Cost-Effective Sensor Systems for Measuring Extracted Chlorophyll-a Concentration

A Thesis Submitted To
The College of Graduate and Postdoctoral Studies
In Partial Fulfillment of The Requirements
For the Degree of Master of Science
In the Department of Electrical and Computer Engineering
University of Saskatchewan
Saskatoon, SK, Canada

By
Rakibul Islam Chowdhury

Permission to Use

In presenting this thesis in partial fulfillment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis. Requests for permission to copy or to make other uses of materials in this thesis/dissertation in whole or part should be addressed to:

Head of the Department of Electrical and Computer Engineering
57 Campus Drive
University of Saskatchewan
Saskatoon, Saskatchewan S7N 5A9
Canada

OR

Dean
College of Graduate and Postdoctoral Studies
University of Saskatchewan
116 Thorvaldson Building, 110 Science Place
Saskatoon, Saskatchewan S7N 5C9
Canada

Disclaimer

Reference in this thesis to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement, recommendation, or favoring by the University of Saskatchewan. The views and opinions of the author expressed herein do not state or reflect those of the University of Saskatchewan and shall not be used for advertising or product endorsement purposes.

Abstract

Chlorophyll-a concentration is one of the most measured metrics in both water quality and plant health monitoring. It is an indicator of algal biomass and provides insight into stressors such as eutrophication and bloom risk. It is also a widely used metric in terrestrial ecosystems as an indicator of photosynthetic activity and nutrient limitation. Most currently used laboratory-based methods for measuring chlorophyll-a exploit spectroscopic techniques and require expensive instrumentation, like spectrophotometer or fluorometer. In addition, the readings are taken inside a black box to avoid optical noise.

The purpose of this thesis is to propose a smart, low-cost, and portable sensor system to measure the concentration of chlorophyll-a in an extracted solution. The goals were achieved using two distinct spectral method. The first approach involves two consumer-grade spectral sensors that read the optical reflectance at 12 discrete wavelengths in visible and near-infrared spectra. The system was tuned for an optimal distance from the sensors to the solution and an enclosure was printed to maintain the distance, as well as to avoid natural light interference. Extracted chlorophyll-a solutions of 52 different concentrations were prepared, and at least 5 readings per sample were taken using the proposed smart sensor system. The ground truth values of the samples were measured in the laboratory using Thermo Nano 2000C. After cleaning the anomalous data, different machine learning models were trained to determine the significant wavelengths that contribute most towards chlorophyll-a measurement. Finally, a decision tree model with 5 important features was chosen based on the lowest Root Mean Square and Mean Absolute Error when it was tested on the validation set. The final model resulted in a mean error of ±0.9 μg/L when applied on the test set. The total cost for the device was around CAD 135.

For the next approach, a rapid system has been proposed using electric impedance spectroscopy (EIS) to measure the concentration of chlorophyll-a, extracted into 95%(v/v) ethanol. Two electrodes accompanied with a high precision impedance converter from Analog Device was used for the development of the sensor. The system was tuned for a fixed electrode orientation, effective area, electrode to electrode distance and excitation voltage by studying different relevant experiments. The proposed sensor was calibrated using the impedance of 95%(v/v) ethanol. Extracted chlorophyll solutions of 60 different concentrations were prepared. At least 5 readings per sample were taken using the proposed system from 1.5 kHz to 7.5 kHz. Samples were then analyzed using standard methods by a spectrophotometer (Genesys20) from Thermo Scientific.

Study of Pearson coefficient, principal component analysis, variance inflation factor and backward elimination were used to identify the significant features for chlorophyll-a measurement using EIS. Finally, a simple linear regression model with 11 important features in the range 2.3kHz to 4.7kHz was chosen based on the lowest Root Mean Square (RMS) and Mean Absolute (MA) Error. The coefficient of determination, R^2 of the fitted model was 0.93. MAE for the final proposed model is $\pm 0.904\ \mu gL^{-1}$ when applied on the test set.

Acknowledgments

At first, all praise goes to the Almighty ALLAH for HIS blessings upon me.

I would like to extend my sincere gratitude towards my supervisor Prof. Khan A. Wahid. Without his continuous support, guidance, and supervision, my program, and this thesis work at the University of Saskatchewan, will not be possible.

I am grateful for the funding support, which made this work possible, from Global Water Future (GWF) at the University of Saskatchewan.

I would like to extend my sincere gratitude and appreciation to our collaborator Dr. Helen Baulch (Associate Professor, School of Environment and Sustainability, University of Saskatchewan, Saskatoon) and her associate Katy Naugent, for helping me designing experimental setup, and providing resources. Moreover, Dr. Helen's valuable feedback in this thesis has been of great help in the assessment of the developed system.

I would like to thank Mohammad Habibullah for his guidance in some technical difficulties. Also, I would like to thank Rinku Basak, Ali Newaz Bahar, and Gazi Ehsan Rahman for the assistance during the experiments. I thank all my friends in Canada and Bangladesh for believing and inspiring me all the time. I want to dedicate all my achievements and hard work to my parents (Abul Khair Chowdhury and Kaniz Fatema), my sister (Raisa Chowdhury) for their love and care. I cannot express in words how much I owe them for tolerating me when I was passing stressful times.

Dedication

Dedicated to my parents, my sister, and my friends in Canada

Table of Contents

List of Tables

List of Figures

List of Abbreviations

µg/L	microgram per Liter
µW/cm^2	microwatt per square centimeters
3D	Three-dimensional
A	Ampere
ADC	Analog to Digital Converter
CAD	Canadian Dollar
cm	centimeter
CRI	Color Rendering Index
EIS	Electrical Impedance Spectroscopy
FWHM	Full-Width Half-Max
g	grams
HAB	Harmful Algal Bloom
I^{2}C	Inter-Integrated Circuit
IoT	Internet of Things
IR	Infrared
KHz	Kilo Hertz
LED	Light Emitting Diode
MAE	Mean Absolute Error
mL	milli-liter
MSE	Mean Squared Error
MUT	Medium Under Test
NIR	Near-Infrared
PFoV	Package Field of View
PC	Principal Component
R^2	Coefficient of Determination
RMSE	Root Mean Absolute Error
UAV	Unmanned Arial Vehicle
USB	Universal Serial Bus
VIS	Visible

Chapter 1: Introduction

Water quality testing is an important part of environmental monitoring. When water quality is poor, it affects not only aquatic life, but the surrounding ecosystem including human as well. Monitoring water quality provides empirical evidence to support decision making on health and environmental issues. However, monitoring water quality in the 21st century is a growing challenge because of the large number of chemicals used in our everyday lives and in commerce that can make their way into our waters. On the other hand, the world population will grow to 9.6 billion by 2050 and global food production need to be increased by 70% to feed the increased population [1]. Hence analyzing plant health can contribute towards the superior quality of vegetables and fruits.

1.1. Different types of Chlorophyll

Chlorophyll is one of the major photosynthetic pigments that is found in almost all plants and algae. There are mainly four types of chlorophylls – chlorophylls a, b, c, and d. Some of them can again be divided into subcategory, for example – chlorophyll c consists of both chlorophyll c1 and c2. However, all kinds of chlorophylls consist of four pyrrole rings containing nitrogen arranged in a ring around a magnesium ion and a long hydrocarbon tail. This hydrocarbon tail is lipid soluble. Chlorophyll-a is found in almost all higher plants, algae, and cyanobacteria; chlorophyll-b is found in higher plants and green algae. Chlorophyll c and d is not found as abundant as the other two kinds. Chlorophyll c is mainly found in diatoms, dinoflagellates, and brown algae whereas chlorophyll d is found only in red algae [2]. However, in higher plants and algae, chlorophyll-a is the principle photosynthetic pigment and the ratio of chlorophyll-a to chlorophyll-b is found 3:1 [3]

1.2. Chlorophyll-a measurement for water quality monitoring

Water quality monitoring has physical properties (e.g. temperature and turbidity), chemical characteristics (pH and dissolved oxygen) and biological indicators (algae and phytoplankton). Chlorophyll-a is one of the major water quality monitoring metrics that provides an estimation of both phytoplankton biomass [4] and trophic status [5][6]. It is a regularly observed physical parameter in different lakes in Canada.

Chlorophyll-a is also an indicator and predictor of harmful algal blooms or HABs. HABs occur when colonies of algae grow out of control and produce toxic or harmful effects on people, fish, shellfish, marine mammals, and birds. The human illness caused by HABs can be debilitating or even fatal. HAB may reduce dissolved oxygen concentration and bring it down as low as to the lethal level for marine lives. Although algal blooms can be natural phenomena, they have expanded as an issue over the last several decades in Canada. Reports of blooms of toxic cyanobacteria in the Great Lakes, for example, have increased significantly since the late 1990s [7]. HAB not only causes human illness but also affects the economy. From 2009 to 2012, blooms caused 22 fish killing events around Vancouver Island, Canada [8].

Canada, with over a million lakes and countless ponds, is faced with the daunting prospect of managing water resources as population expands and the north experiences accelerated warming. At present, management policies to mitigate or respond to algal blooms vary considerably across provinces, and local agencies are typically ill-prepared particularly for toxic blooms [9]. However, initiatives have been taken to minimize the adverse effect by forecasting algal bloom. As chlorophyll-a is one of the main photosynthetic pigment of algae, among many physical parameters, forecasting of HAB requires continuous monitoring of chlorophyll-a in the water. In fact, monitoring chlorophyll-a levels is a direct way of tracking algal growth. Surface waters that have high chlorophyll-a conditions are typically high in nutrients, generally phosphorus and nitrogen. These nutrients cause the algae to grow or bloom. So, a rise in chlorophyll-a concentration triggers alarm for potential algal bloom risk. To calibrate the in-situ instruments to measure chlorophyll-a, samples are collected from the lakes and validated against the lab-based methods.

In general, the amount of chlorophyll-a in a collected water sample is used as a measure of the concentration of suspended phytoplankton. The use of the measurement of phytoplankton as an indicator of water quality is described in Section 10200 (A) of Standard Methods for the Examination of Water and Wastewater.

1.3. Chlorophyll-a measurement in plant science

Deficiencies in any of major nutrients like nitrogen, phosphorus, potassium, calcium, sulfur, magnesium, and iron can manifest as a reduction in leaf chlorophyll-a content. A depletion in overall leaf chlorophyll-a content reduces the amount of solar radiation that can be absorbed which

in turn limits the efficiency of corresponding photosynthetic processes thus lowering primary photosynthetic production. The measurement of chlorophyll-a content can therefore yield important information relating to the presence of biotic stress factors [10] and also of abiotic issues such as light, drought and pigment inhibiting herbicide damage [11][12].

1.4. Chlorophyll-a measurement in food engineering

Chlorophyll-a and its derivatives are largely used in food industry and medicine [13]. Copper-chlorophyll food colorings have been produced and used since 1926 [14]. In addition, chlorophyll-a is also used in cosmetic products and in pharmaceutical industry. Additionally, it has antioxidant and antimutagenic properties [15]. On top of that, chlorophyll-a is also used a natural dying agent [16].

1.5. Research objective

The primary objective of this thesis is to develop a low-cost, portable, and rapid system to monitor extracted chlorophyll-a concentration. The following research objectives were set to meet the goal of developing an affordable system:

- To design and develop a low-cost, portable, and rapid sensors for measuring extracted chlorophyll concentration by exploring its physical and chemical change with increase of concentration.
- To prepare extracted chlorophyll-a samples to train and test developed sensor.
- To develop a spectral data processing pipeline including outlier detection, feature reduction, normalization and modeling for data acquired using developed sensor.
- To compare the feasibility of the proposed system in terms of accuracy and cost over the existing methods and devices.

1.6. Thesis organization

The chapters are organized as follows:

Chapter 2 reviews the literature related to existing methodologies and recent advances in chlorophyll concentration measurement. Moreover, different recent applications of electrical impedance spectroscopy have been discussed.

Chapter 3 provides the details background of data processing. The chapter elaborately explains the theory and equations of data preprocessing, normalization, outlier detection, validation methods and evaluation metrics.

Chapter 4 and Chapter 5 describes the experimental setup and result analysis of two independent experiments to fulfil the research objectives. *Chapter 4* highlights the experiment using optical reflectance spectroscopy while *Chapter 5* deals with extracted chlorophyll-a measurement using electrical impedance spectroscopy.

Finally, *Chapter 6* concludes the findings of this research work and provides direction for future investigations and improvements for the proposed system.

Chapter 2: Literature Review

2.1. Methods for extracting chlorophyll-a

Chlorophyll-a is an organic, complex, and photosynthetic compound which is bound to plant or aquatic cell. Being an organic compound, it does not dissolve into water or any other non-organic solvent. However, it is easily dissolved into organic solvents like ethanol, acetone, ether and chloroform [15].

The use of the solvent differs from one laboratory to another. The expression used for calculating the concentration of extracted chlorophyll also differs based on the solvent. The 95% (v/v) ethanol extraction method is faster and not as expensive as the acetone method, since the extraction and homogenization of the chloroplasts is performed in hot ethanol instead of cold acetone [17]. In [15] and [18] researchers shown that 95% (v/v) ethanol outperforms other common polar solvents to extract chlorophyll from plant leaf. However, for chlorophyll-a extraction from aquatic macrophytes, 100% dimethyl sulfoxide (DMSO) yielded slightly more efficient than methanol or acetone depending on the species examined. Chlorophyll-a extracted into DMSO appeared to be more stable over time than ethanol [19].

2.2. Methods for measuring chlorophyll-a

Chlorophyll can be determined by titration of the bound magnesium with tetra-acetic acid [20]. However, this procedure cannot differentiate between chlorophyll-a and chlorophyll-b. Also, it requires physical separation of the chlorophylls from the other cellular constituents. As a result, it is not as convenient as other methods of measuring chlorophyll-a [21].

The current methodologies of chlorophyll-a measurement mainly rely on its spectroscopic characteristics: light absorption [22][23][24] or fluorescence [25][26][27]. Methods to measure chlorophyll-a in water can be classified into three major branches: (a) in situ Fluorescence (b) Satellite or Unmanned Aerial Vehicle (UAV) Imaging and (c) lab-based methods including spectrophotometry, fluorometry and HPLC. Of these, (a) and (b) are non-destructive, and (c) is a destructive method to measure Chlorophyll-a in water samples.

Many commercial water quality monitoring sensors (such as microFlu by TriOS, Germany; FluoProbes by InterChim, France; YSI 6131 by Yellow Springs Instruments Inc., USA) use fluorescence to measure chlorophyll-a in vivo. However, for the fluorescence method, the total

fluorescence is only 1-2% of absorbed light, resulting in very weak signal for detection by existing sensors [28]. Moreover, the method yields relative values as the loss of light energy is permanent [29]. Frequent field calibration and/or validation are required with this method [30].

Satellite and airborne sensors, or UAVs, can also be used to quantify chlorophyll-a content. They have narrow-band capabilities [31], although satellites can cover large areas more quickly. Both satellite and airborne sensors have disadvantages. They are expensive and have low temporal and spatial resolution which is less than ideal for producing detailed pictures of the algae present. Not only are there technological limitations, but there can also be environmental ones such as cloud cover impedance [32]. However, for quantitative analysis, almost all the imaging methods require calibration based on discrete water samples [33]. So measurement using either remote sensing [34][35] or in-situ probes [36][37] is relative, and must be calibrated against the discrete spectroscopic measurement [38][4].

The spectrophotometric method is widely used in laboratories for validation of airborne methods. Chlorophyll-a is extracted in an organic solvent (e.g. acetone or ethanol) and optical absorbance of the solution is measured using a spectrophotometer. Degradation of the chlorophyll-a solution, which can be caused by exposure to light on the passage of time, can create difficulties when using this method [39].

Standard methods for measuring chlorophyll-a (e.g., ISO 10260:1992), include extraction of chlorophyll-a from the leaf or algae in a solvent, such as 95% (v/v) ethanol. The optical density (OD) is measured at 665nm and 750nm after filtration and extraction of the chlorophyll solution, followed by corrections for degradation products, often done by acidifying the solution [40]. In fluorometric methods, the emission peak is measured at around 685nm [41], with the wavelengths varying based on the extractant used.

Measuring OD at different wavelengths precisely requires individual detectors and precise optical sources or filter. In addition, sometimes, the response is so small that the emitted response gets scattered. Also, optical noise and interferences are common and are difficult to model, since the noise varies temporally. To minimize this noise, in laboratory-based methods, OD is measured inside a black box. Chlorophyll analysis by high performance liquid chromatography is a precise methodology [42]. However, it is a costly and time-consuming procedure.

Some recent studies show promise for measuring chlorophyll A in water. Filippo Attivissimo *et al.* proposed a portable optical sensor for sea water, with a high sensitivity of 0.2

μg/L [43]. Estimation accuracy was improved significantly in remote sensing using gaussian process regression in [44] and [45]. In addition, smartphone-based methods are rapidly advancing, for example, fluorometry has been used to enable smartphone based monitoring of pH [46]. Apps have also been developed for measuring chlorophyll-a in plant tissue [47].

2.3. Usage of electrical impedance spectroscopy in sensor development

Electrical Impedance Spectroscopy (EIS) is frequently used to study the resistance, inductance and capacitance properties of a sample subjected to an applied electric field of fixed or changing frequency [48]. It has wide range of application in biological analysis [49]. It has been used to measure leaf nitrogen concentration [50], the moisture level in foods [51] and to assess the ripening of fruits [52]. To assess the freshness of banana, EIS was performed during different ripening state in [53]. Researchers attached an Ag/AgCl electrode and measured impedance spectra by 4294A Impedance Analyzer over a frequency range of 50Hz to 1MHz. The impedance magnitude, phase angle, real and imaginary part varied markedly with the change of ripening stage.

Many other researchers have reported methodologies based on electrical impedance measurements to determine plant physiological status, such as nitrogen nutrition stress in tomato leaves [54] nitrogen status estimation in lettuce [55], pH measurement in citrus food [56] and tea leaf growth [57]. In [58], authors also used electrical impedance spectroscopy to measure the water content in ethanol-water mixtures.

EIS is less sensitive to environmental variables than other available noninvasive methods. It is a fast, easily implementable, radiation-free and inexpensive method which could be an attractive alternative to optical spectroscopy for various applications [50].

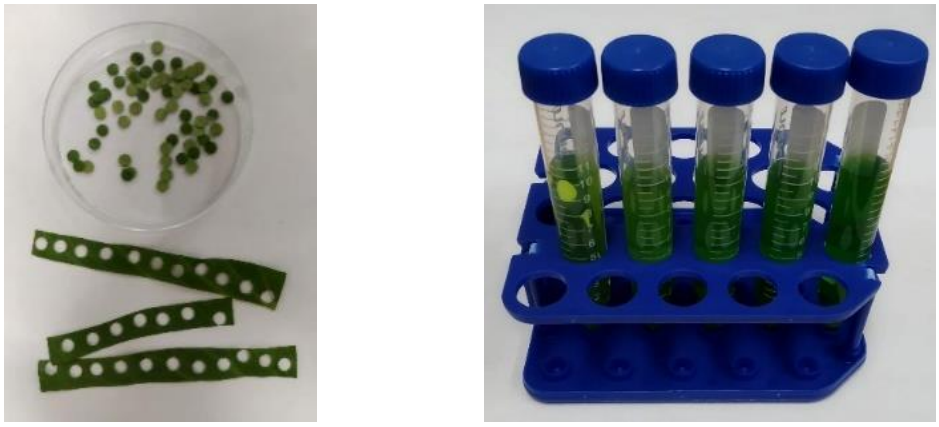
Chapter 3: Sample Preparation and Concentration Measuring

To develop a model that predicts any physical parameter, satisfactory number of samples is required. The number of samples is satisfactory when they include all variations of that parameter within the measuring range. For regression analysis, the actual values of that physical parameter need to be known beforehand. This is also required to evaluate the developed model. In the following sections, how samples were prepared, their distribution and how chlorophyll-a concentrations in the samples were measured have been described.

3.1. Extracting chlorophyll-a

Green leaves are common and natural source of chlorophyll-a. Green lettuce and spinach are commonly found leaves which are full of chlorophyll-a and found almost all the year round. So, both lettuce and spinach leaves were used to extract chlorophyll-a. In all subsequent experiments 95% (v/v) ethanol ($CH_3 – OH$) was used as the polar solvent to extract chlorophyll-a from plant leaves.

To prepare a homogeneous distribution of chlorophyll-a solution, leaves were first cut into equal pieces using paper punch machine. To maintain homogeneous distribution of chlorophyll-a, only leaf lamina was considered, and midrib and veins were avoided as much as possible. These equal pieces of leaf lamina were then submerged into 95% (v/v) ethanol for a day for at least 8 hours or more. Finally, the leaf pieces were filtered out and extracted chlorophyll-a into ethanol was collected. The whole process has been shown in Figure 3-1.



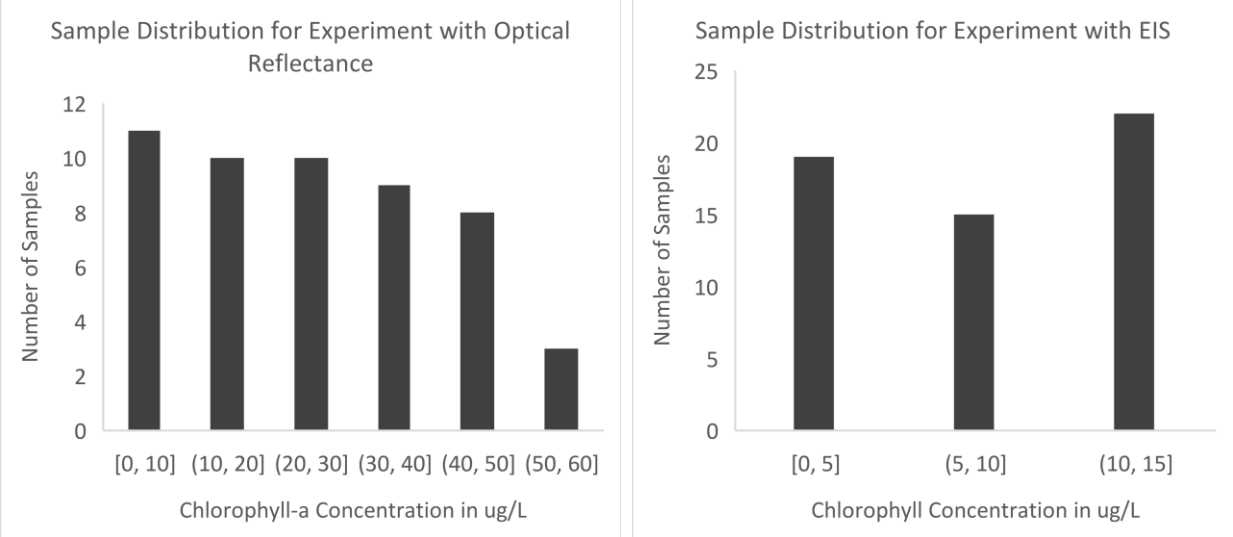
(a) (b)

Figure 3 1: The process of extracting chlorophyll from plant leaf (a) leaf cut into equal pieces, (b) leaf submerged into ethanol,

(c) (d)

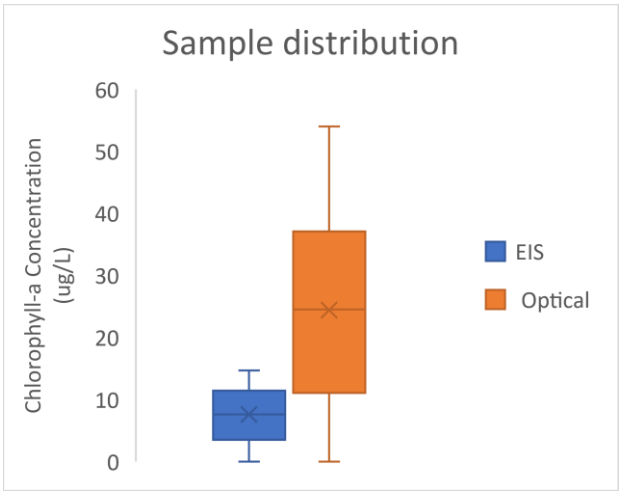
Figure 3-1: (c) filtered leaf after extraction, (d) prepared solution in the cuvette.

Samples were prepared separately at different times for two different experiments using optical reflectance and electrical impedance. In the experiments, the range of chlorophyll-a measurement using optical spectroscopy was found wider than that using electrical impedance spectroscopy. The sample distributions for each experiment have been shown in Figure 3-3



(a) (b)

Figure 3-2: Sample distribution for (a) experiment-1 (b) experiment-2



(c)

Figure 3-3: (c) Sample distribution for two different experiments

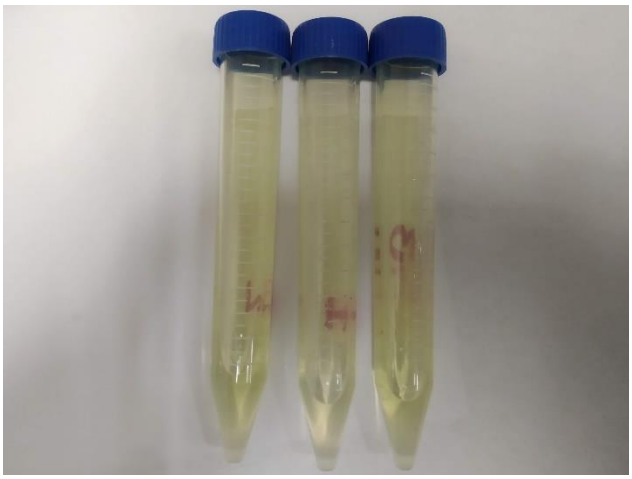
The Box plot in Figure 3-3 shows the range of concentration of the samples for two different experiments. From the histogram shown in Figure 3-3. it is evident that for both experiments, the samples are evenly distributed within respective range.

Chlorophyll-a concentration extracted from algae was also measured in this experiment. In this regard, sample water was collected from different spots of Pike Lake, Saskatoon, SK, Canada. The collected water was then filtered overnight for algae using a P5 quality filter paper from Fisher Scientific. The substrate was then submerged into 95% (v/v) ethanol. Ethanol extracted chlorophyll-a from the collected algae. However, this data was not a part of model development. The process has been illustrated in Figure 3-5.

(a) (b) (c)

Figure 3-4: Chlorophyll extraction from lake water algae (a) Sample water collection from Pike Lake (b) water filtration for separating algae from water (c) filtered algae

(d)



(e)

Figure 3-5: (d) algae submerged into 95% (v/v) ethanol (e) extracted chlorophyll for three different samples

3.2. Concentration measurement of the sample

To label the solutions with actual chlorophyll-a content (referred as, the ground truth value), the traditional lab-based spectrophotometric method has been used. The ground truth value of chlorophyll-a concentration was calculated using the trichromatic method explained in [59] and stated in (3-1)

$$Chl\ A\ (\mu g L^{-1}) = \frac{(13.7*(A_{665}-A_{750})-5.76*(A_{649}-A_{750}))*s}{V*\rho} \qquad (3\text{-}1)$$

where, A_{665} = Absorbance at 665 nm,

A_{669} = Absorbance at 669 nm,

A_{750} = Absorbance at 750 nm,

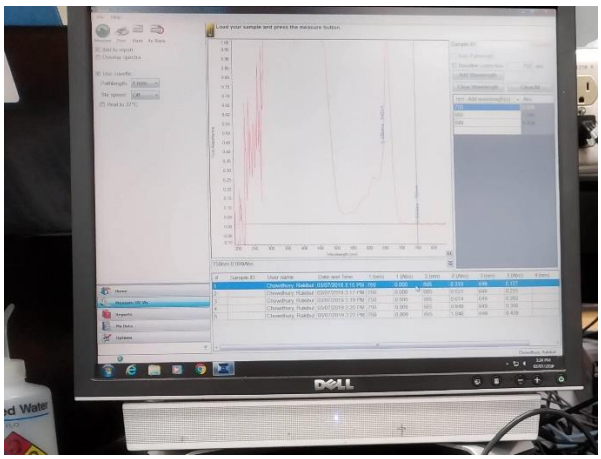
s = solvent extract volume (mL),

V = Sample Volume (L),

ρ = path length of cell (cm)

Thermo Scientific Genesys 20 or Thermo Nanodrop 2000c spectrophotometer (shown in Figure *3-6* (a)) was used to measure the absorbance based on availability to use. Both are industry grade precise spectrophotometer and was used to measure the optical density (OD) at 649, 665, and 750nm. If the OD at 665nm was not in the linear range (between 0.2 and 0.8), the solution was diluted, and the OD of the diluted solution was multiplied by the dilution ratio.

(a)



(b)

(c)

Figure 3-6: Spectrophotometer used for measuring ground truth value of the sample (a) NanoDrop2000c (b) Graphical user interface for NanoDrop 2000c (c) GenSys 20, both from Thermo Scientific

During the ground truth value measurements, samples were carried in a standard 1mm cuvette. Because chlorophyll can fluorophore, to ensure that there was no photobleaching or self-shading effect, the sample was kept in darkness for at least 15 minutes between consecutive readings for both the proposed method and the spectrophotometric method.

Chapter 4: Data Processing

Before developing any model, the dataset was closely monitored for any outlier. Also, the data was shuffled, and the features were brought in the same scale. Out of many predictors or features, few were chosen to predict the dependent variable based on different appropriate methodologies. Different models were tried on the dataset. The developed models were evaluated based on prefixed metrics. The whole data processing part was mainly divided into preprocessing, model development and evaluation. This has been briefly described in following sections.

4.1. Data pre-processing

Optical and electrical impedance spectra reading was taken carefully so that there remains no missing value in the dataset. For all samples, at least 2 readings were taken for both optical and electrical impedance method. The data was saved in a .csv file. For a better optimized model, the obtained data was passed through certain stages.

4.1.1. Data shuffling and feature scaling

To make sure that the model is a generalized one, the obtained data was shuffled before splitting into test and training set. For best accuracy of the model, it's always recommended that training data should have all flavours of data which is ensured by shuffling. Also, by shuffling the data, it was ensured that each data point creates an "independent" change on the model, without being biased by the same points before them. For repeatability of the result, the shuffling function was seeded with same integer all the time.

The prepared dataset contains features highly varying in magnitudes and range. But since, most of the machine learning algorithms use Euclidian distance between two data points in their computations, this is a problem.

If left alone, these algorithms only take in the magnitude of features. The features with high magnitudes will weigh in a lot more in the distance calculations than features with low magnitudes. To supress this effect, all features were required to bring to the same level of magnitudes. This was achieved by scaling the features.

Z-score normalization or standardization was used to scale the features. standardization (or Z-score normalization) means centering the variable around mean and standardizing the variance at 1. The procedure involves subtracting the mean of each observation and then dividing by the standard deviation according to (4-1).

$$x' = \frac{x - \bar{x}}{\sigma} \qquad (4\text{-}1)$$

Where, $\bar{x}$ is mean value of all the values for feature x,

σ is their standard deviation

x' is the scaled value of the observation, x

The result of standardization is that the features will be rescaled so that they will have the properties of a standard normal distribution with mean, μ = 0 and standard scaler, σ = 1.

4.1.2. Outlier detection

Despite many precautions before measuring optical reflectance, the physical sensors can be susceptible to erroneous readings. When the model is trained by these erroneous readings, the accuracy of the model degrades substantively. To improve accuracy, the erroneous readings must be excluded before training and testing the model.

One of the methods to find erroneous readings is to plot the values of one feature against another. This kind of plot is known as pair plot. If there is a strong correlation between the values of two features, one single trendline is expected in their pair plot. The other readings that does not fit into that trendline may cause error. However, formal attention is required before identifying any reading as outlier because they might content important information.

4.2. Feature selection or reduction

In any data analysis, not all features are of same importance. Features that does not contain sufficient information cause both ambiguity and computational complexity. The irrelevant features

may also cause degradation of the model. For this reason, reducing the number of features before analysis is important.

4.2.1. Pearson correlation coefficient

Correlation is a technique for investigating the relationship between two quantitative, continuous variables. Pearson's correlation coefficient is a measure of the strength of the association between the two variables. Given a pair of random variables (X, Y), the Pearson Correlation Coefficient (ρ) is calculated by (4-2).

$$\boldsymbol{\rho_{X,Y}} = \frac{\boldsymbol{cov(X,Y)}}{\boldsymbol{\sigma_X \sigma_Y}} \qquad (4\text{-}2)$$

Where, $\boldsymbol{\rho_{X,Y}}$ is the Pearson coefficient of X and Y

cov(X, Y) is covariance between X and Y,

and σ_X and σ_Y are the standard deviation of X and Y respectively.

Covariance is another measure of the relationship between two random variables. It measures, to what extent the variables change together. However, the metric does not assess the dependency between variables. Unlike the correlation coefficient, it is measured in units using (4-3)

$$\boldsymbol{cov(X,Y)} = \frac{\sum(\boldsymbol{X}-\bar{\boldsymbol{X}})(\boldsymbol{Y}-\bar{\boldsymbol{Y}})}{\boldsymbol{n}} \qquad (4\text{-}3)$$

Where cov(X,y) is the covariance between X and Y,

n is number of populations,

and $\bar{\boldsymbol{X}}$ ***and*** $\bar{\boldsymbol{Y}}$ are the mean value of X and Y.

Pearson's correlation coefficient has a value between +1 and −1. A value of +1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

4.2.2. Principal component analysis

Given a collection of points in higher dimensional space, a "best fitting" line can be defined as one that minimizes the average squared distance from a point to the line. The next best-fitting line can be similarly chosen from directions perpendicular to the first. Repeating this process yields an orthogonal basis in which different individual dimensions of the data are uncorrelated. These basis vectors are called principal components, and several related procedures principal component analysis (PCA) [60]. Hence PCA is defined as an orthogonal linear transformation that transforms

the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. It is a mathematical tool to reduce dimensionality of large dataset. It transforms a large set of variables into a smaller one that is easier to explore and visualize but still contains most of the information in the large set. It is especially helpful for the dataset where there is multicollinearity between variables.

PCA is a feature extraction technique. PCA was performed over impedance magnitude over the predefined frequency range where the device shows maximum sensitivity to minimize the number of features during analysis.

4.2.3. Variance inflation factor

In multivariate regression, "multicollinearity" refers to predictors that are correlated with other predictors. Multicollinearity occurs when the model includes multiple factors that are correlated not just to the response variable, but also to each other. In other words, it results when there are factors that are a bit redundant.

Multicollinearity increases the standard errors of the coefficients. Increased standard errors in turn means that coefficients for some independent variables may be found not to be significantly different from 0. In other words, by overinflating the standard errors, multicollinearity makes some variables statistically insignificant when they should be significant.

Variance inflation factor (VIF) is a measure of the amount of multicollinearity in a set of multiple regression variables. Mathematically for a predictor *i*, it can be calculated as (4-4)

$$\boldsymbol{VIF_i = \frac{1}{1 - R_i^2}} \qquad (4\text{-}4)$$

Where R_i^2 – value is obtained by regressing the predictor i on the remaining predictors.

Theoretically, VIF value can range from 1 to infinity. When VIF for a feature is equal to 1, then it represents that, the value at that particular feature cannot be predicted using other remaining features i.e. the particular feature is independent. A feature with a large VIF represents that, the particular feature is not independent to other remaining features. The general rule of thumb is that features whose VIF exceeds 10 are sign of serious multicollinearity that requires correction. Also, as we remove features with high VIF, VIF for other features decreases.

4.2.4. Backward elimination

Wrapper methods are general purpose feature selection algorithms that search the space of feature subsets, testing the performance of each subset using a fixed learning algorithm. The feature subset that gives the best performance is selected for the final use. Backward elimination is a simple wrapper method for feature reduction.

For backward elimination, the null hypothesis is – there is no relationship between the dependent variable, Y and predictor, X_i. So, if the null hypothesis is rejected, the alternative hypothesis is – there is relationship between the target variable, Y and predictor or feature, X_i.

P-value evaluates the probability that the sample data support the null hypothesis to be true. It measures how compatible the data are with the null hypothesis. A high p-value for a feature indicates that the null hypothesis is correct and therefore there is no relationship between the target variable and the feature. The estimated coefficient come from a normal distribution around zero and can be discarded. The term significance level (α) is used to refer to a pre-chosen probability. If p-value $\leq \alpha$ then the null hypothesis is rejected and therefore the alternative hypothesis is accepted. Otherwise if p-value of a feature is higher than the significance level α, then we accept the null hypothesis, i.e. there is no relationship between the target variable and that feature. As a rule of thumb, the significance level is usually set at 5% or 0.05.

In backward elimination method, at first step, a model is fit on the whole dataset including all independent variables or features. The p-value for all independent variables are then calculated. The feature with the highest p-value is eliminated from the dataset. This process continues until the p-value of all features is less than the significant level.

4.3. Regression models

In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the ‘outcome’ or ‘target’ variable) and one or more independent variables (often called 'predictors', 'covariates', or 'features'). When the number of target variables to be predicted is more than one it is called multivariate regression analysis and otherwise univariate regression analysis. Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning.

4.3.1. Linear regression

Linear regression is a linear approach to modeling the relationship between a response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression. For more than one explanatory variable, the process is called multiple linear regression. The equation for a multiple univariate linear regression model is given in (4-5)

$$y = a + bX \qquad (4\text{-}5)$$

where y is the dependent or target variable and X is a matrix of explanatory variables.

a and b are two parameters that are optimized to fit the data on the model. a is called bias and b is coefficient matrix

For multiple linear regression model, ordinary least squares are used to estimate the values of the coefficients. Given a regression line through the data, ordinary least square calculates the distance from each data point to the regression line, square them and sum all of the squared errors together. The process then seeks to minimize the residuals. This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients.

If the independent variables or the features are in same scale or have been scaled prior to fitting into a model, the coefficient of the linear regression model provide a basis for a crude feature importance score. The higher the coefficient, the more important that feature is.

4.3.2. Decision tree (DT)

Decision tree builds regression models in the form of a tree structure. It breaks down the dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. Finally, the model results in a tree with decision nodes and leaf nodes. A decision node is a node that has two or more branches, each representing values for the attribute tested. Leaf node is the terminating node that represents a decision on the numerical target.

Iterative Dichotomiser 3 or ID3 algorithm employs a top-down, greedy search through the space of possible branches with no backtracking to build a decision tree. To partition the data into subsets, standard deviation is calculated as a metric to measure homogeneity.

4.3.3. Random Forest (RF)

Ensemble learning is the process by which multiple models are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the performance of a model or reduce the likelihood of an unfortunate selection of a poor one.

A Random Forest is an ensemble technique capable of performing regression with the use of multiple decision trees and Bootstrap Aggregation, commonly known as bagging. Bagging, in the Random Forest method, involves training each decision tree on a different data sample where sampling is done with replacement. It combines multiple decision trees in determining the final output rather than relying on individual decision trees. The trees in random forests are run in parallel. There is no interaction between these trees while building the trees. While outputting, it averages the prediction of each tree.

As each tree draws a random sample from the original data set when generating its splits, adding a further element of randomness prevents overfitting.

4.4. Model evaluation

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents the data and also how well the chosen model is going to work for completely unknown dataset. Evaluating model performance with the data used to develop it is not an acceptable norm because it can easily generate overoptimistic and overfitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set which is unknown to the model to evaluate model performance.

4.4.1. Hold out method

In this method, the mostly large dataset is randomly divided into three subsets – training set, validation set and test set.

Training set is the subset of the dataset that is used to build the predictive models by learning the parameters by iterating through this subset.

Validation set is a subset of the dataset that is used to assess the performance of model built using the data in training set. It provides a test platform to fine tune model’s parameters and to

select the best performing one. However, if the model has no tunable parameter, validation set is not necessary in that case

Test set is an unseen dataset which is a subset of the original dataset but kept aside from the beginning to test the likely future performance of the model. If the model overperforms in the training phase and poorly performs when applied on test data, the chance is higher for the model to be overfitted to the training data.

4.4.2. K-fold cross validation

When only a limited amount of data is available, to achieve an unbiased estimate of the model performance *k*-fold cross-validation is used [61]. In *k*-fold cross-validation, the data is divided into *k* subsets of equal size. Models are built *k* times, each time leaving out one of the subsets from training and use it as the test set.

In this procedure, each observation in the data sample is assigned to an individual group and stays in that group for the rest of the process. Therefore, each sample is given the opportunity to be used in the hold set once and used to train the model k-1 times. The result of a k-fold cross-validation run is often summarized with the mean of the individual model scores. It is also a good practice to include a measure of the variance of the scores, such as the standard deviation.

The procedure has a single parameter called k that refers to the number of subsets that the dataset is to be split into. The value of k is chosen such that each train and test group of data samples is large enough to be statistically representative of the broader dataset. A general rule of thumb is to set this value to 5 or 10.

4.4.3. Evaluation metrics

Model evaluation metrics are used to explain the performance of model. The performance of the cross validation was evaluated in terms of primary metrics stated in [62]. They are the root mean square error (RMSE), mean squared error (MSE), mean absolute error (MAE), and coefficient of determination (R^2).

4.4.3.1. Mean absolute error

The average taken between the original values and predicted values is called mean absolute error (MAE) and is calculated using (4-6)

$$\boldsymbol{MAE} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} |\boldsymbol{y_i} - \widehat{\boldsymbol{y}}| \qquad (4\text{-}6)$$

Where $\widehat{\boldsymbol{y}}$ is the predicted value by the model for the observation with true value y_i and n is the sample size.

MAE values from 0 to infinity. MAE does not provide any direction of error, i.e. whether the model is overfitted or underfitted.

4.4.3.2. Root mean square error

In general, RMSE indicates how well the concentration of data points around the regression line. It is calculated using (4-7).

$$\boldsymbol{RMSE} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \widehat{\boldsymbol{y}})^{\mathbf{2}}} \qquad (4\text{-}7)$$

Where $\widehat{\boldsymbol{y}}$ is the predicted value by the model for the observation with true value y_i and n is the sample size.

Its value can range from 0 to infinite but usually higher than MAE. The smaller value represents that the data are close to the fitted line. As the error is squared before summed, it penalizes for the positive and negative error in the same way. RMSE is very sensitive to the outliers and its value is heavily influenced by them.

4.4.3.3. Standard deviation

Standard deviation, σ is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values are closely concentrated to the expected value while a high standard deviation indicates that the values are spread out over a long range. Mathematically standard deviation is calculated using (4-8)

$$\boldsymbol{\sigma} = \sqrt{\frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{x_i} - \overline{\boldsymbol{x}})^{\mathbf{2}}} \qquad (4\text{-}8)$$

Where, $\overline{\boldsymbol{x}}$ is the mean of all samples x_i, and n is the number of observations.

4.4.3.4. Co-efficient of determination

Coefficient of determination or R^2 is a statistical measure of how close the data point is fitted to the regression line. Mathematically it is calculated using (4-9)

$$\boldsymbol{R^2} = \mathbf{1} - \frac{\boldsymbol{SS_{res}}}{\boldsymbol{SS_{tot}}} = \mathbf{1} - \frac{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \widehat{\boldsymbol{y}})^{\mathbf{2}}}{\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} (\boldsymbol{y_i} - \overline{\boldsymbol{y}})^{\mathbf{2}}} \qquad (4\text{-}9)$$

Where, $\hat{\boldsymbol{y}}$ and $\overline{\boldsymbol{y}}$ are the predicted and mean value of *y*, n is the number of observations.

Its value usually ranges from 0 to 1 where 0 indicates none of the variability of the response data around its mean and 1 show that model explains all the variability of the response data around its mean. However, mathematically the value of R^2 can also be negative. When the fit of the model is worse than the fit of a horizontal line (through the mean value), the sum of squares from the model (SS_{res}) becomes larger than the sum of squares from the horizontal line (SS_{tot}). As a result, there ratio becomes grater than 1 and R^2 becomes negative. Such negative value implies that, a horizontal line through the mean value of the data explains the dataset better than the model.

4.4.3.5. Adjusted R^2

The adjusted R^2 is a modified version of R^2 that has been adjusted for the number of predictors in the model. Adding more predictors or features to a regression model tends to increase the R^2 value, which tempts to add even more. But this may cause overfitting and can return an unwarranted high R^2. Adjusted R^2 is used to determine how reliable the correlation is and how much is determined by the addition of independent variables. It is calculated using (4-10)

$$\boldsymbol{adj - R^2 = 1 - (1 - R^2)[\frac{n-1}{n-(k+1)}]} \qquad (4\text{-}10)$$

Where R^2 is the coefficient of determination for *n* observations and with *k* features

The value of adjusted R^2 usually ranges from 0 to 1 but it is always lower than R^2 value.

Chapter 5: Experimental setup and result analysis of experiment – 1

As observed in naked eyes in differently concentrated chlorophyll-a solutions, the color of the solution varies from light to deep green as the concentration is increased. So, in the first experiment, reflectance from the solutions have been measured at different wavelengths as an attempt to capture the color and they were correlated with the actual chlorophyll-a concentration in the solution. In the following sections of this chapter, the experimental setup, data acquisition, and result of this experiment have been discussed.

5.1. Hardware setup

A portion of photons incident on any subject is transmitted, absorbed, and/or reflected, maintaining the law of conservation of energy, and is stated by (5-1)

$$\rho + \alpha + \tau = 1 \qquad (5\text{-}1)$$

where ρ = reflectance,

α = absorbance, and

τ = transmittance.

Reflectance is the amount of flux reflected by a surface, normalized by the amount of flux incident on it, and for a wavelength λ, it is measured as (5-2).

$$\boldsymbol{p(\lambda) = \frac{Gr(\lambda)}{Gi(\lambda)}} \qquad (5\text{-}2)$$

where $p(\lambda)$ = reflectance at wavelength λ,

$Gr(\lambda)$ = reflected energy at wavelength λ, and

$Gi(\lambda)$ = incident energy at wavelength λ.

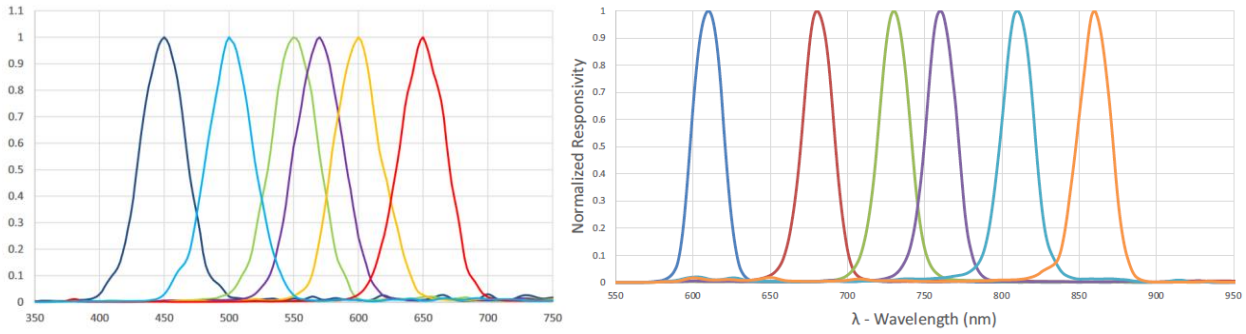
When excited by any wavelength in the spectrum, the reflectance of the subject varies with the electromagnetic energy of the source. Different surface features reflection or absorbance of electromagnetic radiation in different ways. The reflectance properties of an object depend on the material and its physical and chemical state, the surface roughness as well as the geometric circumstances (e.g., the incidence angle of the light). Therefore, the spectral reflectance pattern for different subjects at different wavelengths is a signature phenomenon.

In its liquid state, turbid water has relatively low reflectance, with clear water having the greatest reflectance [63].

5.1.1. Sensors

To measure the reflectance in the visible and near-infrared spectra, the highly integrated commercial sensors, AS7262 and AS7263 of Austria Mikro Systeme (AMS) were used, respectively.

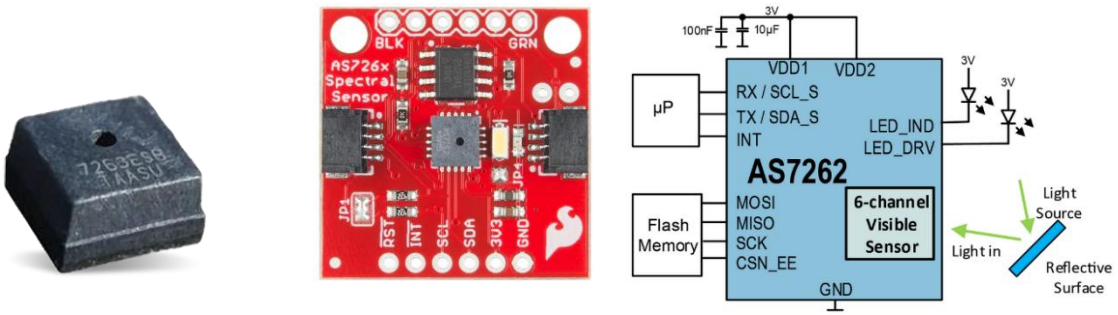
AS7262 delivers 6-channel multi-spectral sensing in the visible wavelengths from approximately 430nm to 670nm with full-width half-max (FWHM) of 40nm, as shown in Figure 5-1(a). According to their datasheet and Figure 5-1(a), six visible channels for AS7262 are 450nm, 500nm, 550nm,570nm,600nm and 650nm. In Figure 5-1(b), the spectral response of the AS7263 is shown. It also delivers 6-channel multi-spectral sensing but in the near-infrared wavelengths from approximately 600nm to 870nm with FWHM of 20nm. According to the datasheet of AS7263 and Figure 5-1(b), the channels for this sensor are 610nm, 680nm, 730nm, 760nm, 810nm and 860nm.



(a) (b)

Figure 5-1: Spectral response (reflectance in visible and near-infrared spectra) of the commercial sensors (a) AS7262 and (b) AS7263 of Austria Mikro Systeme (AMS).

The operating voltage for both the sensors is from 2.7V to 3.6V. They receive signal from and sends spectral data to microcontroller using I2C interface. The sensors output the energy at each channel in $\mu W/cm^2$ unit. The field of view of the sensors is ±20 degree. In this study, spectral breakout boards from SparkFun (SEN-14347 and SEN-14351) were used. The physical sensor, the breakout board and the block diagram for AS7262 have been shown in Figure 5-2.



(a) (b) (c)

Figure 5-2: (a) AS7262 spectral sensor (b) AS7262 Sensor and excitation LED integrated into breakout board from SparkFun (c) block diagram for breakout board

As the I^2C address of both the sensors are same, an I2C multiplexer from SparkFun (TCA9548A) was used to control both the sensors with single processor. The multiplexer can control 8 devices with similar I2C address.

5.1.2. Excitation source led

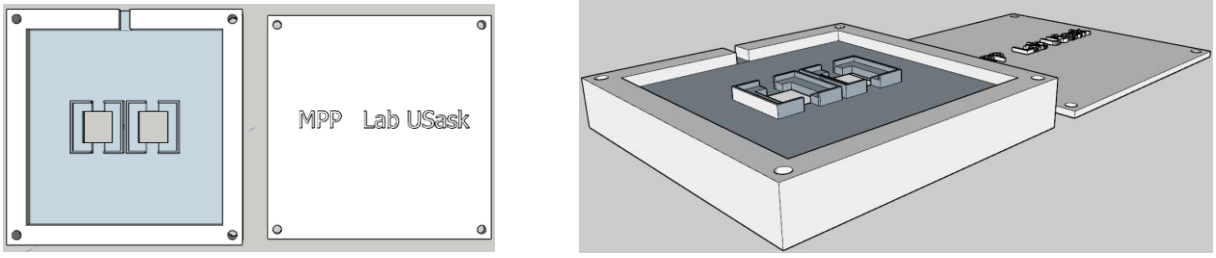
An integrated LED driver with a programmable current was provided with an AS726x for electronic shutter applications. The LED driver can sink current from 12.5 mA to 100mA. The built-in LED driver was used to operate the source LEDs. For measuring reflectance in visible and NIR spectra, part number L130-5780HE1400001 and L130-2790001400001 from Lumileds were used, respectively, as excitation LEDs. They are industry standard LEDs for uniform light. Their specifications are presented in Table 5-1. As shown, the lower color temperature of the light source for the NIR spectra signifies that the spectral peak of this warm-colored light is closer to infrared and this light source emits significant infrared radiation. The Color Rendering Index (CRI) in Table I indicates how accurate the light source is at rendering color when compared to the reference light source. The higher the CRI, the better the color-rendering ability. As a rule of thumb, light sources with a CRI of 80 to 90 are usually considered good at color rendering.

Table 5-1: Specifications for Excitation LEDs

	Light Source (Visible Spectra)	Light Source (NIR Spectra)
Part Number from Lumileds	L130-5780HE1400001	L130-2790001400001
Color Temperature	5700K	2700K
Minimum Color Rendering Index (CRI)	80	90
Typical Luminous Flux at 60mA	27	19
Typical Viewing Angle	116°	116°

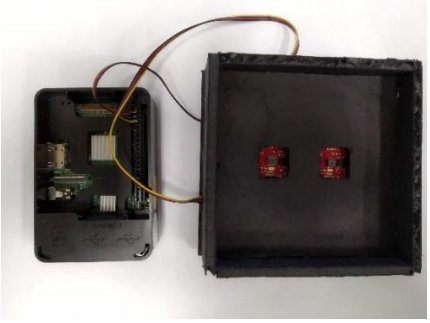
5.1.3. Enclosure

To avoid ambient light, the sensor and sample were placed inside an enclosure. In the experiment, the distance between the sample and the sensors was varied by 0.5 cm, starting at 0.5 cm and up to 3.5 cm to determine the optimal distance between them. The optimal distance was chosen where the sensor returned the strongest reflectance pattern for a fixed sample without any noise or interference, and that distance was kept constant while taking all readings. The height of the enclosure was set to this optimal distance. The rectangular enclosure has an area which is just large enough to cover a standard size (150mm x 15mm) petri dish. The 3D model of the enclosure from a different viewing angle is shown in Figure 5-4(a) and (b). Sketchup 2018 was used for 3D modeling. Figure 5-4(c) shows the finished device.



(a) (b)

Figure 5-3: The 3D model (a) top view, (b) isometric view,



(c)

Figure 5-4: (c) physical device.

5.1.4. Processing unit

While collecting data, Raspberry Pi (RPi) 3 Model B was used for controlling the sensors and storing the data primarily at its micro SD card. RPi is a popular microprocessor that is widely used in image processing [64], IoT systems [65], automations and in many other fields.

The RPi3 has a 1.2 GHz quad-core processor and 1 GB LPDDR2 RAM. It has 40 GPIO pins some of which were used to communicate with LCD display and the sensors. It has a micro SD port to load its operating system where data can be stored primarily. Due to its low computation power, data were collected and analyzed in a separate computer. A simple graphical user interface (GUI) was developed using tkinter library of Python 3.6 and shown in Figure 5-5. After analysis, the trained model was loaded into RPi to test the model on different samples.

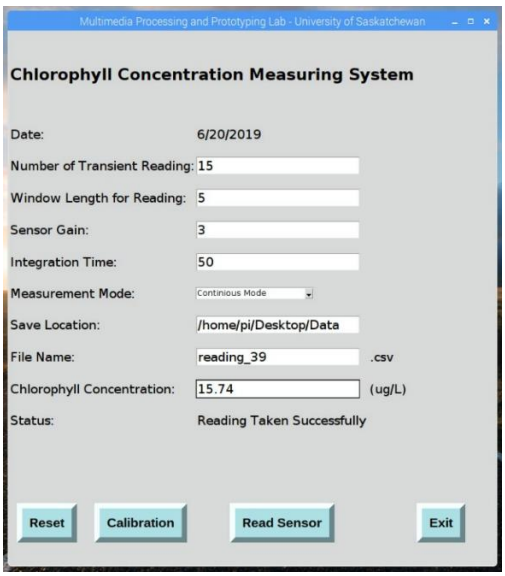
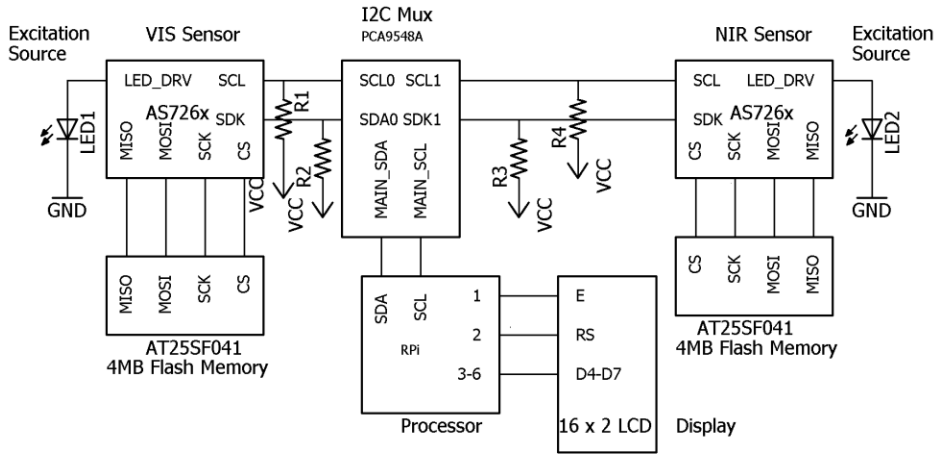


Figure 5-5: Graphical user interface for spectral data collection using AS7262 and AS7263

5.1.5. Power and display

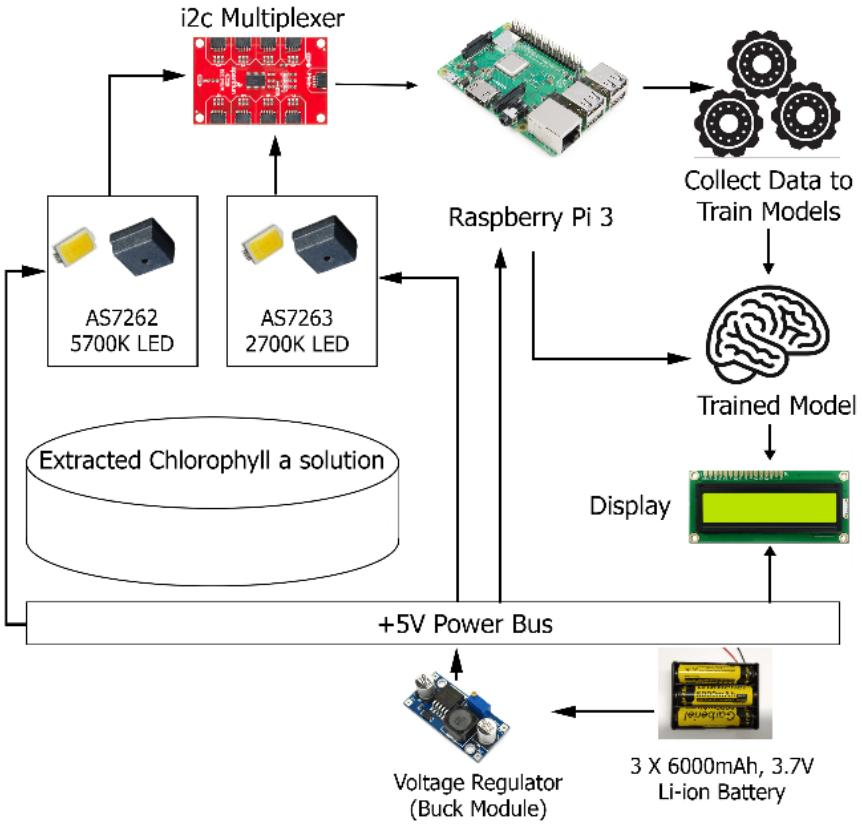
While selecting options for power up the whole system, portability was kept in mind. To supply continuous power to the sensors and processing unit, 3 6000mAh, 3.7V Lithium ion batteries were used. To regulate the voltage to the operating voltage of the sensors and processing unit, a buck module was used. Everything was powered up from this +5V power bus.

To show debugging messages a 16 X 2-character LCD display was used in the system. The schematic of the sensors' interfacing circuit and flow diagram for the setup is shown in Figure 5-7(a) and (b), respectively.



(a)

Figure 5-6: (a) Schematic of interfacing sensors



(b)

Figure 5-7: (b) Flow diagram of the setup

5.1.6. Spectral data collection

To determine the incident energy per cm^2 at a certain wavelength Gi (λ), a reflector (as shown in Figure 5-8) was placed under the sensor which then measured the reflected incident light. Next, the sample was placed under the sensor in similar conditions, and the sensor recorded the data at those wavelengths as Gr (λ). Finally, the reflectance was calculated using (5-2).

Figure 5-8: Reflector used to measure incident energy

A dataset of spectral reflectance at twelve different wavelengths (six at the visible range and six at the NIR spectra) of 52 chlorophyll solutions with varying concentrations was prepared. At least 5 separate readings using the proposed device were taken for each solution to ensure that the data do not overfit. To ensure that light was reflected from the sample and not from the bottom of the container, the sample was placed in a petri dish whose bottom was painted black with an anti-reflective mask. In addition, the volume of the sample was kept constant at 50mL for every reading.

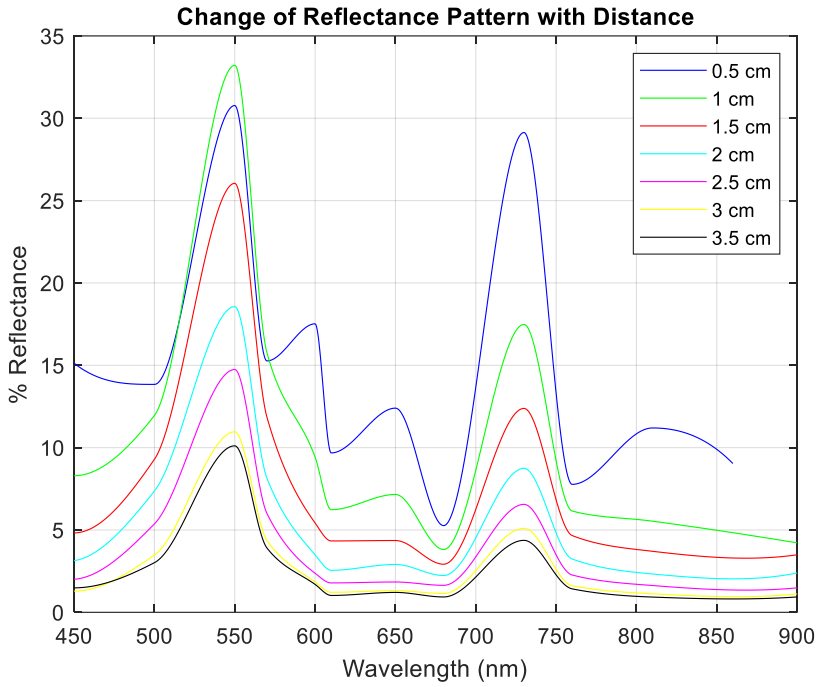
5.2. Result analysis

In optical spectral analysis, first different parameters (e.g. distance between sensor and sample, transient effect etc.) for experimental setup was tuned for optimum result. Later with the tuned experimental setup, reflectance spectra reading was taken and analyzed. Finally, different characteristics (e.g. repeatability, specificity etc.) of the sensor have been measured. The results have been described in the subsections.

5.2.1. Effect of changing distance

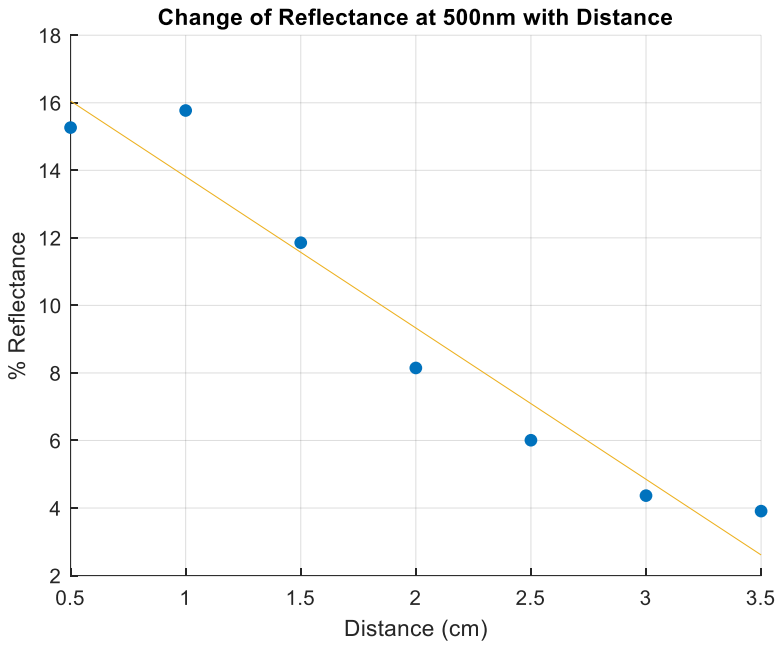
The reflectance pattern changes with the change of distance between the sample solution and the sensor. The pattern for the same sample is shown in Figure 5-10(a). It can be seen that as the distance between the chlorophyll solution and the sensors was increased while keeping other parameters constant, the signal became weak, but the reflectance pattern remained the same. Figure

5-10(b) shows the study of the reflectance at 500nm as an increment of distance. This study confirms that for our device, reflectance at a fixed wavelength changes almost linearly with an increase in distance. As can be seen from Figure 5-10, the highest signal without any noise is attained when the distance between the sensor and the sample remained at 1.5 cm. Shorter distances produced a noisy signal. When the sensor was placed farther away, the reflectance magnitude deteriorated, but could be retrieved by tuning the sensor gain according to the distance. In all the subsequent readings, the distance was kept constant at 1.5 cm.



(a)

Figure 5-9: (a) Reflectance signal pattern with changing distance,



(b)

Figure 5-10: (b) change of reflectance at 500nm with increasing distance.

Figure 5-10(a) shows that the reflectance at 1cm is greater than that of 0.5cm for lower wavelengths. This has been further investigated and depicted using Figure 5-11.

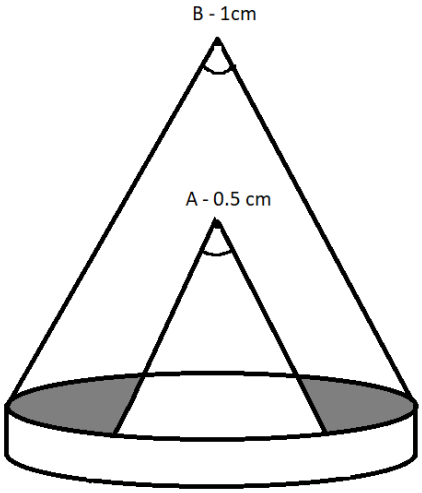


Figure 5-11: PFoV for sensor at 0.5cm and 1cm

In Figure 5-11, sensors are kept at A and B points, which are 0.5cm and 1cm away from the solution surface respectively. According to the datasheet, the sensors have a package field of view (PFoV) of 20^{o}. When the sensors were kept at point A, as shown in Figure 5-11, the sensors

were unable to capture the reflection from the gray area. So, reflected energy from the gray area was missing in the reading at 0.5cm. On the other hand, when sensors were kept at point B, the sensors were exposed to the reflection from the full surface. As a result, reflectance at 1cm is greater than that at 0.5cm.

However, a closer look at Figure 5-10(a) exposes that the reflectance at 1cm is high at lower wavelengths. The probable reason might be Raman scattering. Raman scattering states that photons at lower wavelength scatters more than that at higher wavelength [66]. When excitation lights at lower wavelength collides with the particle of the chlorophyll solution, it scatters more and creates high probability of going out of FoV of the sensors. The higher wavelength photons do not scatter much and so majority of them are captured even at 0.5cm.

5.2.2. Transient analysis

Figure 5-12 shows the recorded instantaneous reflectance reading of the solution at 450nm. The recorded signal showed some transient effect at the first few readings. For other wavelengths, the transient effect lasted for the first 8 to 13 readings of the sensor. Hence, the first 20 readings were omitted and the average of the next ten readings was recorded.

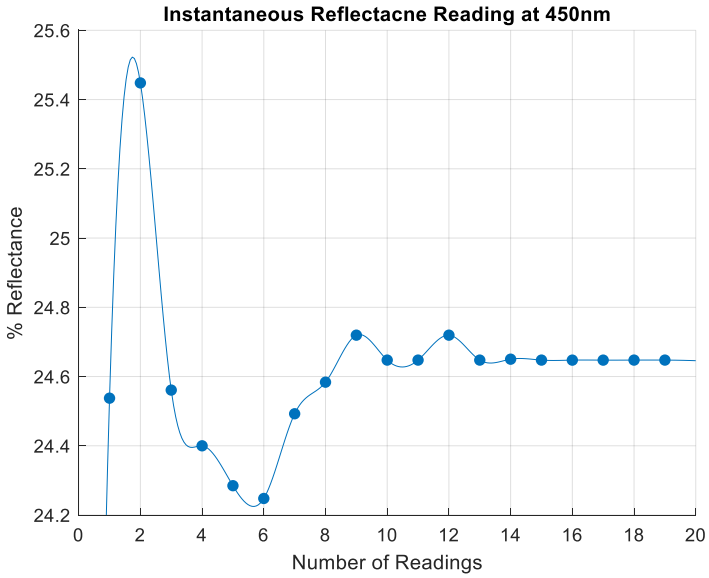


Figure 5-12: Instantaneous reflectance reading at 450nm

5.2.3. Specificity of the sensor

A sensor system is considered specific if it responds only to one kind of component [67]. Numerous interferences can exist in this type of analysis. These include particulate constituents that attenuate light (i.e., turbidity) and dissolved constituents such as dissolved organic matter. In addition, chlorophylls b & c, and pheophytin may interfere with analyses, as has been illustrated in standard lab-based methods, although these limitations have not precluded widespread use of spectroscopic methods in limnology and oceanography.

To examine the specificity of the sensor and assess the impacts of turbidity on measurements, 5mg of clay was mixed with 50mL of extracted solution. Clay was chosen as an interfering substance because it is widely available, and it simulates similar worst case scenarios when chlorophyll is extracted from a water sample in a high turbidity environment such as an erosive river or field. Readings were taken before and after adding clay to the solution. The experimental reflectance spectra shown in Figure 5-13 indicates that the proposed system is not specific to chlorophyll-a only. The turbidity of the solution caused significantly higher reflection at visible wavelengths than that of the original chlorophyll-a solution.

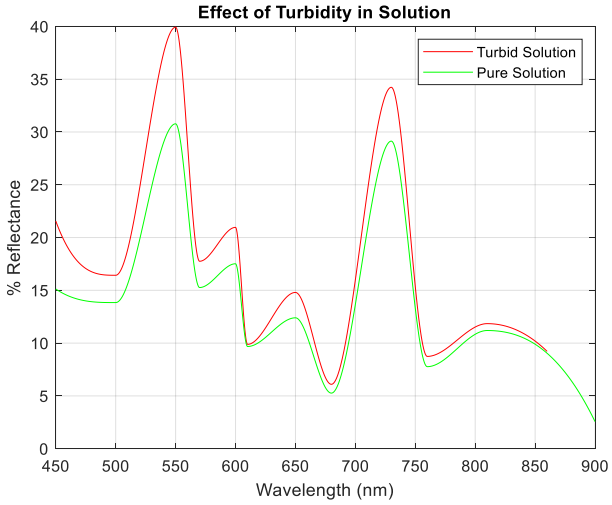


Figure 5-13: Effect of turbidity in reflectance pattern.

Given the proposed sensor is an optical reflectance sensor, any soluble or insoluble substance that changes the color of the solution will affect the reflectance spectra, and hence, the detection value. This is consistent with known interferences for other optical methods for measuring chlorophyll-a concentration. Only the fluorescent sensors are partially specific.

As the proposed sensor is not specific, extraction using 95%(v/v) ethanol solution and filtering through at least P5 graded filter paper is recommended while preparing chlorophyll solution to be measured. This will remove the effect of interfering substances in the medium.

5.2.4. Reflectance spectra analysis

The reflectance pattern was recorded for 52 chlorophyll solutions with varying concentrations. Some of the reflectance patterns with increasing chlorophyll concentration are shown in Figure 5-14. As described previously and shown in Figure 5-1, the sensor provided reflectance readings at 12 discrete wavelengths. The full spectrum was achieved by using interpolation by a 'pchip algorithm'. Figure 5-14 reveals variations in the percentage reflectance at 500nm, 550nm, 700nm, and 750nm, with changing chlorophyll-a concentration. To illustrate this variation more clearly, in Figure 5-15, the percentage reflectance at individual wavelengths was plotted against the chlorophyll concentration. At first glance, Figure 5-15 suggests that the reflectance at 550nm, 570nm, 680nm, 730nm, and 760nm might be significant wavelengths for measuring the concentration of the sample, because for these wavelengths, the reflectance varies significantly and at an almost constant rate. For other wavelengths, this variation is either not regular or so small that it might be undetectable by the model. However, more formal analyses were performed to further explore optimal wavelengths.

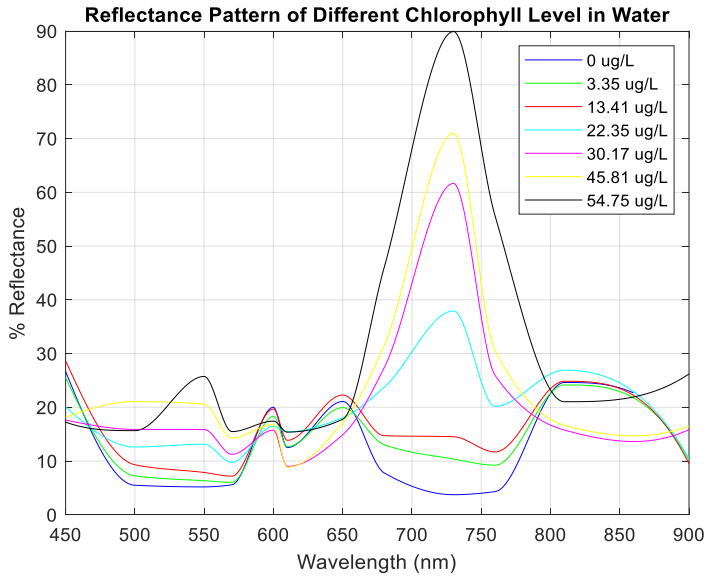
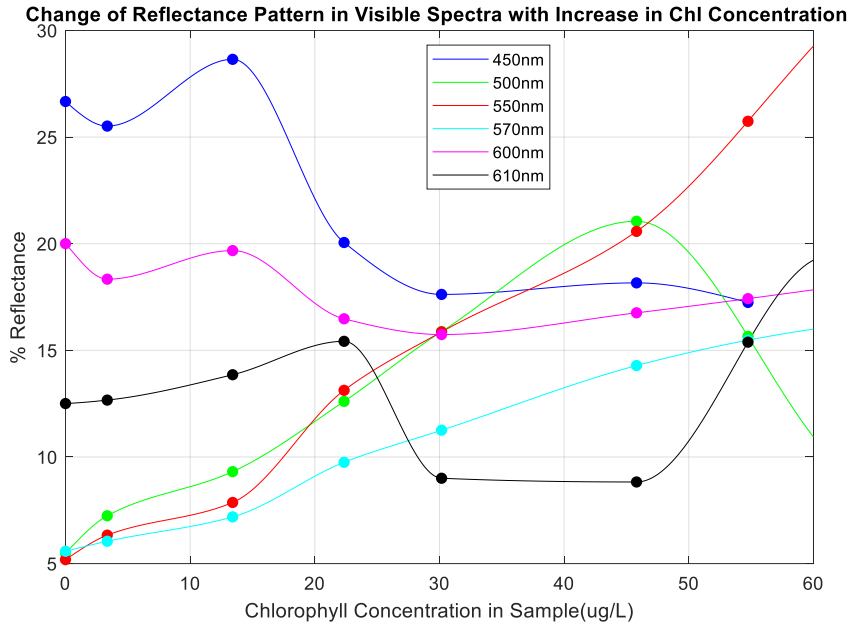
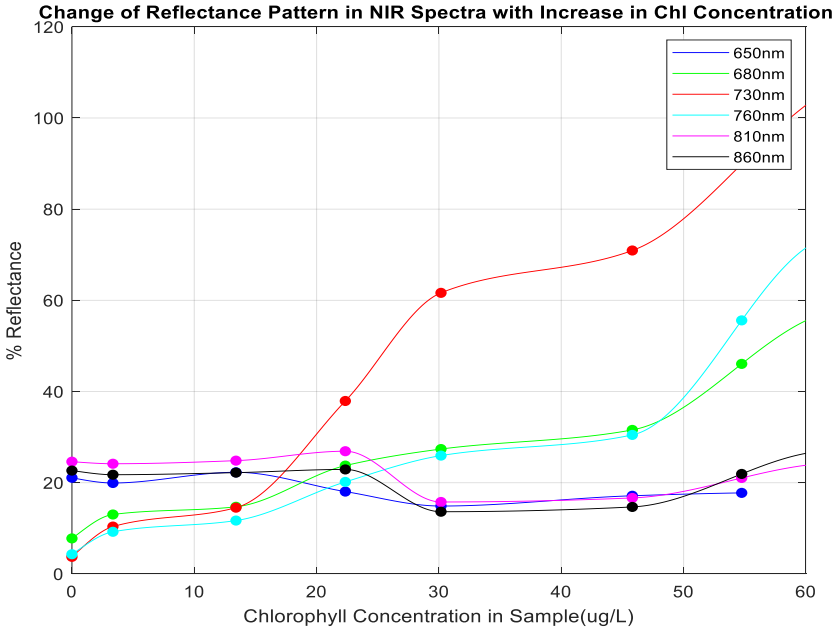


Figure 5-14. Reflectance pattern with increasing concentration of chlorophyll.



(a)



(b)

Figure 5-15. Change of reflectance at (a) visible (b) NIR spectra with increase in chlorophyll concentration in sample.

5.2.5. Analysis of Pearson coefficient

To better understand the correlation between reflectance at a certain wavelength and the chlorophyll concentration in solution, a heatmap of the Pearson coefficient between any two of the features is shown in Figure 5-16. It shows that the measurements of reflectance at 500nm, 550nm, and 570nm are strongly positively correlated with each other, and those at 810nm and 860nm are positively correlated. Those at 550nm and 860nm have the lowest correlations with measured chlorophyll concentrations of the sample, and hence, these two wavelengths have been omitted from the analysis. Although 600nm and 610nm are very close in the spectrum, their measurements of reflection are not strongly correlated with each other. This is probably because two different sensors were involved in taking the readings at these two different wavelengths. Reflectance at 600nm was captured using the AS7262 while that at 610nm was read by the AS7263.

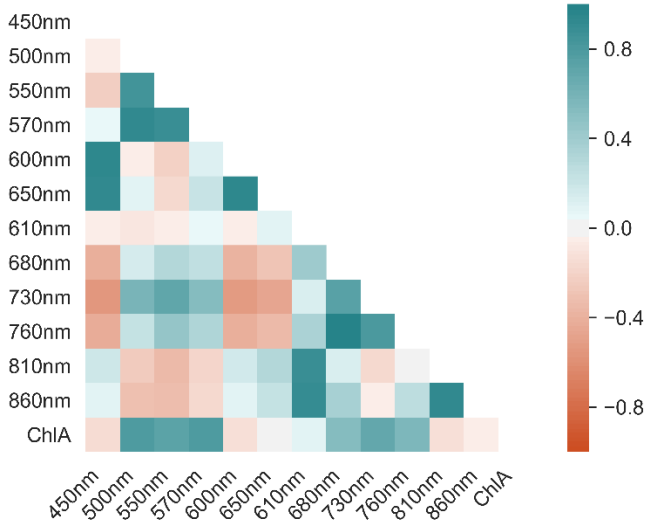


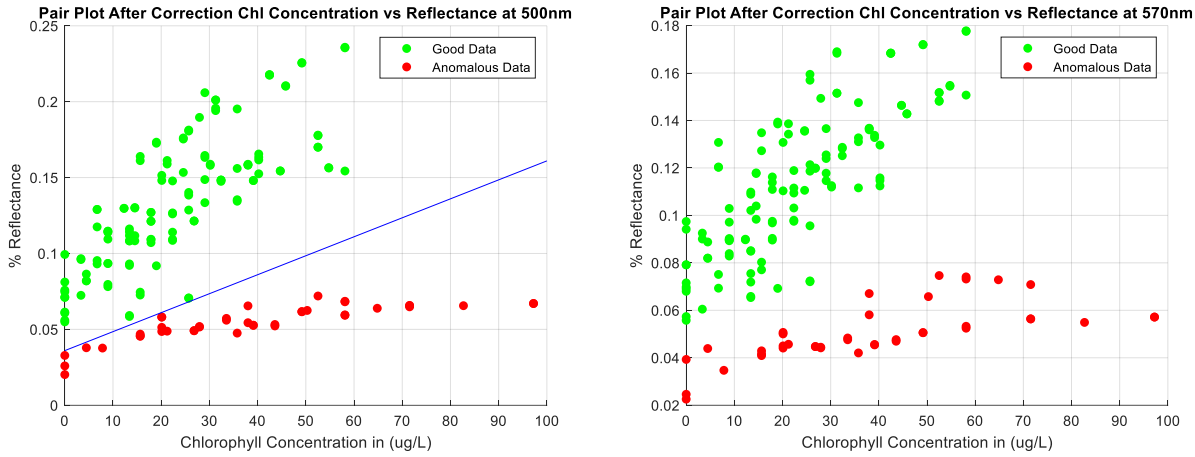
Figure 4-6

Figure 5-16. Heatmap of Pearson coefficient between any two features

5.2.6. Outlier detection

From the Pearson coefficient, it is already known that, there is a strong correlation (>0.8) between the concentration of chlorophyll and the percentage reflectance at 500nm or 570nm. When this correlation is visualized using a pair plot between the percentage reflectance at a wavelength and chlorophyll concentration, a single trendline is expected. However, when a pair plot with chlorophyll was plotted for reflectance at 500nm and 570nm in Figure 5-17(a) and (b)

respectively, the data clustered into two groups. Most of the readings formed a trendline, while the other group of data was some distance away creating an anomaly. Linear thresholding was applied to the pair plot between reflectance at 500nm and the chlorophyll concentration that separated the data into two different groups. From the analysis, data residing below the thresholding line in Figure 5-17(a) appeared to be anomalous and was removed before further analysis was done. The thresholding points were set manually so that it is just sufficient to separate two clusters, and eventually the model accuracy improved significantly. Figure 5-17 also reveals that the proposed system does not perform well beyond the chlorophyll concentration of 60 μgL^{-1}, as all the readings beyond this range were found to be erroneous. However, these more concentrated solutions can be measured by initially diluting the sample and then multiplying the reading by the dilution ratio.



(a) (b)

Figure 5-17: Pair plots of reflectance at (a) 500nm, (b) 570nm and chlorophyll concentration with identified good and anomalous data.

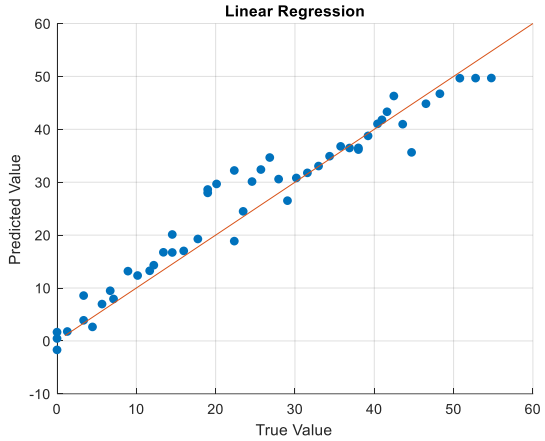
5.2.7. Model performance evaluation

After cleaning the data, the rest of the dataset was divided into a training set, validation set, and test set in a 6:2:2 ratio. It was confirmed that both the test and validation set contain all the variances of the dataset.

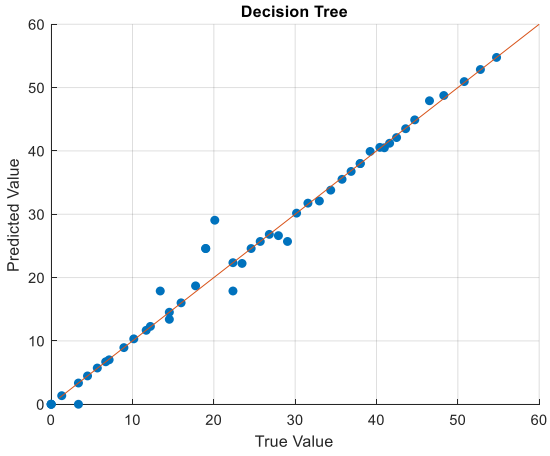
The training set was used to train three different machine learning models: a univariate multiple linear regression model, decision tree model, and random forest model. The best-trained model was selected as the one that outperformed the others when applied to the validation set. The test set was put aside at the very beginning so that it could be used to test the performance of the finally selected model on an unknown dataset.

To evaluate the performance of a model, the Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were used as matrices. The plot of ground truth values (measured using ThermoFisher NanoDrop 2000C) vs predicted value of the model ideally should be a straight line through the midpoint at a slope equal to one. A high R^2 value of this plot is another indicator of a good model.

In Figure 5-19, the plots of ground truth value vs predicted value are shown for different machine learning models. The red line through the midpoint in each figure is not the best-fitted line, but rather indicates the characteristic of an ideal model. The graph shows that the decision tree model predicts chlorophyll concentration closer to the ideal value than the other two models.

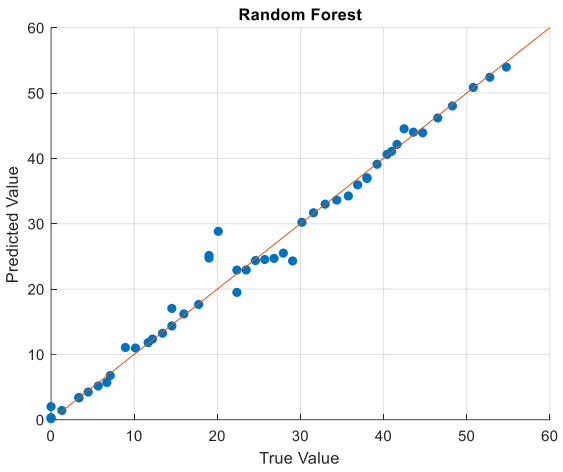


(a)



(b)

Figure 5-18: True vs Prediction plot for (a) linear regression model, (b) decision tree model,



(c)

Figure 5-19: True vs Prediction plot for (c) random forest model.

The values of the model evaluation metrics are shown in Table 5-2. It can be observed that, among the three trained models, the decision tree algorithm resulted in the minimum RMSE and MAE and maximum R^2 value for the True vs Predicted value plot. As a result, the decision tree model was chosen as the best performing one.

Table 5-2: Value of model evaluation matrices

	Linear Regression	Decision Tree	Random Forest
RMSE	5.23	1.81	1.89
MAE	3.41	0.86	0.97
R^2	0.87	0.965	0.964
adj R^2	0.86	0.961	0.960

5.2.8. Feature reduction

While analyzing the feature importance in the decision tree model, it was found that only five wavelengths played a major role in predicting the chlorophyll concentration in the sample. The comparison of these five predictors is shown in Figure 5-20. The two most significant wavelengths are 500nm and 680nm. The first one signifies the physical color of the solution. The traditional lab-based absorbance method, which has been used to measure chlorophyll content for a number of years, measured the absorbance at 685nm and 735nm [68][69]. Hence it is not surprising that reflectance at 680nm played an important role in our model to predict the chlorophyll concentration. Reflectance at 760nm was found to be significant probably because of the correction of the glass petri dish.

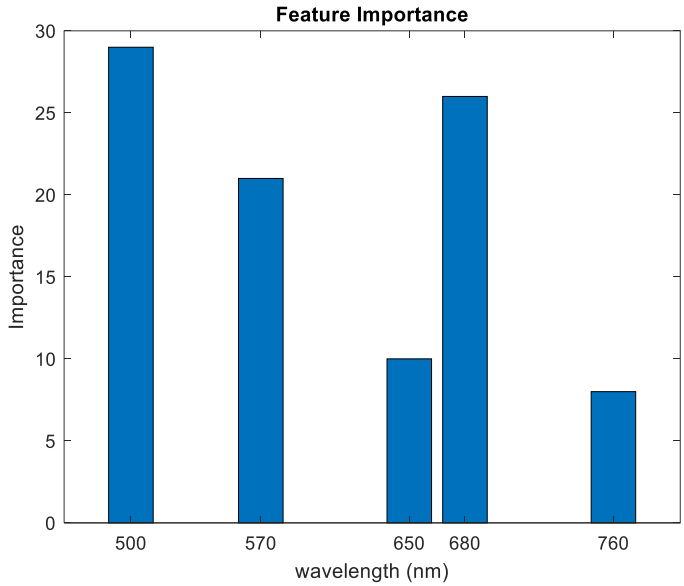


Figure 5-20: Importance of top five predictors

The same machine learning models were again trained from scratch, using only five extracted predictors. The evaluation matrices of the new models are presented in Table 5-3. It can be observed that even after keeping only five predictors, the results did not deteriorate substantively. Comparing Table 5-2 and Table 5-3 shows that, after removing the other seven wavelengths and keeping only the important five, the MAE for the decision tree decreases only by 6.97%.

Table 5-3: Evaluation of Models with Reduced Features

	Linear Regression	Decision Tree	Random Forest
RMSE	5.89	1.88	2.02
MAE	4.25	0.92	1.30
R^2	0.841	0.961	0.957
adj R^2	0.833	0.959	0.955

When tested on the chlorophyll solutions which were extracted from environmental water samples, the MAE was 1.06 μgL^{-1} and RMSE was 1.91 μgL^{-1}.

5.2.9. Repeatability of the sensor

To measure repeatability of this system, the percentage reflectance readings were taken using the proposed sensors on the same chlorophyll solution at one-minute intervals for an hour. Each set of readings was then passed through the proposed model to predict chlorophyll-a concentration. Thus, a total of 60 predictions were recorded for the same sample using the proposed model. The true value of the chlorophyll sample was then measured using the traditional laboratory-based absorption method. The mean of the 60 readings, their standard deviation, and standard error of the mean are shown in Table 5-4.

Table 5-4: Evaluation of Models with Reduced Features

No of readings, n	Measured using proposed system			True value - measured using absorption based method (μgL^{-1})	Error (μgL^{-1})
	Mean (μgL^{-1})	Standard deviation, σ (μgL^{-1})	Standard error of the mean (μgL^{-1})		
60	11.815	0.37	0.05	12.26	- 0.445

The mean value measured using the proposed model is the average of 60 predictions of the same sample. The standard deviation indicates the spread and variability of the predictions. The low standard deviation in this experiment indicates that the predictions are closely concentrated at the mean, which eventually results in high repeatability of the system. The standard error of the mean is a measure of how precise the mean prediction is. The lower value of it represents the high confidence level of an accurate mean. Error is the deviation of the mean of predictions, predicted by the proposed system from the true value. From Table IV it is evident that the mean prediction value is very close to the true value, and the error is within the reported mean error of the system, which is $\pm 0.9\ \mu gL^{-1}$.

5.3. Comparison with existing sensor

Table 5-5 draws a comparison of the proposed system with commercially available sensors and the model proposed in [70]**.**

Table 5-5: Comparison of the proposed system with commercial and literature-developed sensors

Topic of comparison	Make and Model			
	ThermoFisher NanoDrop 2000c	YSI 6025	Model proposed in [70]	The Proposed System
Method of detection	Absorption by extracted solution	In situ fluorescence	Phase fluorometer	Reflection from extracted solution
Weight	2 Kg	~ 5 Kg	-	0.7 Kg
Price/Cost (CAD)	$13,300		$1000	$150
Accuracy	± 0.05 μgL^{-1}	± 0.1 μgL^{-1}	± 4.46 μgL^{-1}	±0.9 μgL^{-1}
Range	Solution can be diluted to measure any range	0 - 400 μgL^{-1}	0 – 890 μgL^{-1}	Solution can be diluted to measure any range
Specificity	Not specific	Partially specific	Partially specific	Not specific
Excitation Source	Xenon Flash Lamp	LED with optical filter	LED with optical filter	2 different LEDs
Detector	CCD Array	Silicon photodiode with optical filter	Silicon photodiode	CMOS silicon with nano-optic filter
Peak Power Consumption	30W	14W	~1W	5W

From Table 5-5 it is observable that, although the proposed system is less accurate than the commercial sensors, it is lightweight, very low-cost, and consumes less power. In addition, it should be noted that while this is a nearly 20-fold decrease in accuracy, it is still sufficient for field and lake applications, where replicate samples often yield greater variation, and chlorophyll concentrations can be extremely high within and among various ecosystems.

5.4. Cost approximation

The cost of the prototype was one of the major concerns in this project. The details of the cost are shown in Table 6-6. As can be seen, the total cost of parts was around $135 (CAD).

Table 5-6: Cost Approximation of a Single Prototype

Item	Quantity	Per Unit Cost (CAD)	Total Cost (CAD)
AS7262	1	$25.95	$25.95
AS7263	1	$25.95	$25.95
Raspberry Pi	1	$39.95	$39.95
Display	1	$15.05	$15.05
Battery with case	1	$11.03	$11.03
Buck Module	1	$16.99	$16.99
Total Cost			$134.92

Chapter 6: Experimental setup and result analysis of experiment – 2

As chlorophyll-a dissolves into organic solvent, the organic solvent molecules form bond with dissolved chlorophyll-a. As the concentration of chlorophyll-a increases, the number of such bonds increases as well. As a result, the organic dipole solvent cannot conduct electricity as the same way as before for all frequencies. This change in electric conductivity has been correlated with the chlorophyll concentration. The experimental setup for this methodology and the result have been depicted in the following sections of this chapter.

6.1. Hardware setup

When plant leaves or algae are submerged into ethanol solution, the chlorophyll inside them makes a heavy dipole complex with it, which decreases the electric conductivity of the ethanol at some frequencies. As electric impedance is opposite to the conductivity; impedance increases when chlorophyll is dissolved into ethanol. This change of impedance is linear within a certain range of chlorophyll concentrations. However, the change of impedance is not uniform at all frequencies, because the complex is not resonated the same way at different frequency of sinusoidal voltage. This property of ethanol has been exploited in the proposed sensor by measuring the impedance before and after adding chlorophyll.

A typical setup of the impedance spectroscopy and our experimental setup is shown in Figure 6-1. Unlike electrochemical impedance spectroscopy, the properties of the electrode-material interface were not the focus of the experiment. The main interest here was on intrinsic electrical properties of the medium under test (MUT); therefore, contributions due to electrode effects had been avoided.

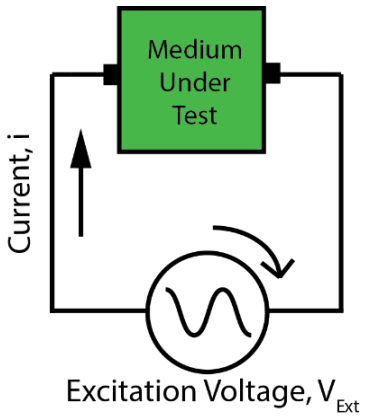


Figure 6-1: A typical setup for electrical impedance spectroscopy (EIS)

Let the impedance experienced by the MUT in Figure 6-1 be $\mathbb{Z}$, and is calculated using (6-1):

$$\mathbb{Z} = \frac{V_{Ext} \angle \theta_v}{i \angle \theta_i} \qquad (6\text{-}1)$$

Where, θ_v and θ_i are voltage and current phase angle, respectively.

As shown in (6-2), the impedance at an angular frequency ω, is a complex number. It has a real part, *Re(ℤ)* that signifies the resistance, R and an imaginary part, *Im(ℤ)* which signifies the reactance, X. Both capacitance and inductance of the medium is responsible for the reactance. The complex $\mathbb{Z}(\omega)$ can also be represented by its magnitude, $|\mathbb{Z}(\omega)|$ and angle, θ_z.

$$\mathbb{Z}(\omega) = |\mathbb{Z}(\omega)| \angle \theta_z = Re(\mathbb{Z}) + Im(\mathbb{Z}) = R + jX(\omega) \qquad (6\text{-}2)$$

The magnitude of the total impedance at an angular frequency ω is |ℤ(ω)| and calculated as in (6-3)

$$|\mathbb{Z}(\omega)| = \sqrt{R^2 + X^2} \qquad (6\text{-}3)$$

When the medium works as a dielectric, for cylindrical electrodes, various effects can be explained using the capacitance equation in (6-4)

$$C = \frac{\epsilon A}{d} = \frac{\epsilon\,(2\pi \mathrm{rl})}{d} \qquad (6\text{-}4)$$

Where, C = capacitance, ϵ = dielectric constant of the medium, A = electrode effective area, d = distance between two electrodes. l = effective length and r = radius of the electrode.

Finally, reactance due to this capacitance is calculated using (6-5)

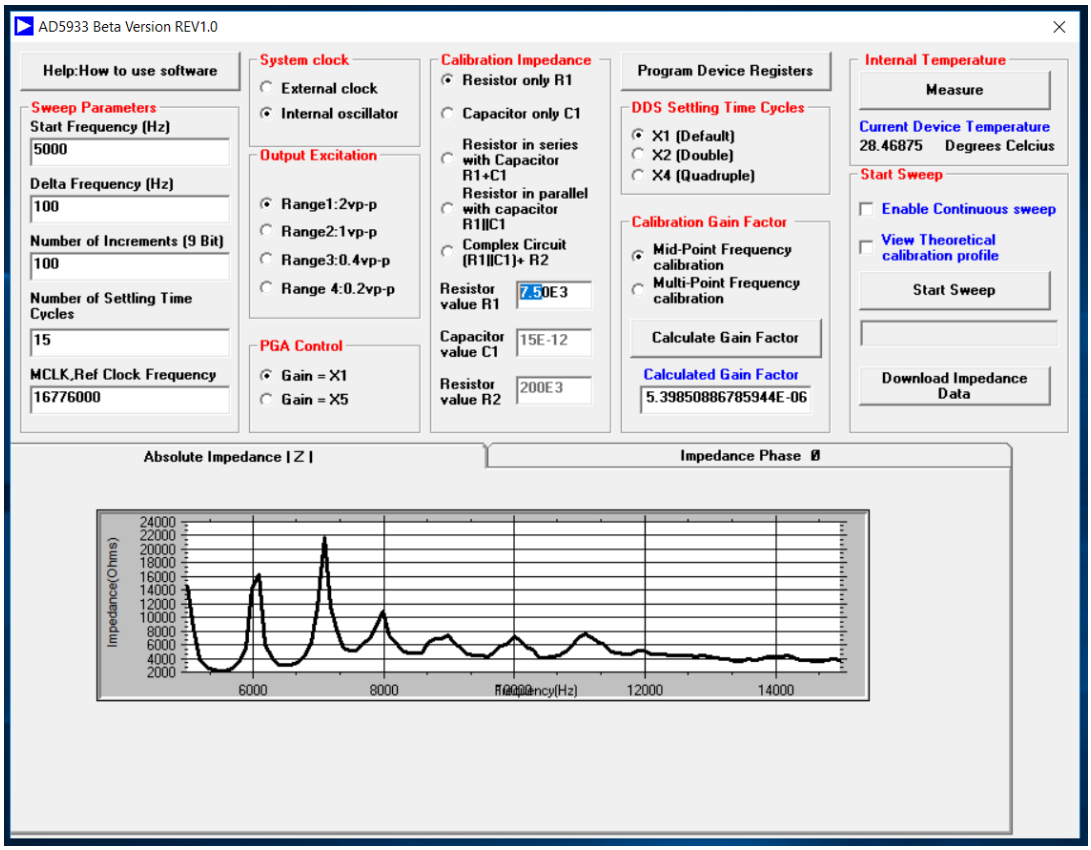
$$X_C = \frac{1}{2\pi f C} \qquad (6\text{-}5)$$

(6-3) – (6-5) have been found useful to explain different experimental results for the proposed sensor.

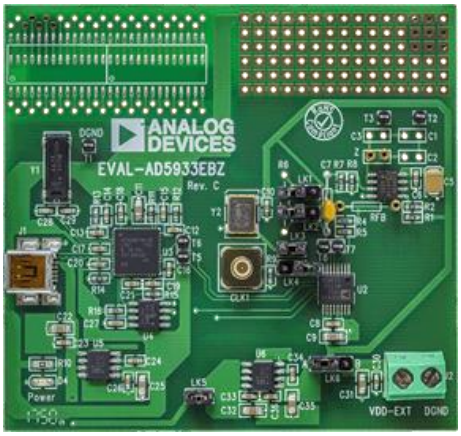
6.1.1. Sensor

In this experiment EVAL-AD5933 Evaluation Board from Analog Device (Figure 6-2(b)), a high precision impedance converter system was used. The device integrates an on-board frequency generator, a 12-bit 1 MSPS analog-to-digital converter (ADC), and an internal temperature sensor. Both the excitation signal and response signal are sampled by ADC and Fourier transformed by on-board DSP engine in order to obtain complex impedance spectrum. The frequency range of AD5933 is from 1 kHz up to 100 kHz. The device has a master clock of 16.77 MHz and supply voltage requirement of 2.7 V to 5.5 V. The device comes in a 16-SSOP package that has a temperature range of -40°C to +125°C. The device offers high accuracy and versatility

that make it suitable for electrochemical analysis, corrosion monitoring, automotive sensors, proximity sensing and bio-impedance measurements. As shown in Figure 6-2(a), it has its own software interface to collect data.



(a)



(a) (b)

Figure 6-2: (a) User interface for Eval-AD5933 from Analog Device (b) Physical Eval AD5933 Device (c) LCR817 from GwINSTEK

The reading of the sensor was verified against a high precision LCR meter (GWINSTEK LCR-817 – Figure 6-2(b)). The LCR device has a built-in signal generator and works in the frequency range from 12 Hz to 10 kHz with 489 steps and 0.05% accuracy.

6.1.2. Electrode

In all experiments, we used tin coated copper electrodes with inductance and Direct Current (DC) resistance of 0.11μH/ft and 0.027Ω/ft, respectively. For all measurements, the electrodes were fully submerged into the solution. Experiments were conducted to find the relationship of impedance reading with the length and diameter of the electrodes, and the distance between two electrodes. The results of the experiments are discussed in the result section.

6.1.3. Impedance spectral data collection

A standard curve constituted of 60 different concentrations of chlorophyll solution was prepared by extracting chlorophyll from lettuce leaf and spinach in 95% (v/v) ethanol ($CH_3 – OH$) to a maximum concentration of 20 μgL^{-1}. The impedance of 95% (v/v) ethanol and extracted chlorophyll in it was measured for various frequencies using the EIS sensor (Eval - AD5933). The solutions were carried in a 10mL standard plastic test tube. Before measuring the impedance of the MUT, the board was calibrated against a known RC circuit. After calibration, the magnitude of the impedance at each frequency point along the sweep was calculated. The experiment setup is shown in Figure 6-3. The excitation voltage was set to 0.2V peak to peak with no DC bias.

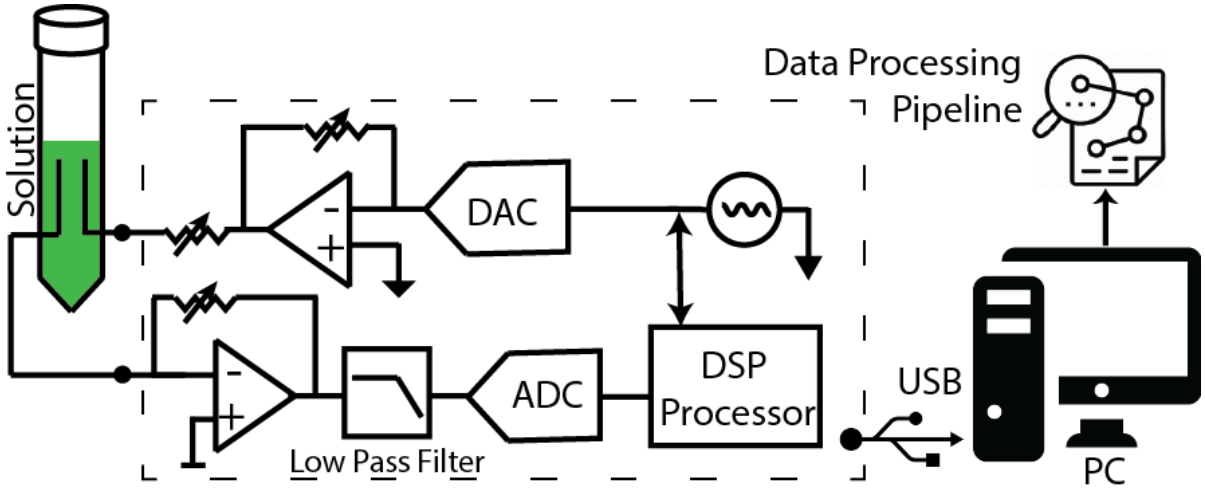


Figure 6-3: Workflow of measuring impedance spectroscopy

To prepare a dataset for development and testing, impedance readings were taken for both extracted chlorophyll solutions and 95% (v/v) ethanol using the same electrode configuration over a frequency sweep of 1.5 kHz to 7.5 kHz. The latter (background readings for ethanol) were then subtracted from the first reading. This avoids the electrode-material interfacing effect, temperature variations, and other variations due to the electrode configuration. The lower boundary of the frequency sweep was limited by the specification of the apparatus used.

6.2. Result analysis

Similar to the experiment -1 explained in Chapter 5: experiments were conducted to fine tune the parameters (e.g. electrode length, diameter, distance between two electrodes etc.) involved with the setup. Once the parameters were finalized, impedance spectra were analyzed to correlate it with known chlorophyll concentration. All the results have been described in the following subsections.

6.2.1. Verification of the readings of impedance sensor

To verify the reading of our proposed EIS sensor, the readings were taken for a parallel RC circuit of R = 1000 Ω (with 1% tolerance) and C = 10 μF. The cross verification was studied against a high precision LCR meter from GwINSTEK (LCR-817), which has a basic accuracy of 0.1%. Readings were taken at regular interval from 1 kHZ to 6 kHz. The theoretical values of impedance in this frequency range have been calculated using an off-the shelf "EIS Spectrum Analyzer" software [71]. The comparison is shown in Figure *6-4*. The plot shows that the variation in the reading using our sensor (AD5933) is close to the value measured using the standard tool (LCR-817), and within 10% of the theoretical value.

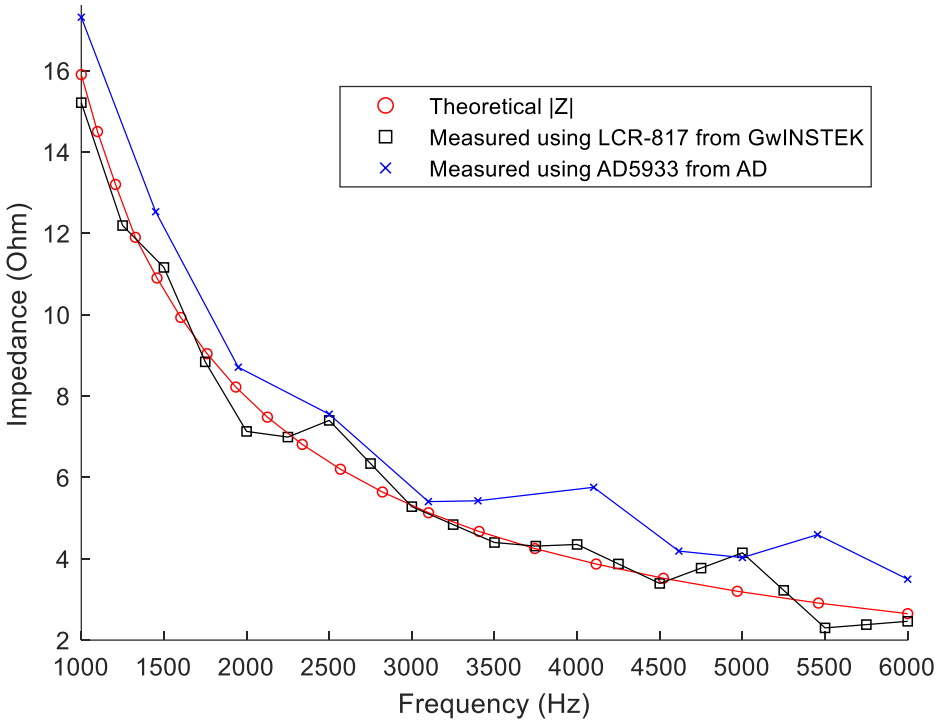


Figure 6-4: Comparison of impedance readings: theoretical, LCR-817 reading, and AD5933 reading

6.2.2. Electrode Configuration

To conduct the experiments, several possible electrode configurations have been considered by varying electrode orientation, electrode effective area, and distance between electrodes. The impedance reading varies for different configurations. To understand how each configuration change the impedance reading and also to choose the best electrode configuration for the experiments, the relationship between different electrode configurations and impedance reading have been studied one at a time for the same chlorophyll solution, keeping all other parameters constant.

At first, two possible electrode orientations were considered - from the top into the solution or submerged into the solution from the bottom, as shown in Figure 6-5. The full impedance spectra for the same chlorophyll solution using both orientations are shown in Figure 6-6.

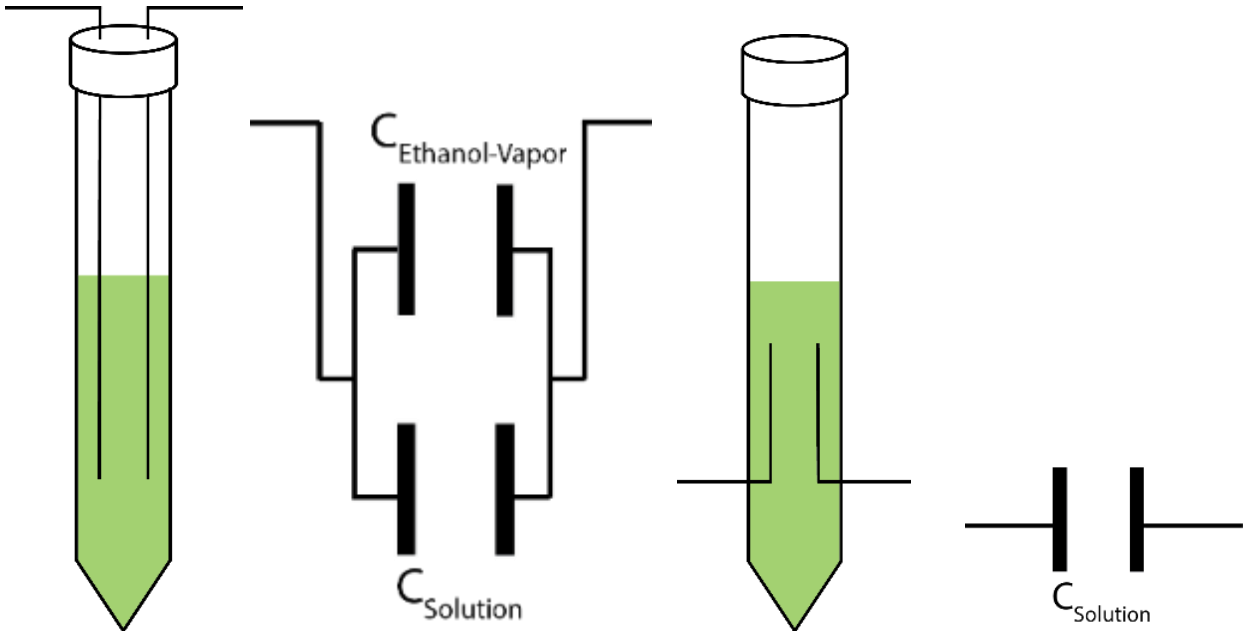


Figure 6-5: Two different considered orientations - (a) electrode from top: creating two capacitors in parallel (b) electrode from the bottom: fully submerged into chlorophyll solution

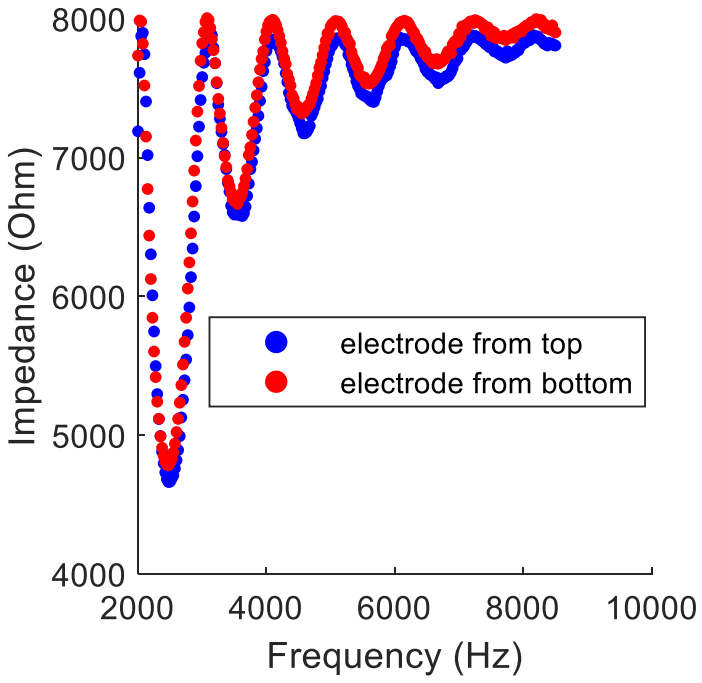


Figure 6-6: Spectra for two different orientations of the electrodes

Figure 6-6 shows that the impedance for orientation (a) is either the same or lower than that of orientation (b) for all frequencies. This occurs because the evaporation rate of ethanol at room temperature is high. As a result, at room temperature ethanol evaporates and fills the empty space on top of the solution. Hence, the upper part of the electrodes is submerged into ethanol vapor while the lower part is submerged into chlorophyll solution. Together they form a parallel combination of capacitors. As a result, the total capacitance increases which eventually causes the impedance to decrease according to (6-5). For orientation (a), the volume of the solution plays an important role in the impedance. It determines the effective area between the electrode and the dielectric of two capacitors in parallel. Also, the evaporation of ethanol depends on temperature. So, the orientation where electrodes are placed from the top is less accurate than that from the bottom. Hence, the orientation (b) was chosen for all subsequent experiments.

The electrodes that were used in the experiments were cylindrical in shape. So, the effective area between the electrodes and the solution can be varied by varying either length or diameter of the electrodes. When effective area was varied by varying one of the variables, it was observed that the shape of the impedance spectra was consistent. It just shifted up or down for the changed parameter. So, in the subsequent experiments, instead of the whole impedance spectra, the effect of changing a parameter on impedance at a fixed frequency (2300Hz) is presented.

In Figure 6-7(a) and Figure 6-7(b) the impedance at 2300Hz with changing electrode length and diameter is presented respectively. It is seen from both the figures that the impedance decreases linearly with increment of length or diameter of electrode. This can be explained by (6-4) and (6-5)**.** As either the length or diameter of the electrodes was increased, the effective area increased linearly as well. This increased the capacitance and eventually decreased the impedance between the electrodes. For subsequent experiments, the length and diameter were fixed to 2.1 cm and 0.033 inch, as this are the closest point to the fitted line.

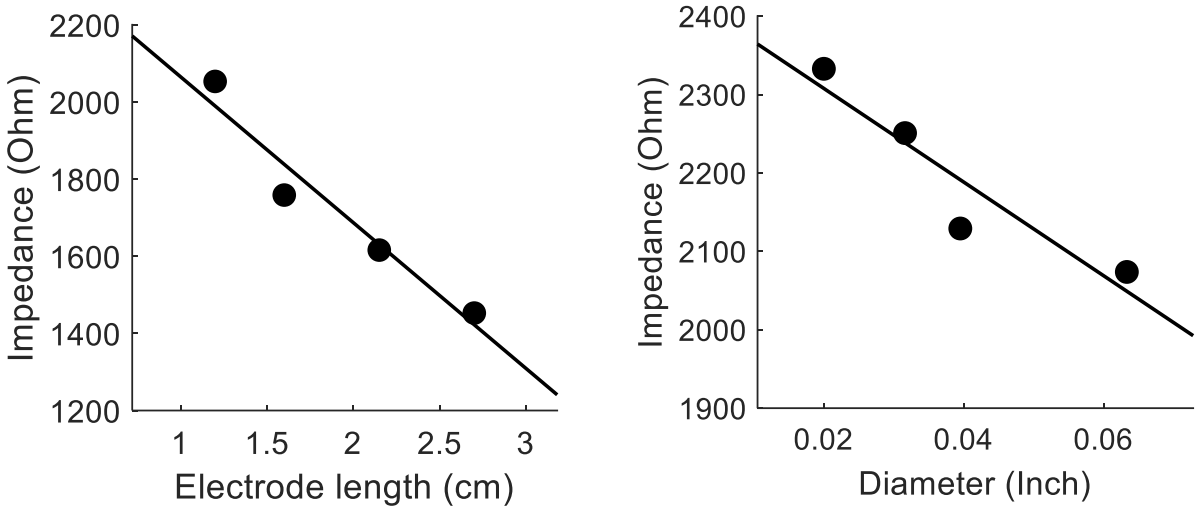


Figure 6-7: Impedance at 2300Hz for varying (a) length (b) diameter of the electrode

Finally, an experiment was conducted by varying the distance between two electrodes from 0.1cm to 0.8cm. The upper boundary was limited by the diameter of experimental test tube. The result of the experiment is shown in Figure 6-8. As the distance between the two electrodes is increased, the capacitance between them decreases. This decreased capacitance increases the reactance and hence increases the magnitude of impedance as well. The distance between the electrodes were kept fixed at 0.3-cm (since like before, this is the closest point to the fitted line in Figure 6-8.)

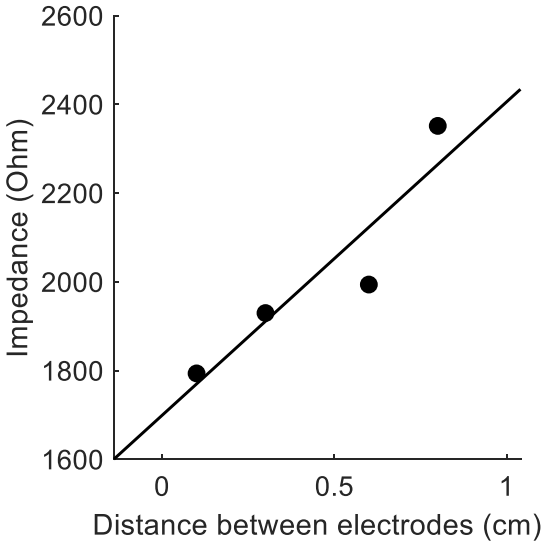
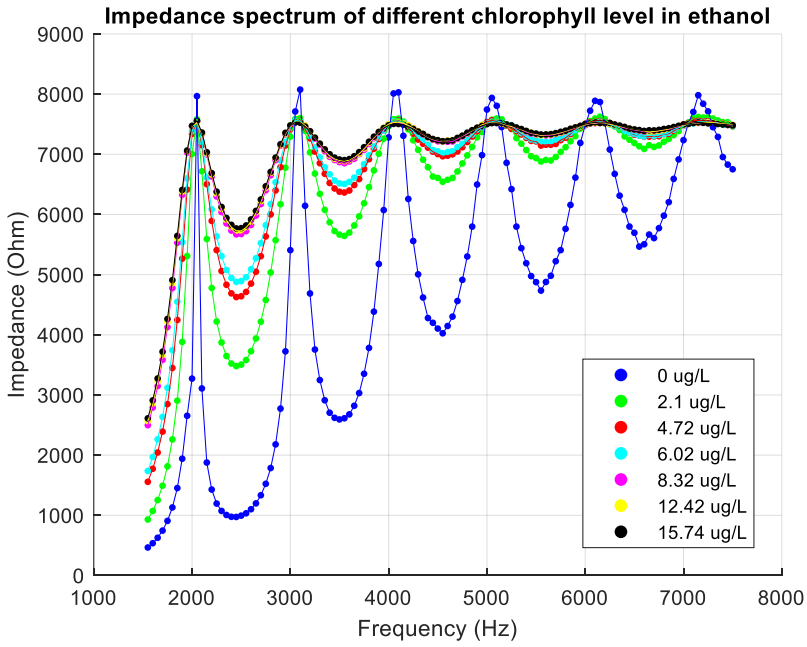


Figure 6-8: Impedance at 2300Hz for varying distance between two electrodes

6.2.3. Impedance spectra analysis

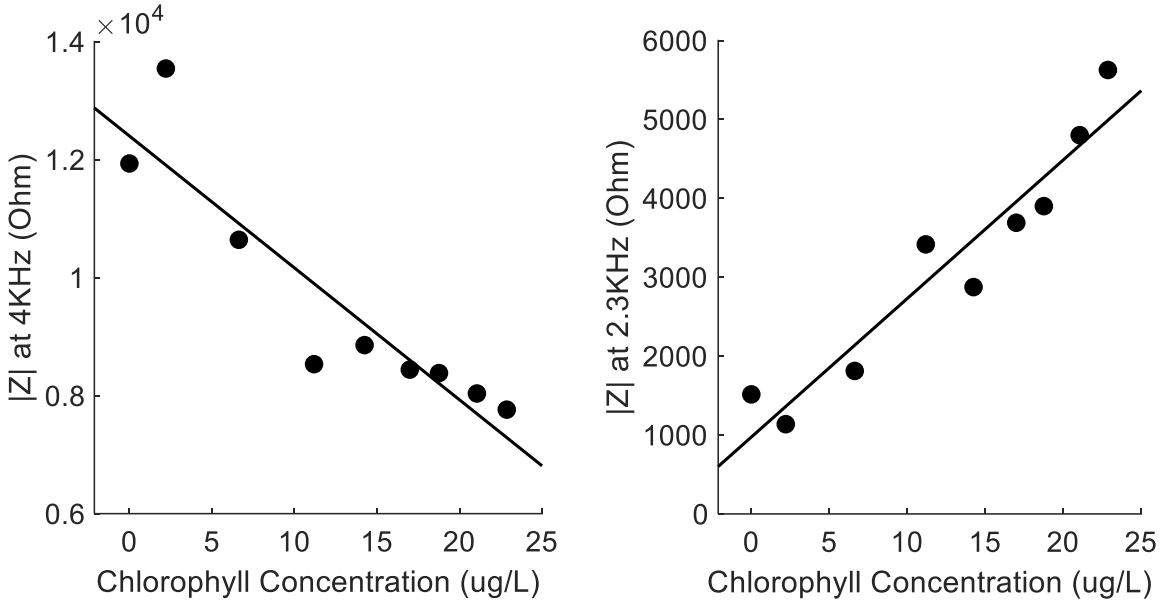
After choosing the optimal values for the experimental setup, impedance readings were taken for 60 different concentrated chlorophyll solutions. The impedance values were recorded at 50Hz interval from 1.5kHz to 7.5kHz. The lower value of the frequency range was constrained by the sensor board specification with 4MHz external oscillation. Among the readings, few impedance spectra were chosen randomly for study and are shown in Figure 6-10(a)**.** This primary study shows that increasing chlorophyll concentration changes the impedance spectra of ethanol. One importance observation from Figure 6-10(a) is that the change of impedance is prominent in lower frequency (2 kHz - 5 kHz), and almost negligible at the higher frequency (more than 5 kHz). This might occur due to the large atomic size of chlorophyll which resonates more at lower frequencies.

To better understand the relationship between the impedance and chlorophyll concentration, impedance at a fixed frequency was studied (shown in Figure 6-10(b) and (c))**.** The figures show that, with increasing chlorophyll concentration, the impedance at 2300Hz and 4000Hz changes almost linearly. The change is proportional for some frequencies while inversely proportional for some others. More formal studies have been presented in the following sections.



(a)

Figure 6-9: (a) Impedance spectra



(b) (c)

Figure 6-10: (b) Impedance at 4kHz with increasing chlorophyll concentration (c) impedance at 2.3kHz with increasing chlorophyll concentration

6.2.4. Principal component analysis

Principal component analysis was applied on the whole dataset. To find out the optimum number of principal components (PC), explained variance was calculated for PC 1 to 10. For each PC, the sum of explained variance has been plotted in Figure 6-11 and the elbow was chosen as the optimum number of principle component. Therefore, 5 principle components were chosen which can explain 99.2% of total variance.

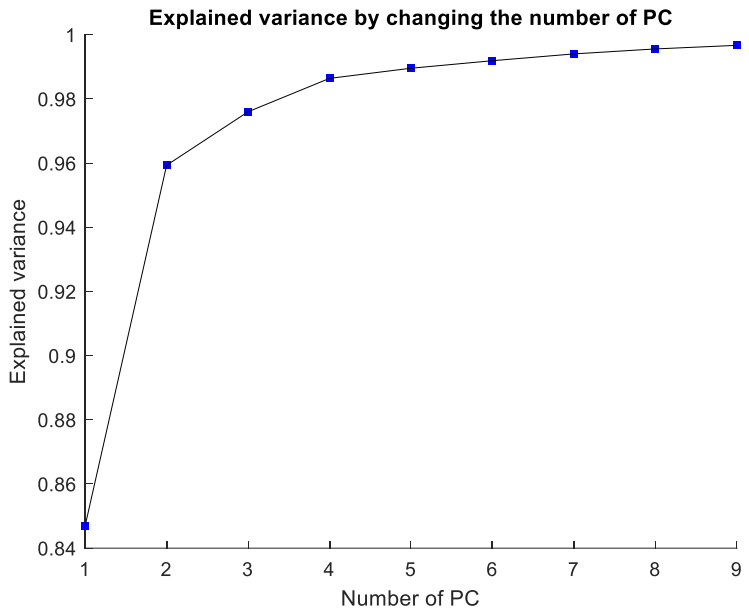


Figure 6-11: Explained variance by changing the number of principle component

A multiple linear regression model was fitted after the principle component analysis and with 5 principle components. The predicted value for the model has been shown against the true value in Figure 6-12 The evaluation metrics have been shown in Table 6-1.

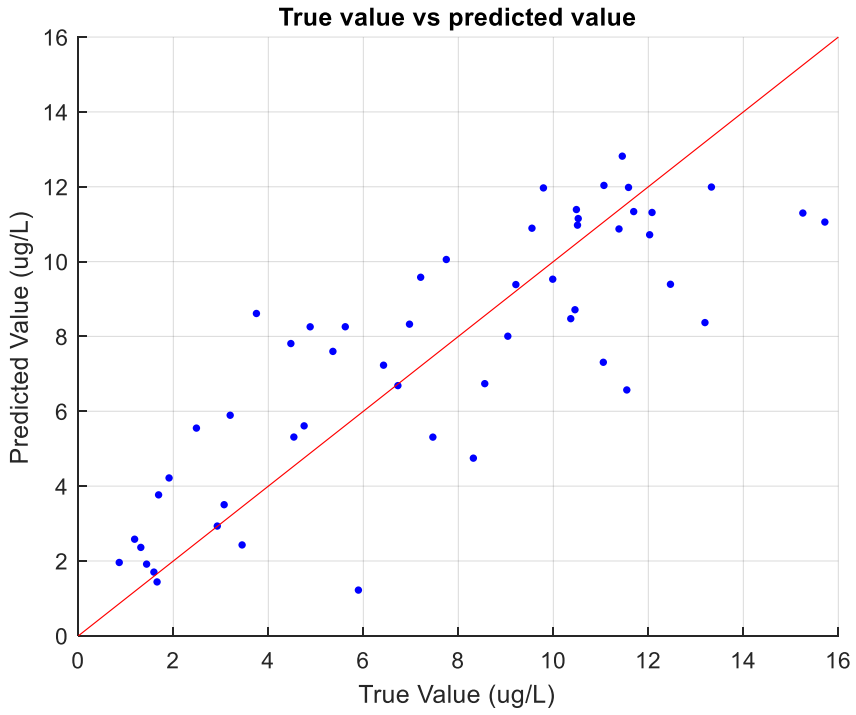


Figure 6-12: True value vs predicted value for linear regression after PCA

Table 6-1: Evaluation metrics for linear regression after PCA

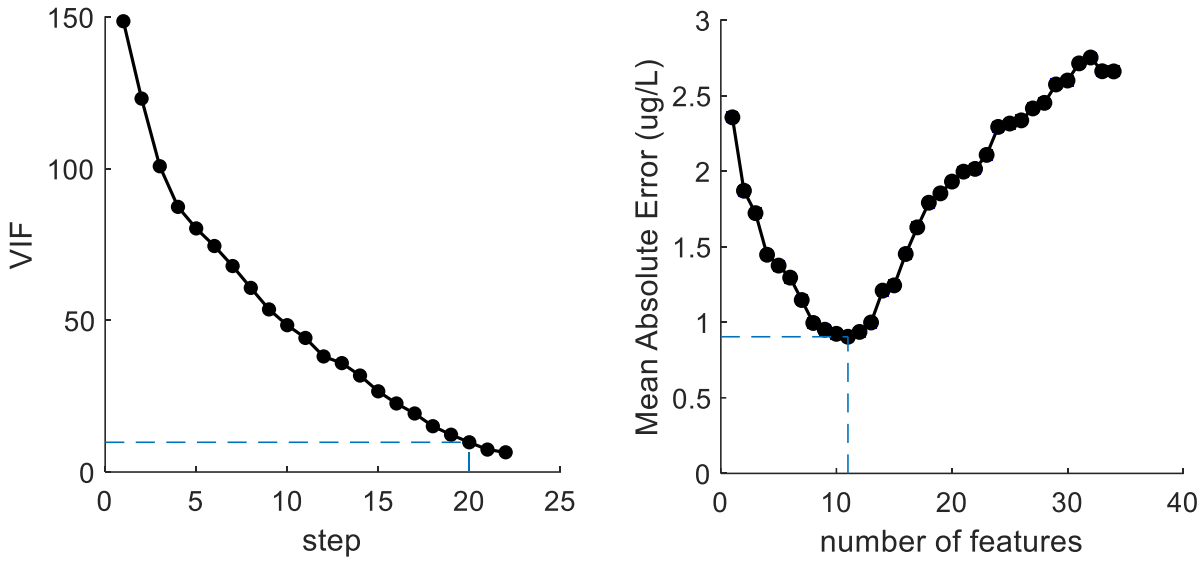
Evaluation metric	Reported value
R^2	0.694
Adj-R^2	0.627
MAE	1.815 μgL^{-1}
RMSE	2.2975 μgL^{-1}

As can be seen from Figure 6-12 and in Table 6-1, there is a large avoidable error in this analysis using PCA. So, features were reduced using further analysis.

6.2.5. Feature reduction

The Pearson correlation coefficient measures the linear correlation between two variables. In this project, Pearson coefficient of the features were studied to understand how each feature are linearly correlated with chlorophyll concentration. Features, whose absolute Pearson coefficient with chlorophyll content is lower than 0.5 was identified as weekly correlated feature with chlorophyll and was removed. This process eliminated 16 least significant features. However, a close absolute value of the remaining features was found which implies that there is strong multicollinearity among them.

To remove multicollinearity and redundancy, features were filtered based on Variance Inflation Factor (VIF). VIF measures how much the variance of an estimated regression coefficient is increased because of collinearity. The low value (< 10) of VIF represents lower collinearity between features. Features with highest VIF were removed at a time and VIF with the remaining features were calculated again. The process was repeated until VIF for all the remaining features were less than 10 which implies significant low collinearity. Figure 6-13(a) shows how VIF decreased gradually by removing the feature of highest VIF at each step.



(a) (b)

Figure 6-13: Feature reduction using: (a) Variance Inflation Factor; (b) using backward elimination

To further reduce the number of features, the dataset was passed through a backward elimination process. In this process, a multivariate linear regression model was fit using all features and p-values of all the features were calculated. Features with a high p-value are not likely to be a meaningful addition to the model because changes in the feature's value are not significantly related to changes in the chlorophyll concentration. We applied a stepwise approach where the feature with the highest p value was eliminated from the feature set and the model was retrained with remaining features. This process continued unless p-value for all features in the model was less than 0.05. This final step returned a total of 11 features. They are – Z_{2300}, Z_{2600}, Z_{2950}, Z_{3200}, Z_{3900}, Z_{4000}, Z_{4150}, Z_{4250}, Z_{4300}, Z_{4400} and Z_{4700}. Here Z_X represents impedance at X Hz. This eventually brought the model to the lowest mean absolute value with minimum number of features (as shown in Figure 6-13(b)). It is important to note that when more than 11 features were removed, the error started to rise again which indicates that the selected set of features is the optimal choice. Figure 6-14 shows the spectra of previously studied chlorophyll solutions in Figure 6-10(a), but only the optimum 11 frequencies are shown. Figure 6-14 shows that the variance of chlorophyll is still visually apparent with only selected 11 features.

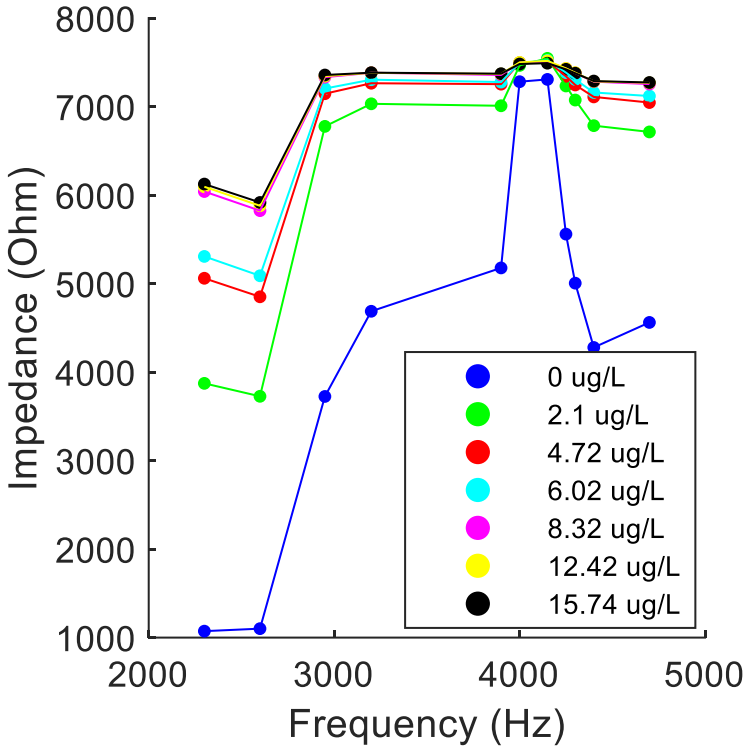


Figure 6-14: Impedance spectra at significant frequencies

6.2.6. Model evaluation

To fit the data into a model, a univariate multiple linear regression model was chosen. To evaluate how well this model will perform on unknown dataset, a 5-fold cross validation was performed on the whole dataset. The results are presented in Table 6-2**.** The low values of root mean square errors for each fold in Table 6-2 signifies that the simple linear regression model will perform with high precision on unknown dataset. Also, the low standard deviation of the RMSE signifies that there is low variance among the experimental results.

Finally, a univariate multiple linear regression model was trained using the whole dataset. The predicted value was plotted against the true value (as shown in Figure 6-15**)**. In the figure, the red line is not the best fitted line, rather it indicates the ideal (1:1) relationship. The evaluation metrics is presented in Table 6-3**.** Here, higher R^2 and adjusted R^2 imply that the model has been a good fit to the dataset. Besides, the low difference between R^2 and adj-R^2 signifies that, all the features chosen to predict the chlorophyll concentration are independent and significant. Moreover, the MAE and RMSE values are small for chlorophyll measurement.

Table 6-2: Evaluation of 5-fold cross validation

Fold No	# of samples in training	# of samples in test	RMSE	Mean RMSE	SD of RMSE
1	56	14	1.05	1.522	0.3269
2			1.58		
3			1.37		
4			1.71		
5			1.90		

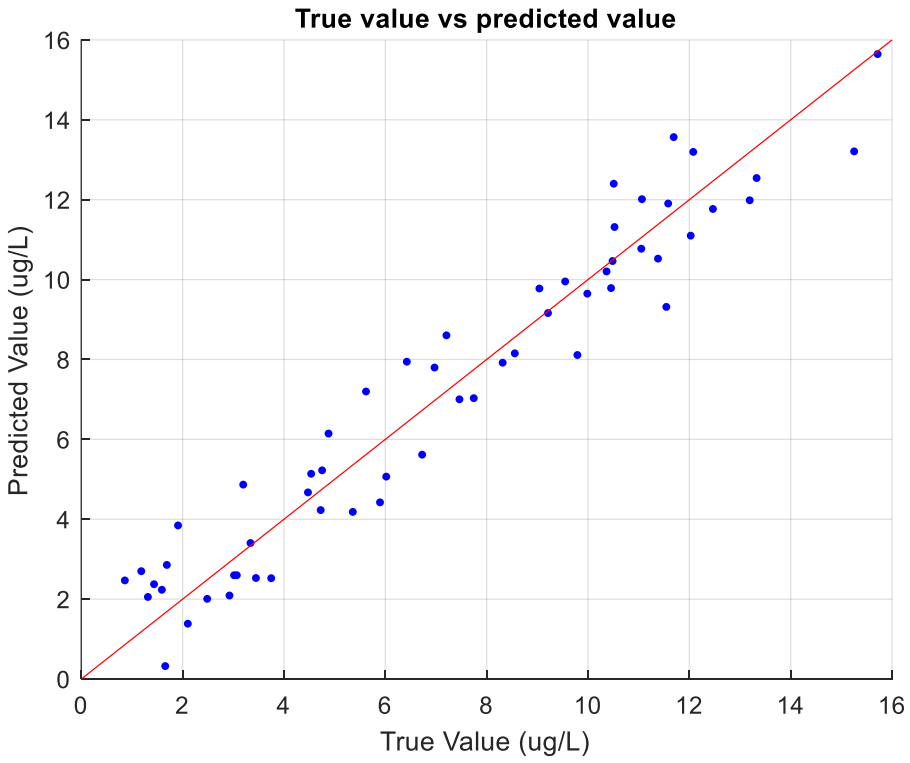


Figure 6-15: True vs predicted value for the proposed model

Table 6-3: Evaluation metrics of the proposed model

Evaluation metric	Reported value
R^2	0.9320
Adjusted R^2	0.9157
MAE	0.904 μgL^{-1}
RMSE	1.05 μgL^{-1}

The residue plot has been shown in Figure 6-16**.** It shows that there is no specific pattern for the residue, rather the error for different chlorophyll concentration is totally random. It signifies that our model has not been overfitted for any specific range. In Figure 6-16 the red line is also not the best fitted curve, rather it indicates the ideal case.

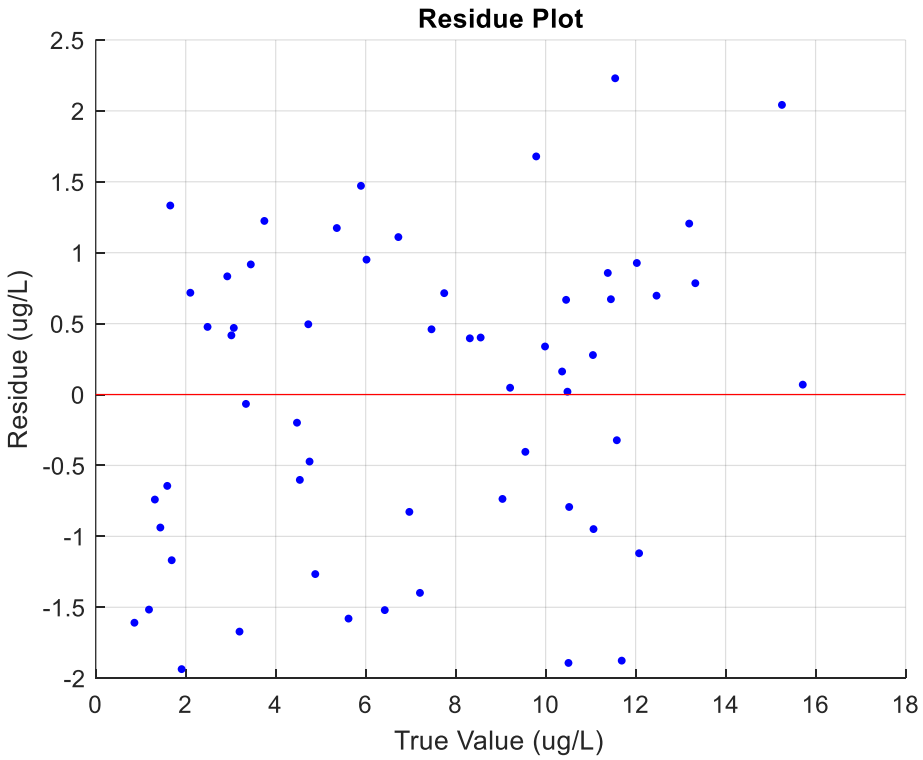


Figure 6-16: Residue plot

6.2.7. Feature importance

Figure 6-17 shows the weight of all 11 features. It is evident that the impedance at 2300Hz, 2600Hz, 3900Hz and 4400Hz plays the most important role for chlorophyll prediction.

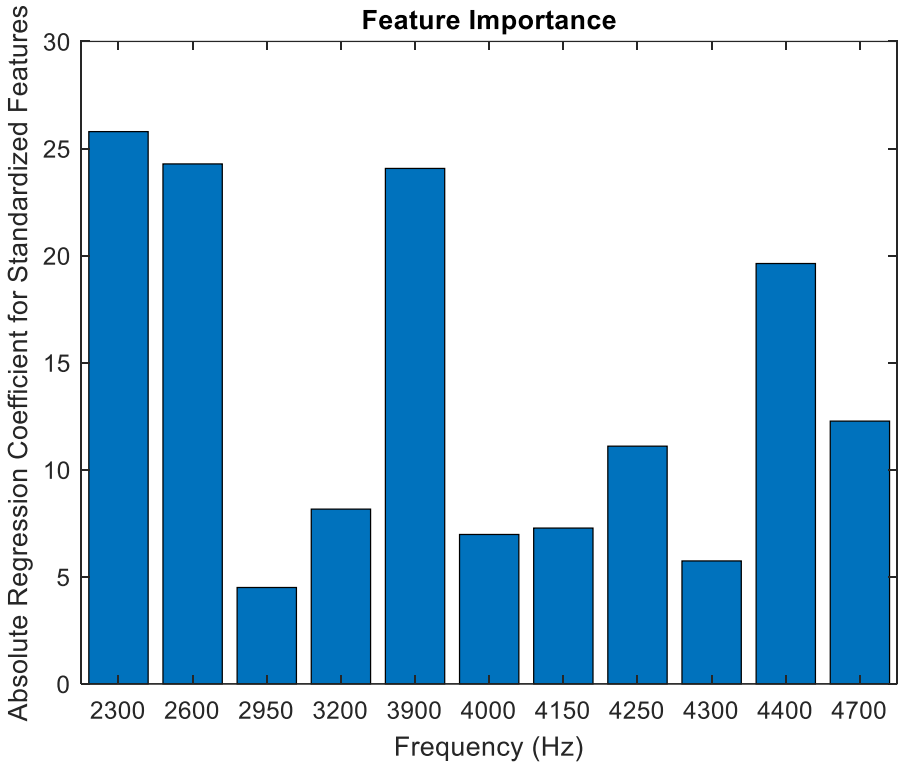


Figure 6-17: Importance of different features in the model

6.2.8. Repeatability

To measure repeatability of our system, impedance reading was taken using our proposed sensors on the same chlorophyll solution for an hour at one-minute interval. Each set of readings were then passed through our proposed model to predict chlorophyll-a concentration. Thus, a total of 60 predictions were recorded for the same sample using our proposed model. The true value of the chlorophyll sample was then measured using traditional laboratory-based absorption method. The mean of the 60 readings, their standard deviation and standard error of the mean have been tabulated in Table 6-4**.**

Table 6-4: Evaluation metric to study the repeatability of the proposed model

No of readings, n	Measured using proposed system			True value - measured using absorption-based method (μgL^{-1})	Error (μgL^{-1})
	Mean (μgL^{-1})	Standard deviation, σ (μgL^{-1})	Standard error of the mean (μgL^{-1})		
60	5.705	0.29	0.037	6.16	0.455

The mean value measured using our proposed model and tabulated in Table *6-4* is the average of 60 predictions of the same sample. Standard deviation of the predictions tells us about the spread and variability of the predictions. The low standard deviation in this experiment indicates that the predictions are closely concentrated to the mean which eventually results in high repeatability of the system. The standard error of the mean is a measure of how precise the mean prediction is. The lower value of it represents the high confidence level of the accurate mean. Error is the deviation of the mean of predictions, predicted by our proposed system from the true value. From Table *6-4* it is evident that the mean prediction value is very close to the true value and the error is within the reported mean error of the system which is $\pm 0.9\ \mu gL^{-1}$.

6.2.9. Sensitivity

The sensitivity of a sensor is defined as the derivative of the output with respect to the stimulus. In other way, it can be defined as the slope of output characteristic graph at a fixed frequency. From Figure 6-10(b), the sensitivity of the proposed sensor at 2300 Hz can be calculated by measuring the slope, which is 175 Ω / (μgL^{-1}). This signifies that for every 1 μgL^{-1} increase or decrease in chlorophyll concentration, the impedance at 2300 Hz increases or decreases by 175 Ω. Therefore, the proposed EIS method exhibits higher sensitivity than traditional methods. However, the sensitivity varies for different frequencies. For example, from Figure 6-10(c), for 4000Hz the sensitivity is 200 Ω / (μgL^{-1}).

6.2.10. Specificity

A sensor system is considered specific if it responds only to one kind of component [67]. To examine the specificity of the proposed sensor and assess the impacts of turbidity and salinity on measurements, 5mg of clay and deionized NaCl were mixed with 50mL of extracted solution. Clay was chosen as an interfering substance because it is widely available and simulates similar worst case scenarios when chlorophyll is extracted from turbid water (such as, erosive river or field). Clay only simulates turbidity and is not soluble in ethanol. On the other hand, NaCl is also not soluble in 100% ethanol. However, as 95% (v/v) ethanol was taken, adding little NaCl may saturate the solution. Readings were taken for the same solution before and after adding clay and NaCl separately, and the spectra is shown in Figure 6-18. It is seen that turbidity does not affect the reading. On the other hand, adding a soluble salt increases the conductivity of the solution. Figure 6-18 shows that after adding NaCl, impedance decreases more at high frequencies than that at low frequencies. This decrease is due to the solubility of NaCl in water. The authors therefore recommend using 100% ethanol for the proposed sensor which might prevent the decrease in reading.

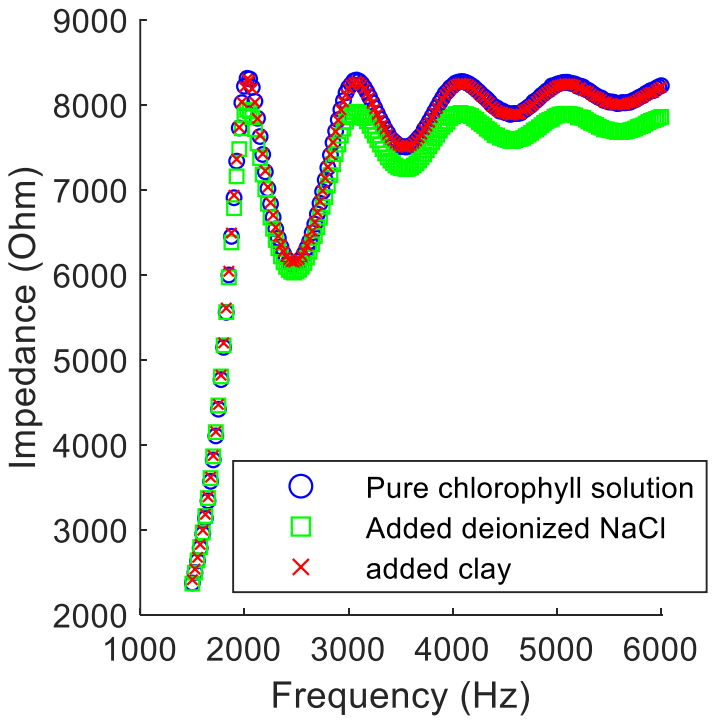


Figure 6-18: Experiment on specificity of the proposed sensor

6.3. Comparison with existing methods

Methods like spectrophotometry, fluorometry and high-performance liquid chromatography (HPLC) have been routinely used in laboratories for decades. In Table 6-5 our proposed EIS method is compared qualitatively with the conventional methods. The table shows that, although HPLC is the most accurate method with the least interference to measure chlorophyll concentration, it is a very expensive and time-consuming process requiring specialized equipment. Our proposed method is a rapid, low-cost alternative to the existing techniques with tolerable error.

Table 6-5: Comparison of the proposed model with current methods

Topic of comparison	***Spectro**** [42][72][73][74]	***Fluorometry*** [42] [73]	***HPLC*** [42][72][74]	***Proposed sensor***
Excitation Source	Light of different wavelength	Light of different wavelength	Separation, followed by fluorometry	Sinusoidal voltage of different frequencies
Measured Response	Optical absorbance	Optical fluorescence	Fluorescence of separated pigment	Electrical impedance
Detector	Photodiode /CMOS[1]	Photodiode /CMOS	Photodiode /CMOS	Electrode
Interference that affects accuracy	Optical interferent (e.g. DOM[2])	Serial dilution & optical interferent (e.g. DOM)	Minimum Interference	Ionic interferent (e.g. DOM)
Accuracy	May overestimate chlorophyll	More accurate than Spectro	Highly accurate	Slightly less than Spectro
Cost	High cost	More Expensive than Spectro	Most Expensive	Low-cost
Operation time	2 – 3 min, needs time to setup for first reading	5 – 6 min, needs time for calibration	15-30 min	1-2 min

*Spectro = Spectrophotometry

[1]CMOS = Complementary Metal Oxide Semiconductor

[2]DOM = Dissolved Organic Material

6.4. Cost of the parts

The cost of the prototypes was one of the major concerns in this project. The details of the cost are shown in Table 6-6. As can be seen, the total cost of parts was around $197 (CAD).

Table 6-6: Cost Approximation of a Single Prototype

Item	Quantity	Per Unit Cost (CAD)	Total Cost (CAD)
Raspberry Pi	1	$39.95	$39.95
Display	1	$15.05	$15.05
Battery with case	1	$11.03	$11.03
Buck Module	1	$16.99	$16.99
Electrode	1	$24.35	$24.35
Eval-AD5933	1	$89.37	$89.37
Total Cost			$197.01

Chapter 7: Conclusion and Future Recommendation

7.1. Conclusion

In this thesis, two different spectral approaches have been proposed to measure extracted chlorophyll concentration. The first approach measures optical reflectance primarily at 12 wavelengths and predicts the concentration of extracted chlorophyll by analyzing reflectance at 5 major wavelengths. The final approach measures electric impedance spectroscopy of the extracted chlorophyll solution from 1500 Hz to 7500 Hz and predicts the concentration of extracted chlorophyll by analyzing impedance at 11 frequencies. Both the approaches have the following features -

- Low-cost – The cost for sensors of both procedures is less than CAD200
- Light weight – both the sensors weighs less than 500g
- Portability – both the sensors are small and can be carried in hand
- Reduced features – the optical spectral sensor takes reflectance at 5 wavelengths and the impedance spectral sensor takes impedance at 11 frequencies into consideration while predicting the concentration of extracted chlorophyll
- Prediction of extracted chlorophyll – Both the sensors can predict continuous chlorophyll concentration with $R^2 > 90\%$, and RMSE $< 2.5\ \mu gL^{-1}$

Overall, the result concludes that the proposed system can estimate extracted chlorophyll concentration with reasonable accuracy which is comparable to existing devices. Although optical spectroscopy performed better than the electrical impedance method to estimate extracted chlorophyll concentration and the range of prediction for electrical impedance spectroscopy is smaller than that of optical spectroscopy, electrical impedance spectroscopy has its own benefits. It is very precise for lower concentration of chlorophyll-a.

7.2. Future recommendation

There are very good potential and considerable scopes and opportunities to explore new possibilities and improve the proposed system to the next level. Some of the recommendations for further works are -

- The proposed optical spectroscopy system only measures reflectance at 12 discrete wavelengths. In future industrial sensors with more channels can be used. In this respect, 18 channel sensor AS7265 (AMS)[75] might be used.
- The device used for measuring electrical impedance spectroscopy (AD5933 evaluation board) can measure only in the kilo Hertz range without any external component. In future frequency divider circuit can be used to measure impedance spectra at the lower frequency range
- In the experiments, around 52 and 60 samples were used. In future more formal study can be conducted using more variations in the chlorophyll concentration.
- The behavior of electrical impedance of chlorophyll at different frequencies can be studied
- Electrical impedance spectroscopy might be used to predict the concentration of Blue Green Algae and pH of water in-situ.

The exploration of the above-mentioned scopes may create new opportunities in the water quality monitoring field

References

[1] K. . Thomson, “World agriculture: towards 2015/2030: an FAO perspective,” *Land Use Policy*, vol. 20, no. 4, p. 375, Oct. 2003.

[2] B. H. Patel, “Natural dyes,” in *Handbook of Textile and Industrial Dyeing*, Elsevier, 2011, pp. 395–424.

[3] S. Jeffrey, “Purification and properties of chlorophyll c from Sargassum flavicans,” *Biochemical Journal*, vol. 86, no. 2, pp. 313–318, Feb. 1963.

[4] J. Gregor and B. Maršálek, “Freshwater phytoplankton quantification by chlorophyll,” *Water Research*, vol. 38 (3), pp. 517–522, 2004.

[5] H. Caspers, “OECD: Eutrophication of Waters. Monitoring, Assessment and Control,” *Internationale Revue der gesamten Hydrobiologie und Hydrographie*, vol. 69, no. 2, pp. 200–200, 1984.

[6] J. Domagalski C. Lin, Y. Luo, J. Kang, S. Wang, L. Brown, and M. Munn, “Eutrophication study at the Panjiakou-Daheiting Reservoir system, northern Hebei Province, People’s Republic of China: Chlorophyll-a model and sources of phosphorus and nitrogen,” *Agricultural Water Management*, vol. 94, no. 1–3, pp. 43–53, Dec. 2007.

[7] J. G. Winter, A. DeSellas, R. Fletcher, L. Heintsch, A. Morley, L. Nakamoto, and K. Utsumi, “Algal blooms in Ontario, Canada: Increases in reports since 1994,” *Lake and Reservoir Management*, vol. 27, no. 2, pp. 107–114, Jun. 2011.

[8] S. Itakura and I. Imai, *Economic impacts of harmful algal blooms on fisheries and aquaculture in western Japan - An overview of internanual variability and interspecies comparison*, no. 47. 2014.

[9] F. R. Pick, “Blooming algae: a Canadian perspective on the rise of toxic cyanobacteria,” *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 73, no. 7, pp. 1149–1158, Jul. 2016.

[10] F. Babani and H. K. Lichtenthaler, “Light-induced and Age-dependent Development ofChloroplasts in Etiolated Barley Leaves as Visualized by Determination of Photosynthetic

Pigments, C02 Assimilation Rates and Different Kinds of Chlorophyll Fluorescence Ratios," *Journal of Plant Physiology*, vol. 148, no. 5, pp. 555–566, Jan. 1996.

[11] D. Pavlovic, B. Nikolic, S. Djurovic, H. Waisi, A. Andjelkovic, and D. Marisavljevic, "Chlorophyll as a measure of plant health: Agroecological aspects," *Pesticidi i fitomedicina*, vol. 29, no. 1, pp. 21–34, 2014.

[12] D. B. Milivojevic and B. . Nikolic, "Effects of diquat on pigment-protein complexes of thylakoid membranes in soybean and maize plants," *Biologia plantarum*, vol. 41, no. 4, pp. 597–600, May 1998.

[13] K. Solymosi and B. Mysliwa-Kurdziel, "Chlorophylls and their Derivatives Used in Food Industry and Medicine," *Mini-Reviews in Medicinal Chemistry*, vol. 17, no. 13, Jul. 2017.

[14] S. Aronoff, "The Chemistry of Chlorophyll (With Special Reference to Foods)," in *Advances in food research*, vol. 4, 1953, pp. 133–184.

[15] A. Hosikian, S. Lim, R. Halim, and M. K. Danquah, "Chlorophyll Extraction from Microalgae: A Review on the Process Engineering Aspects," *International Journal of Chemical Engineering*, vol. 2010, pp. 1–11, 2010.

[16] R. Syafinar, N. Gomesh, M. Irwanto, M. Fareq, and Y. M. Irwan, "Chlorophyll Pigments as Nature Based Dye for Dye-Sensitized Solar Cell (DSSC)," *Energy Procedia*, vol. 79, pp. 896–902, Nov. 2015.

[17] R. M. M. Roijackers, "A comparison between two methods of extracting chlorophyll-a from different phytoplankton samples," *Hydrobiological Bulletin*, vol. 15, no. 3, pp. 179–183, Dec. 1981.

[18] S. Lan, L. Wu, D. Zhang, C. Hu, and Y. Liu, "Ethanol outperforms multiple solvents in the extraction of chlorophyll-a from biological soil crusts," *Soil Biology and Biochemistry*, vol. 43, no. 4, pp. 857–861, Apr. 2011.

[19] D. F. Spencer and G. G. Ksander, "Comparison of Three Methods for Extracting Chlorophyll from Aquatic Macrophytes," *Journal of Freshwater Ecology*, vol. 4, no. 2, pp. 201–208, Dec. 1987.

[20] J. B. Conant, B. F. Chow, and E. M. Dietz, “Studies in the Chlorophyll Series. XIV. Potentiometric Titration in Acetic Acid Solution of the Basic Groups in Chlorophyll Derivatives 1,” *Journal of the American Chemical Society*, vol. 56, no. 10, pp. 2185–2189, Oct. 1934.

[21] L. P. Vernon, “Spectrophotometric Determination of Chlorophylls and Pheophytins in Plant Extracts,” *Analytical Chemistry*, vol. 32, no. 9, pp. 1144–1150, Aug. 1960.

[22] L. P. Vernon, “Spectrophotometry Determination of Chlorophylls and Pheophytins in Plant Extracts,” *Analytical Chemistry*, vol. 32, no. 9, pp. 1144–1150, 1960.

[23] R. J. Ritchie, “Consistent Sets of Spectrophotometric Chlorophyll Equations for Acetone, Methanol and Ethanol Solvents,” *Photosynthesis Research*, vol. 89, no. 1, pp. 27–41, Jul. 2006.

[24] R. I. Chowdhury, K. A. Wahid, K. Nugent, and H. Baulch, “Design and Development of Low-Cost, Portable, and Smart Chlorophyll-a Sensor,” *IEEE Sensors Journal*, vol. 20, no. 13, pp. 7362–7371, Jul. 2020.

[25] O. Holm-Hansen, C. J. Lorenzen, R. W. Holmes, and J. D. H. Strickland, “Fluorometric Determination of Chlorophyll,” *ICES Journal of Marine Science*, vol. 30, no. 1, pp. 3–15, Dec. 1965.

[26] A. M. . Pinto, E. von Sperling, and R. . Moreira, “Chlorophyll-a determination via continuous measurement of plankton fluorescence:,” *Water Research*, vol. 35, no. 16, pp. 3977–3981, Nov. 2001.

[27] U. Schreiber, J. F. Müller, A. Haugg, and R. Gademann, “New type of dual-channel PAM chlorophyll fluorometer for highly sensitive water toxicity biotests,” *Photosynthesis Research*, vol. 74, no. 3, pp. 317–330, 2002.

[28] G. H. Krause and E. Weis, “Chlorophyll Fluorescence and Photosynthesis: The Basics,” *Annual Review of Plant Physiology and Plant Molecular Biology*, vol. 42, no. 1, pp. 313–349, Jun. 1991.

[29] K. Maxwell and G. N. Johnson, “Chlorophyll fluorescence—a practical guide,” *Journal of Experimental Botany*, vol. 51, no. 345, pp. 659–668, Apr. 2000.

[30] H. M. Baulch, M. A. Turner, D. L. Findlay, R. D. Vinebrooke, and W. F. Donahue, "Benthic algal biomass — measurement and errors," *Canadian Journal of Fisheries and Aquatic Sciences*, vol. 66, no. 11, pp. 1989–2001, 2009.

[31] T. Kutser, L. Metsamaa, N. Strömbeck, and E. Vahtmäe, "Monitoring cyanobacterial blooms by satellite remote sensing," *Estuarine, Coastal and Shelf Science*, vol. 67, no. 1–2, pp. 303–312, Mar. 2006.

[32] S. Shang, Z. Lee, G. Lin, C. Hu, L. Shi, Y. Zhang, X. Li, and J. Wu, "Sensing an intense phytoplankton bloom in the western Taiwan Strait from radiometric measurements on a UAV," *Remote Sensing of Environment*, vol. 198, pp. 85–94, Sep. 2017.

[33] C. Kislik, I. Dronova, and M. Kelly, "UAVs in Support of Algal Bloom Research: A Review of Current Applications and Future Opportunities," *Drones*, vol. 2, no. 4, p. 35, Oct. 2018.

[34] H. K. Lichtenthaler, "Chlorophyll Fluorescence Signatures of Leaves during the Autumnal Chlorophyll Breakdown," *Journal of Plant Physiology*, vol. 131, no. 1–2, pp. 101–110, Nov. 1987.

[35] I. Moya, "A new instrument for passive remote sensing1. Measurements of sunlight-induced chlorophyll fluorescence," *Remote Sensing of Environment*, vol. 91, no. 2, pp. 186–197, May 2004.

[36] L. Zeng and D. Li, "Development of In Situ Sensors for Chlorophyll Concentration Measurement," *Journal of Sensors*, vol. 2015, pp. 1–16, 2015.

[37] C. W. Proctor and C. S. Roesler, "New insights on obtaining phytoplankton concentration and composition from in situ multispectral Chlorophyll fluorescence," *Limnology and Oceanography: Methods*, vol. 8, pp. 695–708, 2010.

[38] A. Aminot and F. Rey, *Standard procedure for the determination of chlorophyll a by spectroscopic methods*, no. March. 2000.

[39] D. Pavlovic, B. Nikolic, S. Djurovic, H. Waisi, A. Andjelkovic, and D. Marisavljevic, "Chlorophyll as a measure of plant health: Agroecological aspects," *Pesticidi i fitomedicina*, vol. 29, no. 1, pp. 21–34, 2014.

[40] ISO 10260:1992, *Water quality - measurement of biochemical parameters - spectrometric determination of the chlorophyll - A concentraion*. Geneva, Switzerland, 1992.

[41] D. J. Suggett, O. Prášil, and M. A. Borowitzka, Eds., *Chlorophyll a Fluorescence in Aquatic Sciences: Methods and Applications*. Dordrecht: Springer Netherlands, 2010.

[42] J. Pinckney, R. Papa, and R. Zingmark, "Comparison of high-performance liquid chromatographic, spectrophotometric, and fluorometric methods for determining chlorophyll a concentrations in estaurine sediments," *Journal of Microbiological Methods*, vol. 19, no. 1, pp. 59–66, Jan. 1994.

[43] F. Attivissimo, C. G. C. Carducci, A. M. L. Lanzolla, A. Massaro, and M. R. Vadrucci, "A Portable Optical Sensor for Sea Quality Monitoring," *IEEE Sensors Journal*, vol. 15, no. 1, pp. 146–153, Jan. 2015.

[44] L. Pasolli, F. Melgani, and E. Blanzieri, "Gaussian Process Regression for Estimating Chlorophyll Concentration in Subsurface Waters From Remote Sensing Data," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, no. 3, pp. 464–468, Jul. 2010.

[45] Y. Bazi, N. Alajlan, and F. Melgani, "Improved Estimation of Water Chlorophyll Concentration With Semisupervised Gaussian Process Regression," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 7, pp. 2733–2743, Jul. 2012.

[46] A. Hossain, J. Canning, S. Ast, P. J. Rutledge, Teh Li Yen, and A. Jamalipour, "Lab-in-a-Phone: Smartphone-Based Portable Fluorometer for pH Measurements of Environmental Water," *IEEE Sensors Journal*, vol. 15, no. 9, pp. 5095–5102, Sep. 2015.

[47] F. Vesali, M. Omid, H. Mobli, and A. Kaleita, "Feasibility of using smart phones to estimate chlorophyll content in corn plants," *Photosynthetica*, vol. 55, no. 4, pp. 603–610, Dec. 2017.

[48] S. Thomas, R. Thomas, A. K. Zachariah, and R. K. Mishra, *Spectroscopic Methods for Nanomaterials Characterization*, vol. 2. Matthew Deans, 2017.

[49] M. Grossi and B. Riccò, "Electrical impedance spectroscopy (EIS) for biological analysis and food characterization: A review," *Journal of Sensors and Sensor Systems*, vol. 6, no. 2, pp. 303–325, 2017.

[50] R. Basak, K. Wahid, and A. Dinh, “Determination of Leaf Nitrogen Concentrations Using Electrical Impedance Spectroscopy in Multiple Crops,” *Remote Sensing*, vol. 12, no. 3, p. 566, Feb. 2020.

[51] M. Islam, K. A. Wahid, A. Dinh, and P. Bhowmik, “Model of dehydration and assessment of moisture content on onion using EIS,” *Journal of Food Science and Technology*, vol. 56, no. 6, pp. 2814–2824, Jun. 2019.

[52] M. Islam, K. Wahid, and A. Dinh, “Assessment of Ripening Degree of Avocado by Electrical Impedance Spectroscopy and Support Vector Machine,” *Journal of Food Quality*, vol. 2018, pp. 1–9, Nov. 2018.

[53] A. Chowdhury, T. K. Bera, D. Ghoshal, and B. Chakraborty, “Studying the electrical impedance variations in banana ripening using electrical impedance spectroscopy (EIS),” in *Proceedings of the 2015 Third International Conference on Computer, Communication, Control and Information Technology (C3IT)*, 2015, pp. 1–4.

[54] M. Q. Li, J. Y. Li, X. H. Wei, and W. J. Zhu, “Early diagnosis and monitoring of nitrogen nutrition stress in tomato leaves using electrical impedance spectroscopy,” *International Journal of Agricultural and Biological Engineering*, vol. 10, no. 3, pp. 194–205, 2017.

[55] R. Muñoz-Huerta, A. Ortiz-Melendez, R. Guevara-Gozalez, I. Torres-Pacheco, G. Herrera-Ruiz, L. Conteras-Medina, J. Prado-Olivarez, and R. Ocampo-Velazquez, “An Analysis of Electrical Impedance Measurements Applied for Plant N Status Estimation in Lettuce (Lactuca sativa),” *Sensors*, vol. 14, no. 7, pp. 11492–11503, Jun. 2014.

[56] J. Juansah, I. W. Budiastra, and K. Dahlan, “The Prospect of Electrical Impedance Spectroscopy as Non-Destructive Evaluation of Citrus Fruit Acidity,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 2, no. 11, pp. 58–64, 2012.

[57] Y. Mizukami, K. Yamada, Y. Sawai, and Y. Yamaguchi, “Measurement of fresh tea leaf growth using electrical impedance spectroscopy,” *Agricultural Journal*, vol. 2, no. 1, pp. 134–139, 2007.

[58] T. P. Vello, R. F. de Oliveira, G. O. Silva, D. H. S. de Camargo, and C. C. B. Bufon, “A simple capacitive method to evaluate ethanol fuel samples,” *Scientific Reports*, vol. 7, no.

1, p. 43432, Apr. 2017.

[59] J. F. G. M. Wintermans and A. De Mots, “Spectrophotometric characteristics of chlorophylls a and b and their phenophytins in ethanol,” *Biochimica et Biophysica Acta (BBA) - Biophysics including Photosynthesis*, vol. 109, no. 2, pp. 448–453, Nov. 1965.

[60] “Principal component analysis.” [Online]. Available:

https://en.wikipedia.org/wiki/Principal_component_analysis. [Accessed: 16-Jul-2020].

[61] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-Validation,” in *Encyclopedia of Database Systems*, L. LIU and M. T. ÖZSU, Eds. Boston, MA: Springer US, 2009, pp. 532–538.

[62] A. Botchkarev, “A New Typology Design of Performance Metrics to Measure Errors in Machine Learning Regression Algorithms,” *Interdisciplinary Journal of Information, Knowledge, and Management*, vol. 14, pp. 045–076, Sep. 2019.

[63] D. S. Bhargava and D. W. Mariam, “Spectral reflectance relationships to turbidity generated by different clay materials,” *Photogrammetric Engineering & Remote Sensing*, vol. 56, no. 2, pp. 225–229, 1990.

[64] A. Mishra and A. Dixit, “Embedded Image Capturing & Digital Converting Process using Raspberry pi System interfacing and Comparision of Generation 2 verses Generation 1 models in Raspberry pi,” *International Journal of Computer Science and Information Technologies*, vol. 6, no. 2, pp. 1789–1801, 2015.

[65] A. Imteaj, T. Rahman, M. K. Hossain, M. S. Alam, and S. A. Rahat, “An IoT based fire alarming and authentication system for workhouse using Raspberry Pi 3,” in *2017 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 2017, pp. 899–904.

[66] A. Campion and P. Kambhampati, “Surface-enhanced Raman scattering,” *Chemical Society Reviews*, vol. 27, no. 4, p. 241, 1998.

[67] L. Zipser, “Selectivity of sensor systems,” *Sensors and Actuators A: Physical*, vol. 37–38, pp. 286–289, Jun. 1993.

[68] H. K. Lichtenthaler, “Chlorophyll Fluorescence Signatures of Leaves during the Autumnal

Chlorophyll Breakdown," *Journal of Plant Physiology*, vol. 131, no. 1–2, pp. 101–110, Nov. 1987.

[69] F. Baret, G. Guyot, and D. Major, "Coupled Fluorescence and Reflectance Measurements to Improve Crop Productivity Evaluation," in *Applications of Chlorophyll Fluorescence in Photosynthesis Research, Stress Physiology, Hydrobiology and Remote Sensing*, H. K. Lichtenthaler, Ed. Dordrecht: Springer Netherlands, 1988, pp. 319–324.

[70] J. Kissinger and D. Wilson, "Portable Fluorescence Lifetime Detection for Chlorophyll Analysis in Marine Environments," *IEEE Sensors Journal*, vol. 11, no. 2, pp. 288–295, Feb. 2011.

[71] P. Johnson, "Progress in chemometrics research," *Journal of Chemometrics*, vol. 19, no. 4, pp. 266–267, Apr. 2005.

[72] Sabina, "The Comparison of Spectrophotometric Method and High-Performance Liquid Chromatography in Photosynthetic Pigments Analysis," *OnLine Journal of Biological Sciences*, vol. 11, no. 2, pp. 63–69, Feb. 2011.

[73] A. Dos Santos *et al.*, "Comparison of three methods for Chlorophyll determination : Spectrophotometry and Fluorimetry in samples containing pigment mixtures and spectrophotometry in samples with separate pigments through High Performance Liquid Chromatography," *Acta Limnologica Brasiliensia*, vol. 15, pp. 7–18, 2003.

[74] J. . W. Louda and P. Monghkonsri, "Comparison of Spectrophotometric and HPLC Estimations of Chlorophylls-a , - b , -c and Pheopigments in Florida Bay Seston," *Florida Scientist*, vol. 67, no. 4, pp. 281–292, 2004.

[75] "SparkFun Triad Spectroscopy Sensor - AS7265x (Qwiic) - SEN-15050 - SparkFun Electronics." [Online]. Available: https://www.sparkfun.com/products/15050. [Accessed: 17-Jul-2020].