. The display consists of capacitive touch technology allowing users to swipe, scroll, select, zoom in, zoom out, and move the cursor. Figure 2.5 shows the Raspberry Pi 10.1-inch touchscreen [91].

Figure 2.5: EVICIV 10.1-inch Touchscreen Display for Raspberry Pi [91].

The distinctive features of the 10.1-inch touchscreen display for the Raspberry Pi 4 Model B are as follows:

- 10.1-inch large screen,
- IPS Panel with 178° ultra-wide view angle,
- 1280 x 800 HD resolution with 60Hz refresh rate,
- 10-fingers touch response display,
- Dual integrated speakers, and
- Blue light filter and glare reduction feature to improve viewing comfort.

(c) Digilent Electronics Explorer board

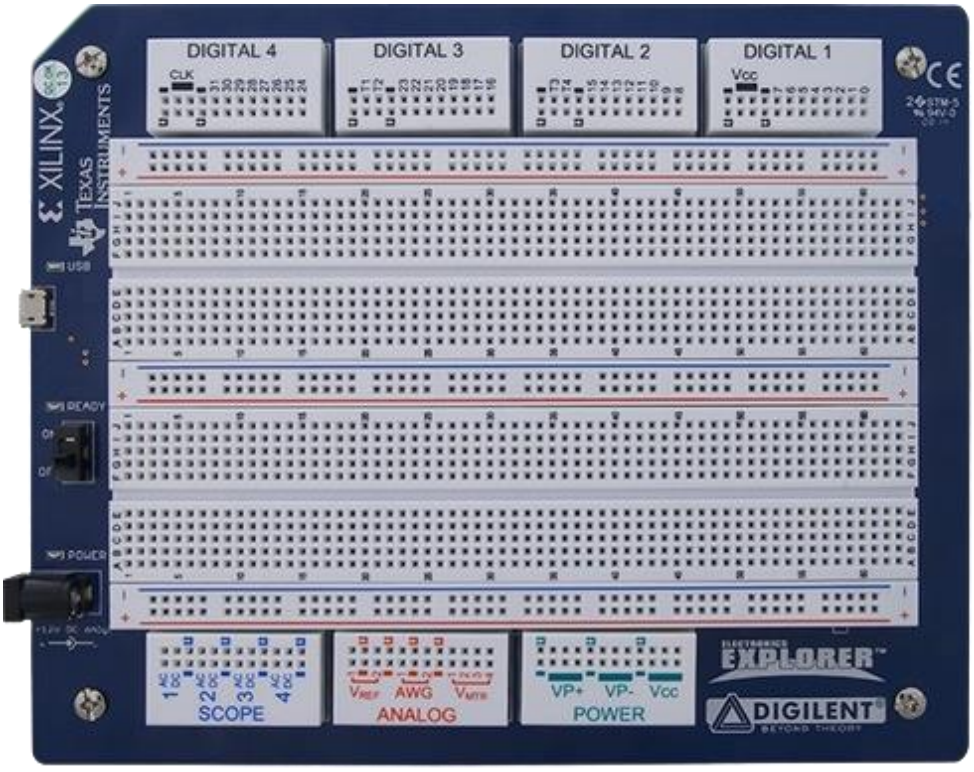

Figure 2.6: Top view of the Digilent Electronic Explorer board [92].

The Electronics Explorer is an all-in-one package for designing and testing analog and digital circuits. It is built around a large, solderless breadboard for quick, straightforward prototyping. The board can be managed and operated using Digilent's WaveForms software. In short, it is an all-in-one USB Oscilloscope, Multimeter, and Workstation [92]. The board comes with the following items:

- USB A to micro–B Cable,
- 12 V external power supply,
- U.S. and E.U. plug adapters, and
- Starter parts kit, including wires, LEDs, resistors, and capacitors.

The Oscilloscope section of the Electronics Explorer consists of 4 channels with a sample rate of 40 MS/s with a bandwidth of 100 MHz and an input voltage range of -20 V to +20 V.

The Arbitrary Waveform generator section of the Electronics Explorer consists of 2 channels with a sample rate of 40 MS/s with a bandwidth of 20 MHz.

The Fixed Power supply section of the Electronics Explorer is a single channel supply with an output voltage of 3.3 V or 5 V and an output current of 2 A.

The Variable Power supply section of the Electronics Explorer consists of a dual channel supply with a positive output voltage between 0 V to 9 V, a negative output voltage between 0 V to -9 V, and an output current of 1.5 A.

2.2 Software

The various software used in this research is described in detail below:

(a) Digilent WaveForms

WaveForms makes acquiring, visualizing, storing, analyzing, producing, and reusing analog and digital signals easy. The software and hardware bring a robust suite of instruments to enable analog and digital on any personal computer. WaveForms application connects to the Electronics Explorer board using the USB interface with full Windows, MacOS, and Linux support (on almost all devices) [93]. The application has two user-controlled power supplies, which vary in capability between devices. Figure 2.7 shows an image containing an example of an electronics explorer board connected to a personal computer using a USB interface.

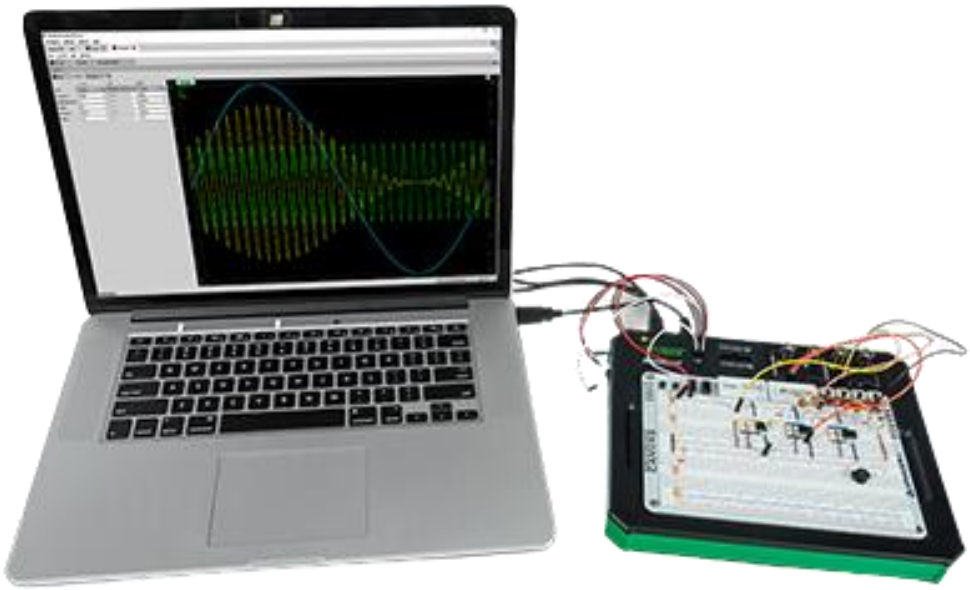

Figure 2.7: An example of an Electronics Explorer board connected to a personal computer using a USB Interface [92].

The oscilloscope offers all functionalities in a benchtop scope, including data acquisition, triggering, and viewing. It provides real-time math channels, X.Y. plots, FFTs, and advanced features. Depending on the device, using the oscilloscope in the application allows mixed signal oscilloscope functionality by adding digital channels and differential or single-ended measurements. Figure 2.8 shows an example of the oscilloscope window in the WaveForms application.

The waveform generator produces sinusoidal, sawtooth, triangular, or user-defined (arbitrary) waveforms. Generating advanced signals like sweeps between user-defined frequency limits and AM or FM-modulated outputs is also possible. Figure 2.9 shows an example of the waveform generator window in the WaveForms application.

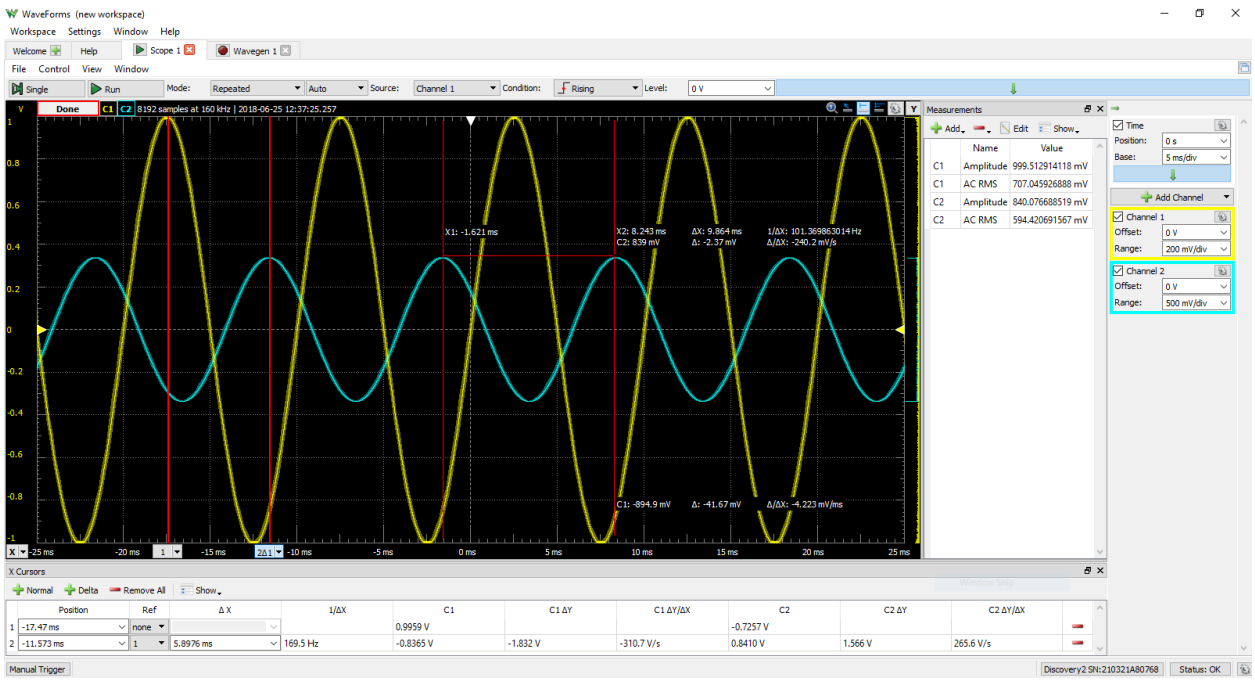

Figure 2.8: An example of the Oscilloscope window in the WaveForms application [93].

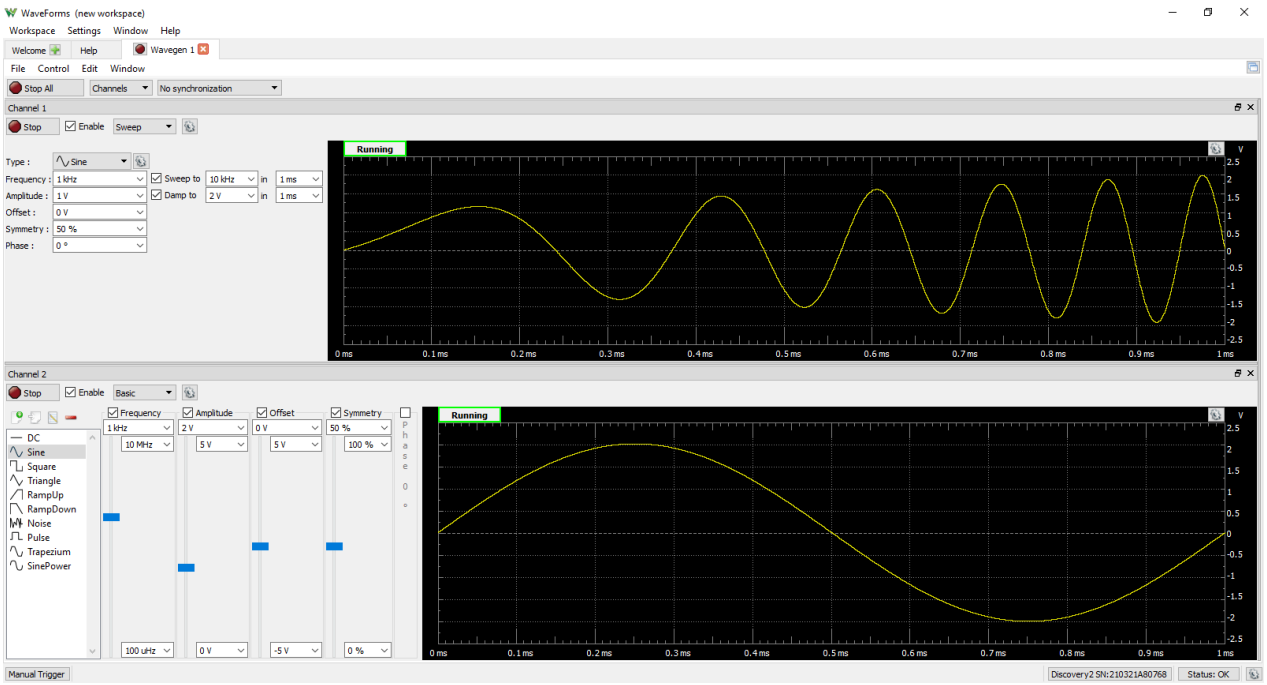

Figure 2.9: An example of the Waveform Generator window in the WaveForms application [93].

The script editor functionality helps automate the functionality of the available instruments using JavaScript. Figure 2.10 shows an example of the script editor window in the WaveForms application.

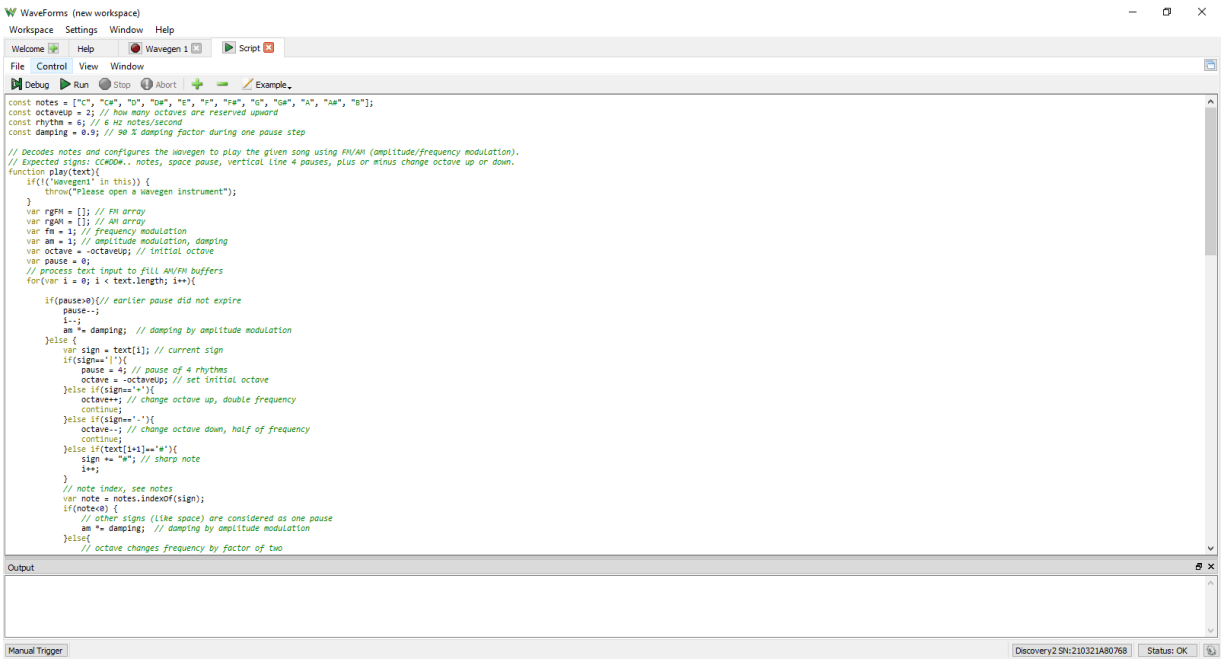

Figure 2.10: An example of the Script editor window in the WaveForms application [93].

(b) Mathworks MATLAB

MATLAB analyzes data, develops algorithms, and creates models combining a tuned desktop environment for iterative analysis and design processes with a programming language expressing matrix and array mathematics [94]. The capabilities of MATLAB include

- Data Analysis: Explore, model, and analyze data.
- Graphics: Visualize and explore data.
- Programming: Create scripts, functions, and classes.
- App Building: Create desktop and web apps.
- Hardware: Connect MATLAB to hardware.
- Parallel Computing: Perform large-scale computations and parallelize simulations.

- External Language Interfaces: Use MATLAB with Python, C/C++, Java, and other languages.
- MATLAB Cloud and Desktop Deployment.

Some of the applications that MATLAB is used for are Control Systems, ML, Signal Processing, Deep Learning, Predictive Maintenance, Test and Measurement, Imagine Processing and Computer Vision, Robotics, Wireless Communications, and other similar applications.

(c) JupyterLab

JupyterLab (jupyter.org) is a flexible web-based interactive development interface for notebooks, code, and data. It enables users to work flexibly, integrated, and extensible with Jupyter notebooks, text editors, terminals, and custom components [95]. It utilizes the same server and document format as the classic Jupyter Notebook. It also offers a model for handling different data formats, understands various file formats, and displays rich kernel output in these formats. The application window can be rearranged so multiple documents and activities are open in the work area using tabs and splitters. It also offers customizable keyboard shortcuts to ease user interface navigation [96].

2.3 Biological and Chemical Materials

2.3.1 Chemical Materials

The chemical materials used in the research are as follows:

- Difco™ Modified mTEC Agar/ m-TEC Agar (Powder)
- Phenol Red (Solution)
- Urea (Powder)
- Tryptone, microbiologically tested (Powder)
- Yeast Extract (Powder)
- Sodium Chloride, Molecular Biology Reagent Grade (Powder)
- Calcium Chloride (Powder)
- Sodium Bicarbonate (Powder)
- 70% Ethanol (Solution), diluted from 100% Ethanol

2.3.2 Biological Materials

The only biological material used in this research is a Primary Raw Sludge generated during the removal of grit, grease, scum, or other insoluble matter from wastewater during treatment. This sludge was collected biweekly or on demand from Jones Island Water Reclamation Facility, jointly operated by Milwaukee Metropolitan Sewerage District (MMSD) and Veolia Water Milwaukee, LLC (MMSD's contracted operator).

2.4 Miscellaneous Materials

The miscellaneous materials used in this research are:

- 15 and 50-mL sterile centrifuge tubes,
- 90 x 15 mm sterile petri dish,
- 1 L conical glass flask,
- 500 mL and 1 L round media storage bottle with screw caps,
- Sterile inoculating loops,
- Electronic Pipette/ Single Channel Pipette,
- Sterile 1000µL Pipette tips,
- 10 mL serological sterile glass pipet,
- 0.45 µm membrane filter discs,
- Thermo Scientific™ Nalgene™ Reusable Filter units,
- Vacuum pump,
- pH, Free and Total Chlorine Photometer,
- Free and Total Chlorine Reagent (Powder),
- 3” x 1/2” x 1/16” Copper coupons,
- Benchtop Conductivity Meter,
- Portable Turbidity Meter,
- Portable Soldering Station with parts,
- Thermo Scientific™ Heratherm Mechanical Convection Oven,
- Class II Biosafety Cabinet, and
- Autoclave

Chapter 3

Experimental Methods

3.1 Standard Operating Procedures

3.1.1 Procedure for Difco™ modified membrane-Thermotolerant *Escherichia coli* (mTEC) Agar Agar plate preparation

- Measure 45.3g of the modified mTEC Agar or the m-TEC Agar and suspend it in 1 L of deionized (DI) water.
- Stir the mixture frequently until the contents are dissolved.
- The solution was subjected to autoclaving at 121 ℃ for 15 – 20 minutes to ensure the mixture was dissolved entirely and sterilized.
- Pour the final solution into desired Petri dishes and allow the medium to solidify.
- Use Petri dishes or store them in a 4 °C refrigeration unit after properly labeling them. These plates are usually used within three months from the date they were made or until they have not been contaminated during a visual inspection.

3.1.2 Procedure for preparation of Lysogeny broth (L.B.) media

The formula for the preparation of Lysogeny broth was first published in 1951 by Bertani about lysogeny [97]. The American Society for Microbiology later published the standard recipe for one liter of L.B. media, adapted from the articles published by Sambrook and Russel, 2001 [98] and Gerhardt et al. 1994 [99]. The standard recipe is as follows:

- Measure 10 g of Tryptone, 5 g of Yeast extract, and 10 g of Sodium chloride (NaCl) and suspend it in 1 L of distilled or deionized water.

- Stir the mixture frequently until most of the contents are dissolved.
- The solution was subjected to autoclaving at 121 ℃ for 15 – 20 minutes to ensure the mixture was dissolved entirely and sterilized.
- Stir the final solution to ensure proper dissolving and transfer the media into a storage bottle.
- After the solution has been cooled down, store the bottle in a four-degree Celsius refrigeration unit after proper labeling.

3.1.3 Procedure for the Culture of *E. coli*

- Streak, modified mTEC or mTEC plates, as shown in Figure 3.1. A new inoculation loop is used for each streak. The inoculation loop is dipped inside the primary raw sludge from Jones Island Water Reclamation Facility at the start of the first streak.

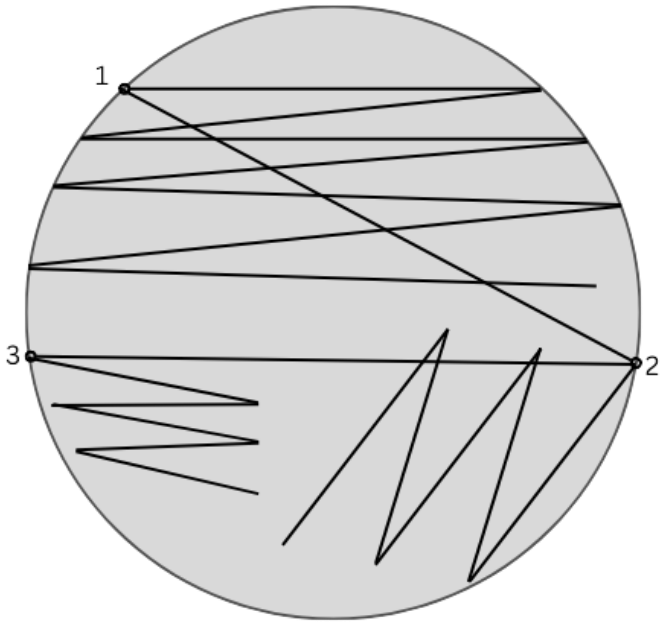

Figure 3.1: An example of how the plates should be streaked.

- After steaking, set the Convection oven (incubator) to 44.5 °C.
- Leave the plates, upside down, inside the incubator for 24 ± 2 hours.
- After 24 hours, transfer 5 mL of L.B. media into three or four 15 mL centrifuge tubes using an electronic pipette fitted with a serological sterile glass pipet.
- Take the plates from the incubator and use an inoculating loop to pick an isolated *E. coli* colony from the streaked plate. The colonies will appear red or magenta on modified mTEC agar plates or yellow-brown/yellow-green on mTEC agar plates.
- Place the colony in the centrifuge tube by turning the loop continuously.
- Set the shaker to 100 rpm for 18 hours and place the shaker inside the incubator with the temperature set to 35 °C.
- Label the centrifuge tubes with the *E. coli* colony, affix them to the shaker plate, and leave them undisturbed for 18 hours.

3.1.4 Preparation of Urea

- Mix 2 g of urea, 0.01 g of phenol red, and 100 mL of deionized water.
- Measure pH and adjust to a pH of 5 ± 0.3 if necessary.
- The solution can either be used immediately or can be stored for later.
- After proper labeling, the prepared solution is stored in a 4 °C refrigerator and should be used within a week.

3.1.5 Estimating the number of *E. coli* colonies

The following method was adapted from the standard method 1603: *Escherichia coli (E. coli)* in water by membrane filtration using modified membrane-thermotolerant *Escherichia coli* agar (modified mTEC) published by the United States Environmental Protection Agency [100].

1. Take eight empty centrifuge tubes and pipette 9 mL of deionized water in each using an electronic pipette attached to a serological sterile glass pipet.
2. Take the inoculated samples from the incubator and transfer 1 mL from this media into the first centrifuge tube.
3. Shake the centrifuge tube using a centrifuge machine.
4. Repeat step 2 to distribute 1 mL from the first centrifuge tube to the second tube and repeat until the last tube has 10 mL in it while all the other seven tubes have 9 mL of mixed dilutions.
5. Rinse the filter units, place a 0.45 µm membrane filter disc on the filter base, grid side up, and attach the funnel. Wet the filter using deionized water and let the water sit there.
6. Transfer 1 mL of the diluted sample from the last tube or using a pipette and uniformly distribute it to the membrane filter. It is recommended that at least three dilutions should be analyzed for a countable plate. This method can also analyze sample volumes of 1 – 100 mL.
7. Rinse the sides of the funnel using deionized water while filtering the samples. Remove the funnel from the filter base after turning the vacuum off.

8. Remove the membrane filter from the base using sterile forceps and carefully place it onto the agar plates to avoid bubble formation. Reseat the membrane filter if bubbles are formed.

9. Store the petri dish, inverted, inside the incubator with the temperature set to 35 °C for around two hours.

10. After two hours of incubation, adjust the incubator's temperature to 44.5 °C and store the petri dish in the incubator for 22 ± 2 hours.

11. Remove the plates from the water bath after the stipulated time and follow one of the two methods below.

1) If the plate was made using the modified mTEC agar, count and record the number of red or magenta colonies.
2) If the plate was made using the mTEC agar, place an absorbent pad on a petri dish and saturate it with 2 mL of urea solution. Transfer the filter from the agar plate onto the saturated absorbent pad and wait 15 – 20 minutes. Count and record the number of yellow, yellow-brown, or yellow-green colonies.

12. Calculate the *E. coli* count in Colony Forming Units (CFU) per 100 mL of the sample using the general formula:

$$E.\ coli \text{ CFU per } 100 \text{ mL} = \frac{\text{The number of } E.\ coli \text{ colonies}}{\text{The volume of sample filtered (mL)}} \times 100$$

13. It is ideal to select a membrane filter with 20 – 80 colonies for optimal results.

14. Record results as *E. coli* CFU per 100 mL of sample.

3.2 Laboratory experiment to evaluate best ultrasound frequency and waveform

The ultrasound sensors used in this experiment are manufactured and sold by Osenon technologies with operating frequencies of 400 kHz, 1 MHz, and 2.5 MHz. Since all the sensors can act as transmitters or receivers, one is placed on one side of the plastic chamber, as seen in Figure 3.2, filled with deionized water. This sensor acts as the transmitter. Another sensor (receiver) is placed on the other side of the plastic chamber.

Figure 3.2: Plastic chamber setup for the experiment to test some aspects of the ultrasound sensor

The transmitter sensors are activated for the 400 kHz ultrasound sensor by providing a fifty-burst input sinusoidal signal with a peak-to-peak voltage of 20 V_{pp}. The test circuit provided by Osenon Technologies is slightly modified to include two sensors, one as the transmitter and the other as the receiver.

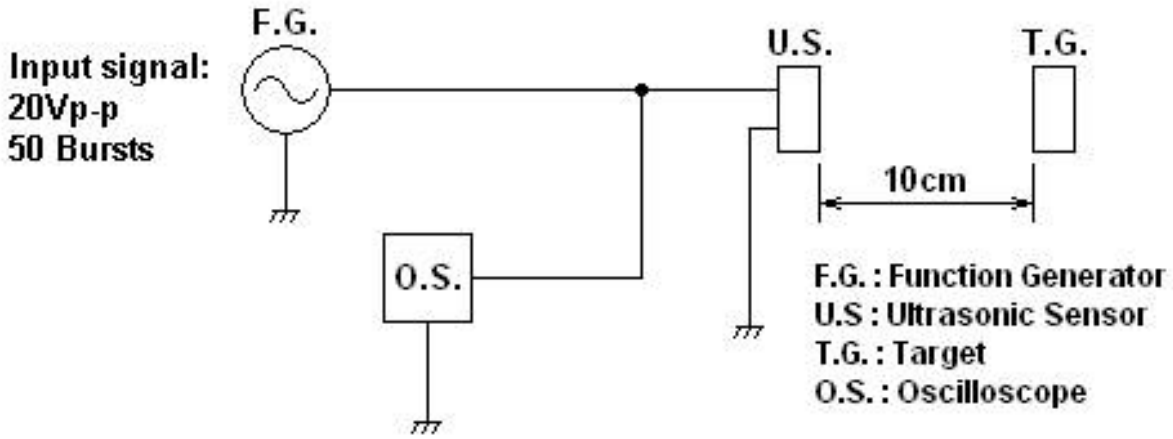

Figure 3.3: Test circuit design for actuating the 400E10TR-1 ultrasound sensor [87].

The transmitter sensors are activated for the 1 MHz ultrasound sensor by providing an eight-burst input sinusoidal signal with a peak-to-peak voltage of at least 3 V_{pp}. The test circuit provided by Osenon Technologies is slightly modified to include two sensors, one as the transmitter and the other as the receiver.

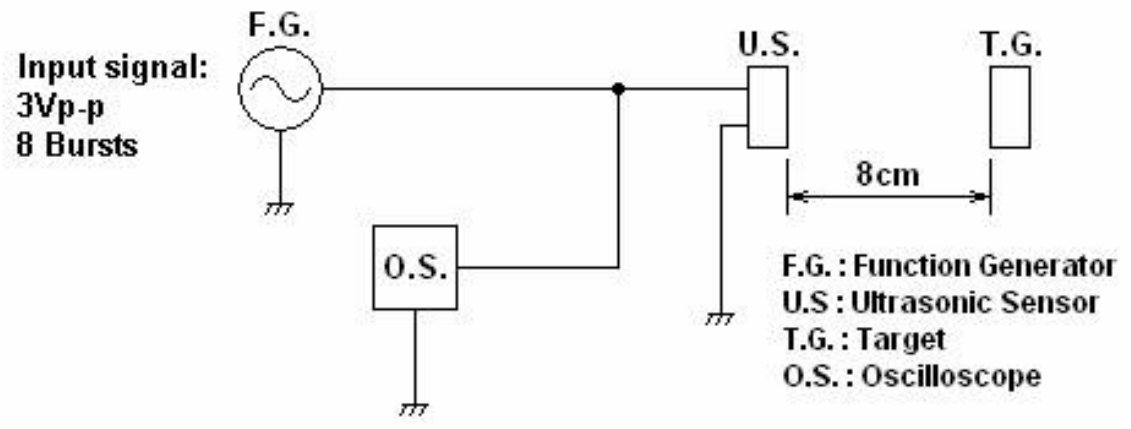

Figure 3.4: Test circuit design for actuating the 1ME21TR-1 ultrasound sensor [86].

The transmitter sensors are activated for the 2.5 MHz ultrasound sensor by providing a continuous input sinusoidal signal with a peak-to-peak voltage of at least 3 V_{pp}. The test circuit provided by Osenon Technologies is slightly modified to include two sensors, one as the transmitter and the other as the receiver.

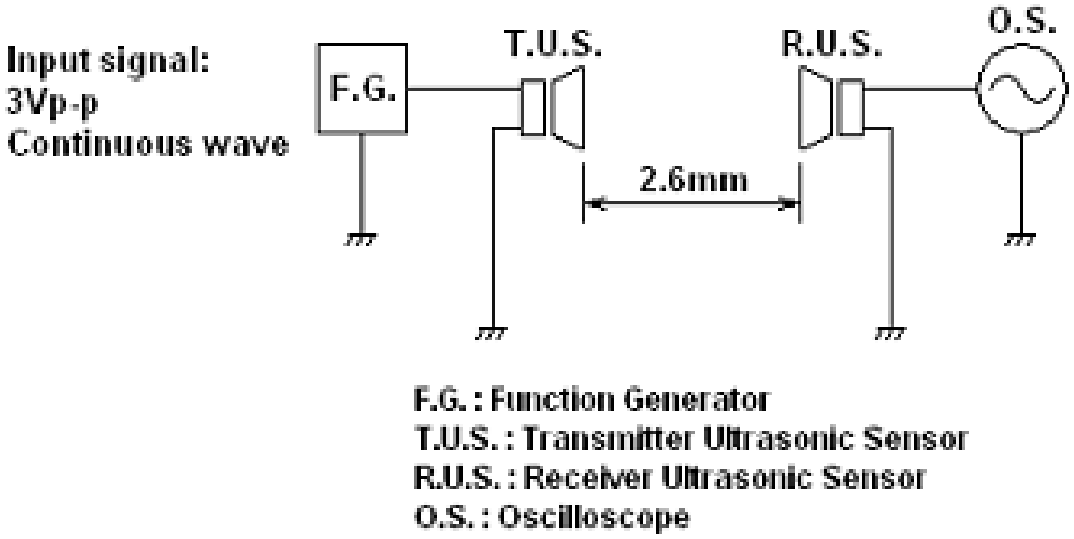

Figure 3.5: Test circuit design for actuating the 2ME20TR-1 ultrasound sensor [87].

The transmitter was connected to a benchtop Function generator with a 10 V output and an eight-burst sinusoidal output waveform. Both the transmitter and the receiver were connected to a benchtop oscilloscope. The two parameters measured in the experiment were time of flight and voltage ratio. The time of flight was measured by calculating the inverse of the distance between the input (transmitted) signal and the output (receiver) signal. The voltage ratio is calculated as the ratio of the output (receiver) voltage to the input (transmitter) voltage. The output waveforms were recorded and analyzed to determine the best ultrasound sensors for the research. In addition, several waveforms - sinusoidal, square, and ramp- were used to excite the 1 MHz ultrasound sensor. The results were recorded to verify whether the ultrasound sensors are most efficient when

using sinusoidal waveforms. Finally, the number of bursts on the wave was varied on the Function generator to verify whether the sensors are most efficient when using eight-burst waveforms.

3.3 Experiment to evaluate the sensor performance in a laboratory-designed pipe loop.

Figure 3.6 shows the schematic of the pipe loop designed at the Global Water Center laboratory. It consists of three piping materials – Copper, Polyvinyl Chloride (PVC), and Cross-linked Polyethylene (PEX). Ultrasound transmitter and receiver were attached to each pipe in the loop using an epoxy. The ultrasound sensors used in this experiment are manufactured and sold by Osenon Technologies with an operating frequency of 1 MHz.

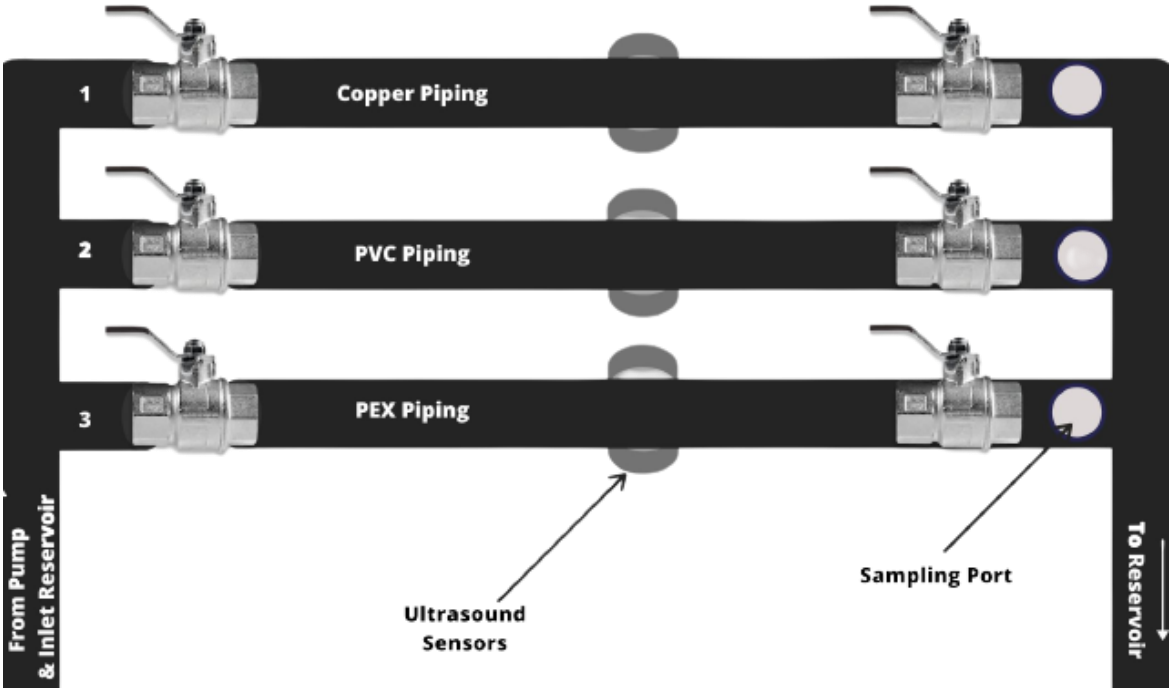

Figure 3.6: Schematic for the pipe loop designed at the laboratory.

The transmitter was connected to the function generator section and the power supply of the Electronics explorer board, while the receiver was connected to the oscilloscope section. Three reservoirs are created using large boxes. Six liters of deionized water were mixed with 5 mL of L.B. media – *E. coli* solution at the final step of the *E. coli* culture and added into the first reservoir. The second reservoir mixed 150 g of NaCl powder with six liters of deionized water. In the final reservoir, 90 g of magnesium sulfate, calcium chloride, and sodium bicarbonate were mixed with six liters of deionized water. Each reservoir was labeled appropriately to distinguish them during the experiment. Based on the experiment, the reservoirs were interchangeably connected to a Dyson pump. The valves of the respective pipes were only opened during the respective experiments.

The Copper pipe was used for studying sensor data during the event of corrosion, and the reservoir labeled "Saline water" was used for experiments in this section of the loop since less than 3% saline water is the most effective for inducing corrosion. The PVC pipe was used for studying sensor data during the event of biofilm deposit, and the reservoir labeled "Biofilm" was used for experiments in this section of the loop since the presence of *E. coli* in water is the most effective for inducing biofilm. The PEX pipe was used for studying sensor data during the event of scaling deposit, and the reservoir labeled "Hard water" was used for experiments in this section of the loop since hard water is the most effective in inducing scaling. The PEX pipe, while resistant to corrosion and mineral build-up, can still have mineral build-up inside the pipe due to the fittings' build-up. In addition, PEX water lines are also susceptible to bacteria based on the liquid used in the system [94]. Figure 3.7 shows the actual pipe loop experiment setup in the laboratory.

The Dyson pump is connected to an ABB variable frequency drive (VFD) to control the flow speed through the pipes. The 90° turn valves for the pipe used for the experiment were turned

to the ON position when in progress, the pump was turned on, and the flow rate was set to 3.5 GPM, the standard flow rate in most commercial and industrial piping systems. The sensor data is collected using the explorer board, processed, and analyzed using MATLAB.

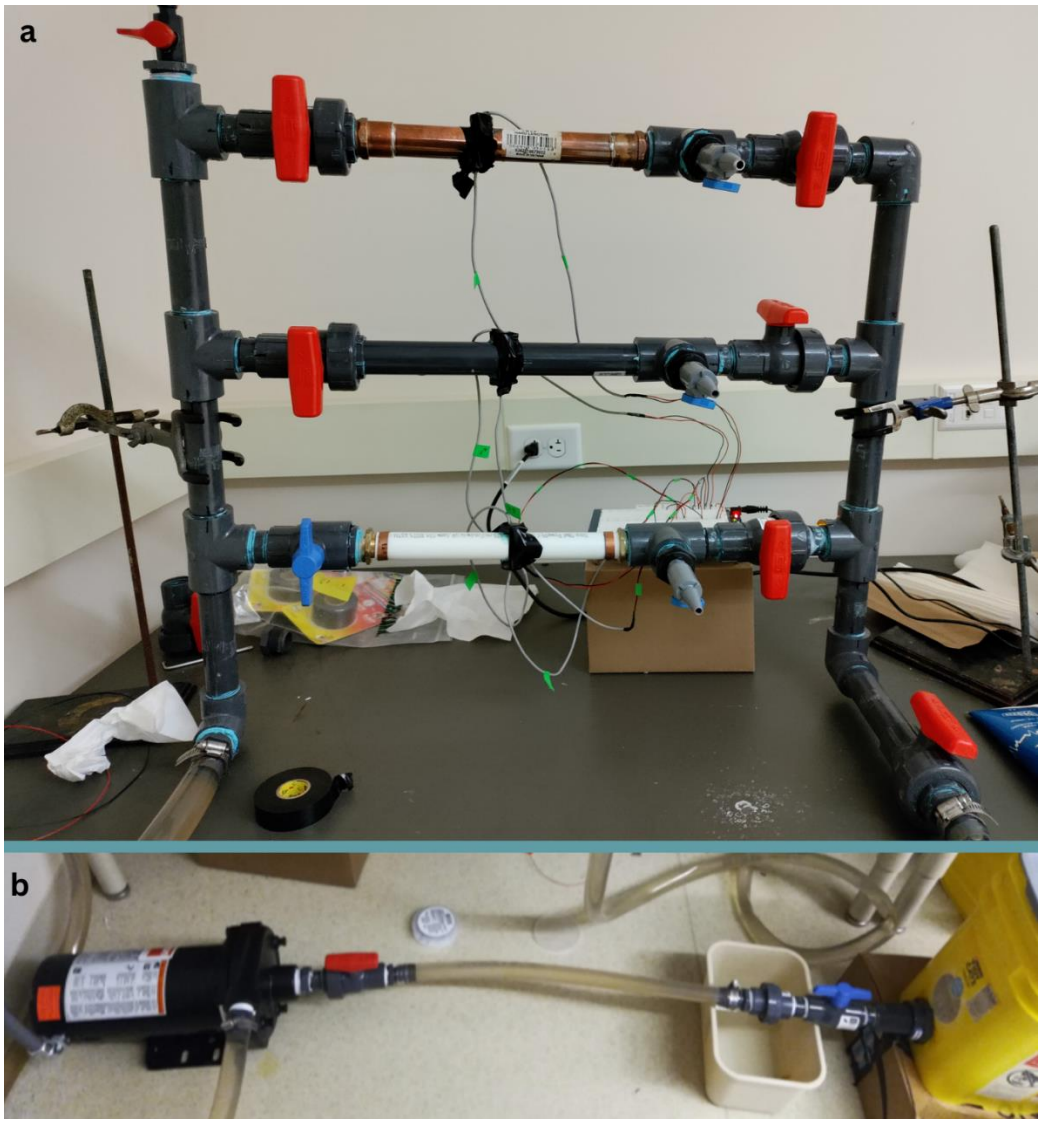

Figure 3.7: (a) Pipe loop set up at the laboratory. From top to bottom, the pipes used are Copper, PVC, and PEX, respectively. (b) Pump and reservoir sections of the pipe loop.

The samples collected from the experiment via the sampling port were subjected to turbidity, conductivity, free chlorine, and total chlorine assessments. In addition to these assessments,

- The samples from the Copper pipe were subjected to a corrosion coupon test where a coupon made of copper is submerged in the sample for 90 days. The coupon's weight after submersion was subtracted from the coupon's weight before submersion and was used to estimate the pipe's corrosion level.
- The samples from the PVC pipe were subjected to a biofilm total plate count test using filter membranes. The sample volumes used for this test are 100 mL, 10 mL, and 1 mL.
- The samples from the PEX pipe were subjected to an orthophosphate test using the Soluble Reactive Phosphorus (SRP) analysis.

The readings or measurements from the experiment were stored in an Excel document and combined with the sensor data. The final Excel document with all the datasets was then analyzed using MATLAB, and graphs were plotted for easier data readability.

3.4 Experiment to evaluate the sensor performance in a pipe loop at the Howard plant.

Figure 3.8 shows the pipe loop experiment associated with Jacobs Engineering at the Howard Avenue Water Treatment Facility in Milwaukee, WI. Ultrasound transmitters and receivers were attached to each pipe in the loop using epoxy and tape to ensure the sensors stayed attached. The ultrasound sensors used in this experiment were manufactured and sold by Osenon Technologies with an operating frequency of 1 MHz. The control pipe with 1.9 mg/L phosphate

was the control used in water distribution systems in Milwaukee to limit corrosion in the piping systems. A second pipe with 0 mg/L phosphate was selected as the next candidate for the ultrasound sensors as it has the highest possibility of corrosion formation. Finally, the pipe with 3.0 mg/L phosphate was selected as the final candidate for ultrasound sensors as it has the highest probability of scaling and biofilm compared to other pipes.

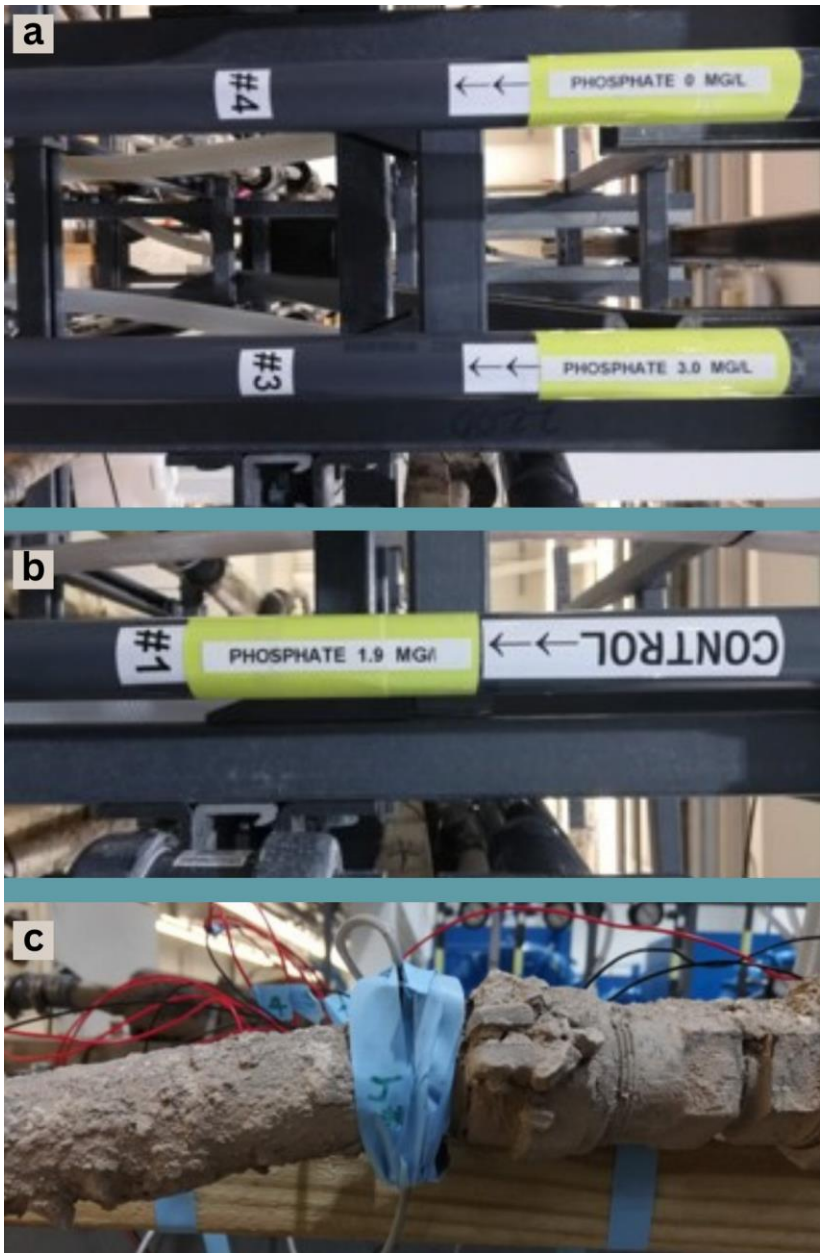

Figure 3.8: (a) Pipe with 0 mg/L phosphate and 3.0 mg/L phosphate. (b) Control pipe with 1.9 mg/L phosphate. (c) An example of how the sensors were attached to the pipe loops at the Howard plant.

Epoxy and tape were used to attach the ultrasound transmitters and receivers to the three pipes. The transmitters were connected to the electronic explorer board's power supply and function generator section. The receivers were connected to the electronic explorer board's oscilloscope section. The data from the sensors were recorded in an Excel document with the WaveForms application installed on the Raspberry Pi. The pipes were subjected to the following tests routinely – pH, temperature (°C), dissolved Oxygen levels (mg/L), total chlorine (mg/L), monochloramine (mg/L), free ammonia (mg/L), nominal turbidity units (NTU), orthophosphate (mg/L PO_4), and nitrite (mg/L). Some of these parameters, including but not limited to dissolved oxygen levels, total chlorine, turbidity, and orthophosphate, were used to estimate the presence of biofilm in the pipes during the analysis. These data were combined with the sensor data in a new Excel document and were analyzed using MATLAB, and graphs were plotted for easier data readability.

3.5 Ground truth experiment to classify various deposits using a Machine Learning algorithm.

Figure 3.9 shows the schematic of the ground truth experiment setup. Ultrasound transmitters and receivers were connected to three plastic containers using super glue to minimize the sensor's air contact between the environment and the container. *E. coli* culture solution was added to the first container, a copper coupon was added to the second container, and a mixture of magnesium sulfate, calcium chloride, and sodium bicarbonate was added to the third container in the second iteration of the ground truth experiment. All three containers were filled with 400 mL of deionized water.

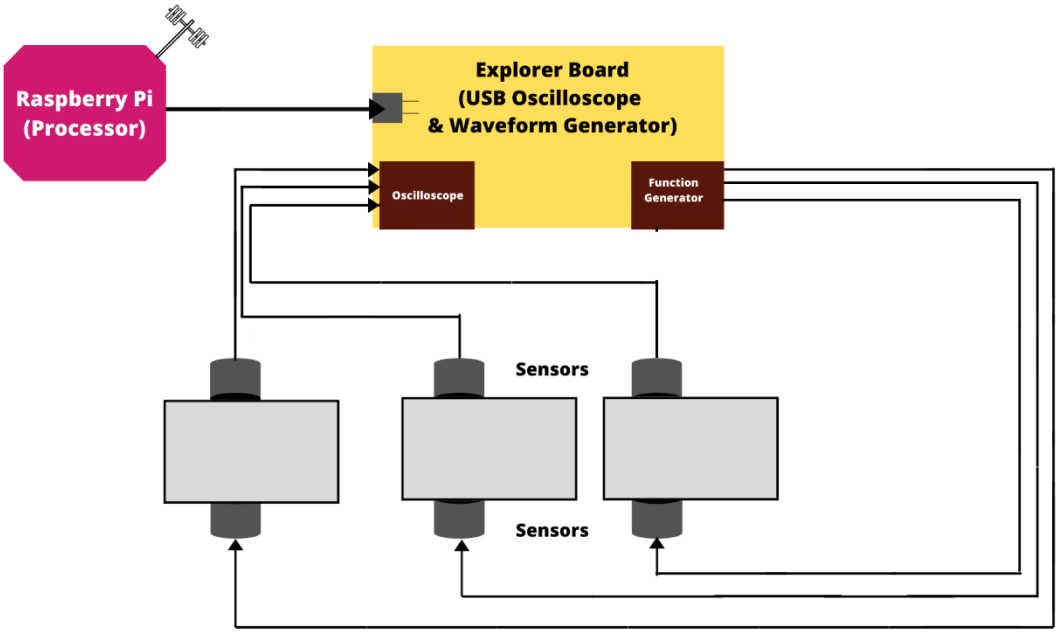

Figure 3.9: Schematic for the ground truth experiment to test a machine learning algorithm.

The containers were placed inside the incubator at 37.5 °C to accelerate bacterial growth. The first container was used to simulate the presence of biofilm, the second container was used to simulate the effect of corrosion in pipes, and the final container was first used to simulate a pipe with no defect. Later in the experiment, a mixture of magnesium sulfate, calcium chloride, and sodium bicarbonate was added to the final container to simulate the effect of scaling in pipes. The transmitters were connected to the electronic explorer board's function generator and power supply section, and the receivers were connected to the electronic explorer board's oscilloscope section. The actual ground truth experiment setup can be seen in Figure 3.10.

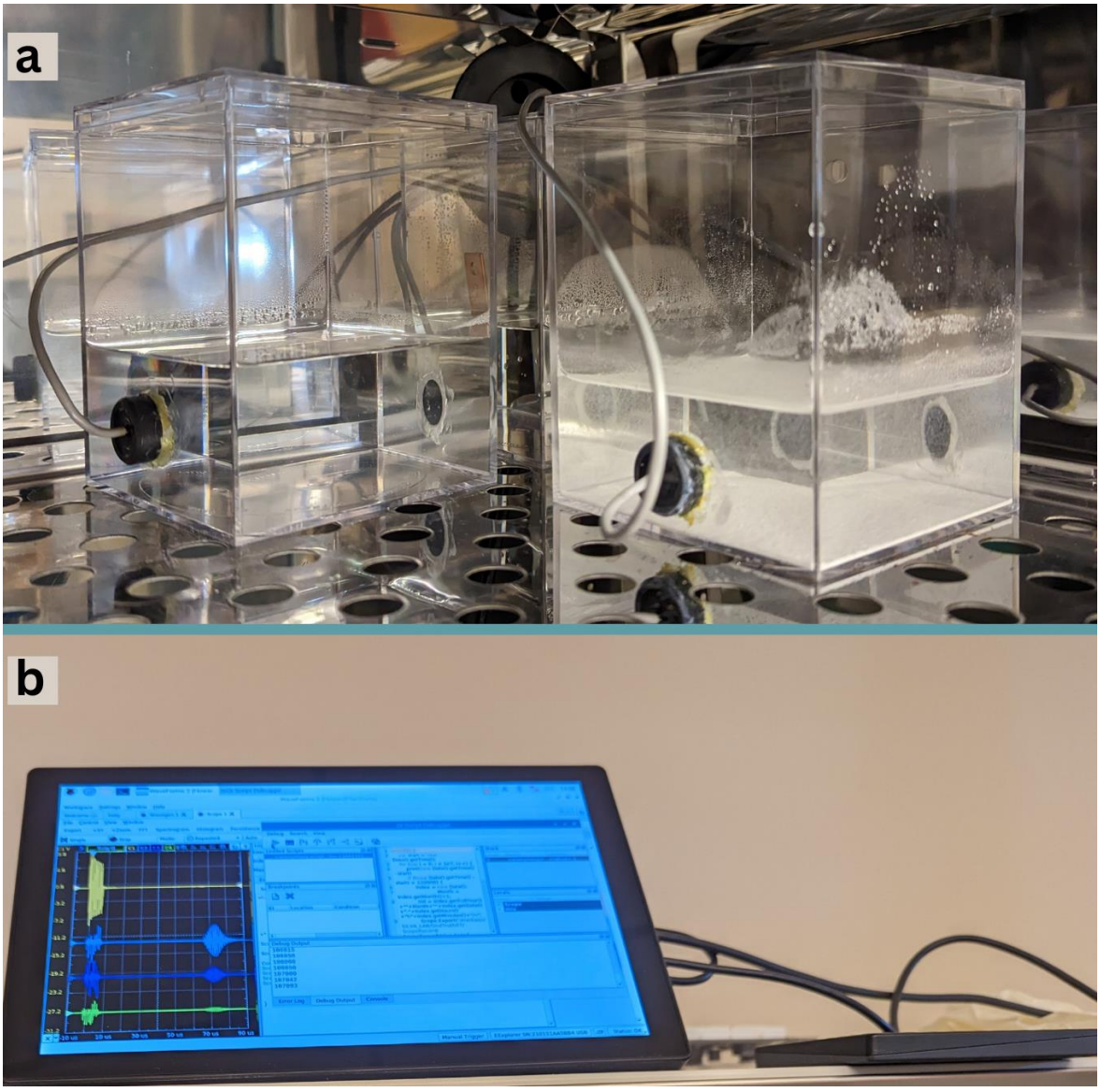

Figure 3.10: (a) Ground truth experiment setup inside the incubator to test machine learning algorithm. (b) Touchscreen display of the WaveForms application in Raspberry Pi.

Figure 3.11 shows oscilloscope readings of the ground truth experiment to test a machine-learning algorithm. The yellow oscilloscope trace indicates the voltage produced by the ultrasound transmitter, and the blue oscilloscope trace indicates the voltage observed on the ultrasound receiver. However, these data had to be treated before extracting the maximum voltage and time of flight values. The first section of the receiver voltage included a noise signal which limits the

effectiveness of using the ultrasound sensor to detect biofilm. When the maximum voltage of the signal received by the sensor is greater than the noise signal, the data is not erroneous. However, if the noise signal is greater than the actual signal, then the maximum voltage and time of flight values reflect the noise signal and not the actual signal, which makes the data erroneous.

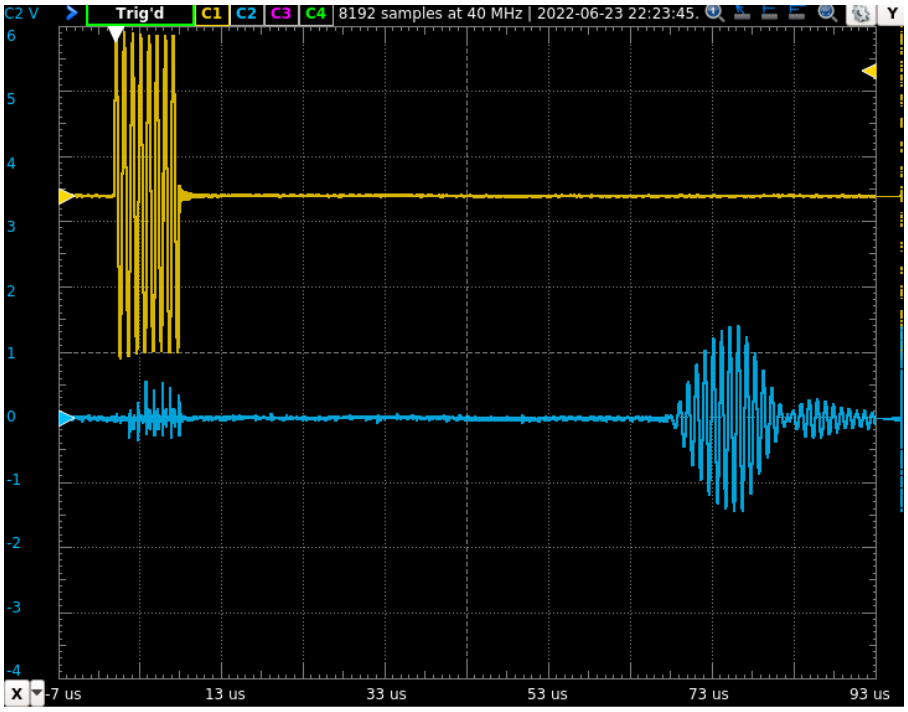

Figure 3.11: The oscilloscope reading from a test experiment to understand the effect of an ultrasound sensor in a test scenario.

Figure 3.12 shows the detailed schematic of the ground truth experiment setup. A waveform generator creates electrical voltage signals in the form of sinusoidal signals, which activates the ultrasound transmitter attached to one side of the pipe surface. The signals absorbed by the ultrasonic receiver are observed using the oscilloscope. The electromagnetic crosstalk causes the noise signal due to the proximity of the wires connected to the waveform generator and oscilloscope. The sensor data was recorded in an Excel document, and the data were processed using MATLAB to select the maximum voltage and time of flight value after rejecting the noise due to crosstalk. The processed data was then rearranged with class labels to test the random forest

ML algorithm. The dataset was separated into a training dataset and a testing dataset. The training dataset was used to train the random forest algorithm, while the testing dataset was used to verify the accuracy of the ML algorithm. Parameters like accuracy, features importance chart, and confusion matrix were used to determine the effectiveness of the random forest algorithm in identifying the type of deposit – No deposit, scaling deposit, biofilm deposit, or corrosion deposit. The accuracy of the ML model is the percentage of correct classifications that it achieves. The feature importance graph refers to a tool that assigns a score to input features based on how useful they are at predicting a target variable. A confusion matrix is a table used to visualize and summarize the performance of a classification algorithm. The model's accuracy was calculated by utilizing testing data that was distinct from the training data and was not observed by the ML model. If the accuracy was less than 95%, the model was re-tuned to make the ML algorithm more effective in classifying the deposits. The ML algorithm was developed and tuned using JupyterLab, a Python-based machine-learning platform.

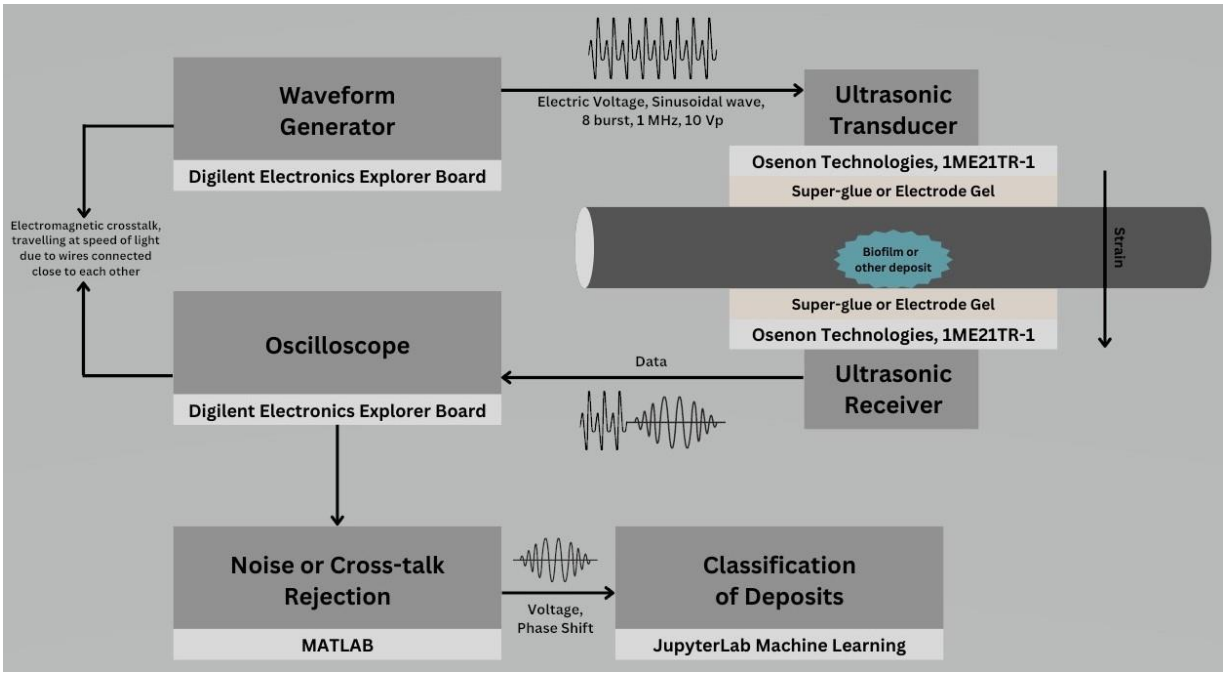

Figure 3.12: Detailed schematic for the ground truth experiment to test a machine learning algorithm.

The hyperparameters used to tune the random forest algorithm are as follows [101]:

- Maximum features (max_features): This parameter is used to randomly select the number of features at each node of the random forest. In this research, the max_features = 1,
- Maximum depth (max_depth): This parameter is used as a stopping criterion to restrict the depth to which a tree can grow. In this research, max_depth= none,
- Minimum samples split (min_samples_split): This parameter indicates the number of data points placed in a node before the node is split. In this research, min_samples_split = 7,
- Minimum samples leaf (min_samples_leaf): This parameter indicates the minimum data points allowed in a leaf node. In this research, min_samples_leaf = 1,
- Number of trees (n_estimators): This parameter indicates the number of trees required in the model. In this research, n_estimators = 200,
- Bootstrap: This parameter indicates the sample number of data points. In this research, bootstrap = true, so the whole data is used for every decision tree,
- Criterion: This parameter measures the quality of splits in a decision tree. In this research, the default gini impurity criterion is used.

3.6 Customer Discovery Process

A business model was tested with the National Science Foundation (NSF) Innovation Corps (I-Corps™) Site of Southeastern Wisconsin, a local chapter of the NSF program, designed to facilitate an invention's transformation into a commercial product through an entrepreneurial training program. The I-Corps training program is widely recognized in the U.S. and internationally and addresses three needs [102].

- Train an entrepreneurial workforce.
- Enable positive economic impact by bringing cutting-edge technologies to market.
- Nurture an innovation ecosystem.

The I-Corps program prepares scientists and engineers to increase research projects' economic and societal impact by extending their focus beyond the laboratory. The I-Corps program was launched in 2011 and believes in experimental learning using the customer discovery process [103]. The program provides access to a mentor network and funding to support a team in the discovery process. The first step in the market survey or customer discovery process was identifying the key partners, value proposition, target customers, and channels. The next step is interviewing industry representatives or potential customers to test the business model or hypothesis. The final process is analyzing the model and making adjustments that help in customer acquisition in the long run.

Chapter 4

Results and Discussion

4.1 Laboratory experiment to evaluate best ultrasound frequency and waveform

Table 10 shows the time of flight and voltage ratio readings obtained from the experiment. The time of flight was observed to be the highest at the ultrasound sensor frequency of 2 MHz and the lowest at the ultrasound sensor frequency of 400 kHz. Considering the time of flight alone, the sensor with an operating frequency of 400 kHz would be ideal. However, the voltage ratio was for the ultrasound sensor with an operating frequency of 1 MHz, and the 400 kHz sensor recorded the lowest voltage reading. From the results of this experiment, the 1 MHz sensor was selected as the best candidate for the research. A low-frequency sensor is less focused and has a greater penetration depth [104]. Due to the less focused low-frequency sensor, the peak voltage measured at the sensor side for the 400 kHz sensor is comparatively low to the 1 MHz sensor. Compared to the 2 MHz sensor, the 1 MHz sensor will have a higher penetration depth indicating a higher voltage ratio.

Table 10: Time of flight and voltage changes of ultrasound sensors of different frequencies.

Frequency (MHz)	**Time of flight (µs)**	**Voltage Ratio, $\frac{\mathbf{Vout}}{\mathbf{Vin}}$ (V)**
0.4	-0.028	0.18
1.0	-0.074	0.55
2.0	0.176	0.16

Table 11 shows the experiment's results where different waveforms were applied to the 1 MHz ultrasound sensor, and the resultant waveform was observed using the oscilloscope. A general rule of thumb is that sinusoidal signals should be used to drive an ultrasound sensor compared to a square wave since the square wave vibrates at the fundamental frequency and at all harmonics, which can result in erroneous data. When a sinusoidal signal drives an ultrasound sensor, it only vibrates at the specified frequency, thus creating a pure excitation [105]. The table below shows that the sinusoidal signal is the best excitation source for the Osenon ultrasound sensors. When the sensors used in this research were excited using a square wave, a waveform was visible, but it could not observe a specific pattern from which data could be analyzed. No waveform was observed on the oscilloscope for a ramp or triangular waveform used as the excitation source. The sinusoidal waveforms exhibited the best behavior of the ultrasound sensors, and the waveforms at the receiver ultrasound sensor were repeated exactly but with voltage attenuation.

Table 11: Experiment with different waveforms applied to the 1 MHz ultrasound sensor.

Waveform Type	**Result**
Sinusoidal	The sinusoidal waveforms were repeated exactly with signal attenuations.
Square	A waveform is visible on the oscilloscope but is not repeated exactly, and patterns cannot be identified.
Ramp/Triangular	No waveform was visible on the oscilloscope.

4.2 Experiment to evaluate the sensor performance in a laboratory-designed pipe loop.

The parameters measured or recorded in this experiment were copper loop voltage, plastic loop voltage, PEX loop voltage, turbidity, conductivity, plate count, total chlorine, and free chlorine. The presence of biofilm is defined by the change in the heterotrophic plate count over time and turbidity and conductivity levels observed in the pipe loop. The increased turbidity in the culture indicates bacterial growth and biomass, as there is a direct relationship between turbidity and the number of cells [106]. High turbidity levels can also promote the likelihood of corrosion [107]. Although separate pipe loops were used to detect the presence of three different parameters, the pipe loop setup did have a single pump source connected to all three reservoirs, which increased the chances of cross-contamination in the pipes. Figure 4.1 shows the results of the pipe loop experiment at the laboratory. The sample number on the X- axis refers to the different samples obtained over several days of the experiment and are not equally spaced.

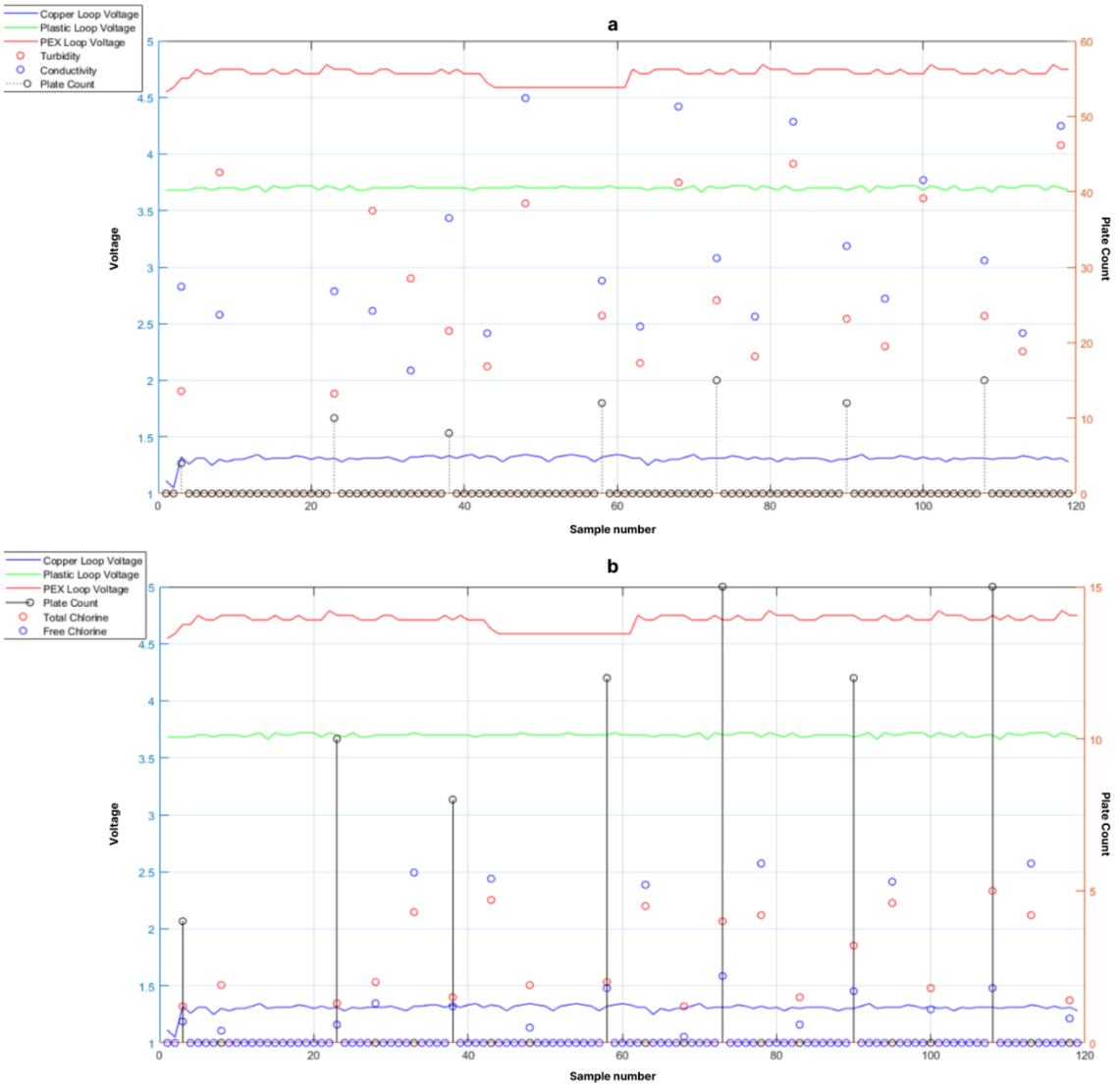

Figure 4.1: Pipe loop experiment designed at the laboratory. (a) Graph showing the loop voltages, turbidity, conductivity, and plate count parameters. (b) Graph showing the loop voltages, total chlorine, free chlorine, and plate count parameters.

It can be seen that with a rise in the turbidity and conductivity, there was a drop in the PEX loop voltage. This trend indicates a ”disturbance” on the inner walls of the pipes. The disturbance detected by the sensors can be caused by biofilm, corrosion, or scaling. The total chlorine and free chlorine measurements were comparatively higher during this interval which strongly suggests the detachment of scaling, biofilm, or a combined deposit. In the PVC pipeline, the heterotrophic plate count (HPC), along with the turbidity and conductivity measurements, indicated a high value

which indicates a high probability of bacterial presence in the water flowing through the pipe. Since the HPC count observed in the samples is high, bacterial cells in the liquid indicate a high possibility of biofilm formation in the pipe loop. Figure 4.2 shows the graphs representing the correlation between the bacterial plate count with the turbidity and conductivity measurements from the samples obtained from the pipe loop. The scatter plot shows an R-squared value of 0.42 for the correlation between turbidity and plate count and 0.69 for the correlation between conductivity and plate count. The scatter plot shows that there is an exponential relation for the laboratory pipe loop experiment between the plate count measurements and the turbidity and conductivity of the samples from the pipe loop.

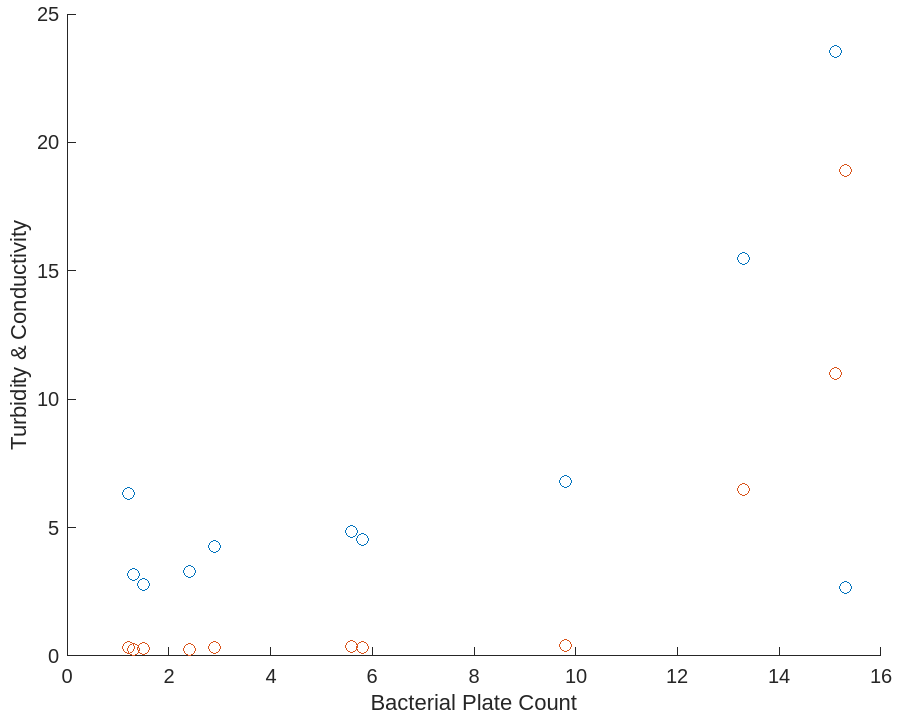

Figure 4.2: Graph showing the change in turbidity and conductivity levels with changes in the bacterial plate count.

Overall, the results suggest that sensor attachment can detect “disturbances” – due to corrosion, scaling, or biofilm. A ground truth experiment was set up to study the sensor attachment's capability to detect and classify deposits – biofilm, corrosion, or scaling.

4.3 Experiment to evaluate the sensor performance in a pipe loop at the Howard plant.

The parameters measured and recorded in this experiment were voltage changes in the pipe with 1.9 mg/L, 3.0 mg/L, or 0 mg/L phosphate-dosed water, turbidity, total chlorine, dissolved oxygen, pH, dissolved oxygen, monochloramine, free ammonia, nitrite, and orthophosphate. It was found from a study of the relationship between dissolved oxygen and biofilm that there is a direct correlation between the two [108]. The rate of biofilm increased with the increase in dissolved oxygen levels. The study also shows that the dissolved oxygen levels limit the presence of ammonia-oxidizing bacteria on the surface of the biofilm, but bacteria still exist at the deeper layer where oxygen is depleted [108]. In a study conducted by Lee et al., it was found that monochloramine, compared with free chlorine, is the most effective in penetrating biofilm or reducing its persistence. However, free chlorine was more effective at deactivating microorganisms near the biofilm source [109]. Figure 4.3 shows the graph with the data captured from the ultrasound sensors and measurements of other parameters correlated in time in the control pipe with a 1.9 mg/L phosphate level. The sample number on the X- axis refers to the different samples obtained over several days of the experiment and are not equally spaced. It can be seen from the zoomed-in version of the graph (Figure 4.3, b) that with a rise in dissolved oxygen, and pH, there is a sharp decrease in the sensor voltage that indicates that the sensor arrangement can be used in detecting “disturbances” in the inner walls of the pipes. The “disturbances” can be due to biofilm, corrosion, or scaling. The presence of monochloramine and free chlorine, even though in smaller quantities, explains the fluctuations in the voltage readings from the ultrasound sensor since they play a role in inactivating or penetrating biofilm, thus causing it to disperse into the liquid rather than attach to the pipe wall. A similar voltage trend can be seen in Figure 4.4, which

records the data from the pipe with a 3.0 mg/L phosphate level, and Figure 4.5, which records the data from the pipe with a 0 mg/L phosphate level. A ground truth experiment was setup to properly understand the ability of the sensor attachment to detect and classify deposits – biofilm, corrosion, or scaling.

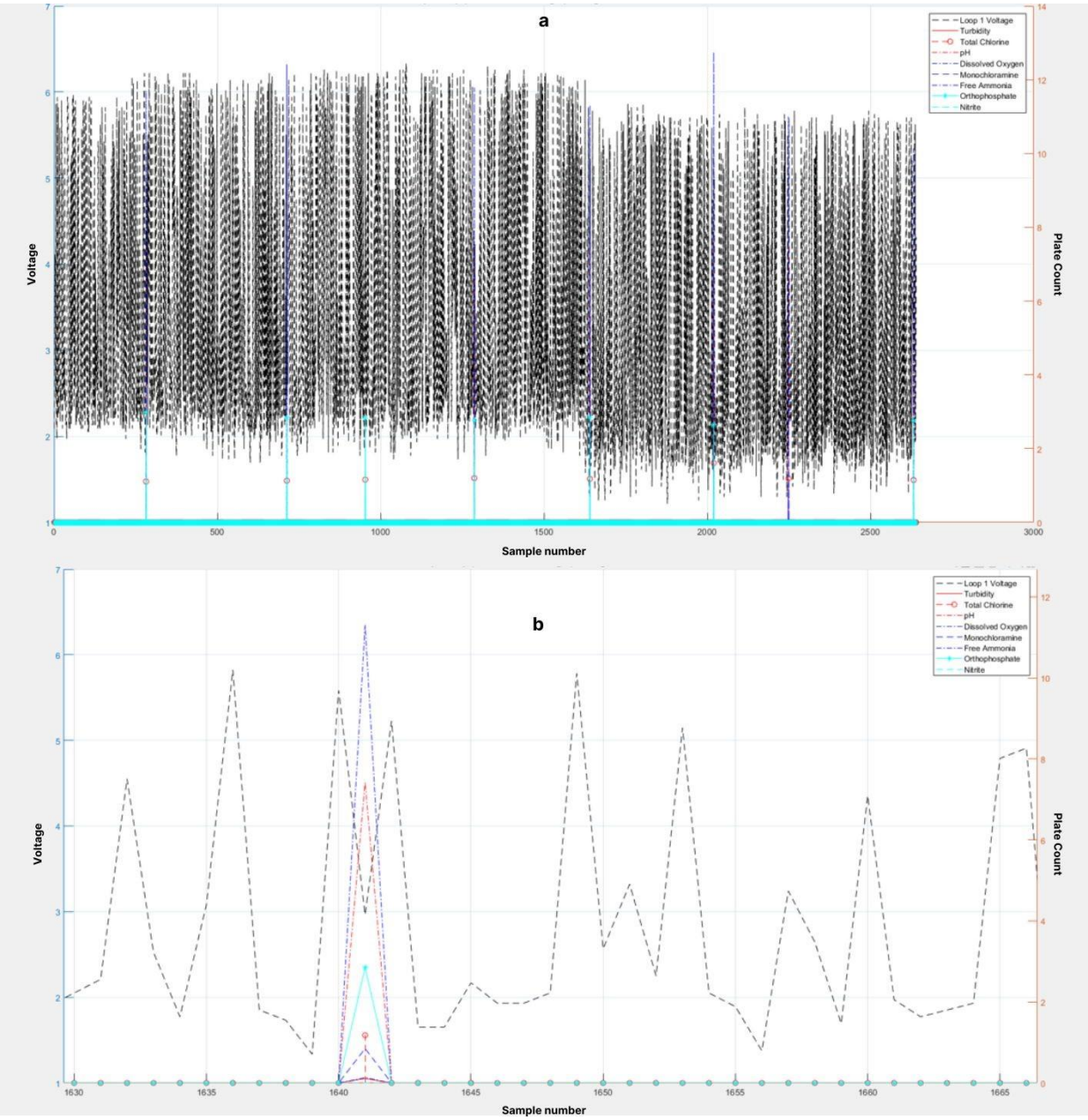

Figure 4.3: Pipe loop experiment setup at the Howard wastewater treatment plant. Parameters were recorded on the pipe with the 1.9 mg/L phosphate level. (a) Graph showing the data of around 2600 sampling points from the experiment. (b) Zoomed-in graph showing a subset of the data for easier readability.

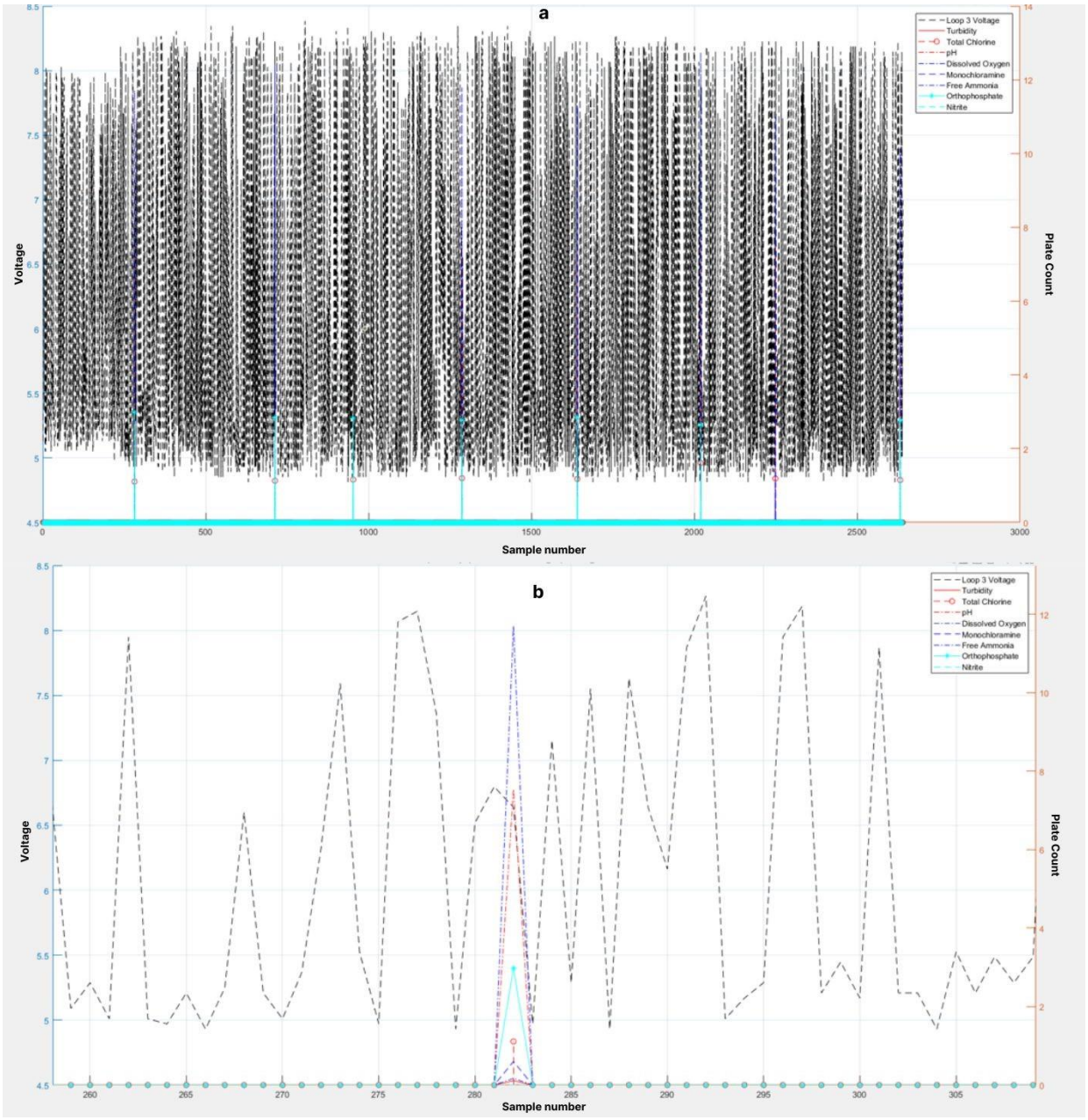

Figure 4.4: Pipe loop experiment setup at the Howard wastewater treatment plant. Parameters were recorded on the pipe with the 3.0 mg/L phosphate level. (a) Graph showing the data of around 2600 sampling points from the experiment. (b) Zoomed-in graph showing a subset of the data for easier readability.

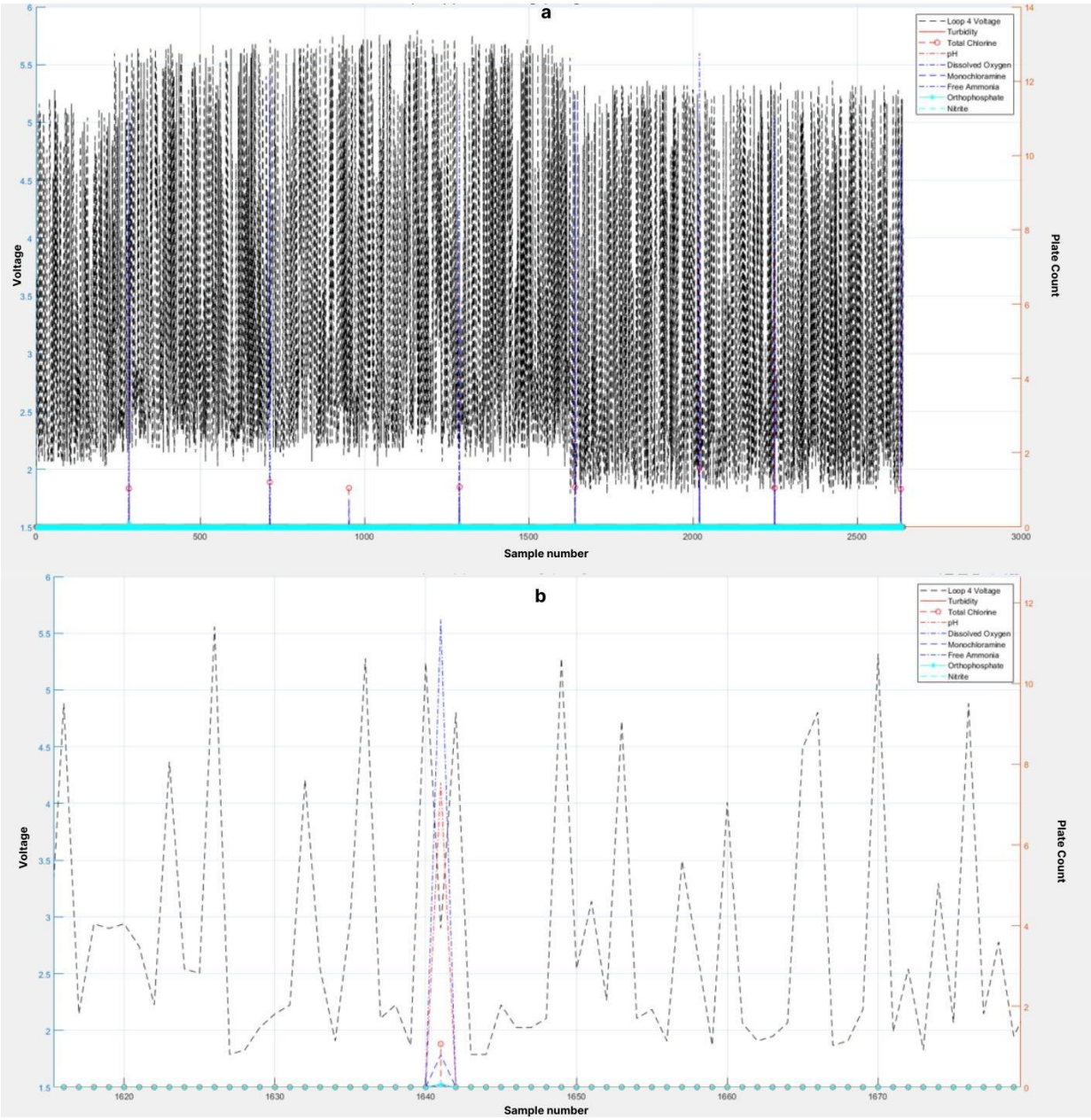

Figure 4.5: Pipe loop experiment setup at the Howard wastewater treatment plant. Parameters were recorded on the pipe with the 0 mg/L phosphate level. (a) Graph showing the data of around 2600 sampling points from the experiment. (b) Zoomed-in graph showing a subset of the data for easier readability.

4.4 Ground truth experiment to classify various deposits using a Machine Learning algorithm.

In this study, the term "biofilm presence" refers to elevated bacterial activity within the plastic container. "Scaling" is defined as the accumulation of minerals caused by hard water, while "corrosion" refers to the presence of metal deposits within the container. To prevent cross-contamination, various containers were employed in the ground truth experiment. Additionally, voltage and time of flight readings from the sensor were obtained prior to adding *E. coli*, corrosion coupon, and hard water to each of the four containers filled with 400 mL of tap water, to better understand the baseline measurements of the test setups before using the data in the ML model.

Table 12: Method for compensation of peak voltage measured across test setups.

Test Setup	**Stimuli**	**Mean peak voltage for tap water (V)**	**Compensation factor (V)**	**Mean peak voltage observed with Stimuli (V)**	**Compensated peak voltage (V)**
1	Tap water	3.2798	0.0000	3.2798	3.2798
2	*E. coli*	3.2498	0.0300	3.2265	3.2565
3	Corrosion Coupon	3.4073	-0.1275	2.3261	2.1986
4	Hard Water	3.2898	-0.0100	3.4078	3.3978

To compensate for the differences from the test setups, a compensation factor for the voltage and time of flight measurements as observed in table 12 and 13 were added to the voltage and time of flight dataset recorded after the addition of various stimuli and the compensated dataset was used to ensure that the ML model was able to identify the effects of the stimuli. Table 12 describes the method for compensation of peak voltages measured across the test setups and table 13 describes the method for compensation of time of flight measured across the test setups. The measurement precision of the peak voltage observations was 0.0025 V.

A. The mean peak voltage recorded on the container where tap water was used as stimuli (test setup 1) was treated as the baseline measurement since no other impurities were added to this test setup, and the setup was left undisturbed. The mean peak voltage recorded was 3.2798 V and the mean time of flight recorded was 84.98 µs.

B. For the test setup 2, where *E. coli* was later added as impurity, the mean peak voltage measured before the addition of stimuli and containing tap water was 3.2498 V and the mean time of flight measured was 85.88 µs. The compensation factor was calculated by substracting the mean value measured on test setup 2 from the mean value measured on test setup 1. The compensation factor was then added to the entire biofilm dataset to compensate for the variation in the initial value. The voltage compensation factor and time of flight compensation factors were calculated as follows:

   Voltage compensation factor = 3.2798 V – 3.2498 V = 0.0300 V

   Time of flight compensation factor = 84.98 µs – 85.88 µs = -0.90 µs

C. For the test setup 3, where corrosion coupon (or copper) was later added as impurity, the mean peak voltage measured before the addition of stimuli and containing tap water was 3.4073 V and the mean time of flight measured was 77.58 µs. The compensation factor was calculated by substracting the mean value measured on test setup 3 from the mean value measured on test setup 1. The compensation factor was then added to the entire corrosion or copper dataset to compensate for the variation in the initial value. The voltage compensation factor and time of flight compensation factors were calculated as follows:

   Voltage compensation factor = 3.2798 V - 3.4073 V = -0.1275 V

   Time of flight compensation factor = 84.98 µs – 77.58 µs = 7.40 µs

D. For the test setup 4, where hard water was later added as impurity, the mean peak voltage measured before the addition of stimuli and containing tap water was 3.2898 V and the mean time of flight measured was 86.18 µs. The compensation factor was calculated by substracting the mean value measured on test setup 4 from the mean value measured on test setup 1. The compensation factor was then added to the entire scaling dataset to compensate for the variation in the initial value. The voltage compensation factor and time of flight compensation factors were calculated as follows:

   Voltage compensation factor = 3.2798 V – 3.2898 V = -0.0100 V

   Time of flight compensation factor = 84.98 µs – 86.18 µs = -1.20 µs

Table 13: Method for compensation of time of flight measured across test setups.

Test Setup	**Stimuli**	**Mean time of flight for tap water (µs)**	**Compensation factor (µs)**	**Mean time of flight observed with Stimuli (µs)**	**Compensated time of flight (µs)**
1	Tap water	84.98	0.00	84.98	84.98
2	*E. coli*	85.88	-0.90	66.40	65.50
3	Corrosion Coupon	77.58	7.40	66.41	73.81
4	Hard Water	86.18	-1.20	64.48	63.28

Figure 4.6 shows a confusion matrix, a performance measurement for ML classification, and a comparison matrix that plots the actual and predicted labels. The confusion matrix indicates that the ML model is able to distinguish biofilm and copper or copper and tapwater with an accuracy score of 100% using a single feature - peak voltage measured.

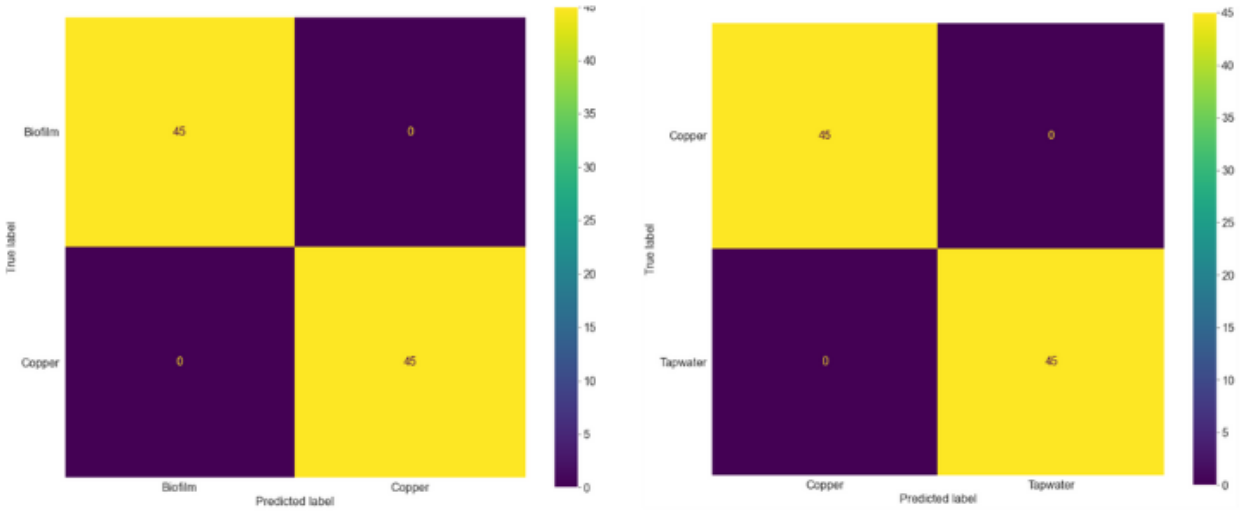

Figure 4.6: The confusion matrix of the machine learning algorithm to classify two types of deposits using peak voltage feature.

Figure 4.7 shows the confusion matrix that indicates the ability of ML model to classify copper and tapwater or biofilm and tapwater with an accuracy score of 100% using a single feature - time of flight measurements.

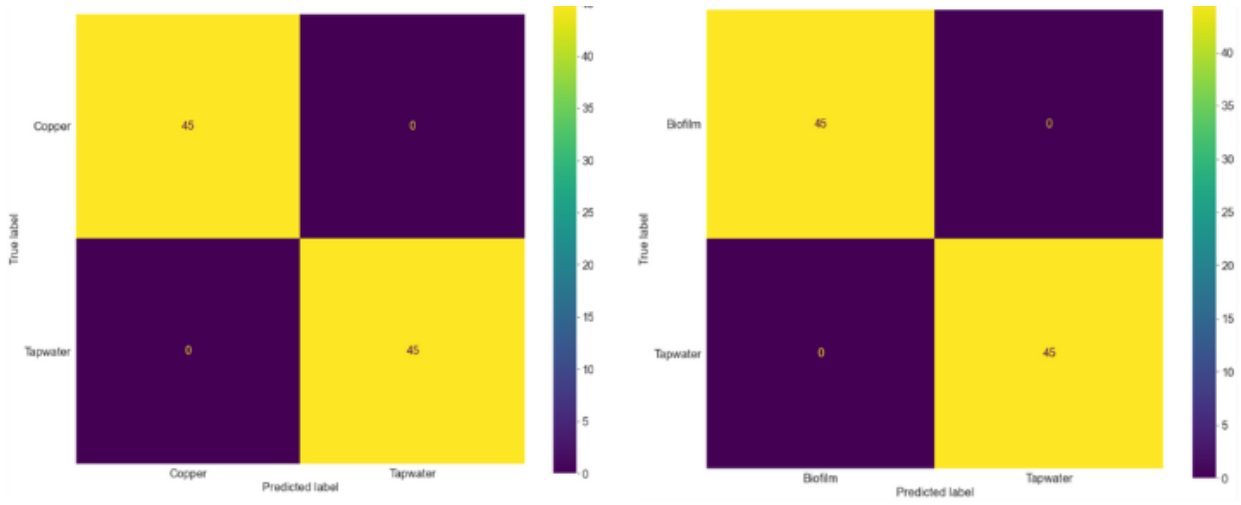

Figure 4.7: The confusion matrix of the machine learning algorithm to classify two types of deposits using peak voltage feature.

The ML algorithm classifies the presence of deposits with the help of changes in voltage and time of flight measurements from the testing data acquired from the sensors. The three classes used in the experiment are biofilm, corrosion, and tap water. Each class has three different features – peak voltage measured, time of flight, and peak voltage ratio. Figure 4.8 shows the confusion matrix for a ML model that classifies the three classes using three features. It can be seen that the algorithm was able to classify the presence of biofilm, corrosion or tapwater correctly except for five tap water samples which the algorithm misclassified as biofilm since the voltage levels were close to each other. The tuned random forest model produced an accuracy of 99.99% for the experiment classifying three different deposits using three features. The model also produced an F1-score of 1.0 indicating a 100% accuracy in class-wise performance.

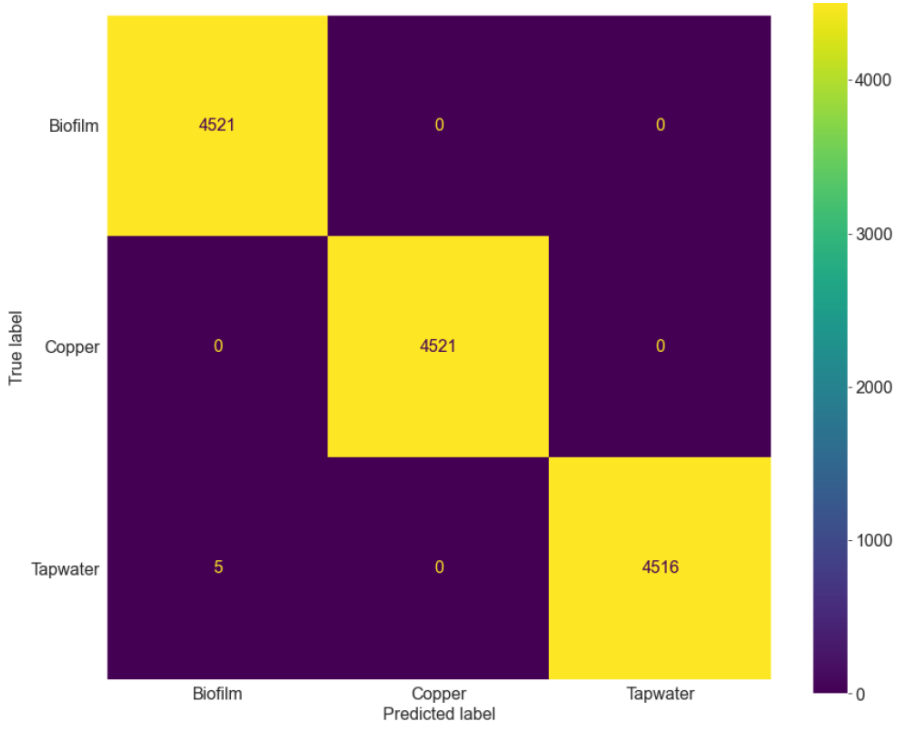

Figure 4.8: The confusion matrix of the machine learning algorithm to classify three types of deposits using three features.

Figure 4.9 indicates the feature importance graph, which ranks the features' importance in classifying three deposits – biofilm, corrosion, and tap water. It can be seen that the peak voltage ratio, which is the ratio of the peak voltage measured on the receiver sensors to the peak voltage measured on the transmitter sensors, is ranked the highest with a feature importance score of around 0.23 followed by the time of flight with a score of around 0.17 and peak voltage measured with a score of around 0.01, indicating that the peak voltage ratio feature was the most used feature by the model in the classification of the three deposits – biofilm, corrosion, and tap water.

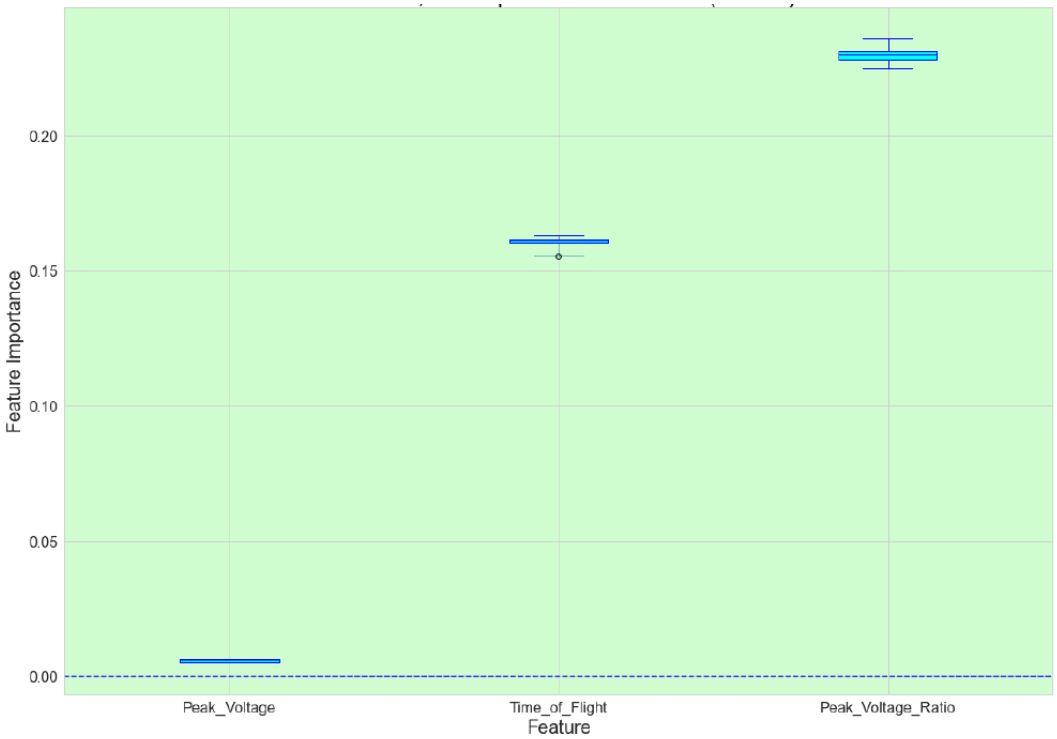

Figure 4.9: Graph showing the feature importance of the machine learning algorithm to classify three types of deposits using three features.

Figure 4.10 shows the confusion matrix for the experiment classifying four deposits using three features. It can be seen that the algorithm was able to classify deposits correctly except for scaling samples which the algorithm misclassified as biofilm since the voltage levels were close to each other. The four classes used in the experiment are biofilm, corrosion, scaling, and tap water.

Each class has three different features – peak voltage measured, time of flight, and peak voltage ratio. The tuned random forest model produced an accuracy of 99.95% for the experiment classifying four different deposits using three features. The model also produced an F1-score of 1.0 indicating a 100% accuracy in class-wise performance.

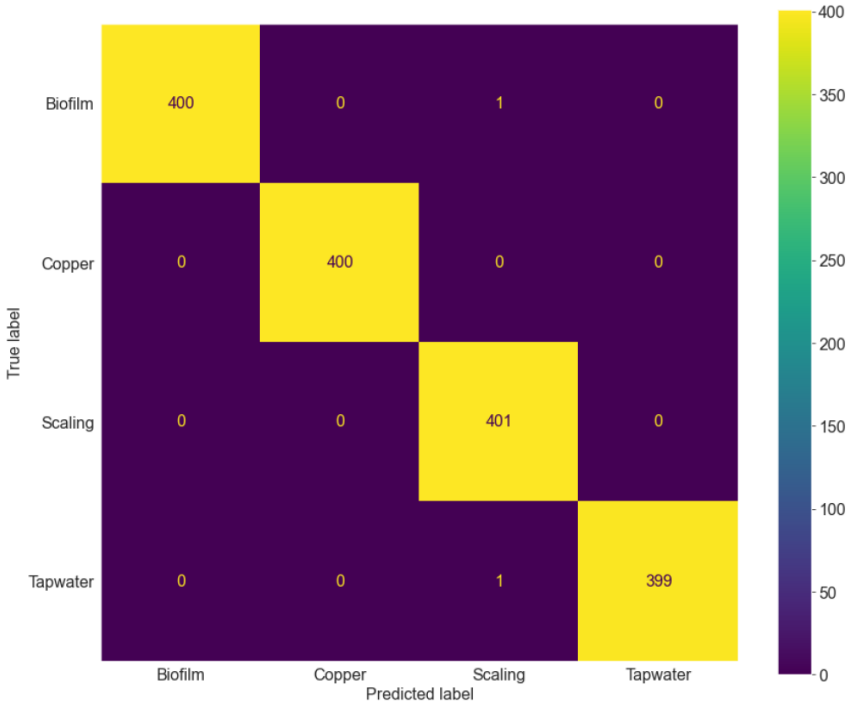

Figure 4.10: The confusion matrix of the machine learning algorithm to classify four types of deposits using three features.

Figure 4.11 indicates the feature importance graph, which ranks the features on their importance in classifying four deposits – biofilm, corrosion, scaling, and tap water. It can be seen that the peak voltage ratio, which is the ratio of the peak voltage measured on the receiver sensors to the peak voltage measured on the transmitter sensors, is ranked the highest with a feature importance score of around 0.44 followed by the time of flight with a score of around 0.01 and peak voltage measured with a score of around 0.01, indicating that the peak voltage ratio feature

was the most used feature by the model in the classification of the three deposits – biofilm, corrosion, and tap water.

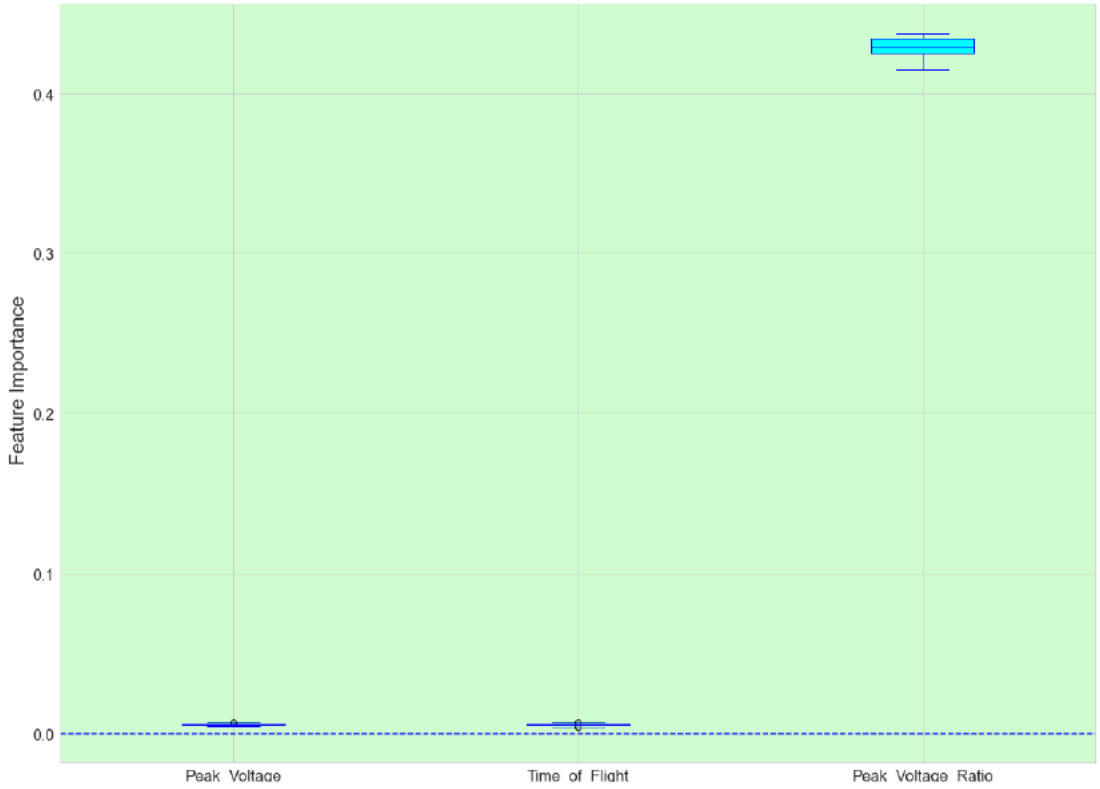

Figure 4.11: Graph showing the feature importance of the machine learning algorithm to classify four types of deposits using three features.

Figure 4.12 shows the confusion matrix for the experiment classifying four deposits using three features. It can be seen that the algorithm was able to classify deposits correctly except for scaling samples which the algorithm misclassified as biofilm since the voltage levels were close to each other. The four classes used in the experiment are biofilm, corrosion, scaling, and tap water. Each class has two different features – peak voltage measured and time of flight. The tuned random forest model produced an accuracy of 99.99% for the experiment classifying four different deposits using two features. The model also produced an F1-score of 1.0 indicating a 100% accuracy in class-wise performance.

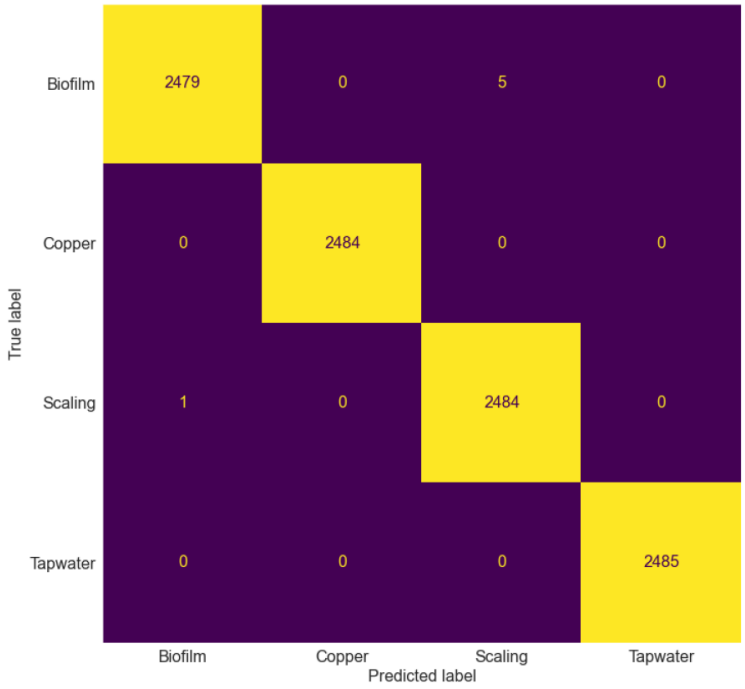

Figure 4.12: The confusion matrix of the machine learning algorithm to classify four types of deposits using two features.

Figure 4.13 indicates the feature importance graph, which ranks the features' importance in classifying four deposits – biofilm, corrosion, scaling, and tap water. It can be seen that the time of flight was ranked the highest, indicating that this feature was the most used by the model in the classification of the four deposits – biofilm, corrosion, scaling, and tap water. In this model, excluding the additional peak voltage ratio was beneficial, which was similar to the peak voltage measured since it attributed to the problem of multi-collinearity, which negatively impacted the classification accuracy. Removing the redundant feature showed that the model's accuracy increased from around 99.95% to 99.99%. The feature importance graph shows that the time of flight measurement was ranked highest with a score of around 0.48 followed by peak voltage measured with a score of around 0.41. From the proof of concept experiments performed earlier in the research, it was demonstrated that the voltage readings varied significantly with deposits or

disturbances in the test chamber. In the previous models, the biofilm deposits and corrosion deposits could be classified properly since the peak voltage ratio was the important feature. In the classification of biofilm and scaling, the time of flight measurements was recognized as the most important feature. However, time of flight measurements are extremely sensitive to variations and additional datasets are necessary to gain a comprehensive understanding of the sensor configuration's efficacy when paired with ML techniques for distinguishing between scaling and biofilm through time of flight measurements.

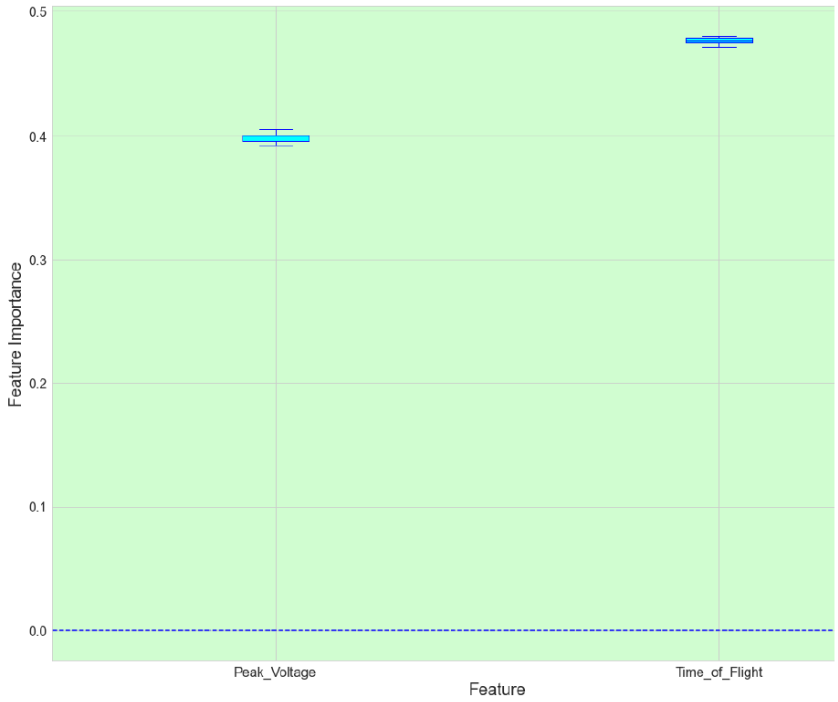

Figure 4.13: Graph showing the feature importance of the machine learning algorithm to classify four types of deposits using two features.

Figure 4.14 shows the confusion matrix for the experiment classifying three pipe structures at the Howard wastewater treatment plant. It can be seen that the algorithm was able to classify the pipe loop correctly except for the minor misclassification of two pipe structures – loop one and loop four. The first loop, labeled loop one, consists of 1.9 mg/L phosphate-dosed water flowing

inside the pipe structure. The second loop, labeled loop three, consists of 3.0 mg/L phosphate-dosed water flowing inside the pipe structure. The third loop, labeled loop four, consists of 0 mg/L phosphate-dosed water flowing inside the pipe structure. Each class has two different features – peak voltage measured and time of flight. The tuned random forest model produced an accuracy of 99.68% for the experiment classifying three pipe structures with the available data.

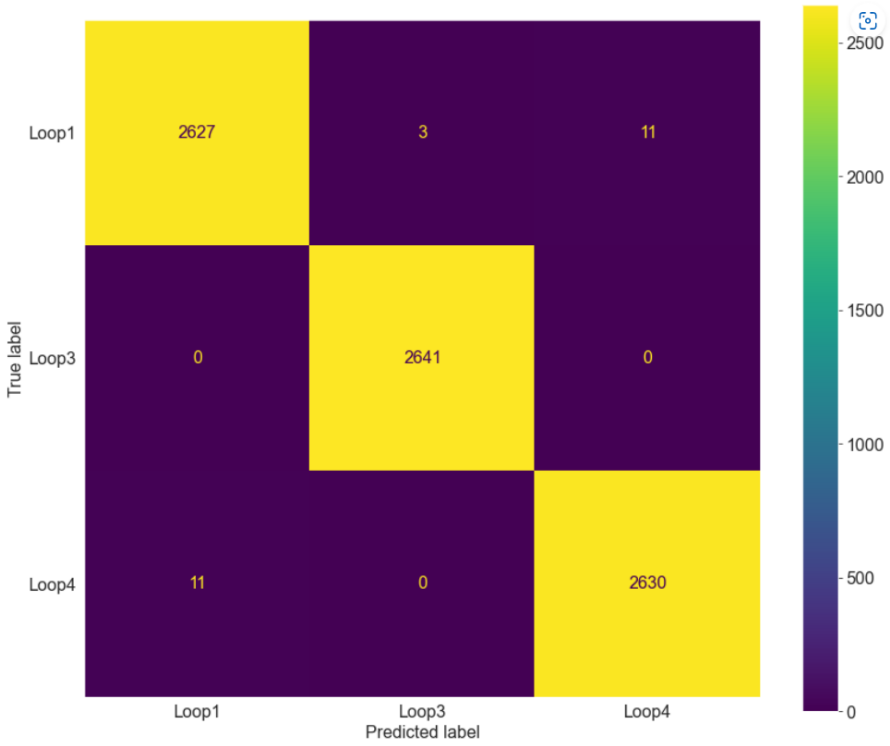

Figure 4.14: The confusion matrix of the machine learning algorithm to classify the three pipe structures at the Howard wastewater treatment plant.

Figure 4.15 shows the confusion matrix for classifying three pipe structures in the laboratory experiment. Each class has two different features – peak voltage measured and time of flight. It can be seen that the algorithm was able to classify the pipe loops correctly except for the minor misclassification of two pipe structures – copper loop and plastic loop. It can also be noted that there was a considerable misclassification of the PEX pipe loop voltages since the voltage levels of the PEX pipe and plastic pipe were similar. However, the misclassification is small

considering the large sample size, and the classification accuracy of the PEX pipe samples was 98.6%. The tuned random forest model produced an overall accuracy of 99.18% for the experiment classifying three pipe structures with the available data.

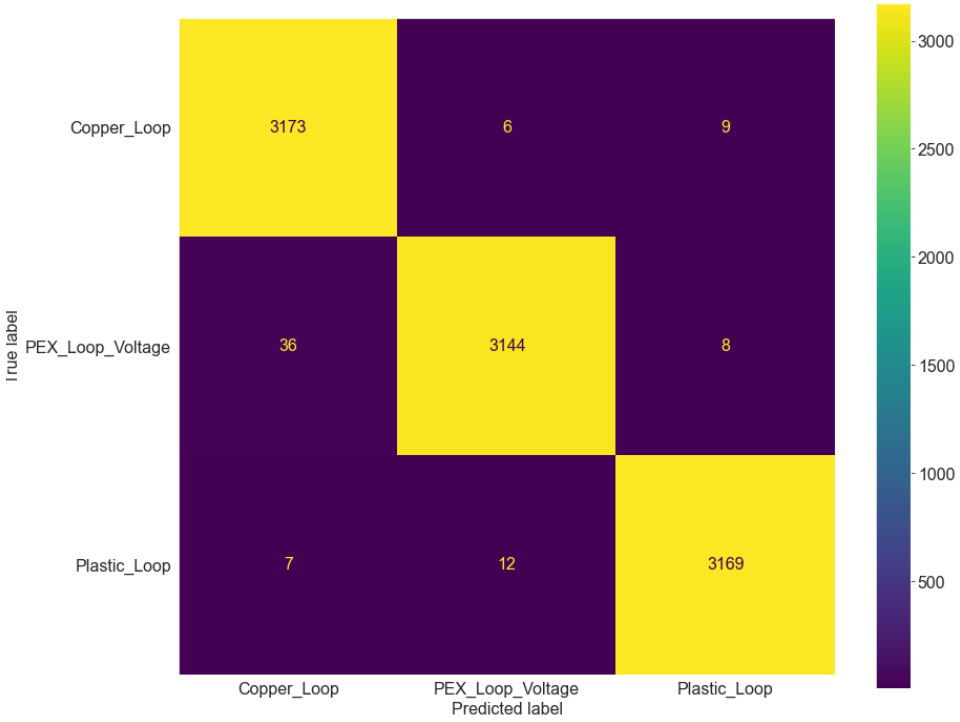

Figure 4.15: The confusion matrix of the machine learning algorithm to classify three pipe structures in the laboratory experiment.

Chapter 5

Customer Discovery Process

5.1 Business Model and Hypothesis

Business Model – A biofilm sensing device placed outside a piping system will provide near real-time results for customers.

Business Model Hypothesis – Detecting biofilm formation in potable water systems will enable maintenance and other management actions to assure public health.

5.2 Customer Discovery Process

It was identified that the key partners of W.R. Tech, an entrepreneurial entity (for the I-Corp program) formed by some authors of this research along with other members of the laboratory and aimed at the manufacturing of a novel-non invasive technique for the detection of biofilm inside pipes, would be the U.S. Department of Agriculture (USDA), U.S. Environmental Protection Agency (EPA), and the Food and Drug Administration (FDA) by enforcing laws or guidelines in the importance of detection and eradication of biofilm in food or water industries. W.R. Tech identified the following value propositions.

- Non-invasive detection technique and does not affect the production line.
- Lower installation costs.
- Help eradicate the biofilm formation at an early stage.
- Elimination of unnecessary cleaning helps reduce production or cleaning costs.
- Improved product safety and fewer recalls due to contamination.

W.R. Tech identified that it would cater to process engineers, operations or production managers, and process supervisors either directly or through partnerships with equipment manufacturers so that the equipment used in the water or food industry is equipped with biofilm detection techniques which would benefit the industries in the long term and improve the quality of water or food-related products.

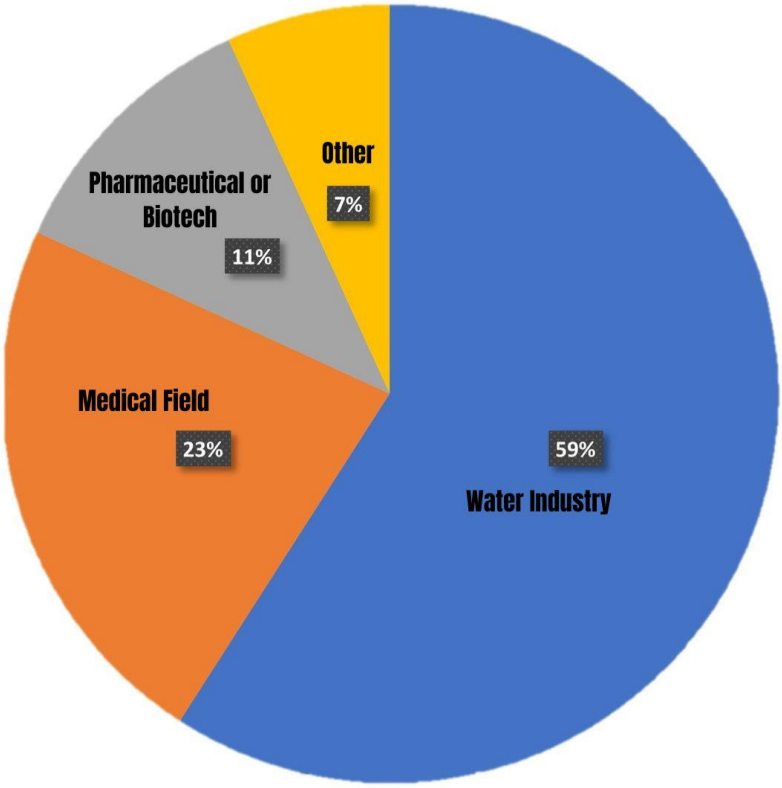

Figure 5.1: A pie chart showing the different industries interviewed during the I-Corps Customer discovery process and their percentage.

Around 44 interviews were conducted in the customer discovery process to understand how effective a biofilm sensor would be in the industry. These interviews were held with professionals or engineers serving as CEOs, R&D Directors, Techincal Directors, Principal Engineers, R&D Engineers, Operations managers, Water quality specialists, Physicians, Dentists, and Business Consultants, to name a few. From Figure 5.1, it can be seen that around 59% of the

interviews conducted by W.R. Tech were with professionals in the Water industry, 23% of the interviews were with professionals in the Medical field, 11% of the interviews were with professionals in the Pharmaceutical or biotech industry, and around 7% of the interviews were with professionals in the other fields like the food industry, schools, and energy generation and distribution.

In the water industry, most of the industry partners mentioned that biofilm formation is a significant problem and that an effective method for detecting biofilm helps monitor water quality and maintenance. Most of the interviewees in the water sector indicated that the sensor technique should have the capacity to be integrated with existing industry data acquisition systems like Supervisory Control and Data Acquisition (SCADA) or Building Automation and Control (BAC) networks. Almost all industries support the hypothesis that detecting water issues can help extend pipe structures' shelf life or longevity and avoids the problem of unwanted flushing. A few companies mentioned that they have preventive methods for preventing biofilm build-up. However, they recognize that biofilm detection can be incorporated into the next generation of sensing technologies. It was also mentioned that corrosion or moisture deposits are also a significant problem, and it would be helpful to have real-time feedback for better maintenance.

In the medical field, while most hospitals practice autoclaving all their instruments or tools for every use, biofilms are a significant problem in the indwelling venus catheter and prosthetic implants. In the health industry, it is crucial to identify biofilm at early stages to help decide on the best treatment. Moreover, biofilm sensors can be best used in plaque control on teeth. Biofilm detection can also be effective for animal care, ensuring that relevant research is not discarded or restarted.

In the pharmaceutical or biotechnology field, it is essential to detect water issues, including biofilm formation upstream of the distribution system, to ensure water quality. Biofilm detection can help ensure that water delivered through the filtration system is safe for drug production. Detecting bacteria can also help ensure that the water chillers or heaters are not a potential source of contamination for patients. Most companies interviewed mentioned hiring professionals to deal with microbe issues since they are difficult to control and are ready to invest in biofilm detection technologies to provide enhanced production control.

In other industries, especially the food industry, the company interviewed recognized that they could save on manufacturing or overhead cost if biofilm can be detected since they usually stop their production at regular intervals for cleaning.

Overall, during the interviews, around 30 of the 44 interviewees mentioned a need for sensors to detect or predict contamination problems before they occur. Almost 40 out of 44 companies mentioned that the sensors or the sensing technique should be affordable. Five companies mentioned actively investing in research and development to solve the biofilm problem and that biofilm sensors would have the potential to be a global technology, and mentioned that they would prefer self-calibrating sensors. At least 6 out of 44 companies wanted any new technology integrated with their current setup to make incorporating technology more straightforward. They also mentioned that laws play a role in the technologies companies adopt since the test process of new technology is expensive. One company mentioned that they had recently invented a biofilm detector but were actively looking for newer technology or using a different sensing technique for the time being. However, they did not reveal the reasons for switching the detection techniques due to proprietary issues. The story arc of the I-Corp program is seen in Figure 5.2.

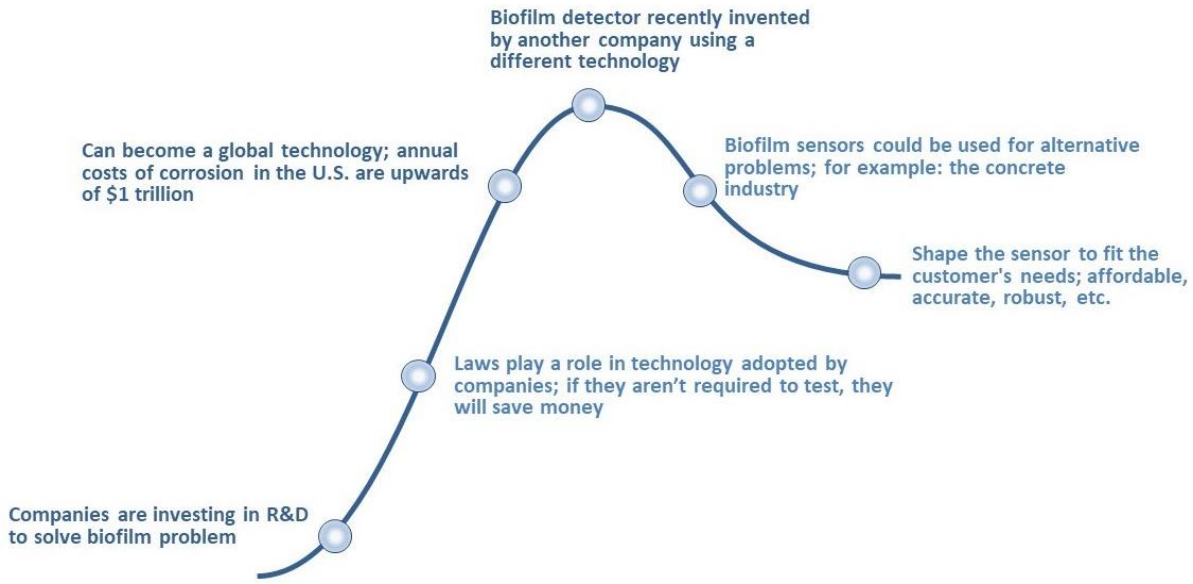

Figure 5.2: A story arc of the critical responses during the customer discovery process.

From the I-Corp program, W.R. Tech was able to conclude that there is an immediate requirement for an effective technique for the detection and that the applications of the technology can be extended to several industries other than the water industry:

- In the medical field, for the detection of bacterial attachment in diagnostic devices that cannot be cleaned easily,
- In the food industry, where detection of bacterial presence is extremely vital to avoid contamination of food and prevent liabilities in the future,
- In pharmaceutical industries, biofilm can cost millions of dollars in liabilities and recall of contaminated products.

Chapter 6

Conclusions and Future Research

It was observed that the 1 MHz ultrasound sensor, in combination with an eight-burst sinusoidal signal, was the most effective in detecting biofilm in closed-wall piping systems. The sensor voltage and time of flight variations were correlated to the deposit inside the piping system. The presence of corrosion or metal deposit inside the piping system dramatically impacts the receiver voltage. Combining the non-invasive technique for detecting biofilm and classifying deposits – no deposit, biofilm deposit, and corrosion deposit using the random forest ML algorithm is an effective method. The random forest ML algorithm can classify the deposits with an overall accuracy of 99.99%.

While the research demonstrates excellent results in detecting biofilm and classifying between biofilm and corrosion in the ground truth experiments, large datasets are still required to reinforce the effectiveness of the combination of the non-invasive detection of sensors and the use of ML in data classification in a real-world environment, especially in the classification of biofilm and scaling or other similar deposits to obtain a classification model with an accuracy of 100%. Another limitation of the current method is that a baseline measurement for sensor data is required before it can be used to predict the presence of biofilm. The baseline data requirement means that for proper detection and classification of biofilm or other deposits, it is necessary first to calculate the voltage and time of flight of the piping structure since the method relies on the change in voltage and time of flight to predict the formation or presence of biofilm or other deposits. Another major limitation of the current sensor arrangement is that it can only detect the deposits in a small cross-section of the pipe loop. If deposits (biofilm, scaling, or corrosion) are formed at a different pipe cross-section, the sensors cannot correctly detect or classify the deposits.

In future research, additional methods need to be explored to improve the performance of the detection method using ultrasound sensors. An effective strategy would be to use digitally encoded signals and study the changes in these signals to predict biofilm detection more accurately. The use of ultrasound sensors on the same side of the pipe utilizing the multiple internal reflection method should be explored to detect the presence of deposits in a long cross-sectional area of the pipe. This dissertation is a strong foundation for this future work, and is part of an invention disclosure filed in June 2023 [1]. Additionally, a consumer study based on cost should be conducted to comprehend the marketability of ultrasound techniques for sensing biofilms and the industry's willingness to invest in such technology. The commercial iteration of the study, which will be developed later, will feature a duo of ultrasound sensors fastened to pipe surface with a clamp and linked to a touch-enabled screen using a cable. While the ultrasound sensors would have to be attached permanently to the pipe surface, the handheld touch screen interface can be used to actuate the different sensors and analyze the various deposits. Through this handheld tablet-like interface, the user can initiate ultrasound readings and receive instantaneous feedback regarding the buildup in the pipeline. Converting the code for the ML algorithm and data processing to enable remote monitoring or integration with the Internet of Things can be done effortlessly using platforms like Microsoft Azure or others with similar capabilities.

References

[1] Y. Adibi, S. Davis, M. Quadros da Silva and R. Patrick O'Day, "Novel non-invasive detection of thin film biofilm and classification of deposits using machine learning.". Patent IP-0009, 06 2023.

[2] World Health Organization, "Drinking-water," World Health Organization, 21 03 2022. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/drinking-water#:~:text=Contaminated%20water%20and%20poor%20sanitation,individuals%20to%20preventable%20health%20risks.. [Accessed 22 03 2023].

[3] U. Szewzyk, R. Szewzyk, W. Manz and K. H. Schleifer, "Microbiological Safety of Drinking Water," *Annual Review of Microbiology*, vol. 54, 2000.

[4] "Health Risks from Microbial Growth and Biofilms in Drinking Water Distribution Systems," United States Environmental Protection Agency, Washington D.C, 2002.

[5] M. Schaechter, NC Engelberg, BI Eisenstein and G Medoff, Mechanisms of microbial disease, 3 ed., Baltimore, Maryland: Williams and Wilkins, 1998.

[6] WG Mackay, LT Gribbon, MR Barer and DC Reid, "Biofilms in drinking water systems – a possible reservoir for Helicobacter pylori," *Journal of Applied Microbiology*, vol. 38, no. 12, pp. 181 - 185, 1998.

[7] S Fass, ML Dincher, DJ Reasoner, D Gatel and JC Block, "Fate of Escherichia coli experimentally injected in a drinking water distribution pilot system," *Water Research*, vol. 30, no. 9, pp. 2215 - 2221, 1996.

[8] DL Swerdlow, BA Woodruff, RC Brady, PM Griffin, S Tippen, HD Donnell Jr, E Geldreich, BJ Payne, A Meyer Jr and JG Wells, "A waterborne outbreak in Missouri of Escherichia coli O157:H7 associated with bloody diarrhea and death," *Annals of Internal Medicine*, vol. 117, no. 10, pp. 812 - 819, 1992.

[9] GJ Kirmeyer, M Friedman, KD Martel, D Howe, M LeChevallier, M Abbaszadegan, M Karim, J Funk and J Harbour, Pathogen intrusion into the distribution system, Denver, Colorado: Amer Water Works Association, 2001.

[10] W. P. Iverson, "Microbial Corrosion, Technical Summary Report Number 1," National Bureau of Standards, Washington, D.C., 1968.

[11] T. Ford and R. Mitchell, The Ecology of Microbial Corrosion, New York: Plenum Press, Advances in Microbial Corrossion .

[12] J.-D. Gu, Corrosion, Microbial, Hong Kong: Elsevier, 2019.

[13] V. S. Agarwala, P. L. Reed and S. Ahmad, "Corrosion Detection and Monitoring - A Review," in *NACE International*, Orlando, Florida, 2000.

[14] W. P. Iverson, "Microbial Corrosion of Metals," *Advances in Applied Microbiology*, vol. 32, pp. 1 - 36, 1987.

[15] H. Zhang, Y. Tian, J. Wan and P. Zhao, "Study of biofilm influenced corrosion on cast iron pipes in reclaimed water," *Applied Surface Science*, vol. 357, pp. 236-247, 2015.

[16] W. Kaempfer and M. Berndt, "Estimation of service life of concrete pipes in sewer networks," *Materials Science*, 1999.

[17] N. Nwagha, "Statistical study on the corrosion of mild steel in saline mediums".

[18] "The effect of sulphuric acid on storage tanks," Sulphuric acid on the Web, 2005. [Online]. Available: https://www.sulphuric-acid.com/TechManual/Storage/storagetanks.htm. [Accessed 02 05 2023].

[19] S. H. Flint, P. J. Bremer and J. D. Brooks, "Biofilms in dairy manufacturing plant - description, current concerns and methods of control," *Biofouling,* vol. 11, no. 1, pp. 81-97, 1997.

[20] O. P. J. Snyder, "HACCP-TQM for retail and food service operations," in *Advances in Meat Research - Volume 10 HACCP in Meat, Poultry and Fish Processing*, London, New York, Blackie Academic & Professional, 1995.

[21] A. C. L. Wong and O. Cerf, "Biofilms: Implications for hygiene monitoring of dairy plant surfaces," *Bulletin of the International Dairy Federation,* vol. 302, pp. 40-44, 1995.

[22] S. J. Forsythe and P. R. Hayes, Food Hygiene, Microbiology and HACCP, vol. 3, Springer, 1998.

[23] R. M. Donlan and J. Carr, "CDC," 2005. [Online]. Available: https://phil.cdc.gov/default.aspx.

[24] T. M. Mosteller and J. R. Bishop, "Sanitizer Efficacy Against Attached Bacteria in a Milk Biofilm," *Journal of Food Protection,* vol. 56, no. 1, pp. 34-41, 1993.

[25] G. D. Christensen, W. A. Simpson, J. A. Younger, L. M. Baddour, F. F. Barrett and D. M. Melton, "Adherence of cogulase negative Staphylococi to plastic tissue cultures:a quantitative model for the adherence of staphylococci to medical devices.," *Journal of Clinical Microbiology,* vol. 22, pp. 996-1006, 1985.

[26] The Editors of Encyclopaedia Britannica, "Tissue Culture," Britannica, 20 07 1998. [Online]. Available: https://www.britannica.com/science/tissue-culture/additional-info#history. [Accessed 03 04 2023].

[27] T. Mathur, S. Singhal, S. Khan, D. J. Upadhyay, T. Fatma and A. Rattan, "Detection of biofilm formation among the clinical isolates of Staphylococci: An evaluation of three different screening methods," *Indian Journal of Medical Microbiology,* vol. 24, no. 1, pp. 25-29, 2006.

[28] Alcibiades, "Wikipedia," 16 02 2006. [Online]. Available: https://en.wikipedia.org/wiki/Chinese_hamster_ovary_cell#/media/File:Cho_cells_adherend2.jpg.

[29] G. D. Christensen, W. A. Simpson, A. L. Bisno and E. H. Beachey, "Adherence of slime–producing strains of Staphylococcus epidermidis to smooth surfaces.," *Infection and Immunity,* vol. 37, pp. 318-326, 1982.

[30] D. J. Freeman, F. R. Falkiner and C. T. Keane, "New method for detecting slime production by coagulase negative staphylococci.," *Journal of Clinical Pathology,* vol. 42, pp. 872-874, 1989.

[31] K. G. Porter and Y. S. Feig, " The use of DAPI for identifying and counting aquatic microflora.," *Limnology and Oceanography,* vol. 25, pp. 943-948, 1980.

[32] "Image Processing and Analysis in Java, NIH," 29 07 2010. [Online]. Available: http://rsb.info.nih.gov/ij/images/.

[33] A. W. Bauer, W. M. M. Kirby, J. C. Sherris and M. Turck, "Antibiotic sensitivity testing by a standardized single disk method," *American Journal of Clinical Pathology,* vol. 45, pp. 493-496, 1966.

[34] T. D. Wikins, L. V. Holdeman, I. J. Abramson and W. E. C. Moore, "Standardized Single-Disc Method for Antibiotic Susceptibility Testing of Anaerobic Bacteria," *Antimicrobial Agents and Chemotherapy,* vol. 1, no. 6, pp. 451-459, 1972.

[35] "U.S. National Oceanic and Atmospheric Administration," 08 12 2022. [Online]. Available: http://oceanexplorer.noaa.gov/explorations/04etta/background/antimicrobial/media/antimicrobial2.html.

[36] P. R. Langer-Safer, M. Levine and D. C. Ward, "Immunological method for mapping genes on Drosophila polytene chromosomes," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 79, no. 14, pp. 4381-4385, 1982.

[37] H. Frickmann, A. E. Zautner, A. Moter, J. Kikhney, R. M. Hagen, H. Stender and S. Poppert, "Fluorescence in situ hybridization (FISH) in the microbiological diagnostic routine laboratory: a review," *Critical Reviews in Microbiology,* vol. 43, no. 3, pp. 263-293, 2017.

[38] T. Ried, "Wikipedia," 02 11 2006. [Online]. Available: https://en.wikipedia.org/wiki/Fluorescence_in_situ_hybridization#/media/File:FISH_(technique).gif.

[39] N. P. Ivleva, M. Wagner, A. Szkola, H. Horn, R. Niessner and C. Haisch, "Label-Free in Situ SERS Imaging of Biofilms," *Journal of Physical Chemistry B,* vol. 114, no. 31, pp. 10184-10194, 2010.

[40] "Nikalyte," [Online]. Available: https://www.nikalyte.com/2022/04/07/how-does-surface-enhanced-raman-spectroscopy-work/.

[41] B. Park, S. Lee, S.-C. Yoon, J. Sundaram, W. Windham, A. J. Hington and K. C. Lawrence, "AOTF hyperspectral microscopic imaging for foodborne pathogenic bacteria detection," in *Proceedings of SPIE 8027, Sensing for Agriculture and Food Quality and Safety III*, Orlando, Florida, 2011.

[42] N. M. Short, Sr., "NASA," [Online]. Available: https://web.archive.org/web/20110703101153/http:/rst.gsfc.nasa.gov/Intro/Part2_24.html.

[43] G. Chen, R. J. Palmer and D. C. White, "Instrumental analysis of microbiologically influenced corrosion," *Biodegradation,* vol. 8, pp. 189-200, 1997.

[44] T. Stricklin, "What Is Hard Water & How To Test For Water Hardness," Spring Well , 15 02 2023. [Online]. Available: https://www.springwellwater.com/3-easy-ways-to-test-for-hard-water/. [Accessed 03 05 2023].

[45] "Nondestructive Evaluation," [Online]. Available: https://www.nde-ed.org/NDETechniques/Ultrasonics/index.xhtml.

[46] I. V. Deutsch, "History of NDT Instrumentation," AIPnD, 2000. [Online]. Available: https://www.ndt.net/article/wcndt00/papers/idn378/idn378.htm. [Accessed 15 10 2018].

[47] S. Davis, "Novel Non-Invasive Technology for the Detection of Thin Biofilm in Piping Systems (Phase - 1)," UWM Digital Commons, Milwaukee, 2019.

[48] "Rail Inspection," NDT Resource Center, 1996. [Online]. Available: https://www.nde-ed.org/EducationResources/CommunityCollege/Ultrasonics/SelectedApps/railinspection.htm. [Accessed 5 2 2018].

[49] "Basic Principles of Ultrasonic Testing," NDT Resource Center, 1996. [Online]. Available: https://www.nde-ed.org/NDETechniques/Ultrasonics/Introduction/description.xhtml. [Accessed 02 02 2018].

[50] D. M. A. Morgan, "Ultrasound (Introduction)," Radiopaedia, [Online]. Available: https://radiopaedia.org/articles/ultrasound-introduction?lang=us. [Accessed 26 10 2018].

[51] A. Aubry and A. Derode, "Multiple Scattering of Ultrasound in Weakly Inhomogeneous Media: Application to Human Soft Tissues," *arXiv.org,* pp. 225-233, 2011.

[52] M. Pakula, "Attenuation and dispersion of ultrasound in cancellous bone. Theory and experiment," *Journal of the Acoustical Society of America,* vol. 140, no. 9, p. 3080, 2016.

[53] M. Dzida, E. Zorebksi, M. Zorebski, M. Zarska, M. Geppert-Rybaczynska, M. Chorazewski, J. Jacquemin and I. Cibulka, "Speed of Sound and Ultrasound Absorption in Ionic Liquids," *Chemical Reviews,* vol. 117, pp. 3883-3929, 2017.

[54] T. Stilson, "Piezoelectric Sensors," Princeton Sound Lab, 17 10 1996. [Online]. Available: https://soundlab.cs.princeton.edu/learning/tutorials/sensors/node7.html. [Accessed 03 04 2023].

[55] Electrical4U, "Piezoelectric Transducer: Applications & Working Principle," Electrical 4 U, 28 10 2020. [Online]. Available: https://www.electrical4u.com/piezoelectric-transducer/. [Accessed 03 04 2023].

[56] CLI, "Wikimedia," 28 01 2011. [Online]. Available: https://commons.wikimedia.org/wiki/File:Piezoelectric_effect.svg.

[57] S. Davis and M. R. Silva, "A Proof-of-Concept Study on Utilizing a Novel Non-invasive Sensor for Detection of Thin Biofilm in Simulated Water Pipes," *Sensing and Imaging,* vol. 22, no. 21, 2021.

[58] S. Davis, N. Salowtiz, L. Beversdorf and M. R. Silva, "The Effect of Various Parameters on a Portable Sensor for the Detection of Thin Biofilms in Water Pipes," *Sensors,* vol. 21, no. 13, 2021.

[59] E. Kujundzic, A. C. Fonseca, E. A. Evans, M. Peterson, A. R. Greenberg and M. Hernandez, "Ultrasonic monitoring of earlystage biofilm growth on polymeric surfaces," *Journal of Microbiological Methods,* vol. 68, no. 3, pp. 458-467, 2007.

[60] H. Shemesh, D. E. Goertz, L. W. M. van der Sluis, N. de Jong, M. K. Wu and P. R. Wesselink, "High frequency ultrasound imaging of a single-species biofilm," *Journal of Dentistry,* vol. 35, no. 8, pp. 673-678, 2007.

[61] S. T. V. Sim, S. R. Suwarno, Y. X. S. Lim, W. X. J. Lim, T. H. Chong and A. G. Fane, "Development of novel acoustic sensor for early detection of biofouling in reverse-osmosis systems," in *Procedia Engineering,* 2012.

[62] F. Seida, C. Flocken, P. Bierganns and M. Schultz, "Device And Method For Detecting Deposits". US Patent US20150000407A1, 26 09 2017.

[63] P. Bierganns and M. M. Broecher, "Device and Method for Detecting and Analyzing Deposits". US Patent US20160076990A1, 30 10 2018.

[64] T. Mitchell, Machine Learning, New York: McGraw Hill, 1997.

[65] A. Samuel, "Some studies in Machine learning using the game of checkers," *IBM Journal of Research and Development,* vol. 3, no. 3, pp. 210-229, 1959.

[66] C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

[67] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, Prentice Hall, 2010.

[68] M. Mohri, A. Rostamizadeh and A. Talwalkar, Foundations of Machine Learning, The MIT Press, 2012.

[69] T. Mitchell, Machine Learning, McGraw Hill, 1997.

[70] M. Studer, G. Ritschard, A. Gabadinho and N. Muller, "Discrepancy Analysis of State Sequences," *Sociological Methods and Research,* vol. 40, no. 3, pp. 471-510, 2011.

[71] S. Milborrow, "Wikimedia," 18 06 2020. [Online]. Available: https://commons.wikimedia.org/wiki/File:Decision_Tree_-_survival_of_passengers_on_the_Titanic.jpg.

[72] J. Gareth, D. Witten, T. Hastie and R. Tibshirani, An Introduction to Statistical Learning, New York: Springer, 2015.

[73] M. Bramer, Principles of Data Mining, London: Springer, 2007.

[74] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning, Springer, 2001.

[75] T. K. Ho, "Random Decision Forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition,* Montreal, 1995.

[76] L. Breiman, "Random Forests," *Machine Learning,* vol. 45, no. 1, pp. 5-32, 2001.

[77] "What is random forest?," IBM, [Online]. Available: https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.. [Accessed 25 04 2023].

[78] V. Jagannath, "Wikimedia," 24 03 2017. [Online]. Available: https://commons.wikimedia.org/wiki/File:Random_forest_diagram_complete.png.

[79] K. M. Andrade, B. P. Menezes Silva, L. R. de Oliveira and P. R. Cury, "Automatic dental biofilm detection based on deep learning," *Jorunal of Clinical Periodontology*, vol. 50, no. 5, pp. 571-581, 2023.

[80] G. Dimauro, F. Deperte, R. Maglietta, M. Bove, F. L. Gioia, V. Reno, L. Simone and M. Gelardi, "A Novel Approach for Biofilm Detection Based on a Convolutional Neural Network," *Electronics*, vol. 9, no. 6, 2020.

[81] M. Artini, A. Patsilinakos, R. Papa, M. Bozovic, M. Sabatino, S. Garzoli, G. Vrenna, M. Tilotta, F. Pepi and R. Ragno, "Antimicrobial and antibiofilm activity and machine learning classification analysis of essential oils from different Mediterranean plants using Pseudomonas aeruginosa," *Molecules*, vol. 23, 2018.

[82] A. Patsilinakos, M. Artini, R. Papa, M. Sabatino, M. Bozovic, S. Garzoli, G. Vrenna, R. Buzzi, S. Manfredini and L. Selan, "Machine learning analyses on data including essential oil chemical composition and in vitro experimental antibiofilm activities against Staphylococcus species," *Molecules*, vol. 24, 2019.

[83] J. Wang, Z. Jiang, Y. Wei, W. Wang, F. Wang, Y. Yang, H. Song and Q. Yuan, "Multiplexed identification of bacterial biofilm infections based on machine-learning-aided lanthanide encoding.," *ACS Nano*, vol. 16, pp. 3300-3310, 2022.

[84] W. Nash, T. Drummond and N. Birbilis, "A review of deep learning AI for corrosion detection," in *Corrosion, NACE International*, Orlando, 2019.

[85] A. G. Wilson, "The Case for Bayesian Deep Learning," 29 01 2020. [Online]. Available: https://doi.org/10.48550/arXiv.2001.10995. [Accessed 30 04 2023].

[86] "1ME21TR-1 Ultrasonic sensor," Osenon, [Online]. Available: http://www.osenon.com/en/cp_view.asp?/43.html. [Accessed 18 03 2023].

[87] "400E10TR-1 Ultrasonic Sensor," Osenon, [Online]. Available: http://www.osenon.com/en/cp_view.asp?/17.html. [Accessed 18 03 2023].

[88] "2ME20TR-1 ultrasonic sensor," Osenon, [Online]. Available: http://www.osenon.com/en/cp_view.asp?/87.html. [Accessed 18 03 2023].

[89] "Raspberry Pi 4 8GB," CanaKit, [Online]. Available: https://www.canakit.com/raspberry-pi-4-8gb.html. [Accessed 18 03 2023].

[90] "Raspberry Pi 4," [Online]. Available: https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/. [Accessed 18 03 2023].

[91] "EVICIV Raspberry Pi 10.1 Inch Touchscreen Display with Rear Housing - 1280x800 Support, Type-C, IPS 178 degree Ultra-Wide View Angle Monitor with Cooling Fan, 10 Finger Capacitive Touch," Amazon, [Online]. Available: https://www.amazon.com/Raspberry-Touch-Screen-Display-Case/dp/B098785YZJ/ref=sr_1_14?crid=2BNIL9KZGI8XR&keywords=raspberry%2Bpi&qid=1645830897&sprefix=raspberry%2Bpi%2Caps%2C87&sr=8-14&th=1. [Accessed 18 03 2023].

[92] "Electronics Explorer," Digilent, [Online]. Available: https://digilent.com/reference/test-and-measurement/electronics-explorer/start. [Accessed 18 03 2023].

[93] "Digital Waveforms," Digilent, [Online]. Available: https://digilent.com/shop/software/digilent-waveforms/. [Accessed 18 03 2023].

[94] "MATLAB," Mathworks, [Online]. Available: https://www.mathworks.com/products/matlab.html. [Accessed 18 03 2023].

[95] "Homepage," Jupyter, [Online]. Available: https://jupyter.org. [Accessed 18 03 2023].

[96] "Overview," JupyterLab, 2018. [Online]. Available: https://jupyterlab.readthedocs.io/en/stable/getting_started/overview.html. [Accessed 18 03 2023].

[97] G. Bertani, "Studies on Lysogenesis I, The Mode of Phage Liberation by Lysogenic Escherichia coli," *Journal of Bacteriology,* vol. 62, no. 3, pp. 293 - 300, 1951.

[98] J. Sambrook and D. Russell, Molecular Cloning: A Laboratory Manual, vol. 3, Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press, 2001.

[99] P. Gerhardt, R. G. E. Murray, W. A. Wood and N. R. Krieg, Methods for general and molecular bacteriology, Washington, D. C: American Society for Microbiology, 1994.

[100] R. K. Oshiro, "Method 1603: Escherichia coli (E. coli) in Water by Membrane Filtration using modified membrane-thermotolerant Escherichia coli agar (Modified mTEC)," United States Environmental Protection Agency, Washington D.C., 2014.

[101] P. Vadapalli, "Random Forest Hyperparameter Tuning: Processes Explained with Coding," upGrad, 23 09 2022. [Online]. Available: https://www.upgrad.com/blog/random-forest-hyperparameter-tuning/. [Accessed 28 04 2023].

[102] "NSF's Innovation Corps (I-Corps™)," National Science Foundation, [Online]. Available: https://new.nsf.gov/funding/initiatives/i-corps#:~:text=The%20U.S.%20National%20Science%20Foundation%27s,transformation%20of%20invention%20to%20impact.. [Accessed 25 04 2023].

[103] "About I-Corps," National Science Foundation, [Online]. Available: https://new.nsf.gov/funding/initiatives/i-corps/about-i-corps. [Accessed 25 04 2023].

[104] C. A. Speed, "Therapeutic ultrasound in soft tissue lesions," *Rheumatology,* vol. 40, no. 12, pp. 1331-1336, 2001.

[105] H. Shekhani, "Piezo SHOCK Show #34: Should I use an oscillator circuit to drive my ultrasonic transducer?," Ultrasonic Advisors, 25 09 2021. [Online]. Available: https://www.ultrasonicadvisors.com/piezo-shock-show-34-should-i-use-an-oscillator-circuit-to-drive-my-ultrasonic-transducer. [Accessed 26 04 2023].

[106] G. Karki, "Measurement of bacterial growth using UV spectrophotometer," One Biology Notes, 23 05 2020. [Online]. Available: https://www.onlinebiologynotes.com/measurement-of-bacterial-growth-using-uv-spectrophotometer/#:~:text=Increased%20turbidity%20in%20a%20culture,and%20thus%20quantifies%20the%20turbidity.. [Accessed 29 05 2023].

[107] "Turbidity," CorrosionPedia, 19 09 2022. [Online]. Available: https://www.corrosionpedia.com/definition/1123/turbidity#:~:text=Higher%20levels%20of%20turbidity%20increase%20the%20likelihood%20of%20corrosion.. [Accessed 29 05 2023].

[108] A. Jang, P. L. Bishop, S. Okabe, S. G. Lee and I. S. Kim, "Effect of dissolved oxygen concentration on the biofilm and in situ analysis by fluorescence in situ hybridization (FISH) and microelectrodes," *Water Science and Technology,* vol. 47, no. 1, pp. 49-57, 2003.

[109] W. H. Lee, D. G. Wahman, P. L. Bishop and J. G. Pressman, "Free Chlorine and Monochloramine Application to Nitrifying Biofilm: Comparison of Biofilm Penetration, Activity, and Viability," *Environmental Science and Technology,* vol. 45, no. 4, pp. 1412-1419, 2011.

[110] "Different Types of Corrosion - Pitting Corrosion," The Corrosion Clinic, 1995. [Online]. Available: https://www.corrosionclinic.com/types_of_corrosion/pitting_corrosion.htm. [Accessed 01 05 2023].

[111] "13 PEX pipes advantages and disadvantages," HPD Construction, 16 08 2021. [Online]. Available: https://www.hpdconsult.com/pex-pipes-advantages-and-disadvantages/. [Accessed 30 05 2023].

[112] "Osenon Technology," [Online]. Available: http://www.osenon.com/en/cp_view.asp?/43.html.

Appendix

Appendix A: MATLAB code for extracting the voltage and time of flight and rejecting cross-talk noise

```
tic
clc; clear;
file = dir('Directory\*.csv');
%i = 0;
for i = 1:size(file,1)
    %i = i+1;
    data = xlsread(file(i).name);
    data = data(:,:);
    [r,c] = size(data);

    [ndata,index] = max(data(:,2));
    voltage(i) = ndata;

    % Tap Water Data
    [ndataw,indexw] = max(data(2000:r,3));
    filen(i) = string(file(i).date);
    indw(i) = indexw;
    voltagew(i) = ndataw;
    phasew(i) = data(indexw,1);

    % Biofilm Data
    [ndatab,indexb] = max(data(2000:r,4));
    indb(i) = indexb;
    voltageb(i) = ndatab;
    phaseb(i) = data(indexb,1);

    % Metal Data
    [ndatam,indexm] = max(data(2000:r,5));
    indm(i) = indexm;
    voltagem(i) = ndatam;
    phasem(i) = data(indexm,1);
    %writeData = struct('f',file(i).name,'v',ndata,'p',index);
    %t(i,:) = struct2table(writeData);
end
t = table(filen', voltage',voltagew',phasew',voltageb',phaseb',voltagem',phasem');
writetable(t,'Write_GndTruth.xlsx');
toc
```

Appendix B: JupyerLab Python code for Machine Learning experiment

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

get_ipython().run_line_magic('matplotlib', 'inline')

plt.rcParams['figure.figsize'] = 10,6
plt.rcParams['text.color'] = 'white'
plt.rcParams['axes.labelcolor'] = 'white'
plt.rcParams['xtick.color']= 'white'
plt.rcParams['ytick.color']= 'white'
plt.rcParams['axes.titlecolor'] = 'white'
plt.rcParams['legend.facecolor'] = 'black'

# # This Model predicts determining the presence of Biofilm, Copper, or Tap Water within the system.

Mats = pd.read_excel('Write_GndTruth_Dec9a.xlsx')
pd.set_option("display.precision", 15)

#Drop Column
Mats = Mats.dropna()
Mats = Mats.drop('Date Time', axis = 1)

#Setup TW Data
TW = Mats.iloc[:,0:3]
TW.columns = ['Input_Voltage','Voltage_Recieved','Phase_Shift']
TW['Voltage_Recieved_Ratio'] = TW['Voltage_Recieved']/TW['Input_Voltage']
TW['Class'] = 'Tapwater'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)

#Setup Biofilm Data
Biofilm = Mats.iloc[:,0:3]
Biofilm.columns = ['Input_Voltage','Voltage_Recieved','Phase_Shift']
Biofilm['Voltage_Recieved_Ratio'] = Biofilm['Voltage_Recieved']/Biofilm['Input_Voltage']
Biofilm['Class'] = 'Biofilm'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)

#Setup Copper Data
Copper = Mats.iloc[:,0:3]
Copper.columns = ['Input_Voltage','Voltage_Recieved','Phase_Shift']
Copper['Voltage_Recieved_Ratio'] = Copper['Voltage_Recieved']/Copper['Input_Voltage']
```

```
Copper['Class'] = 'Copper'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)

# Setup Entire Dataframe with all classes
Mats_GTruth = pd.concat([TW,Biofilm , Copper], axis = 0 , join = 'outer',ignore_index=True)

#Setup TestData for model
TestData = Mats_GTruth.drop('Input_Voltage',axis = 1)

# sns.scatterplot(data = TestData, x = 'Class', y = 'Voltage_Recieved_Ratio',
#                 hue = 'Class').set(title = 'Voltage Recieved Ratio v Class')

# sns.scatterplot(data = TestData, x = 'Class', y = 'Phase_Shift',
#                 hue = 'Class').set(title = 'Time of flight v Class')

# sns.pairplot(data = TestData,hue = 'Class', height = 3)

# sns.scatterplot(data= TestData,x = 'Phase_Shift',y = 'Voltage_Recieved',hue = 'Class')
# #The data separates the two types of Tap Water.
# # Would expect the model to perform very well based on this.

# sns.boxplot(data = TestData).set(title = "BoxPlot for Features")

# sns.boxplot(data = TestData, x = 'Class',y = 'Voltage_Recieved').set(title = "BoxPlot for
Features")

# sns.boxplot(data = TestData, x = 'Class',y = 'Voltage_Recieved_Ratio').set(title = "BoxPlot for
Features")

# sns.scatterplot(data = TestData, x = 'Class',y = 'Phase_Shift', hue ='Class').set(title = "BoxPlot
for Features")

# # Model 1 - Base Model - no adjustments

from sklearn.metrics import confusion_matrix,accuracy_score,ConfusionMatrixDisplay
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
RandomForestClassifier

X = pd.get_dummies(TestData.drop('Class',axis = 1))
y = TestData['Class']
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.3, random_state=101,shuffle
= True
                                    ,stratify =y)
```

```
# Splitting the data set into a training and a test set. The test set size is 30% of the data.
X_holdout_set, X_validation, y_holdout_set, y_validation = train_test_split(X_test, y_test,
test_size=0.5, random_state=101,
                                                    shuffle = True, stratify =y_test)

baseRfc = RandomForestClassifier(oob_score=True,bootstrap = True)

baseRfc.fit(X_train,y_train)

basePreds = baseRfc.predict(X_test)
accuracy_score(y_test,basePreds)

baseRfc.oob_score_

from sklearn.metrics import
accuracy_score,auc,roc_curve,confusion_matrix,ConfusionMatrixDisplay,classification_report,p
lot_confusion_matrix,accuracy_score

from sklearn.inspection import permutation_importance
test_results = permutation_importance(
   baseRfc, X_holdout_set, y_holdout_set, random_state=101, n_jobs=2
)
train_results = permutation_importance(
   baseRfc, X_train, y_train, random_state=101, n_jobs=2
)
sorted_importances_idx = train_results.importances_mean.argsort()
train_importances = pd.DataFrame(
   train_results.importances[sorted_importances_idx].T,
   columns=X.columns[sorted_importances_idx],
)
test_importances = pd.DataFrame(
   test_results.importances[sorted_importances_idx].T,
   columns=X.columns[sorted_importances_idx],
)

import matplotlib.pyplot as plt
labels = X.columns

def box_plot(data, edge_color, fill_color):
   bp = ax.boxplot(data, patch_artist=True, labels = labels)

   for element in ['boxes', 'whiskers', 'fliers', 'means', 'medians', 'caps']:
      plt.setp(bp[element], color=edge_color)

   for patch in bp['boxes']:
      patch.set(facecolor=fill_color)
```

```
        return bp

plt.rcParams.update({'font.size': 16})
plt.rcParams.update({'figure.figsize': (16.0, 12.0)})
for name, importances in zip(["train", "test"], [train_importances, test_importances]):

    features = X.columns
    x = np.arange(len(features))

    sns.set_style('dark')
    sns.set_style('whitegrid')
    ax = importances.plot.box(whis=10)

    ax.set_facecolor("#1CC4AF")

    ax.patch.set_facecolor('palegreen')
    ax.patch.set_alpha(0.45)


    plt.xlabel('xlabel', fontsize=14)
    ax.set_title(f"Permutation Importances for Biofilm,\nCorrosion, and Tap Water
classification({name} set)",fontdict={'fontsize': 24,},color = 'black')
    ax.set_ylabel("Feature Importance",fontdict = {'fontsize':20 })
    ax.set_xlabel("Feature",fontdict = {'fontsize':20 })

    bp1 = box_plot(importances, 'blue', 'cyan')
    ax.axhline(y=0, color="b", linestyle="--")

ax.figure.tight_layout()

plt.rcParams.update({'figure.figsize': (14.0, 12.0)})
plt.rcParams.update({'font.size': 16})
ax.set_title(f"Permutation Importances for Biofilm, Scaling,\nCorrosion, and Tap Water({name}
set)",fontdict={'fontsize': 20})
sns.set_style('dark')
ConfusionMatrixDisplay.from_predictions(y_test,basePreds)
print(classification_report(y_test,basePreds))

# # Model 2: 4- classes and 3 features Tuned

Mats = pd.read_excel('Write_GndTruth_02_19_23.xlsx')
pd.set_option("display.precision", 15)

#Drop Column
Mats = Mats.dropna()
Mats = Mats.drop('Date Time', axis = 1)
```

```
#Setup TW Data
TW = Mats.iloc[:,0:3]
TW.columns = ['Input_Voltage','Voltage_Received','Phase_Shift']
TW['Voltage_Received_Ratio'] = TW['Voltage_Received']/TW['Input_Voltage']
TW['Class'] = 'Tapwater'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)

#Setup Biofilm Data
Biofilm = Mats.iloc[:,0:3]
Biofilm.columns = ['Input_Voltage','Voltage_Received','Phase_Shift']
Biofilm['Voltage_Received_Ratio'] = Biofilm['Voltage_Received']/Biofilm['Input_Voltage']
Biofilm['Class'] = 'Biofilm'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)

#Setup Copper Data
Copper = Mats.iloc[:,0:3]
Copper.columns = ['Input_Voltage','Voltage_Received','Phase_Shift']
Copper['Voltage_Received_Ratio'] = Copper['Voltage_Received']/Copper['Input_Voltage']
Copper['Class'] = 'Copper'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)
#Setup Scaling Data
Scaling = Mats.iloc[:,0:3]
Scaling.columns = ['Input_Voltage','Voltage_Received','Phase_Shift']
Scaling['Voltage_Received_Ratio'] = Scaling['Voltage_Received']/Scaling['Input_Voltage']
Scaling['Class'] = 'Scaling'
Mats = Mats.drop(Mats.iloc[:,1:3],axis = 1)

# Setup Entire Dataframe with all classes
Mats_GTruth = pd.concat([TW,Biofilm , Copper, Scaling], axis = 0 , join =
'outer',ignore_index=True)

#Setup TestData for model
TestData = Mats_GTruth.drop('Input_Voltage',axis = 1)

X = pd.get_dummies(TestData.drop('Class',axis = 1))
y = TestData['Class']
X_train, X_test, y_train, y_test = train_test_split( X, y, test_size=0.3, random_state=101,shuffle
= True
                                    ,stratify =y)
# Splitting the data set into a training and a test set. The test set size is 30% of the data.
X_holdout_set, X_validation, y_holdout_set, y_validation = train_test_split(X_test, y_test,
test_size=0.5, random_state=101,
                                                shuffle = True, stratify =y_test
```

```
rfc.oob_score_

RandomForestClassifier()

rfcf = RandomForestClassifier(max_features=1,
                              max_depth = None,
                              criterion = 'gini',
                              min_samples_split=7,
                              min_samples_leaf = 1,
                              n_estimators =200,
                              bootstrap = True,
                              oob_score=True,
                              random_state=101)
rfcf.fit(X_train,y_train)
fpreds =rfcf.predict(X_holdout_set)

sns.set_style('dark')
ConfusionMatrixDisplay.from_predictions(y_holdout_set,fpreds)
print(classification_report(y_holdout_set,fpreds))
print(rfcf.oob_score_)

test_results = permutation_importance(
   rfcf, X_holdout_set, y_holdout_set, random_state=101, n_jobs=2
)
train_results = permutation_importance(
   rfcf, X_train, y_train, random_state=101, n_jobs=2
)
sorted_importances_idx = train_results.importances_mean.argsort()

train_importances = pd.DataFrame(
   train_results.importances[sorted_importances_idx].T,
   columns=X.columns[sorted_importances_idx],
)
test_importances = pd.DataFrame(
   test_results.importances[sorted_importances_idx].T,
   columns=X.columns[sorted_importances_idx],
)

import matplotlib.pyplot as plt
labels = X.columns

def box_plot(data, edge_color, fill_color):
   bp = ax.boxplot(data, patch_artist=True, labels = labels)

   for element in ['boxes', 'whiskers', 'fliers', 'means', 'medians', 'caps']:
      plt.setp(bp[element], color=edge_color)
```

```
    for patch in bp['boxes']:
        patch.set(facecolor=fill_color)

    return bp

plt.rcParams.update({'font.size': 16})
plt.rcParams.update({'figure.figsize': (16.0, 12.0)})
for name, importances in zip(["train", "test"], [train_importances, test_importances]):

    features = X.columns
    x = np.arange(len(features))

    sns.set_style('dark')
    sns.set_style('whitegrid')
    ax = importances.plot.box(whis=10)

    ax.set_facecolor("#1CC4AF")

    ax.patch.set_facecolor('palegreen')
    ax.patch.set_alpha(0.45)

    plt.xlabel('xlabel', fontsize=14)
#     ax.set_title(f"Permutation Importances for Biofilm, Scaling,\nCorrosion, and Tap Water
classification({name} set)",fontdict={'fontsize': 24,},color = 'black')
    ax.set_ylabel("Feature Importance",fontdict = {'fontsize':20 })
    ax.set_xlabel("Feature",fontdict = {'fontsize':20 })

    bp1 = box_plot(importances, 'blue', 'cyan')
    ax.axhline(y=0, color="b", linestyle="--")

ax.figure.tight_layout()

plt.barh(feature_names, rfcf.feature_importances_)

from sklearn.inspection import permutation_importance
#    import shap

print(f"RF train accuracy: {rfcf.score(X_train, y_train):.3f}")
print(f"RF test accuracy: {rfcf.score(X_test, y_test):.3f}")
print(f"RF Holdout test accuracy: {rfcf.score(X_holdout_set, y_holdout_set):.3f}")
print(f"RF Validation test accuracy: {rfcf.score(X_validation, y_validation):.3f}")
```

```
rfcf = RandomForestClassifier(max_features=1,
                              max_depth = None,
                              criterion = 'gini',
                              min_samples_split=7,
                              min_samples_leaf = 1,
                              n_estimators =200,
                              bootstrap = True,
                              oob_score=True,
                              random_state=101)
rfcf.fit(X_train,y_train)
train_preds =rfcf.predict(X_train)
test_preds =rfcf.predict(X_test)
holdout_preds = rfcf.predict(X_holdout_set)
#  X_holdout_set, y_holdout_set

plt.rcParams.update({'figure.figsize': (16.0, 12.0)})
plt.rcParams.update({'font.size': 16})
sns.set_style('dark')
ConfusionMatrixDisplay.from_predictions(y_holdout_set,holdout_preds)
print(classification_report(y_holdout_set,holdout_preds))

test_results = permutation_importance(
   rfcf, X_holdout_set, y_holdout_set, random_state=101, n_jobs=2
)
train_results = permutation_importance(
   rfcf, X_train, y_train, random_state=101, n_jobs=2
)
sorted_importances_idx = train_results.importances_mean.argsort()

train_importances = pd.DataFrame(
   train_results.importances[sorted_importances_idx].T,
   columns=X.columns[sorted_importances_idx],
)
test_importances = pd.DataFrame(
   test_results.importances[sorted_importances_idx].T,
   columns=X.columns[sorted_importances_idx],
)

import matplotlib.pyplot as plt
labels = X.columns

def box_plot(data, edge_color, fill_color):
   bp = ax.boxplot(data, patch_artist=True, labels = labels)

   for element in ['boxes', 'whiskers', 'fliers', 'means', 'medians', 'caps']:
      plt.setp(bp[element], color=edge_color)
```

```
    for patch in bp['boxes']:
        patch.set(facecolor=fill_color)

    return bp

plt.rcParams.update({'font.size': 16})
plt.rcParams.update({'figure.figsize': (16.0, 12.0)})
for name, importances in zip(["train", "test"], [train_importances, test_importances]):

    features = X.columns
    x = np.arange(len(features))

    sns.set_style('dark')
    sns.set_style('whitegrid')
    ax = importances.plot.box(whis=10)

    ax.set_facecolor("#1CC4AF")

    ax.patch.set_facecolor('palegreen')
    ax.patch.set_alpha(0.45)

    plt.xlabel('xlabel', fontsize=14)
#     ax.set_title(f"Permutation Importances for Biofilm, Scaling,\nCorrosion, and Tap Water
classification({name} set)",fontdict={'fontsize': 24,},color = 'black')
    ax.set_ylabel("Feature Importance",fontdict = {'fontsize':20 })
    ax.set_xlabel("Feature",fontdict = {'fontsize':20 })

    bp1 = box_plot(importances, 'blue', 'cyan')
    ax.axhline(y=0, color="b", linestyle="--")

ax.figure.tight_layout()

final_model_3_feat_biofilm.fit(X,y)

import joblib

joblib.dump(final_model_3_feat_biofilm,'V2_final_model_3_feat_biofilm.pkl')

list(X.columns)

joblib.dump(list(X.columns),'col_names_3_feat_biofilm.pkl')
```

```
# # Loading the Model

new_columns = joblib.load('col_names_3_feat_biofilm.pkl')

new_columns

loaded_model = joblib.load('V2_final_model_3_feat_biofilm.pkl')

NewData = pd.read_csv('12_9_22_GTruth_Cleaned.csv')
pd.set_option("display.precision", 15)
NewData = NewData.drop('Unnamed: 0',axis = 1)
NewData
len(actual_class)

NewData['Class'].value_counts()

pred_cols = list(NewData.columns.values)[:-1]
class_col = list(NewData.columns.values)[-1:]
actual_class = NewData[class_col]
pred = pd.Series(loaded_model.predict(NewData[pred_cols]))

sns.set_style('dark')
ConfusionMatrixDisplay.from_predictions(actual_class,pred)
print(classification_report(actual_class,pred))

f1Score = f1_score(actual_class,pred, average = 'weighted')*100

recall = recall_score(actual_class,pred, average = 'weighted')*100

precision = precision_score(actual_class,pred,average ='weighted' )*100

accuracy = accuracy_score(actual_class,pred)*100

metrics = ({ 'accuracy': accuracy, 'f1Score': f1Score ,'precision':precision, 'recal':recall})

mets = pd.Series(metrics)
df2 = mets.to_frame()
df2.set_axis(['Percentage'], axis='columns', inplace=True)
df2.index.name="NewData Set"
df2 = df2.round({'Percentage':4})
dfnewdata_metrics = df2.style.format(precision = 4)
dfnewdata_metrics
```