

**DETECTION OF**  
**CHLOROPHYLL A**  
**IN WATER USING**  
**OPTICAL SENSORS**

# Detection of Chlorophyll A In Water Using Optical Sensors

A Dissertation for

Course code and Course Title: ELE-625 Project

Credits: 16

Submitted in partial fulfilment of Master's Degree

M.Sc. in Electronics

by

**MR. ATUL F GAUNKAR**

Seat No.: 22P0360002

ABC ID: 578414177388

PRN: 201905587

Under the Supervision of

**DR. MARLON SEQUIERA**

School of Physical and Applied Sciences  
Electronics



**GOA UNIVERSITY**  
**MAY 2024**

Examined by:



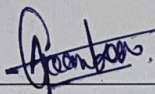
Seal of the School



### DECLARATION BY STUDENT

I hereby declare that the data presented in this Dissertation report entitled, **Detection of Chlorophyll A In Water Using Optical Sensors** is based on the results of investigations carried out by me in the M.Sc. Electronics at the School of Physical and Applied Sciences, Goa University under the Supervision of Dr/Prof. Marlon Sequiera and the same has not been submitted elsewhere for the award of a degree or diploma by me. Further, I understand that Goa University or its authorities will be not responsible for the correctness of observations / experimental or other findings given the dissertation.

I hereby authorize the University authorities to upload this dissertation on the dissertation repository or anywhere else as the UGC regulations demand and make it available to any one as needed.



---

Mr. Atul F Gaunkar

Seat no: 22P0360002

Electronics Discipline

School of Physical & Applied Sciences

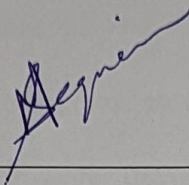
Date: 14/05/2024

Place: Goa University



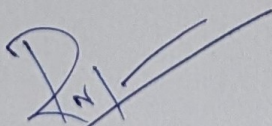
## COMPLETION CERTIFICATE

This is to certify that the dissertation report **Detection of Chlorophyll A In Water Using Optical Sensors** is a bonafide work carried out by Mr. Atul F Gaunkar under my supervision in partial fulfilment of the requirements for the award of the degree of Masters in the Discipline Electronics at the School of Physical and Applied Sciences, Goa University.



Dr. Marlon Sequiera  
Electronics Discipline  
School of Physical & Applied Sciences

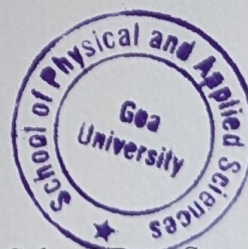
Date: 13/05/2024



Signature of Dean of the School/HoD of Dept

Date:

Place: Goa University



School/Dept Stamp

## **PREFACE**

This dissertation represents the culmination of several months of research hardwork, and it is with great pleasure that I present it to the academic community.

Throughout this journey, I have been fortunate to receive support and guidance from many individuals and institutions, to whom I owe immense gratitude.

The dissertation is organized as follows: Chapter 1 provides an introduction to the topic, including its significance and relevance. Chapter 2 reviews the existing literature on Detection of chlorophyll A in water using various methods, synthesizing key findings and identifying gaps in the literature. Chapter 3 outlines the methodology employed in this study, detailing the research design, data collection methods, and analytical approach. Finally, Chapter 4 presents the empirical findings and analysis.

## **ACKNOWLEDGEMENT**

Many hours of hard work and sincere efforts have gone into the success of this project. All this would have not been possible without the timely assistance and guidance of some very important people. First and foremost, I would like to thank God for giving me the strength, courage and determination to complete our project.

I express my deep gratitude towards all the teaching faculty Prof. Rajendra S. Gad, Dr. Jivan S. Parab, Dr. Marlon Sequeira, Dr. Narayan Vetrekar, Dr. Aniketh Gaonkar and Dr. Sandip Gawali and other non-teaching staff for their support and guidance throughout the completion of the project. Their knowledge and suggestions have been crucial in helping me plan, carry out and execute the project.

I would like to express my gratitude to my family and close friends for their unfailing support and inspiration and motivation during the project.

I owe the success of our entire project to all people mentioned above.

## Table Of Content

Sr No.	Title	Page no
<b>Chapter 1</b>	<b>INTRODUCTION.....</b>	<b>01</b>
1.1	Chlorophyll A.....	02
1.2	Background.....	12
1.3	Aim and Objectives.....	13
1.4	Hypotheses.....	14
1.5	Scope.....	15
<b>Chapter 2</b>	<b>LITERATURE SURVEY.....</b>	<b>16</b>
2.1	Literature reviews.....	17
2.2	Tabular reviews.....	30
2.3	Literature conclusion.....	38
<b>Chapter 3</b>	<b>METHODOLOGY.....</b>	<b>39</b>
3.1	Samples.....	40
3.2	Data Acquisition.....	45
3.3	Pre-Processing.....	51
3.4	Data Partitioning.....	54
3.5	Machine Learning.....	56
<b>Chapter 4</b>	<b>ANALYSIS AND CONCLUSION.....</b>	<b>67</b>
4.1	Analysis and Results.....	68
4.2	Conclusion.....	72
4.3	Future work.....	73

<b>Chapter 5</b>	<b>REFERENCES.....</b>	<b>74</b>
5.1	Paper References.....	75
5.2	References link.....	80
5.3	Appendix.....	81



**Title of figures**

<b>Fig.no</b>	<b>Title</b>	<b>page no</b>
1.	Algae culture water.....	02
2.	Wavelength graph of Chlorophyll A.....	04
3.	Diagram of Secchi stick depth.....	08
4.	Different Types of Algae Samples.....	40
5.	Sample in Cuvette.....	40
6.	Secchi stick Measurement.....	45
7.	Setup of Spectrophotometer.....	46
8.	Block Diagram of Spectrophotometer.....	47
9.	Interaction of Light with the Sample.....	48
10.	Smoothened Spectra.....	52
11.	Support Vector Regression.....	57
12.	Random Forest Regression.....	59
13.	KNN Regression.....	62
14.	Decision tree Regression.....	65
15.	Graph of Support Vector Regression.....	69
16.	Graph of Partial Square Least Regression.....	70
17.	Graph of Decision Tree Regression.....	70
18.	Graph of Random Forest Regression.....	71
19.	Graph of KNN Regression.....	71

**Title of tables**

<b>Table no.</b>	<b>Title</b>	<b>page no</b>
1.	Trophic state levels.....	11
2.	Younger algae dilution.....	41
3.	Moderate algae dilution.....	42
4.	Older algae dilution.....	43
5.	Regressors and their parameters.....	66
6.	Results of all regressors.....	69

**Abbreviations used**

Chl-A	Chlorophyll A
UV-VIS	Ultraviolet-visible
SDD	Secchi disk depth
SVR	Support Vector Regression
KNR	K Nearest Regression
PLSR	Partial Least Square Regression
PCA	Principal Component Analysis
SG	Savitsky Golay
RFR	Random Forest Regression
DTR	Decision Tree Regression
ML	Machine Learning

**“DETECTION OF CHLOROPHYLL A IN WATER USING OPTICAL  
SENSORS”**

**Atul F Gaunkar**

**Dept. of Electronics**

**Goa University**

**Abstract:**

This work has developed a method for detection of chlorophyll A in water using UV-VIS spectroscopy with the side of machine learning algorithms. A comparative study is done on the efficacy of the different machine learning model to detect the concentration of Chl-A. Chlorophyll a is essential for the survival of plants and algae, as it is directly involved in the process that allows them to produce the energy they need to grow and reproduce. It is also an important indicator of the trophic state of water bodies, as its concentration is directly related to the amount of algae present. Chlorophyll a absorbs light most effectively in the blue and red regions of the electromagnetic spectrum, with peak absorption occurring in the blue (around 430-450 nm) and red (around 660-680 nm) wavelengths. It is necessary to detect Chlorophyll a in water because it serves as a reliable indicator of the presence and abundance of algae and other photosynthetic organisms. Machine learning model for prediction of Chl-A was build and the achieved RMSE are 0.25, 0.57, 0.44, 0.64 and 0.27 for SVR, PLSR, DTR, RFR and KNR respectively.



# CHAPTER-1

## 1. INTRODUCTION

### 1.1: CHLOROPHYLL A

The green pigment known as chlorophyll A [Chl-A] is present in cyanobacteria, algae, and plants. It is essential to the mechanism by which these organisms turn light energy into chemical energy—a process known as photosynthesis. Chlorophyll A is a useful biomarker of phytoplankton biomass in aquatic environments, giving important details regarding primary productivity and water quality.

The amounts of phytoplankton, the tiny algae that make up the base of the aquatic food chain, is indicated by the quantity of chlorophyll A in the water. Algal blooms caused by nutrient-rich environments, such as sewage pollution or agricultural runoff, can produce high amounts of chlorophyll A. Excessive algal blooms can disrupt aquatic life, lower oxygen levels in the water, and produce toxic circumstances that are harmful to people and other creatures, even though certain blooms are normal and necessary for the health of the ecosystem.

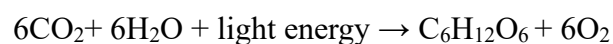
Observing chlorophyll levels for evaluating the health of ecosystems and water quality, A levels in water bodies are essential. Scientists can detect nutrient contamination, monitor changes in primary production, and put protective measures against dangerous algal blooms into place by detecting the quantities of chlorophyll A. Furthermore, information on chlorophyll A assists in making management choices that preserve aquatic ecosystem balance and safeguard water resources.



Fig 1: Algae Culture Water

Photosynthesis is the fundamental process through which autotrophic organisms, such as plants, algae, and cyanobacteria, harness light energy to convert carbon dioxide and water into organic compounds, mainly sugars, releasing oxygen as a byproduct. Chlorophyll-a is the primary pigment responsible for capturing light energy during the light-dependent reactions of photosynthesis. Located within the thylakoid membranes of chloroplasts, chlorophyll-a absorbs light most efficiently in the blue and red regions of the electromagnetic spectrum, with minimal absorption in the green region, giving plants their green colour.

The process of photosynthesis begins when chlorophyll-a molecules absorb photons of light energy. This energy is then used to drive a series of chemical reactions, leading to the conversion of carbon dioxide and water into glucose and oxygen. The formula for photosynthesis can be summarized as:



Chlorophyll-a plays a central role in this process by transferring the absorbed light energy to reaction centers, where it is used to drive the synthesis of ATP (adenosine triphosphate) and NADPH (nicotinamide adenine dinucleotide phosphate), which are energy-rich molecules used to power the subsequent dark reactions of photosynthesis.

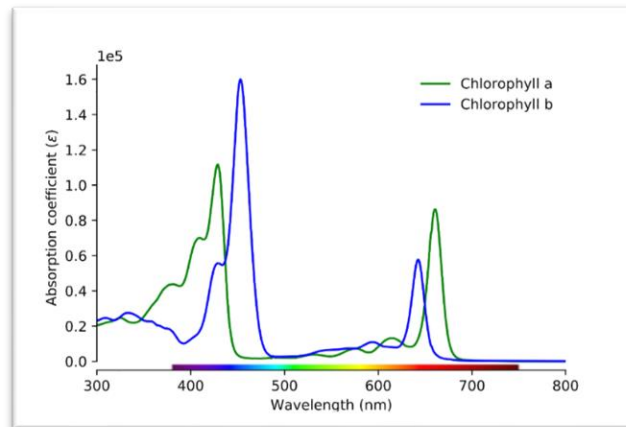


Fig 2: Wavelength graph of chlorophyll A

### Importance of Chlorophyll-a Measurement

Chlorophyll-a serves as a proxy for primary productivity, which is the rate at which autotrophic organisms convert inorganic carbon ( $\text{CO}_2$ ) into organic matter through photosynthesis. Therefore, monitoring chlorophyll-a concentrations provides valuable insights into the overall productivity and health of aquatic ecosystems.

### Chlorophyll-a in Aquatic Environments

In aquatic environments, chlorophyll-a is primarily found in phytoplankton, microscopic algae that inhabit the upper layers of water bodies. Phytoplankton are the base of the aquatic food chain, and their abundance and distribution influence the entire aquatic ecosystem. Therefore, changes in chlorophyll-a concentrations can have significant impacts on the structure and function of aquatic ecosystems.

### Factors Influencing Chlorophyll-a Concentrations

Chlorophyll-a concentrations in aquatic environments are influenced by various factors, including nutrient availability, light intensity, temperature, and water movement. Nutrients, particularly nitrogen and phosphorus, are essential for phytoplankton growth and chlorophyll-a production. Excess nutrient inputs, often resulting from human activities such as agriculture and urbanization, can lead to eutrophication, characterized by an overgrowth of phytoplankton and increased chlorophyll-a concentrations. In



addition to nutrients, light availability is a critical factor influencing chlorophyll-a concentrations. Light is essential for photosynthesis, and changes in light intensity can affect phytoplankton growth rates and chlorophyll-a production.

### **Monitoring Chlorophyll-a Concentrations**

The measurement and monitoring of chlorophyll-a concentrations in aquatic environments are essential for assessing water quality, detecting eutrophication, and understanding ecosystem dynamics. Various methods are used to quantify chlorophyll-a, including spectrophotometry, fluorometry, and remote sensing.

#### **I. Spectrophotometer method:**

The spectrophotometric method is one of the most commonly used techniques for the determination of chlorophyll-a concentration in water bodies. It involves the extraction of chlorophyll-a from water samples followed by spectrophotometric analysis of the chlorophyll extract.

Here is a detailed explanation of the spectrophotometric method:

- **Sample Collection:**

Water samples are collected from the study site using a sampling bottle. The samples should be collected from the desired depth to ensure representative results.

- **Filtration:**

The water samples are filtered to collect the chlorophyll-containing particles. Typically, a filter with a pore size of 0.7 to 1.0  $\mu\text{m}$  is used for this purpose. The filtrate is discarded, and the filter paper with the chlorophyll-containing particles is retained for further analysis.

- **Extraction:**

The chlorophyll is extracted from the filter paper using a solvent such as acetone, methanol, or ethanol. The extraction is typically performed in a dark room to minimize chlorophyll degradation due to light exposure. The solvent is added to the filter paper, and the chlorophyll is allowed to extract for a specified period (usually 24 hours) with occasional shaking to facilitate extraction.

- Spectrophotometric Analysis:

After extraction, the chlorophyll extract is analyzed using a spectrophotometer. The absorbance of the chlorophyll extract is measured at two specific wavelengths: around 665 nm (absorption peak of chlorophyll-a) and 750 nm (reference wavelength). The choice of wavelengths may vary depending on the spectrophotometer used.

- Calculation of Chlorophyll-a Concentration:

Chlorophyll-a concentration is calculated using the following formula:

$$\text{chl} - a \left[ \frac{\text{mg}}{\text{m}^3} \right] = \left[ \frac{A * V}{m * l} \right]$$

Where:

A = Absorbance of the sample at 665 nm

V = Volume of the extract (in litre)

m = Slope of the calibration curve (in mg/L per unit of absorbance)

ℓ = Pathlength of the cuvette (typically 1 cm)

**SDD: Secchi stick depth**

Secchi stick measurements, also known as Secchi disk depth measurements, are a simple and widely used method for estimating water transparency or turbidity in aquatic environments such as lakes, rivers, reservoirs, and oceans. It is named after its inventor, Angelo Secchi, who was an Italian scientist and priest in the 19th century. This method plays a significant role in offering valuable insights into the transparency of water, which is crucial for understanding the well-being of aquatic ecosystems.

It is made up of a black or white circular disk, usually 20 centimetres in diameter, attached to a calibrated pole. To conduct a measurement using the Secchi Stick, the disk is slowly lowered into the water until it is no longer visible, and then raised until it reappears. The depth at which the disk disappears is recorded as the Secchi depth. The principle behind the Secchi Stick measurement is based on the fact that as light passes through water, it gets scattered and absorbed by suspended particles, dissolved substances, and living organisms like algae and plankton. The higher the concentration of these impurities in the water column, the lower the Secchi depth, indicating reduced water transparency.

Researchers, resource managers, and legislators may use this data to make well-informed choices on water quality and conservation by learning important details about the dynamics and health of aquatic ecosystems.

$$Chl\ a = e^{[2.997 - 1.47 \ln[SDD]]}$$

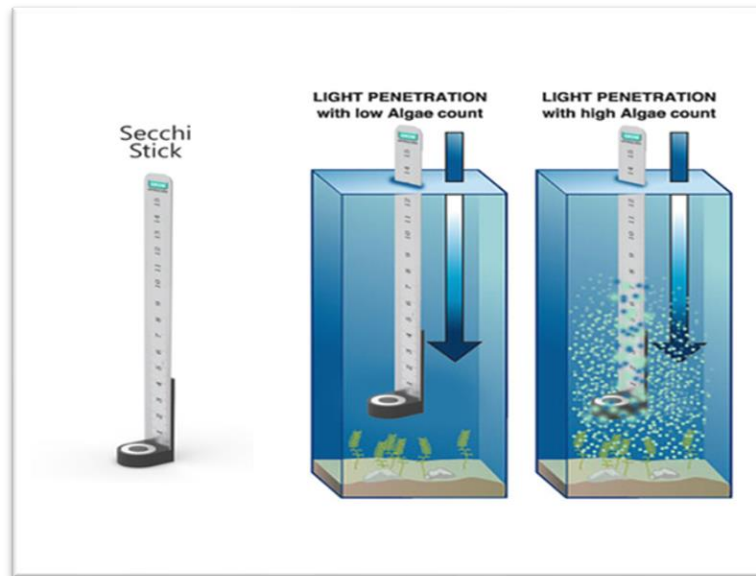


Fig 3: Diagram of Secchi stick depth

## II. Fluorometric Method:

The fluorometric method is based on the measurement of chlorophyll-a fluorescence emitted by chlorophyll-a molecules when excited by light of a specific wavelength. This method offers higher sensitivity and lower detection limits compared to spectrophotometric methods.

- **Extraction:** Chlorophyll-a is extracted from water samples using a solvent such as acetone or ethanol.
- **Analysis:** The chlorophyll extract is excited with a light source at a specific wavelength (e.g., 430 nm), and the fluorescence emitted by chlorophyll-a is measured at a longer wavelength (e.g., 670 nm) using a fluorometer.
- **Calculation:** Chlorophyll-a concentration is calculated based on the fluorescence intensity of the chlorophyll extract.



### **III. High-Performance Liquid Chromatography (HPLC):**

HPLC is a more advanced technique for chlorophyll-a determination that offers high sensitivity and precision. It involves the separation of chlorophyll-a from other pigments present in the water sample using a chromatographic column, followed by quantification of chlorophyll-a based on its retention time and peak area.

- **Extraction:** Chlorophyll-a is extracted from water samples using a solvent such as acetone or methanol.
- **Separation:** The chlorophyll extract is injected into an HPLC system equipped with a chromatographic column and a detector (e.g., UV-Vis detector).
- **Analysis:** Chlorophyll-a is separated from other pigments present in the sample based on its retention time and peak area.
- **Quantification:** Chlorophyll-a concentration is determined by comparing the peak area of the chlorophyll-a peak to that of a standard chlorophyll-a solution.

### **IV. Remote Sensing:**

Remote sensing techniques, such as satellite imagery and hyperspectral imaging, offer a non-invasive approach for monitoring chlorophyll-a concentrations over large spatial scales.

- **Data Acquisition:** Remote sensing data, such as Landsat or Sentinel satellite imagery, is acquired for the study area.

- **Data Processing:** Algorithms are applied to the remote sensing data to derive chlorophyll-a concentrations based on the spectral reflectance of the water surface.
- **Chlorophyll-a Mapping:** Chlorophyll-a concentrations are mapped spatially using the processed remote sensing data.

### **Relation between trophic state of water body and chlorophyll-A**

The trophic state of water bodies is a critical indicator of their overall health and ecological balance. Trophic state refers to the level of productivity in a body of water, primarily determined by the amount of nutrients, such as phosphorus and nitrogen, present in the water. These nutrients are essential for the growth of algae and other aquatic plants. Among the various indicators used to assess the trophic state of water bodies, chlorophyll a concentration is one of the most widely used and reliable measures.

It is commonly known that the trophic condition of water and the amount of chlorophyll an are related. Chlorophyll a content is often low in oligotrophic, or low-nutrient, water bodies, such as deep, clear lakes. This is due to the fact that the growth of algae and other aquatic plants is constrained by the restricted supply of nutrients. As a result, there is often little aquatic creature biomass and the water is clean with high visibility.

The growth of algae and other aquatic plants is encouraged by high amounts of nutrients, especially phosphorus and nitrogen, found in eutrophic, or high-nutrient, water bodies. Because of this, eutrophic water bodies have a substantially greater concentration of chlorophyll a than oligotrophic water bodies. Dense algal blooms may result from these elevated amounts of chlorophyll a, which may have substantial effects on the environment, the economy, and public health.

**Oligotrophic Waters:** These are low-nutrient, clear waters, often found in deep lakes or reservoirs. Oligotrophic waters have low concentrations of chlorophyll a due to limited nutrient availability, resulting in low algal productivity. The water tends to be clear with good visibility and supports a relatively low biomass of aquatic organisms.

**Mesotrophic Waters:** These waters have moderate nutrient levels and productivity. Mesotrophic waters fall between oligotrophic and eutrophic waters in terms of nutrient levels and chlorophyll a concentrations. They often have a moderate amount of algae and aquatic plant growth.

**Eutrophic Waters:** Eutrophic waters are high in nutrients, particularly phosphorus and nitrogen. These nutrients promote the growth of algae and other aquatic plants, leading to high chlorophyll a concentrations and high primary productivity. Eutrophic waters may experience algal blooms, reduced water clarity, and decreased oxygen levels, which can negatively impact aquatic life and water quality.

Trophic State	Description	Nutrients level	Chlorophyll a concentration ( $\mu\text{g/L}$ )
Oligotrophic	Low-nutrient, clear water	Low	< 2
Mesotrophic	Moderate nutrient levels, moderate productivity	Moderate	2 - 20
Eutrophic	High-nutrient, high productivity	High	> 20

Table no.1: Trophic state levels

## **1.2: BACKGROUND**

The rapid increase in population all over the world has demanded an increase in the yield of agro products. This has in turn resulted in indiscriminate use of fertilizers. As a result, nutrients such as nitrogen and phosphorus are discharged into water bodies, causing eutrophication. Eutrophication is the process by which a water body becomes overly nutrient-rich, resulting in abundant growth of simple plant life such as algae. Because of this, eutrophic water bodies have a substantially greater concentration of chlorophyll a than oligotrophic water bodies. Dense algal blooms may result from these elevated amounts of chlorophyll a, which may have substantial effects on the environment, the economy, and public health.

In this work, we determine the concentration of chlorophyll-A using an optical approach, specifically a spectrophotometer with machine learning. Monitoring chlorophyll a concentration in water is essential for assessing water quality, ecosystem health, and the effectiveness of management efforts. By tracking changes in chlorophyll a concentration over time, scientists and resource managers can identify trends, diagnose problems, and take action to protect and restore water bodies for future generations.

### **1.3: AIM AND OBJECTIVE**

- Prediction of chlorophyll-a concentration in water bodies leveraging machine learning algorithms.
- Comparative study of different Machine Learning (ML) models.
- Assess the quality of water bodies using chlorophyll-A concentration.

#### **1.4: HYPOTHESES**

- We aim to establish a strong correlation between chlorophyll-a concentration and optical sensor readings, which will enable accurate and reliable detection of chlorophyll-a in a variety of aquatic environments.
- We expect our approach to exhibit high sensitivity and specificity in detecting chlorophyll-a, enabling accurate measurement of chlorophyll-a levels across a broad range.
- We anticipate our methodology will allow for real-time and in-situ monitoring of chlorophyll-a, providing valuable insights into temporal and spatial variations in water quality.
- We aim to develop cost-effective, efficient, and reliable techniques for detecting chlorophyll-a in water. This will aid in environmental monitoring, resource management, and the protection of aquatic ecosystems.



### **1.5: SCOPE**

The identification of chlorophyll-a in water is crucial because it serves as a major indication of ecosystem health and water quality. Algae and cyanobacteria contain the pigment chlorophyll-a, whose concentration in water bodies gives important information about nutrient levels, algal biomass, and possible dangers of eutrophication and toxic algal blooms. It is feasible to evaluate the trophic status of water bodies, pinpoint pollution sources, track changes in water quality over time, and provide guidance for management and conservation initiatives by precisely measuring the quantity of chlorophyll-a. Thus, it is crucial to create effective and trustworthy techniques for detecting chlorophyll-a in water in order to monitor the environment, manage resources, and safeguard aquatic ecosystems. By developing a reliable and affordable method for the detection and measurement of chlorophyll-a in water and by leveraging machine learning algorithm and optical sensing techniques to increase the accuracy and efficiency of chlorophyll-a monitoring in a range of aquatic environments, this project seeks to advance this goal.

## CHAPTER-2

## **2.0: LITERATURE SURVEY**

### **2.1: LITERATURE REVIEWS**

- In 2022, Yang et.al., the author utilize ZY1-02D hyperspectral satellite subdivision to estimate the chlorophyll-a concentration of Baiyangdian Lake. The study area was chosen to be the Baiyangdian Nature Reserve in northern China, which contains a typical inland lake and wetland. By analyzing the correlation between the spectral reflectance of the ZY1-02D hyperspectral image and the chlorophyll-a concentration, a quantitative hyperspectral model of the chlorophyll-a concentration was established. The results showed that the estimation of the chlorophyll-a concentration of Baiyangdian Lake based on the hyperspectral Fluorescence Line Height (FLH) model was ideal, with an R<sup>2</sup> value of 0.78. The study provides new ideas and technical support for monitoring inland water environments.[1]
- In 2019, S. Yadav et. al aimed to estimate the concentration of chlorophyll-a in freshwater Lake Biwa and the coastal water of Wakasa Bay in Japan using satellite images and a spectral decomposition algorithm. Results show that both Landsat-8/OLI and Sentinel-2A/MSI sensors provided accurate data for coastal water, but Sentinel-2A yielded better results for the lake water. The obtained results can be useful for evaluating primary productivity in both freshwater and coastal water bodies.[2]
- In 2002, Éva Párista et.al., suggests using 90% (v/v) ethanol for chlorophyll extraction and measurement. Using a succession of ethanol/water solutions with decreasing water content, the chlorophyll concentrations of cultures of *Synechococcus elongatus* Nägeli and *Scenedesmus acutus* Meyen were

ascertained. Additionally, the study examined extraction yields using acetone, methanol, and 90% ethanol. The results of the trials using *Synechococcus elongatus* Nägeli and *Scenedesmus acutus* Meyen indicated that the maximum extraction yield was obtained with 100% methanol, which was followed by 90% ethanol and acetone.[3]

- Pawan Kumar and associates investigated Renuka Lake in the Lesser Himalaya region of India for similar purposes using surface water samples. Renuka Lake was classified as hyper-eutrophic. The study's conclusion emphasized the need to control nutrient enrichment to restore Renuka Lake.[4]
- In 2023, Mathilde de FLEURY et.al., studied water bodies in the semi-arid Sahelian region and were detected using a deep learning model built on the U-Net architecture. However, the outcomes demonstrated that a significant obstacle and potential source of inaccuracy for numerous current algorithms and databases is the identification of aquatic vegetation. Water identification using a deep learning model based on U-Net architecture that uses spectral information, thresholding, and classifications 98% accuracy in identifying Sahelian water bodies was attained. effectively categorized various optical water kinds to aid in the knowledge of eco-hydrology.[5]
- In 2023, Arias-Rodriguez et.al., created a global dataset of lake characteristics by combining data from various water monitoring programs with harmonized Landsat-8 and Sentinel-2 data. This dataset is then utilized for time series analysis, water quality maps for lakes across continents, and model training for global water quality prediction. The study models water quality using machine learning techniques like random forest regression (RFR) and extreme learning machine (ELM). In an effort to enhance the models, the writers also look into

more feature engineering. For SDD, TURB, and BOD, trained models obtained relatively strong correlations. Random forest regression (RFR) and the extreme learning machine (ELM) performed better.[6]

- In 2022, Barraza-Moraga et.al., proposed the study to determine whether the Sentinel-2 MSI sensor is suitable for estimating Chl-a in a lake located in the central region of Chile. It also suggests an empirical method that applies multiple linear regressions, compares the effectiveness and performance of L1C and L2A products and divides the equations created using spring-summer and fall-winter data. Spectral band multiple linear regressions. L1C and L2A product comparison.  $R^2 > 0.87$  indicates a strong algorithmic association with Chl-a. The obtained spatial distribution of the concentrations of Chl-a in Lanalhue Lake.[7]
- In 2022, Pompêo et.al., evaluated the quality of surface water using satellite photos by counting the number of Cyanobacteria cells per milliliter (cyano), measuring light penetration using the Secchi disk technique (SD), and measuring the concentration of Chlorophyll a (chl a). Atmospheric correction for Case 2 Regional Coast Color (C2RCC) in satellite and Sentinel-2 imagery solid chl a and SD estimates for reservoirs in the Cantareira System. Chl a and cyano have a strong association in the Broa reservoir.[8]
- In 2022, Wang et.al., established the inversion model of Chl-a concentration using single-band and band-ratio approaches based on in-situ hyperspectral data that corresponds to bands on Sentinel-2. one-band technique. band-ratio approach. In inland lakes, the band-ratio technique works better for retrieving Chl-a and Remote sensing monitoring of Chl-a in northeast China is feasible.[9]

- In 2022, Jiarui Shi et.al., compared how well Sentinel-2 and Gaofen-6 satellite sensors estimate the quantities of chlorophyll-a in tiny bodies of water. The usefulness of remote sensing for tracking water quality measures, such as chlorophyll-a, has been shown in earlier studies. However, there aren't many direct comparisons between various satellite sensors—especially when it comes to smaller bodies of water. By assessing both sensors' capabilities and applicability for tracking chlorophyll-a concentrations in smaller aquatic habitats, the study seeks to close this gap. Sentinel-2 and Gaofen-6 combined sensors were compared to estimate chlorophyll-a. For estimate, four machine learning models and five semi-empirical models were employed. Water bodies' chlorophyll-a concentrations are estimated using a combination of Sentinel-2 and Gaofen-6. The most accurate model is the extreme gradient boosting tree model.[10]
- In 2018, Malahlela et.al., used Landsat 8 OLI data to map the amounts of chlorophyll-a in the Vaal Dam, which is affected by algae and cyanobacteria. Prior studies have shown how useful remote sensing is for tracking the amounts of chlorophyll-a in aquatic bodies, especially when using data from Landsat. Few research, meanwhile, have explicitly examined how algae and cyanobacteria affect water quality using Landsat data. In order to close this gap, this work uses Landsat 8 OLI images to provide a thorough characterization of the chlorophyll-a concentrations in the Vaal Dam. Utilizing Landsat OLI data, stepwise logistic regression (SLR) was conducted. For analysis, vegetation indicators sensitive to chl-a were constructed. The range of overall accuracy was 65% to 83%. The differential vegetation index and chl-a have a positive association.[11]



- In 2022, Cadondon et.al., examined the determination of chlorophyll-a pigment in *Spirulina* using a portable pulsed LED fluorescence lidar system as a means of tracking algal development. The study tackles the requirement for real-time, non-destructive monitoring methods in algae culture. The study illustrates the efficacy of the lidar system in precisely detecting chlorophyll-a concentrations using *Spirulina* as a model organism. With possible implications in the environmental monitoring and algal cultivation industries, this research advances the development of effective monitoring tools for algal growth. Transportable LED fluorescence lidar device with pulses. Common techniques include adjusted chlorophyll-a concentration, optical density at 680 nm, and EEM fluorescence chlorophyll-a pigment at 680 nm. Chlorophyll-a has been precisely measured using a portable LED fluorescence lidar device, and the F680/F700 lidar ratio is correlated with the concentration of *spirulina*. [12]
- In 2020, Kovalevskaya et.al., proposed a study suggesting the inadjustment to the spectrophotometric technique for quantifying chlorophyll-a in water bodies' suspended particles. Chlorophyll-a analysis is frequently performed using spectrophotometric techniques, however different approaches can produce different outcomes. By altering the current procedure, the research seeks to increase the precision and dependability of chlorophyll-a readings. The work advances the precision of chlorophyll-a analysis, which is essential for evaluating the water quality in aquatic environments, by improving the spectrophotometric approach. Spectrophotometric technique with nuclear filters. Filters should be briefly dried for 30 minutes at 50–55°C. Chlorophyll in lab cultures and naturally occurring plankton is compared. Chlorophyll a has been measured spectrophotometrically using nuclear filters. [13]

- In 2012, Wang et.al., proposed a portable measuring tool with an emphasis on fluorescence detection that is intended for the identification of chlorophyll-a in aquatic environments. The creation of transportable tools for measuring chlorophyll-a is essential for monitoring water quality in real time. Compared to conventional techniques, the equipment delivers better sensitivity and accuracy by using fluorescence detection. This study contributes to the field of environmental management and water quality monitoring by addressing the demand for practical, on-site chlorophyll-a measuring instruments. Detector with two optical lenses and LED illumination. The connection between fluorescence intensity and concentration is measured using a spectrophotometer. It is advised to use a new dual optical device to test the content of chlorophyll-a. For low quantities of chlorophyll-a, good linear consistency was found.[14]
- In 2020, Markogianni et.al., used Landsat data to estimate the concentrations of chlorophyll-a in Greece's inland water bodies. The use of remote sensing techniques has shown promise in the monitoring of water quality metrics like chlorophyll-a. On the other hand, little study has been done expressly on Greek inland water bodies. This work intends to close this knowledge gap and offer insightful information on chlorophyll-a concentrations in Greek inland waterways using Landsat data. With potential implications in environmental management and conservation efforts in Greece, the research advances remote sensing techniques for monitoring water quality. step-by-step examination of multiple regression (MLR). Analyzing principal components (PCA). developed models with precise measurements for both man-made and natural lakes. The

suggested structure encourages ongoing assessment of the quality of lake water.[15]

- Lu Wang et.al., used hyperspectral imaging technology to evaluate the impact of season models on the accuracy of Chl-a estimation in outdoor aquaculture ponds in 2015. After removing pigment with 90% ethanol, water samples were collected from the ponds of the Freshwater Fish Seed Breeding Center in Guangzhou, China, and then tested for Chl-a content ( $\mu\text{g/L}$ ) using a standard spectrophotometer technique. Seasonal variations in chlorophyll-a concentrations are as follows: 1.119–32.216  $\mu\text{g/L}$  in winter, 28.097–53.360  $\mu\text{g/L}$  in spring, 100.162–196.403  $\mu\text{g/L}$  in summer, and 49.241–83.933  $\mu\text{g/L}$  in autumn.[16]
- In 2010, Jieying Xiao and Zijing Guo suggested a study to ascertain the concentration of chlorophyll A in waters. A total of 1000 milliliters of water were sampled from three artificial lakes. The hot ethanol spectrophotometric technique was used to measure the level of chlorophyll-A concentration and Analysis of hyperspectral reflectance data. Utilizing reflectance data, identify the concentration of chlorophyll-a. Water bodies in Shijiazhuang City has low concentrations of chlorophyll-a. Hyperspectral reflectance measurements were used to indirectly detect the concentration of chlorophyll-a. [17]
- In 2023, Ashwini Mudaliar et.al examined how to evaluate the quantity of cyanobacterial chlorophyll-a in two wetland habitats using multi-temporal Sentinel-2 photos as a water quality indicator. The ability of remote sensing, especially with Sentinel-2 data, to track chlorophyll-a concentrations in different aquatic bodies has been shown by earlier study. Water samples from the marshes of Wadhwana and Timbi were randomly selected. extraction of

dissolved oxygen using the American Public Health Association's recommended technique. The study found a correlation between declining water quality and a decrease in dissolved oxygen and chlorophyll a content. Wetlands' dissolved oxygen levels may be efficiently mapped using Sentinel-2 data. [18]

- In 2019, Pan, Y et.al., introduced a submersible in-situ chlorophyll fluorescence detection device that is intended for extremely accurate and instantaneous monitoring of chlorophyll content in aquatic settings. Chlorophyll fluorescence in the water column is immediately measured by the system using fluorescence detection technology, which allows for quick and precise observations without the need to collect and analyze samples. accuracy-focused optical route, fluorescence signal modulation, and filtering optimization. Longer in-situ continuous detection duration with low power consumption. Coefficient of correlation between linear response and R over 0.999. Measuring time in situ up to three months with an accuracy of 0.02µg/L. [19]
- In 2021, Basak, R et.al., developed a novel method for estimating chlorophyll-a concentration in algae species using Electrical Impedance Spectroscopy (EIS). They successfully demonstrated that EIS could be used as a rapid and non-destructive technique for chlorophyll-a estimation in algae. By measuring the impedance spectra of different algae species, they were able to establish a correlation between impedance values and chlorophyll-a concentration. The results showed that the proposed EIS method had a high sensitivity and accuracy for chlorophyll-a estimation, making it a promising alternative to traditional methods. This novel approach has the potential to significantly improve the efficiency and accuracy of chlorophyll-a monitoring in aquatic environments. [20]

- In 2023, Avantika Latwal et.al utilized Sentinel-2 satellite imagery to detect and map water and chlorophyll-a spread for assessing water quality in inland water bodies. They successfully demonstrated the effectiveness of Sentinel-2 imagery for monitoring water quality parameters. By employing spectral indices and machine learning algorithms, they were able to accurately detect water and estimate chlorophyll-a concentration. The results showed that Sentinel-2 imagery could provide valuable information for assessing the spatial and temporal dynamics of water quality in inland water bodies. This approach offers a cost-effective and efficient solution for large-scale water quality monitoring, with significant potential for supporting environmental management and conservation efforts. [21]
- In 2018, Watanabe, F., Alcântara et.al optimized a semi-analytical algorithm for estimating chlorophyll-a concentration in productive inland waters. They utilized remote sensing data from Landsat 8 OLI imagery and field measurements to develop and validate the algorithm. The results demonstrated that the optimized algorithm provided accurate estimates of chlorophyll-a concentration, with good agreement between satellite-derived and field-measured chlorophyll-a values. The study highlighted the potential of remote sensing techniques for monitoring water quality in productive inland waters, offering a cost-effective and efficient method for large-scale chlorophyll-a monitoring and environmental management. [22]
- In 2017, Yang, Z., Reiter et.al investigated the use of ratio-based Near-Infrared (NIR) to Red (Red) indices for estimating chlorophyll-a concentrations in diverse water bodies. They developed and tested several NIR/Red indices using Landsat 8 OLI imagery and field measurements of chlorophyll-a concentration.

The results showed that the NIR/Red indices provided accurate estimates of chlorophyll-a concentration across different types of water bodies, including lakes, reservoirs, and rivers. The study demonstrated the effectiveness of remote sensing techniques for monitoring chlorophyll-a concentrations in diverse aquatic environments, offering a valuable tool for water quality assessment and management. [23]

- In 2012, Matthews et.al developed an algorithm for detecting trophic status (chlorophyll-a concentration), cyanobacterial dominance, surface scums, and floating vegetation in both inland and coastal waters using Sentinel-2 satellite imagery. The algorithm, based on machine learning techniques and spectral indices, achieved high accuracy in identifying these water quality parameters. The results demonstrated the effectiveness of the algorithm in providing comprehensive assessments of water quality, including eutrophication status and the presence of cyanobacterial blooms and surface pollutants. This approach offers a valuable tool for monitoring and managing water quality in diverse aquatic environments, supporting environmental conservation and management efforts. [24]
- In 2004, Gregor et.al conducted a comparative analysis of in vitro, in vivo, and in situ methods for quantifying freshwater phytoplankton using chlorophyll-a concentration. They found that the in-situ method, which involved direct measurement of chlorophyll-a fluorescence in water samples using a fluorometer, provided the most accurate and reliable results compared to the in vitro and in vivo methods. The in-situ method demonstrated strong correlations with the in vitro method, which involved chlorophyll extraction from water samples in the laboratory, and outperformed the in vivo method, which relied

on chlorophyll fluorescence measurements in vivo using a portable fluorometer. The findings suggest that the in-situ method is the most suitable for accurate and efficient quantification of freshwater phytoplankton and chlorophyll-a concentration in natural water bodies. [25]

- In 2001, Giardino et.al investigated the feasibility of detecting chlorophyll concentration, Secchi disk depth, and surface temperature in a sub-alpine lake using Landsat imagery. They found that Landsat imagery could effectively detect chlorophyll concentration, Secchi disk depth, and surface temperature in the sub-alpine lake with reasonable accuracy. The results showed strong correlations between Landsat-derived values and field measurements for chlorophyll concentration ( $R^2 = 0.76$ ), Secchi disk depth ( $R^2 = 0.83$ ), and surface temperature ( $R^2 = 0.87$ ). The study demonstrated the potential of Landsat imagery for monitoring water quality parameters in sub-alpine lakes, providing valuable insights into the spatial and temporal dynamics of water quality in these environments. [26]
- In 2023, Karimian et.al developed a novel framework to predict chlorophyll-a concentrations in water bodies using multi-source big data and machine learning algorithms. They integrated data from various sources, including satellite imagery, meteorological data, and water quality parameters, to train machine learning models for chlorophyll-a prediction. The results showed that the framework achieved high accuracy in predicting chlorophyll-a concentrations, with a coefficient of determination ( $R^2$ ) of 0.85. The study demonstrated the potential of big data and machine learning approaches for accurate and efficient monitoring of chlorophyll-a concentrations in water bodies, providing valuable tools for water quality assessment and management. [27]

- In 2008, Randolph et.al., proposed a study to determine if cyanobacteria in turbid and productive waters might be identified by hyperspectral remote sensing employing optically active pigments such as phycocyanin and chlorophyll-a. They discovered that cyanobacterial biomass in murky waters could be precisely identified and measured when using hyperspectral remote sensing in conjunction with spectral analysis of phycocyanin and chlorophyll-a. Strong correlations were seen between the values obtained from remote sensing and laboratory measurements of chlorophyll-a and phycocyanin, with coefficients of determination ( $R^2$ ) of 0.82 and 0.79, respectively. This work shows how hyperspectral remote sensing may be used to track cyanobacterial blooms in murky aquatic settings, offering important information for managing and assessing water quality. [28]
- In 2018, Ansper, A., & Alikas, K. assessed the viability of extracting chlorophyll-a concentrations from water bodies using Sentinel-2 Multi-Spectral Instrument (MSI) data. Using Sentinel-2 MSI data, they created and verified an algorithm for retrieving chlorophyll-a, and then they compared the outcomes with measurements made in-situ. With a coefficient of determination ( $R^2$ ) of 0.88, the study showed that Sentinel-2 MSI data could reliably estimate chlorophyll-a concentrations in different aquatic bodies. The results indicate that Sentinel-2 MSI data may be utilized efficiently for WFD reporting, offering important insights for managing and assessing water quality. [29]
- In 2006, Duan, H., Zhang, Y., Zhang, B. et al. assessed chlorophyll-a concentration and trophic state for Lake Chagan using Landsat Thematic Mapper (TM) imagery and field spectral data. They developed regression models to estimate chlorophyll-a concentration from Landsat TM data and



found a strong correlation between satellite-derived and field-measured chlorophyll-a concentrations, with a coefficient of determination ( $R^2$ ) of 0.82. Additionally, they used trophic state indices calculated from chlorophyll-a concentrations to classify Lake Chagan as mesotrophic. The study demonstrated the effectiveness of Landsat TM imagery for monitoring chlorophyll-a concentration and trophic state in Lake Chagan, providing valuable insights for water quality assessment and management. [30]

## 2.2: TABULAR REVIEWS

Paper no.	Abstract	Model	Results
[1]	In 2022, Yang et.al., ZY1-02D hyperspectral satellite subdivision was utilized to estimate the chlorophyll-a concentration. By analyzing the correlation between the spectral reflectance of the ZY1-02D hyperspectral image and the chlorophyll-a concentration, a quantitative hyperspectral model of the chlorophyll-a concentration was established.	Hyperspectral Fluorescence Line Height (FLH) model	$R^2 = 0.78$
[2]	In 2019, S. Yadav et. al aimed to estimate of chlorophyll concentration in lake water and coastal water using Landsat-8 and Sentinel-2A satellite imagery.	Spectral decomposition algorithm for chlorophyll-a estimation using Landsat-8/OLI and Sentinel-2A/MSI satellite images	Landsat-8/OLI and Sentinel-2A/MSI sensors provided accurate data for coastal water.
[3]	In 2002, Éva Párista et.al., The study critically examined the ISO:10260, 1992 standard, which recommends using 90% (v/v) ethanol for chlorophyll extraction and measurement.	Acetone, methanol, and 90% ethanol	Maximum extraction yield obtained with 100% methanol, followed by 90% ethanol and acetone
[4]	Pawan Kumar proposed a study focusing on understanding the	Assessment of trophic status and	Classification of Renuka Lake as

	trophic status of Renuka Lake and identifying factors limiting its water quality.	identification of limiting factors in Renuka Lake	hyper-eutrophic based on water quality parameters.
[5]	In 2023, Mathilde de FLEURY et.al.,proposes a method for mapping water bodies in the Sahel region using Sentinel-2 Multispectral Instrument (MSI) images and a U-Net convolutional neural network (CNN).	Utilized a deep learning model based on the U-Net architecture	98% accuracy
[6]	In this paper, the authors create a global dataset of lake characteristics by combining data from various water monitoring programs with harmonized Landsat-8 and Sentinel-2 data. This dataset is then utilized for time series analysis, water quality maps for lakes across continents, and model training for global water quality prediction.	Random Forest Regression (RFR) and Extreme Learning Machine (ELM).	Strong correlations for SDD, TURB, and BOD
[7]	In 2023, Arias-Rodriguez et.al., examined a study whether the Sentinel-2 MSI sensor is suitable for estimating Chl-a	Random Forest Regression (RFR) and Extreme Learning Machine (ELM). Multiple Linear Regression (MLR).	High accuracy for RFR and ELM.  L1C and L2A gives $R^2 > 0.87$ .

[8]	In 2022, Pompêo et.al., suggested a study focusing on assessing water quality in Brazilian reservoirs using Sentinel-2 imagery.	C2RCC atmospheric correction.	Chla and cyano have a strong association in the Broa reservoir.
[9]	Wang et al. established an inversion model to estimate chlorophyll-a (Chl-a) concentration in inland lakes using Sentinel-2 bands based on in-situ hyperspectral data..	using Sentinel-2 bands	Band-ratio approach showed better performance for chlorophyll-a (Chl-a) estimation in inland lakes compared to the single-band technique.
[10]	In 2022, Jiarui Shi et.al., compared how well Sentinel-2 and Gaofen-6 satellite sensors estimate the quantities of chlorophyll-a in tiny bodies of water.	Comparison of machine learning and semi-empirical models for chlorophyll-a estimation using fused Gaofen-6 and Sentinel-2 sensor data	The extreme gradient boosting tree model was found to be the most accurate for estimating chlorophyll-a concentrations in small water bodies using data from both Sentinel-2 and Gaofen-6 sensors
[11]	This study aimed to map chlorophyll-a concentrations in the Vaal Dam, impacted by cyanobacteria and algae, using Landsat 8 OLI data.	Mapping model for chlorophyll-a concentrations using Landsat 8 OLI data in the Vaal Dam	Successful mapping of chlorophyll-a concentrations in the Vaal Dam

			using Landsat 8 OLI data.
[12]	In 2022, Cadondon et.al., examined the determination of chlorophyll-a pigment in Spirulina	Portable Pulsed LED Fluorescence Lidar System	Successful measurement of chlorophyll-a pigment in Spirulina
[13]	This study aimed to modify the spectrophotometric method for the determination of chlorophyll-a in the suspended matter of water bodies.	Modified spectrophotometric method	Improved accuracy and efficiency achieved in the determination of chlorophyll-a in suspended matter of water bodies using the modified spectrophotometric method.
[14]	This study introduces a portable measurement instrument designed for the measurement of chlorophyll-a in water bodies, supporting fluorescence detection.	Portable measurement instrument for chlorophyll-a measurement in water bodies supporting fluorescence detection	Successful development of a portable measurement instrument for chlorophyll-a measurement in water bodies supporting fluorescence detection.
[15]	This study aimed to estimate chlorophyll-a concentrations in inland water bodies in Greece using Landsat data.	Multiple regression (MLR) and principal component	Successful estimation of chlorophyll-a concentrations in Greece's inland

		analysis (PCA) models	water bodies using Landsat data.
[16]	Lu Wang (2015) utilized hyperspectral imaging technology to evaluate the impact of seasonal variations on the accuracy of chlorophyll-a (Chl-a) estimation in outdoor aquaculture ponds.	Using hyperspectral imaging technology	Winter: 1.119–32.216 µg/L Spring: 28.097–53.360 µg/L Summer: 100.162–196.403 µg/L Autumn: 49.241–83.933 µg/L
[17]	The study conducted by Jieying Xiao and Zijing Guo in 2010 aimed to determine the concentration of chlorophyll-a in waters.	Utilization of hyperspectral reflectance measurements for indirect detection of chlorophyll-a concentration in water bodies	Water bodies in Shijiazhuang City were found to have low concentrations of chlorophyll-a.
[18]	Ashwini Mudaliar et al. investigated the evaluation of cyanobacterial chlorophyll-a quantity in two wetland habitats using multi-temporal Sentinel-2 images as a water quality indicator	Using multi-temporal Sentinel-2 images and dissolved oxygen mapping in wetland habitats	Correlation observed between declining water quality, decreased dissolved oxygen, and chlorophyll-a content in wetland habitats. Sentinel-2 data proved effective in efficiently mapping dissolved oxygen levels in wetlands.

[19]	The research team introduced a submersible in-situ chlorophyll fluorescence detection device designed for highly accurate and instantaneous monitoring of chlorophyll content in aquatic settings.	Submersible in-situ chlorophyll fluorescence detection device	$R \Rightarrow 0.999$
[20]	The authors developed a novel method for estimating chlorophyll-a concentration in algae species.	Electrical Impedance Spectroscopy (EIS)	High sensitivity and accuracy observed for chlorophyll-a estimation
[21]	Avantika Latwal et al. utilized Sentinel-2 satellite imagery to detect and map water and chlorophyll-a spread for assessing water quality in inland water bodies.	Sentinel-2 satellite imagery	Successful detection and mapping of water and chlorophyll-a spread using Sentinel-2 satellite imagery.
[22]	Watanabe, F., Alcântara et al. optimized a semi-analytical algorithm for estimating chlorophyll-a concentration in productive inland waters.	using Landsat 8 OLI imagery	The optimized algorithm provided accurate estimates of chlorophyll-a concentration,
[23]	Yang, Z., Reiter et al. investigated the use of ratio-based Near-Infrared (NIR) to Red (Red) indices for estimating chlorophyll-a concentrations in diverse water bodies.	Ratio-based Near-Infrared (NIR) to Red (Red) indices using Landsat 8 OLI imagery	NIR/Red indices provided accurate estimates of chlorophyll-a concentration across different types of water bodies.

[24]	Matthews et al. developed an algorithm for detecting trophic status (chlorophyll-a concentration), cyanobacterial dominance, surface scums, and floating vegetation in both inland and coastal waters using Sentinel-2 satellite imagery.	Sentinel-2 satellite imagery	High accuracy
[25]	Gregor et al. conducted a comparative analysis of in vitro, in vivo, and in situ methods for quantifying freshwater phytoplankton using chlorophyll-a concentration.	chlorophyll-a concentration	in-situ method is the most suitable for accurate and efficient quantification of freshwater phytoplankton and chlorophyll-a concentration in natural water bodies.
[26]	Giardino et al. investigated the feasibility of detecting chlorophyll concentration, Secchi disk depth, and surface temperature in a sub-alpine lake using Landsat imagery.	using Landsat imagery	Chlorophyll concentration: $R^2 = 0.76$ Secchi disk depth: $R^2 = 0.83$ Surface temperature: $R^2 = 0.87$
[27]	Karimian et al. developed a novel framework to predict chlorophyll-a concentrations in water bodies using multi-source big data and machine learning algorithms.	multi-source big data and machine learning algorithms	High accuracy in predicting chlorophyll-a concentrations, with a coefficient



			of determination (R <sup>2</sup> ) of 0.85
[28]	The purpose of this study was to determine if cyanobacteria in turbid and productive waters might be identified by hyperspectral remote sensing employing optically active pigments such as phycocyanin and chlorophyll-a.	Hyperspectral remote sensing model and optically active pigments such as phycocyanin and chlorophyll-a.	chlorophyll-a and phycocyanin, with coefficients of determination (R <sup>2</sup> ) of 0.82 and 0.79, respectively.
[29]	The authors of this study assessed the viability of extracting chlorophyll-a concentrations from water bodies using Sentinel-2 Multi-Spectral Instrument (MSI) data.	Sentinel-2 Multi-Spectral Instrument (MSI) data	R <sup>2</sup> = 0.88
[30]	This study used field spectral data and Landsat Thematic Mapper (TM) images to evaluate the trophic condition and concentration of chlorophyll-a	Regression models	R <sup>2</sup> = 0.82.

### **2.3: LITERATURE REVIEWS CONCLUSION**

The Insight from some of the research papers are about the detection of chlorophyll A in water and machine learning. It highlights the main aspects like, how to determine the concentration of chlorophyll A and how concentration of chlorophyll A helps to monitor the environment, manage resources, and safeguard aquatic ecosystems. By taking the highlights from the research papers, the detection of chlorophyll A in water and trophic state index which includes the different types of sensors was implemented.

In the case of machine learning implementation, some research papers talk about the process of data acquisition and machine learning. Different methods like deep learning, hybrid machine learning, convolution neural networks, regressions were implemented by different researchers. Different image processing algorithm like Landsat imagery, Sentinel-2 satellite imagery, etc were implemented

## CHAPTER-3

### 3.0: METHODOLOGY

#### 3.1: SAMPLES

Samples of pure algae were brought in from Goa University's Botany Laboratory. Three distinct age groups of algae, including younger, moderate, and older, were represented in the samples. Immediately spectra were removed using the below dilution.

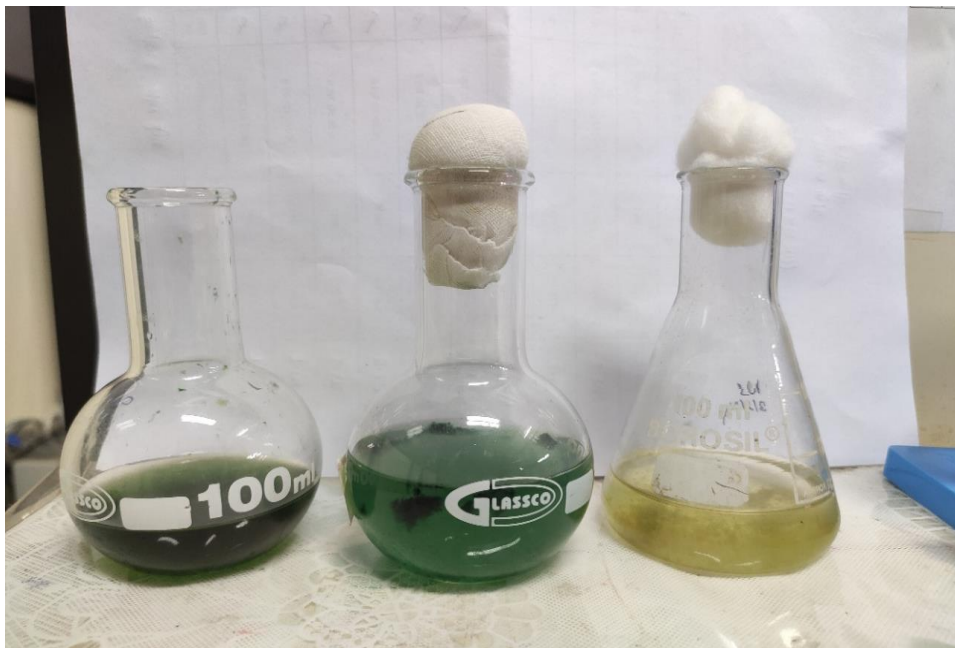


Fig 4: Different Types of Algae Samples



Fig 5: sample in cuvette

Pure algae samples were diluted with distilled water in the following dilution concentration [%v/v]: 0%, 90%, 81%, 72.9%, 65.61%, 59.04%, 53.144%, 47.82%, 43.04%, 38.74%, 34.86%. using below formula:

$$\frac{9\text{ml} * \text{algae solution } \%}{10\text{ml}} = \text{conc.}$$

Secchi depth of solution was found using Secchi stick. A Secchi stick measurement was taken by lowering the Secchi stick vertically into the water until it can no longer be seen and the depth was noted down.

Concentration of chl-A of each sample were obtained using the standard formula

$$\text{Chl } a = e^{[2.997 - 1.47 \ln[SDD]]}$$

Secchi depth of solution: 2.0 cm = 0.02 m

Formula: chl-a conc =  $e [ 2.997 - 1.47 \ln[SDD]]$

$$= e[2.997 - 1.47 \ln(0.02)] = 6296.02 \mu\text{g/L} = 6.296 \text{ mg/L}$$

Sr No	Algae Sample	Water	Dilution %	Concentration
1	10 ml of 100%	0 ml	0 %	6.296
2	9 ml of 100 %	1 ml	90 %	5.666
3	9 ml of 90 %	1 ml	81 %	5.099

4	9 ml of 81 %	1 ml	72.9 %	4.589
5	9 ml of 72.9 %	1 ml	65.61 %	4.130
6	9 ml of 65.61 %	1 ml	59.049 %	3.717
7	9 ml of 59.049 %	1 ml	53.144 %	3.345
8	9 ml of 53.144 %	1 ml	47.829 %	3.010
9	9 ml of 47.829 %	1 ml	43.046 %	2.709
10	9 ml of 43.046 %	1 ml	38.742 %	2.439
11	9 ml of 38.742 %	1 ml	34.867 %	2.194

Table no. 2: Younger Algae dilution

Secchi depth of solution: 2.5 cm = 0.025m

Formula: Chl-a Conc =  $e [2.997 - 1.47 \ln [SDD]]$

=  $e [2.997 - 1.47 \ln (0.025)] = 4535.32 \mu\text{g/l} = 4.535 \text{ mg/L}$

Sr No	Algae Sample	Water	Dilution %	Concentration
1	10 ml of 100%	0 ml	0 %	4.535
2	9 ml of 100 %	1 ml	90 %	4.081
3	9 ml of 90 %	1 ml	81 %	3.673

4	9 ml of 81 %	1 ml	72.9 %	3.306
5	9 ml of 72.9 %	1 ml	65.61 %	2.975
6	9 ml of 65.61 %	1 ml	59.049 %	2.677
7	9 ml of 59.049 %	1 ml	53.144 %	--
8	9 ml of 53.144 %	1 ml	47.829 %	--
9	9 ml of 47.829 %	1 ml	43.046 %	--
10	9 ml of 43.046 %	1 ml	38.742 %	--
11	9 ml of 38.742 %	1 ml	34.867 %	--

Table no. 3: Moderate algae dilution

Secchi Depth of Solution: 1.7 cm = 0.017m

Formula: chl-a conc =  $e [2.997 - 1.47 \ln(\text{SDD})]$

$$= e[2.997 - 1.47 \ln(0.017)] = 7995.04 \mu\text{g/L} = 7.995 \text{ mg/L}$$

Sr No	Algae Sample	Water	Dilution %	Concentration
1	10 ml of 100%	0 ml	0 %	7.995
2	9 ml of 100 %	1 ml	90 %	7.199
3	9 ml of 90 %	1 ml	81 %	6.475

4	9 ml of 81 %	1 ml	72.9 %	5.828
5	9 ml of 72.9 %	1 ml	65.61 %	5.245
6	9 ml of 65.61 %	1 ml	59.049 %	4.720
7	9 ml of 59.049 %	1 ml	53.144 %	4.248
8	9 ml of 53.144 %	1 ml	47.829 %	3.823
9	9 ml of 47.829 %	1 ml	43.046 %	3.441
10	9 ml of 43.046 %	1 ml	38.742 %	3.097
11	9 ml of 38.742 %	1 ml	34.867 %	2.787

Table no.4: Older algae dilution



### 3.2: DATA ACQUISITION

Reference method to find the chlorophyll A concentration:

#### **Secchi-stick measurement**

In aquatic environments, the Secchi stick depth, also called the Secchi depth, is a measurement of the transparency of the water. A Secchi stick, which consists of a circular disk that is either black or white and fastened to a pole that has been calibrated, is used to measure it. The Secchi stick is gradually lowered into the water until the disk is no longer visible, and then it is lifted till it reappears, in order to determine the Secchi depth. The Secchi depth is the depth at which the disk vanishes from vision.

A measure of water transparency is the Secchi depth; deeper depths correspond to cleaner water, whereas shallower depths correspond to more turbid water. Light penetration is affected by several variables, including the quantity of dissolved chemicals, suspended particles, and live organisms in the water column. In order to evaluate the spatial and temporal fluctuations in water purity, Secchi depth measurements are usually made at regular intervals and places within a water body.

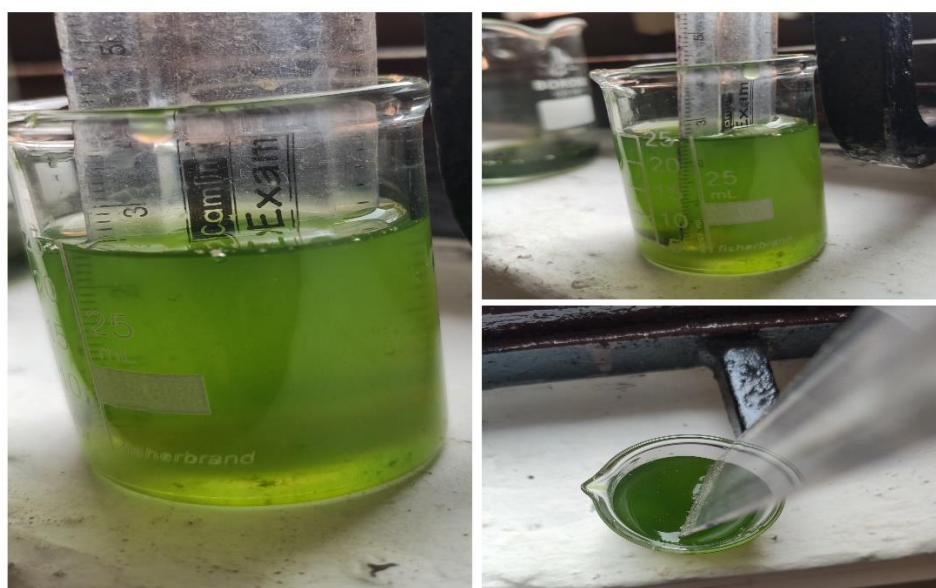


Fig 6: Secchi stick measurement

### **Data Acquisition Using Jasco v-770 spectrophotometer**

The Jasco V-770 spectrophotometer was used to acquire spectra of all samples. The samples were placed in a quartz cuvette of optical length of 10 mm and absorbance range 200-800 nm and then the cuvette was kept in the sample holder. The UV-VIS spectra of the algae samples were acquired in absorbance mode. The UV-VIS bandwidth and range were set to 20 nm and 200-800 nm respectively. Every sample was measured in parts of 11, where each part was measured 3 times which made up to 3 spectra per concentration.



Fig 7: Setup of Spectrophotometer

### **Absorbance Spectroscopy**

Absorbance spectroscopy is a fascinating technique that plays a crucial role in measuring the absorption of light by a substance as a function of wavelength. It is an

essential tool in various fields like chemistry, biochemistry, and physics, where it is used for quantitative analysis of substances.

The principle behind absorbance spectroscopy is that every chemical compound absorbs, transmits, or reflects light over a certain range of wavelengths. When light passes through a sample, the amount of light absorbed is proportional to the concentration of the absorbing substance in the sample and the path length of the light through the sample.

Absorbance spectrophotometers consist of a light source, a monochromator to isolate a specific wavelength of light, a sample holder, and a detector (photodetector). They are commonly used for measuring absorbance in the ultraviolet (UV) and visible (VIS) regions of the electromagnetic spectrum.

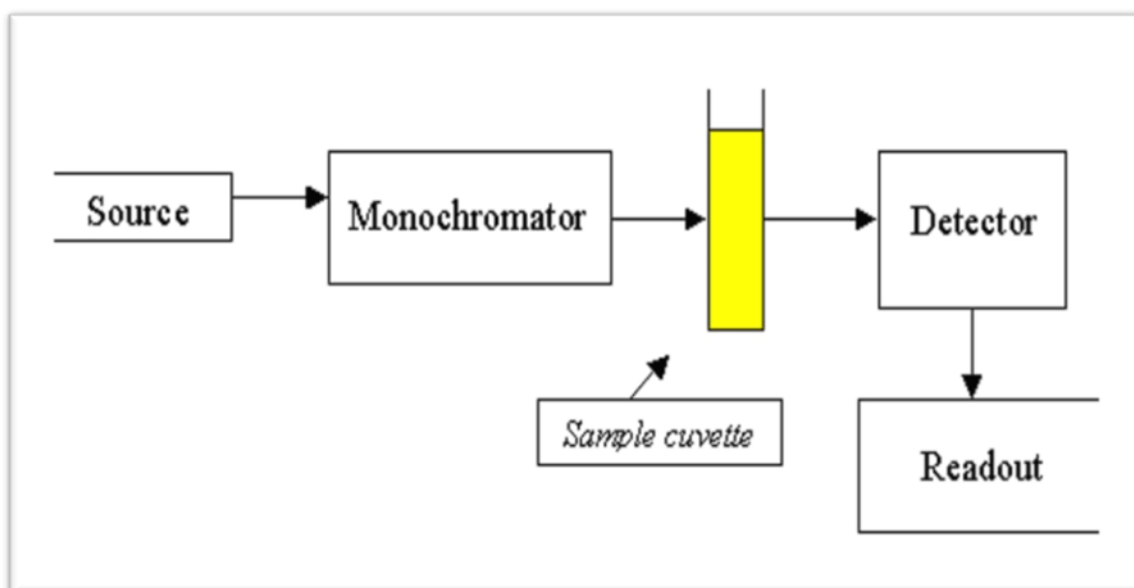


Fig 8: Block Diagram of Spectrophotometer

The block diagram of the Spectrophotometer shows a light source that emits a broad spectrum of light covering the desired wavelength range. Common light sources include tungsten-halogen lamps, deuterium lamps (for the UV range), and xenon

lamps (for the visible and near-infrared range), a monochromator which is a crucial component that separates the incoming polychromatic light into its constituent wavelengths, a sample cuvette for the sample and a detector measures the intensity of light transmitted through the sample as a function of wavelength.

### 3.2.1: Operating Principle of Spectrophotometer: Beer Lambert Law

Spectroscopy is based on the interaction between light and matter. When molecules absorb IR radiation, transitions occur from a ground vibrational state to an excited vibrational state. If the frequency of the light matches the frequency of the vibration of the bonds in the molecule, the molecule absorbs some of the light.

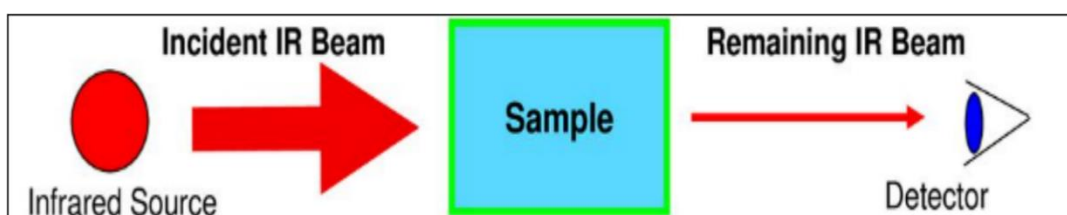


Fig 9: Interaction of Light with the Sample

For each wavelength of light passing through the spectrometer, the intensity of the light passing through the reference cell is measured. This is usually referred to as  $I_0$  - that's  $I$  for Intensity.

The intensity of the light passing through the sample cell is also measured for that wavelength - given the symbol,  $I$ . If  $I$  is less than  $I_0$ , then the sample has absorbed some of the light (neglecting reflection of light off the cuvette surface). A simple bit of math is then done in the computer to convert this into something called the absorbance of the sample - given the symbol,  $A$ . The absorbance of a transition depends on two external assumptions.

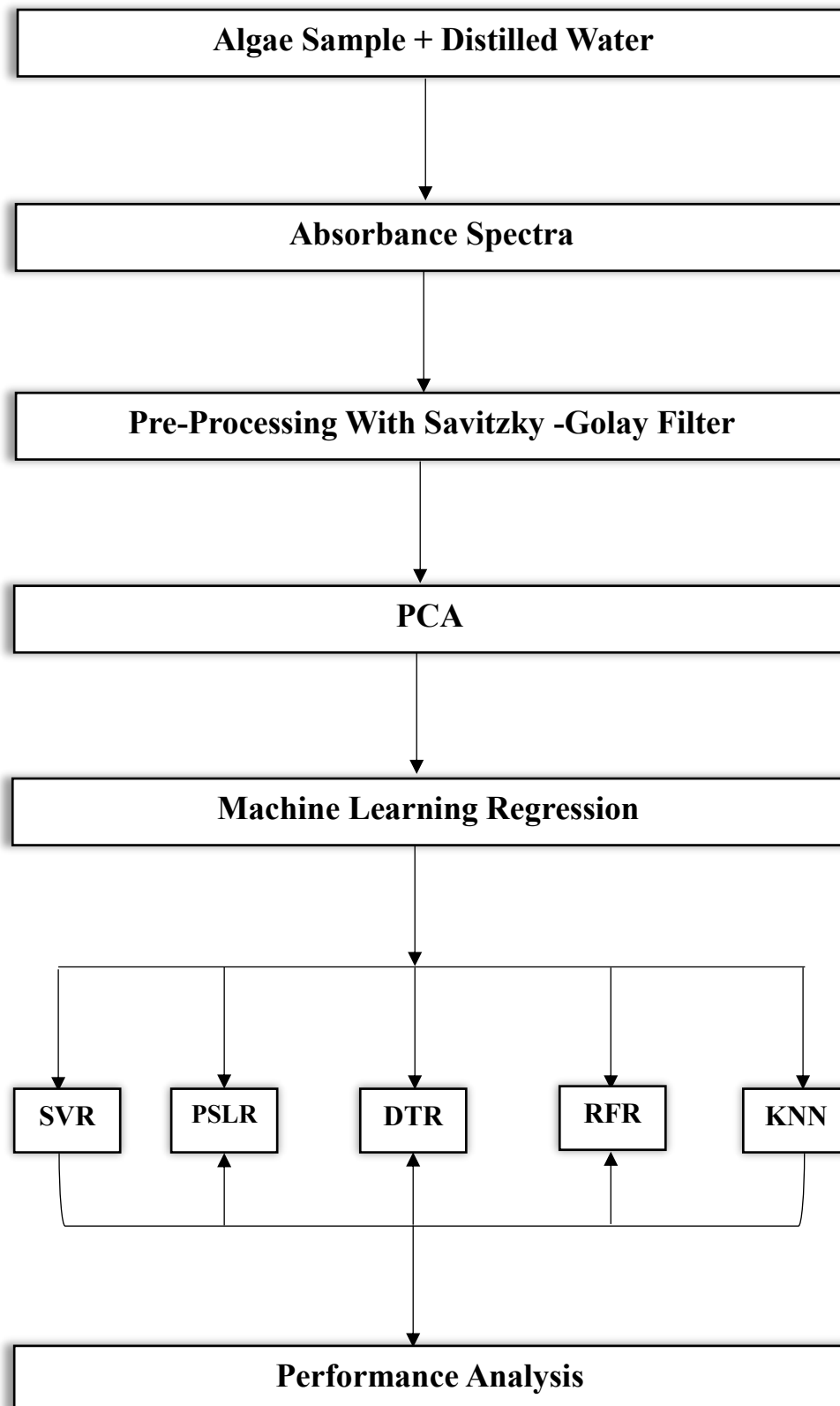
- The absorbance is directly proportional to the concentration ( $c$ ) of the solution of the sample used in the experiment.
- The absorbance is directly proportional to the length of the light path ( $l$ ), which is equal to the width of the cuvette.
- The equation can be written as:

$$A \propto cl$$

- This proportionality can be converted into equality by including a proportionality constant ( $\epsilon$ ).

$$A = \epsilon cl$$

Where  $A$  is the value of absorbance,  $\epsilon$  is the Molar absorbance coefficient,  $c$  is molar concentration and  $l$  is optical path length

**FLOWCHART:**

### **3.3: PREPROCESSING**

#### **3.3.1: Savitzky-Golay filter**

The Savitzky-Golay filter was named after its inventors Abraham Savitzky and Marcel J.E. Golay. This digital filtering method is used for smoothing and differentiating data and is highly effective in reducing noise from data sets that contain random variations or fluctuations. It is an invaluable tool that is primarily used in signal processing, chromatography, spectroscopy, and time series analysis. The Savitzky-Golay filter is a method of smoothing data points that relies on fitting subsets of adjacent points with a low-degree polynomial using linear least squares. Unlike some other smoothing techniques that can change or distort the important features of the data, the Savitzky-Golay filter maintains these features, such as peak heights and widths, while simultaneously reducing noise. This ability to preserve the underlying signal's integrity is one of the key advantages of the Savitzky-Golay filter. The effectiveness of a filter is determined by two main factors: the window size and the polynomial order. The window size determines how many data points are included in the local regression. Bigger window sizes produce smoother results but may over smooth the data, while smaller window sizes may not adequately reduce noise. The polynomial order sets the degree of the polynomial used for the local regression. Higher polynomial orders allow the filter to capture more complex variations in the data but may lead to overfitting. Choosing the right values for these parameters is essential for achieving the desired balance between noise reduction and preservation of signal features. The Savitzky-Golay filter has many advantages, one of which is its versatility. Not only can it be used to smooth data, but it can also be utilized for numerical differentiation of noisy data. This means that it can effectively differentiate a noisy signal, which can help extract meaningful information from data sets that would otherwise be difficult to accurately

interpret. This capability is especially useful in fields like spectroscopy, where the precise identification and quantification of spectral features are vital. The Savitzky-Golay filter is widely used in various scientific and engineering applications. For instance, in chromatography, it helps to decrease baseline noise and enhance the detection of analyte peaks in chromatograms. Similarly, in spectroscopy, the filter is used to eliminate spectral artifacts and enhance the signal-to-noise ratio, resulting in more accurate and dependable spectral analysis. In time series analysis, the filter is used to smooth out data and identify trends, patterns, and anomalies. The Savitzky-Golay filter is an effective method for smoothing and differentiating data, especially when there is noise present. However, it has some limitations. The filter's performance relies heavily on the choice of window size and polynomial order, which may require some trial and error to select optimal values. Moreover, the filter may not work well when the data contains complex or rapidly changing patterns. In the end, the Savitzky-Golay filter is a very useful tool in many technical and scientific domains. It can successfully reduce noise while maintaining the data's key characteristics. By carefully choosing the right criteria, researchers and analysts may use the filter to enhance the accuracy and dependability of their data analysis and interpretation.

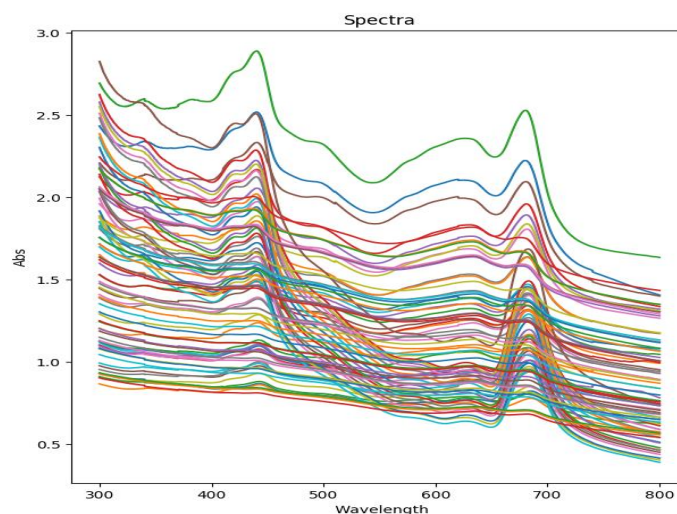


Fig 10: Smoothened Spectra



### 3.3.2: Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique used in data analysis and machine learning. Its primary goal is to simplify complex datasets while retaining the most important information.

In PCA, the algorithm identifies the directions (principal components) along which the data varies the most. These components are orthogonal to each other, meaning they are uncorrelated. The first principal component explains the largest amount of variance in the data, followed by the second component, and so on.

PCA is useful for several purposes:

1. **Dimensionality Reduction:** By representing data using a smaller number of principal components, PCA reduces the dimensionality of the dataset while preserving as much variance as possible.
2. **Visualization:** PCA can help visualize high-dimensional data in a lower-dimensional space, making it easier to explore and interpret.
3. **Feature Extraction:** PCA can be used to extract the most important features from a dataset, which can then be used for further analysis or modelling.
4. **Noise Reduction:** PCA can help remove noise and redundant information from the data, leading to better performance in downstream tasks.

### **3.4: DATA PARTITIONING**

#### **3.4.1: Cross Validation**

Cross-validation is a technique widely used in machine learning to evaluate the performance of a predictive model on an independent dataset. It involves dividing the original dataset into subsets, where the model is trained on one subset and evaluated on the remaining subset. This process is repeated multiple times, with each fold serving as both the training and validation set. The most common form of cross-validation is k-fold cross-validation, where the dataset is divided into k equal-sized folds.

Cross-validation provides several benefits, including a better estimate of performance, reduced variance, and optimization of hyperparameters. By averaging the performance over multiple iterations, cross-validation provides a more reliable estimate of the model's performance. It also helps reduce the variance in performance estimates by using multiple validation sets. Additionally, cross-validation is often used to tune the hyperparameters of a model by selecting the values that result in the best average performance across the folds. Overall, cross-validation is a critical tool for assessing and optimizing machine learning models to ensure that they generalize well to new and unseen data.

K-fold cross-validation is a widely used technique in machine learning to evaluate the performance of a predictive model. Its main objective is to determine how well a model trained on a specific dataset can generalize to new, unseen data. The process involves randomly partitioning the original dataset into k equal-sized subsets, or folds, where K is chosen to be 10 and training and evaluating the model k times, with a different fold serving as the validation set each time, while the remaining k-1 folds are used for training. By using every data point for both training and validation, k-fold cross-

validation provides a more reliable estimate of the model's performance. After training and evaluating on all  $k$  folds, the performance metrics are averaged to obtain an overall assessment of the model's performance. This method helps to reduce variance in performance estimates, compared to a single train-test split, and provides a more accurate representation of the model's ability to generalize to new data. It is frequently used to assess and compare different models, as well as to tune hyperparameters to optimize model performance.

### **3.5: MACHINE LEARNING**

Machine learning is a branch of artificial intelligence (AI) that focuses on the development of algorithms and statistical models that enable computers to perform tasks without explicit programming. Instead, machine learning algorithms use data to learn and improve their performance over time.

#### **3.5.1: SUPPORT VECTOR REGRESSION [SVR]**

Support Vector Regression (SVR) is a powerful machine learning technique used for regression tasks. Unlike traditional regression methods that aim to minimize error directly, SVR focuses on finding a hyperplane in a high-dimensional space that best represents the relationship between input variables and the target variable. This hyperplane is determined by support vectors, which are data points closest to the hyperplane and influence its position. SVR aims to minimize the margin of error within a certain threshold, known as the epsilon-insensitive tube, while also penalizing deviations outside this tube. This allows SVR to handle non-linear relationships between variables by using kernel functions to map the input data into a higher-dimensional space where a linear relationship may exist. SVR is particularly useful in scenarios where the data is noisy or exhibits non-linear patterns, making it a versatile tool in various fields such as finance, economics, and engineering. Its ability to effectively handle complex data relationships and its robustness against outliers make it a popular choice for regression tasks.

1. **Margin of Tolerance:** SVR introduces the concept of a margin of tolerance around the predicted value. Instead of minimizing error directly, SVR seeks to ensure that most of the data points fall within this margin.

2. Support Vectors: SVR identifies a subset of training data points, known as support vectors, which are crucial for defining the regression function. These support vectors lie either on the margin boundaries or within the margin itself.
3. Kernel Trick: SVR often employs a kernel function to transform the input features into a higher-dimensional space, where the data might be more separable. Common kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels.
4. Regularization Parameter: SVR includes a regularization parameter ( $C$ ) that controls the trade-off between maximizing the margin and minimizing the error on the training data. Higher values of  $C$  prioritize minimizing training error, while lower values prioritize maximizing the margin.
5. Epsilon Parameter: Another parameter in SVR is  $\epsilon$  (epsilon), which defines the width of the margin of tolerance. Larger values of  $\epsilon$  allow more data points to fall outside the margin, potentially resulting in a wider margin and a less complex model.

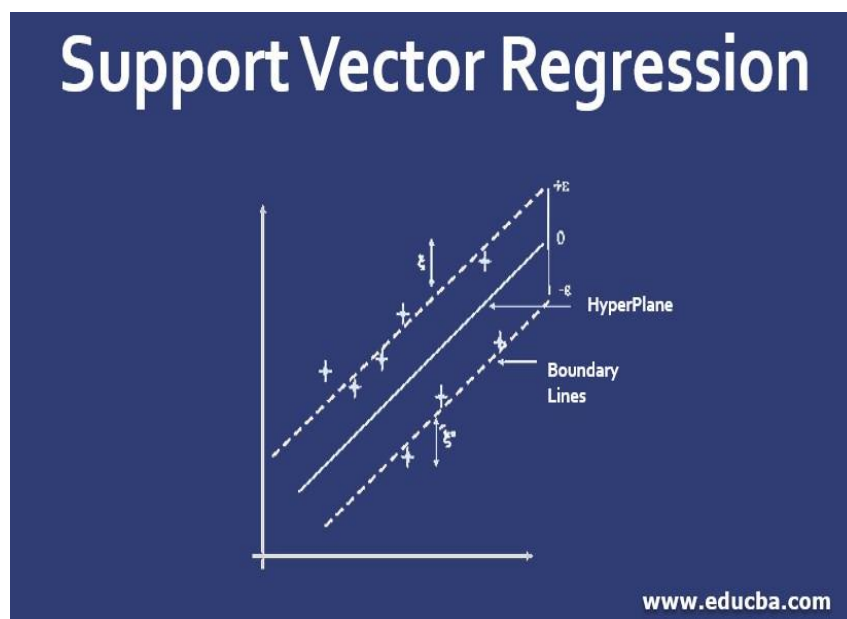


Fig 11: Support Vector Regression

### 3.5.2: RANDOM FOREST REGRESSION [RFR]

Random Forest Regression is a powerful machine learning technique used for regression tasks. It belongs to the ensemble learning family and is based on the Random Forest algorithm, which is an ensemble of decision trees. In Random Forest Regression, multiple decision trees are trained on random subsets of the training data and using random subsets of features. During prediction, each tree in the forest independently predicts the target variable, and the final prediction is obtained by averaging or taking the majority vote of these individual predictions. This ensemble approach helps to reduce overfitting and improve the model's generalization performance. Random Forest Regression is robust to outliers, handles high-dimensional data well, and provides feature importance scores, which can be useful for understanding the relationship between input variables and the target variable. It is widely used in various fields such as finance, healthcare, and marketing for predicting continuous it a popular choice for regression tasks.

1. Ensemble Learning: A Random Forest Regressor consists of a collection of decision trees, where each tree is trained independently on a random subset of the training data and a random subset of the features. This randomness helps to reduce overfitting and improves the model's robustness.
2. Bootstrap Aggregating (Bagging): Random Forest employs a technique called bootstrap aggregating or bagging, where each tree is trained on a bootstrapped sample of the original training data. This sampling with replacement ensures diversity among the trees, leading to a more robust ensemble model.
3. Random Feature Selection: In addition to sampling data points, Random Forest also randomly selects a subset of features at each split in the decision tree. This feature

randomness further enhances the diversity among the trees and reduces the correlation between them.

4. Prediction: Once the ensemble of decision trees is built, predictions for new data points are made by aggregating the predictions of individual trees. For regression tasks, the final prediction is often the average (or another aggregation) of the predictions from all the trees.

5. Interpretability and Versatility: While Random Forests may not be as interpretable as individual decision trees, they offer high predictive accuracy and can handle both numerical and categorical features without requiring feature scaling. They are also less sensitive to outliers and noise in the data.

6. Hyperparameter Tuning: Random Forests have several hyperparameters that can be tuned to optimize performance, such as the number of trees in the ensemble, the maximum depth of each tree, and the size of the random feature subsets.

7. Applications: Random Forest Regressors are widely used in various domains, including finance, healthcare, and ecology, for tasks such as predicting stock prices, estimating patient outcomes, and analysing ecological data.

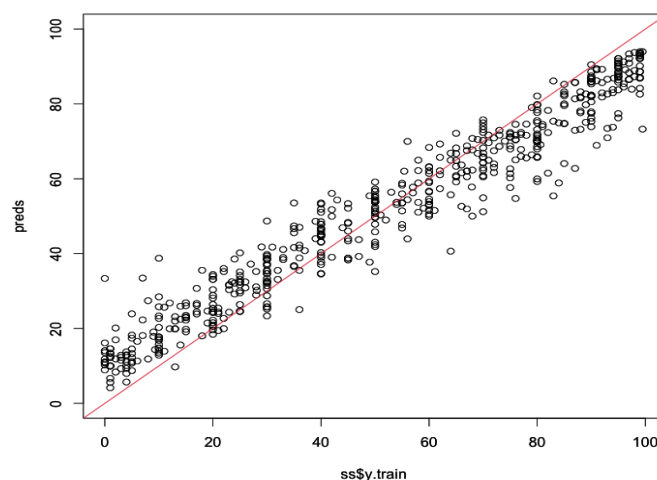


Fig 12: Random Forest Regression

### 3.5.3: K-NEAREST NEIGHBOR REGRESSION [KNN]

K-Nearest Neighbors (KNN) regression is a straightforward yet effective non-parametric algorithm used for regression tasks. Unlike traditional regression techniques, KNN regression doesn't assume a functional form for the relationship between variables. Instead, it relies on the principle that similar data points should have similar target values. In KNN regression, when a prediction is needed for a new data point, the algorithm looks at the K nearest data points in the feature space, based on some distance metric (usually Euclidean distance), and averages their target values to predict the target value for the new data point. The choice of K, the number of neighbors to consider, is a critical parameter in KNN regression. A smaller K value leads to more complex models with higher variance but lower bias, while a larger K value results in smoother predictions with lower variance but higher bias. KNN regression is intuitive, easy to implement, and doesn't make strong assumptions about the underlying data distribution. However, its performance can be sensitive to the choice of K and the distance metric used, and it can be computationally expensive, especially for large datasets. Despite these limitations, KNN regression remains a popular choice for regression tasks, particularly in situations where the data is noisy or lacks a clear functional form, and interpretability is less of a concern.

1. Neighbor-based Approach: KNN Regressor operates on the principle that similar data points tend to have similar target variable values. It does not explicitly learn a model but rather memorizes the training data to make predictions.

2. Parameter K: The "K" in KNN refers to the number of nearest neighbors to consider when making a prediction. A higher value of K considers more neighbors, potentially resulting in smoother predictions but may overlook local patterns. Conversely, a lower



value of  $K$  focuses on fewer neighbors, capturing more local variations but may be sensitive to noise.

3. Distance Metric: KNN Regressor typically uses a distance metric, such as Euclidean distance or Manhattan distance, to measure the similarity between data points in the feature space. The choice of distance metric can influence the algorithm's performance and should be selected based on the characteristics of the data.

4. Prediction: To make a prediction for a new data point, KNN Regressor identifies the  $K$  nearest neighbors based on the chosen distance metric and averages their target variable values to obtain the predicted value. In regression tasks, this average is often the mean value of the target variable among the nearest neighbors.

5. Non-parametric Nature: KNN Regressor is a non-parametric algorithm, meaning it does not assume any specific form for the underlying data distribution. Instead, it relies solely on the training data during prediction, making it flexible and adaptable to different types of data distributions.

6. Scalability and Efficiency: While KNN Regressor is conceptually simple and easy to implement, it can be computationally expensive, especially for large datasets or high-dimensional feature spaces. Techniques such as KD-trees or ball trees can be employed to improve the algorithm's efficiency.

7. Hyperparameter Tuning: The choice of  $K$  and the distance metric are critical hyperparameters in KNN Regressor. Cross-validation techniques can be used to select optimal values for these hyperparameters based on the performance on a validation set.

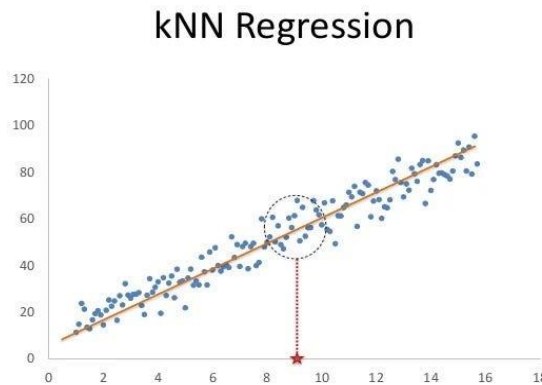


Fig 13: KNN Regression

#### **3.5.4: PARTIAL LEAST SQUARES REGRESSION (PLSR)**

Partial Least Squares Regression (PLSR) is a powerful statistical technique used in machine learning for building predictive models, especially when dealing with datasets with a high number of correlated independent variables. Unlike traditional linear regression, which assumes that predictors are uncorrelated, PLSR can handle multicollinearity by extracting a set of orthogonal components that explain the maximum variance in both the independent and dependent variables. By doing so, PLSR effectively reduces the dimensionality of the dataset while capturing the most relevant information for predicting the target variable. This makes PLSR particularly useful in situations where there are many predictors and limited sample sizes. Additionally, PLSR is robust to noise and outliers, making it suitable for dealing with noisy data. Consequently, PLSR finds applications in various fields such as chemometrics, economics, and marketing, where it is used for modeling complex relationships and making accurate predictions.

1. **Dimensionality Reduction:** PLSR aims to reduce the dimensionality of the predictor variables while still preserving their relationship with the response variable. It achieves

this by extracting a small number of latent variables, or components, that explain the maximum covariance between the predictor variables and the response variable.

2. Iterative Process: PLSR iteratively constructs these latent variables by maximizing the covariance between the predictor variables and the response variable in each component. Unlike Principal Component Analysis (PCA), which focuses solely on explaining the variance of the predictor variables, PLSR considers both the variance and the covariance.

3. Simultaneous Modelling: PLSR builds the latent variables in a way that optimally predicts the response variable while also considering the predictor variables. This simultaneous modelling approach allows PLSR to handle multicollinearity effectively, making it suitable for situations where predictors are highly correlated.

4. Prediction and Interpretation: Once the latent variables are constructed, PLSR can be used for prediction by regressing the response variable on these latent variables. Additionally, PLSR provides insights into the relationships between the predictors and the response, making it valuable for interpretation.

5. Parameter Tuning: PLSR includes parameters such as the number of components to extract and the scaling method, which can affect the performance of the model. Cross-validation techniques are often used to determine the optimal number of components and other tuning parameters.

### **3.5.5: DECISION TREE REGRESSION [DTR]**

Decision trees are powerful and versatile models used in machine learning for both classification and regression tasks. They work by recursively partitioning the data into subsets, with each partition based on the value of a particular feature. This process continues until a stopping criterion is met, such as a maximum tree depth or the

minimum number of samples in a leaf node. Decision trees are easy to interpret and understand, making them particularly useful for explaining the logic behind a model's predictions. They can handle both numerical and categorical data, and are robust to outliers. However, decision trees are prone to overfitting, especially when the tree is deep or the dataset is noisy. To address this issue, techniques such as pruning and ensemble methods like Random Forests and Gradient Boosting are often used. Decision trees find applications in various domains including finance, healthcare, and marketing, where they are used for tasks such as customer segmentation, risk assessment, and medical diagnosis.

1. **Tree Structure:** The decision tree is constructed recursively by splitting the data into subsets based on the values of features. At each node of the tree, a decision is made regarding which feature to split on and what threshold to use for the split. This process continues until a stopping criterion is met, such as reaching a maximum tree depth or minimum number of samples per leaf.
2. **Splitting Criteria:** The algorithm selects the best feature and threshold for splitting the data at each node based on a splitting criterion, typically aiming to minimize variance or mean squared error in the resulting subsets. Common splitting criteria include mean squared error, mean absolute error, and variance reduction.
3. **Predictions:** Once the decision tree is built, predictions for new data points are made by traversing the tree from the root node to a leaf node. The predicted value for a data point is typically the average (or another aggregation) of the target variable in the leaf node to which it belongs.

4. Interpretability: Decision trees are highly interpretable models, as they can be visualized graphically, allowing users to understand the decision-making process and identify important features for prediction.
5. Overfitting: Decision trees are prone to overfitting, especially when the tree depth is not properly constrained. Regularization techniques such as pruning or limiting the maximum depth of the tree are commonly used to mitigate overfitting.
6. Ensemble Methods: Decision trees can be further improved by using ensemble methods such as Random Forests or Gradient Boosting, which train multiple decision trees and combine their predictions to achieve better performance and generalization.

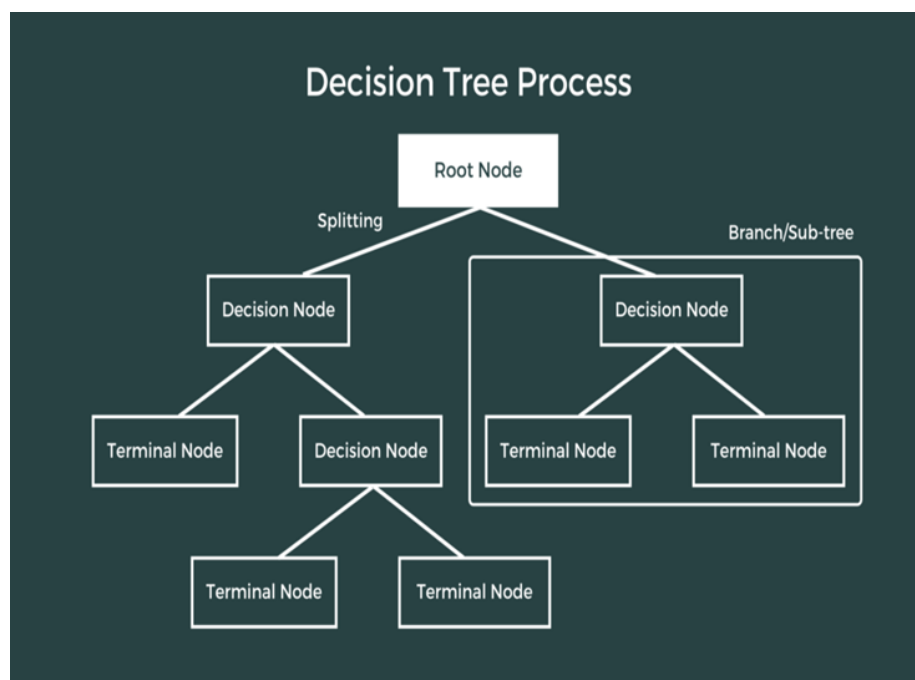


Fig 14: Decision tree Regression

### 3.6: REGRESSIONS AND THEIR PARAMETERS

Regressors	Parameters
SVR	C=1.0, epsilon=0.2
DTR	max_depth=50
RFR	max_depth=2, random_state=0
PSLR	n_components=10
KNN	n_neighbors=1, p= 1, weights = 'distance', algorithm = 'auto'

Table no. 5: Regressors and their parameters

## CHAPTER-4

## 4.0: ANALYSIS AND CONCLUSIONS

### 4.1: ANALYSIS AND RESULTS

The project consists of two parts which includes data collection and machine learning. Detection of chlorophyll A in water makes suitable for environmental monitoring, resource management, and the protection of aquatic ecosystems.

In this project, we delve in the analysis and comparative study of various ML Regressor for prediction of Chlorophyll A concentration using Absorbance spectroscopy. All of the investigations were carried out using the Google-Colab online platform. The python 3.11.3 version implemented the analytical statistics, ML models and graphics. The dataset was split into two parts: train and test, using 80:20 split that is standard for all performance analyses. K-fold splits was 10 and SG filter was applied for the dataset.

For the current study, R2 and RMSE values were used to determine the performance of all the models.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{Predicted - Actual}{n}}$$

It was observed that the absorbance of every diluted sample varied while the structure of the spectra remained similar. Actual v/s Predicted graph of the results obtained by the algorithms that are PLSR, SVR, DTR, RFR and KNR are shown in figure 15, figure 16, figure 17, figure 18 and figure 19 respectively.

A summary of the results obtained by the algorithms are shown in Table 6. Average RMSE obtained after the 10 -fold cross validation was be 0.25, 0.57, 0.44, 0.64 and 0.27 for SVR, PLSR, DTR, RFR and KNR respectively.



Algorithm	Best Fold [n=10]				Average RMSE
	R2	MSE	MAE	RMSE	
SVR	0.99	0.01	0.08	0.09	0.25
PSLR	0.95	0.07	0.20	0.26	0.57
DTR	0.97	0.03	0.09	0.19	0.44
RFR	0.88	0.15	0.31	0.39	0.64
KNN	0.0	0.0	0.0	1.0	0.27

Table no. 6: Results of all regressors

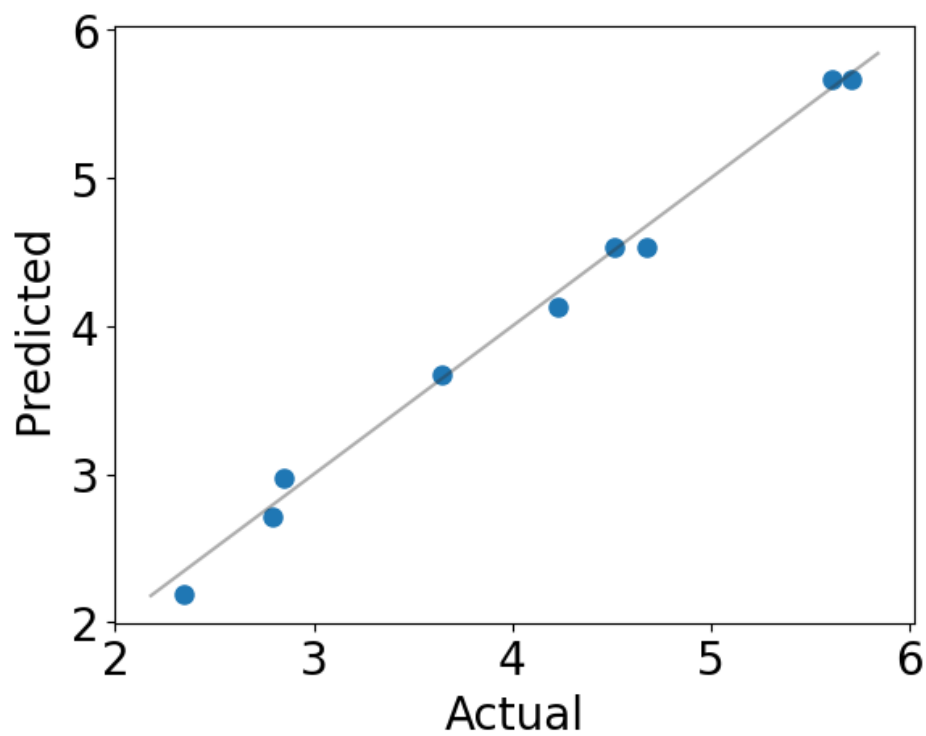
**Graphs of Best folds:**

Fig 15: Graph of Support Vector Regression

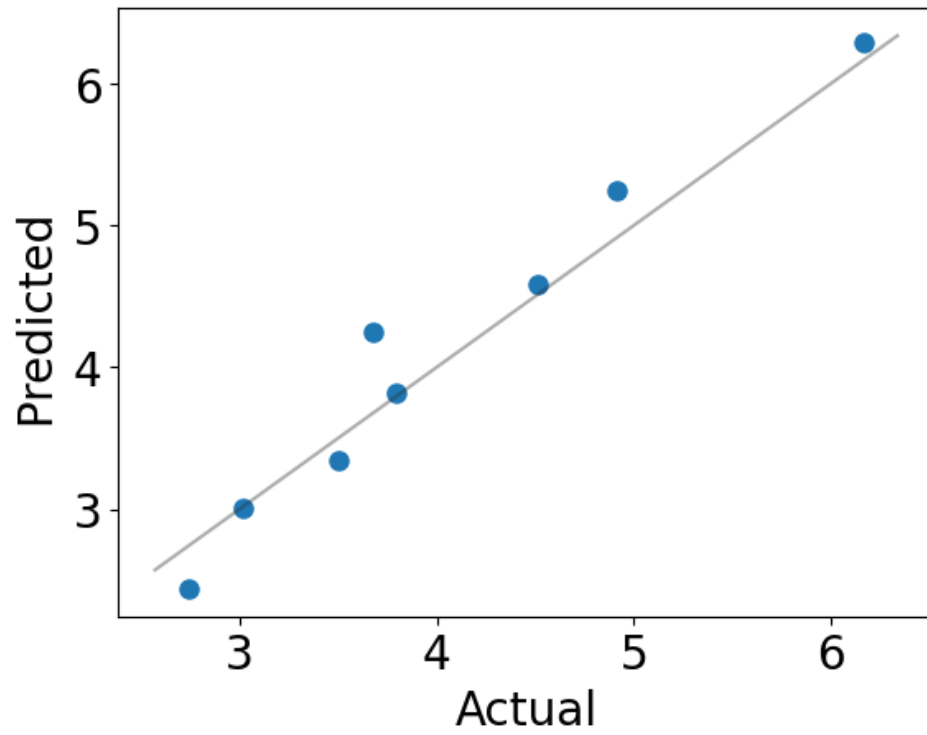


Fig 16: Graph of Partial Square Least Regression

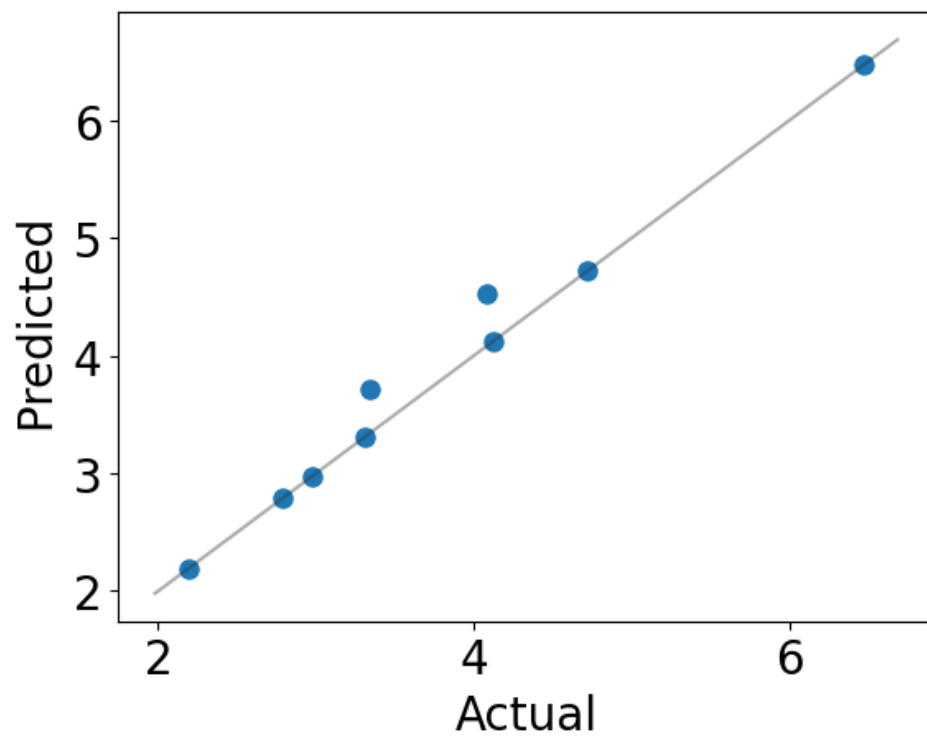


Fig 17: Graph of Decision Tree Regression

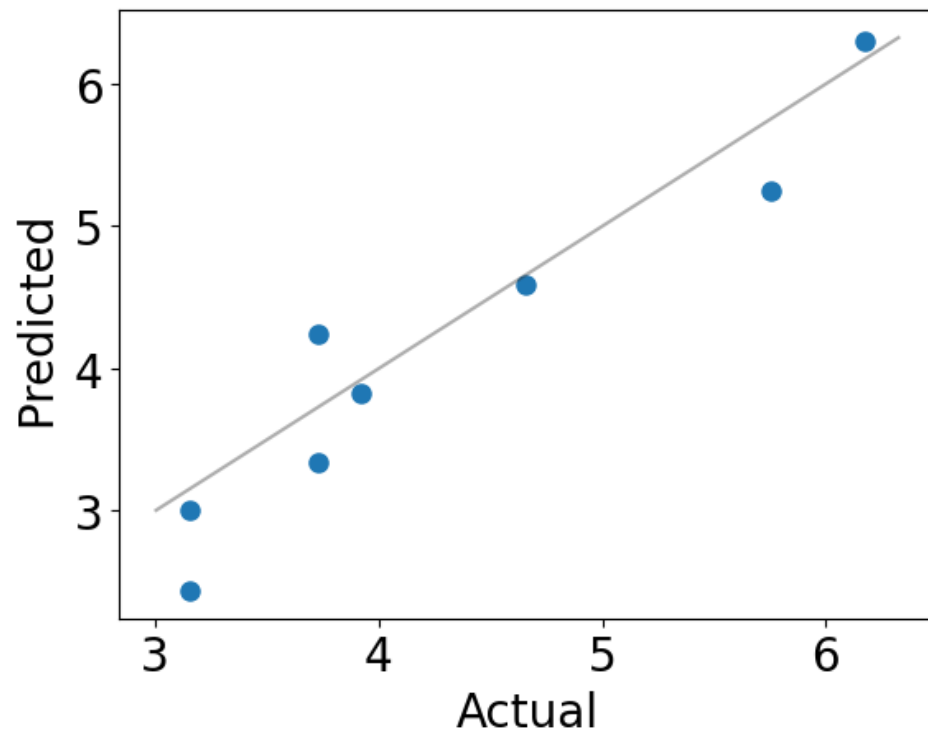


Fig 18: Graph of Random Forest Regression

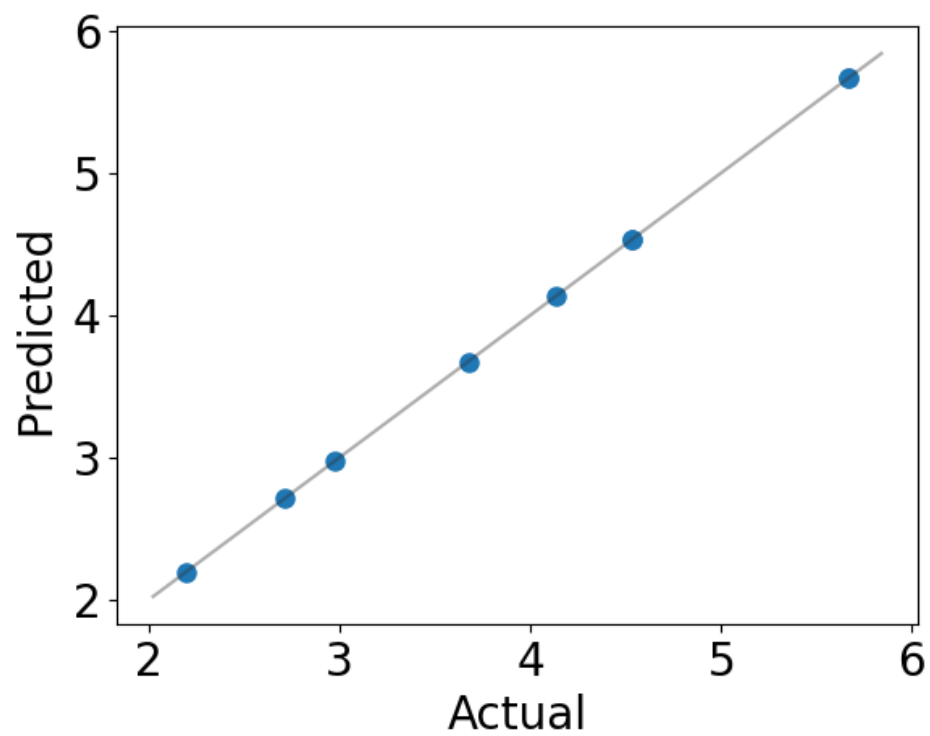


Fig 19: Graph of KNN Regression

## 4.2: CONCLUSION

In conclusion, The Detection of chlorophyll A In Water Using a spectrophotometer project was a challenging yet rewarding experience for me. Using a spectrophotometer, the amount of chlorophyll-a in water samples was effectively determined in this investigation. Chlorophyll-a was selectively detectable because of its absorbance spectra, which showed distinctive peaks in the visible spectrum's red and blue areas. A linear link between the content of chlorophyll-a in the water samples and its absorbance at particular wavelengths was found by using the Secchi stick measurement.

The results show that spectrophotometric analysis is a useful method for measuring chlorophyll-a in water samples. This technique provides an efficient, quick, and economical way to measure chlorophyll-a levels, which are crucial for water quality monitoring and aquatic ecosystem health assessments.

Overall, this project was successful for determining the chlorophyll A concentration.

### **4.3: FUTURE WORK**

Now as the detection of chlorophyll A in water using optical sensors was successfully done, now its time to take the project to a next level. The next part of the project will be developing a portable device for in-situ method. For this there will be a need to build a driver circuits and will have to buy high power LEDs and Laser Diode. And we have to test the circuit once it is built. And after that we will have to collect data and create dataset and same machine learning can be applied. And then we can determine the chlorophyll A concentration.

## CHAPTER-5

## 5.0: REFERENCES

### 5.1: Paper References

- [1] Yang, Z., Gong, C., Ji, T., Hu, Y., & Li, L. (2022). Water Quality Retrieval from ZY1-02D Hyperspectral Imagery in Urban Water Bodies and Comparison with Sentinel-2. *Remote Sensing*, 14(19), 5029. <https://doi.org/10.3390/rs14195029>
- [2] Yadav, S., Yamashiki, Y., Susaki, J., Yamashita, Y., & Ishikawa, K. (2019). CHLOROPHYLL ESTIMATION OF LAKE WATER AND COASTAL WATER USING LANDSAT-8 AND SENTINEL-2A SATELLITE. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLII-3/W7, 77–82. <https://doi.org/10.5194/isprs-archives-xlii-3-w7-77-2019>
- [3] Párista, É., Ács, É., & Böddi, B. (2002). *Hydrobiologia*, 485(1/3), 191–198. <https://doi.org/10.1023/a:1021329602685>
- [4] Kumar, P., Mahajan, A. K., & Meena, N. K. (2019). Evaluation of trophic status and its limiting factors in the Renuka Lake of Lesser Himalaya, India. *Environmental Monitoring and Assessment*, 191(2). <https://doi.org/10.1007/s10661-019-7247-0>
- [5] Mathilde de FLEURY, Laurent Kergoat, Brandt, M., Rasmus Fensholt, Ankit Kariryaa, Gyula Mate Kovács, Stéphanie Horion, & Grippa, M. (2023). Sentinel-2 MSI for mapping Sahelian water bodies using a U-Net network. <https://doi.org/10.5194/egusphere-gc8-hydro-22>
- [6] Arias-Rodriguez, L. F., Ulaş Firat Tüzün, Duan, Z., Huang, J., Ye Tuo, & Disse, M. (2023). Global Water Quality of Inland Waters with Harmonized Landsat-8 and Sentinel-2 Using Cloud-Computed Machine Learning. *Remote Sensing*, 15(5), 1390–1390. <https://doi.org/10.3390/rs15051390>
- [7] Barraza-Moraga, F., Alcayaga, H., Pizarro, A., Félez-Bernal, J., & Urrutia, R. (2022). Estimation of Chlorophyll-a Concentrations in Lanalhue Lake Using Sentinel-

2 MSI Satellite Images. Remote Sensing, 14(22), 5647.

<https://doi.org/10.3390/rs14225647>

[8] Pompêo, M., Viviane Moschini-Carlos, Marisa Dantas Bitencourt, Sòria-Perpinyà, X., Vicente, E., & Delegido, J. (2022). Water Quality Assessment Using Sentinel-2 Imagery Estimating Chlorophyll A, Secchi Disk Depth, and Cyanobacteria Cell Number in Brazilian Reservoirs. <https://doi.org/10.3390/blsf2022014047>

[9] Shi, X., Gu, L., Jiang, T., & Jiang, M. (2022). Retrieval of chlorophyll-a concentration based on Sentinel-2 images in inland lakes. <https://doi.org/10.1117/12.2631480>

[10] Shi, J., Shen, Q., Yao, Y., Li, J., Chen, F., Wang, R., Xu, W., Gao, Z., Wang, L., & Zhou, Y. (2022). Estimation of Chlorophyll-a Concentrations in Small Water Bodies: Comparison of Fused Gaofen-6 and Sentinel-2 Sensors. Remote Sensing, 14(1), 229–229. <https://doi.org/10.3390/rs14010229>

[11] Malahlela, O. E., Oliphant, T., Tsoeleng, L. T., & Mhangara, P. (2018). Mapping chlorophyll-a concentrations in a cyanobacteria- and algae-impacted Vaal Dam using Landsat 8 OLI data. South African Journal of Science, 114(9/10). <https://doi.org/10.17159/sajs.2018/4841>

[12] Cadondon, J. G., Ong, P. M. B., Vallar, E. A., Shiina, T., & Galvez, M. C. D. (2022). Chlorophyll-a Pigment Measurement of Spirulina in Algal Growth Monitoring Using Portable Pulsed LED Fluorescence Lidar System. Sensors, 22(8), 2940. <https://doi.org/10.3390/s22082940>

[13] Kovalevskaya, R. Z., Zhukava, H. A., & Adamovich, B. V. (2020). Modification of the Method of Spectrophotometric Determination of Chlorophyll A in the Suspended Matter of Water Bodies. Journal of Applied Spectroscopy, 87(1), 72–78. <https://doi.org/10.1007/s10812-020-00965-9>



- [14] Wang, C., Li, D., Zhang, L., Ding, Q., & Fu, Z. (2012). A Portable Measurement Instrument for the Measurement of Water Body Chlorophyll-a in the Support of Fluorescence Detection. IFIP Advances in Information and Communication Technology, 484–494. [https://doi.org/10.1007/978-3-642-27275-2\\_54](https://doi.org/10.1007/978-3-642-27275-2_54)
- [15] Markogianni, V., Kalivas, D., Petropoulos, G. P., & Dimitriou, E. (2020). Estimating Chlorophyll-a of Inland Water Bodies in Greece Based on Landsat Data. Remote Sensing, 12(13), 2087. <https://doi.org/10.3390/rs12132087>
- [16] Wang, L., Pu, H., & Sun, D.-W. (2016). Estimation of chlorophyll-a concentration of different seasons in outdoor ponds using hyperspectral imaging. Talanta, 147, 422–429. <https://doi.org/10.1016/j.talanta.2015.09.018>
- [17] Xiao, J., & Guo, Z. (2010). Detection of chlorophyll-a in urban water body by remote sensing. <https://doi.org/10.1109/iita-grs.2010.5602990>
- [18] Mudaliar, A., & Pandya, U. (2023). Assessment of Cyanobacterial Chlorophyll A as an Indicator of Water Quality in Two Wetlands Using Multi-Temporal Sentinel-2 Images. <https://doi.org/10.3390/ecws-7-14252>
- [19] Pan, Y., & Qiu, L. (2019). A Submersible in-Situ Highly Sensitive Chlorophyll Fluorescence Detection System. IOP Conference Series: Materials Science and Engineering, 677, 022065. <https://doi.org/10.1088/1757-899x/677/2/022065>
- [20] Basak, R., Wahid, K. A., & Dinh, A. (2021b). Estimation of the Chlorophyll-A Concentration of Algae Species Using Electrical Impedance Spectroscopy. Water, 13(9), 1223. <https://doi.org/10.3390/w13091223>
- [21] Avantika Latwal, Shaik Rehana, & Rajan, K. S. (2023). Detection and mapping of water and chlorophyll-a spread using Sentinel-2 satellite imagery for water quality assessment of inland water bodies. Environmental Monitoring and Assessment, 195(11). <https://doi.org/10.1007/s10661-023-11874-7>

- [22] Watanabe, F., Alcântara, E., Imai, N., Rodrigues, T., & Bernardo, N. (2018). Estimation of Chlorophyll-a Concentration from Optimizing a Semi-Analytical Algorithm in Productive Inland Waters. *Remote Sensing*, 10(2), 227. <https://doi.org/10.3390/rs10020227>
- [23] Yang, Z., Reiter, M., & Munyei, N. (2017). Estimation of chlorophyll-a concentrations in diverse water bodies using ratio-based NIR/Red indices. *Remote Sensing Applications: Society and Environment*, 6, 52–58. <https://doi.org/10.1016/j.rsase.2017.04.004>
- [24] Matthews, M. W., Bernard, S., & Robertson, L. (2012). An algorithm for detecting trophic status (chlorophyll-a), cyanobacterial-dominance, surface scums and floating vegetation in inland and coastal waters. *Remote Sensing of Environment*, 124, 637–652. <https://doi.org/10.1016/j.rse.2012.05.032>
- [25] Gregor, J., & Maršálek, B. (2004). Freshwater phytoplankton quantification by chlorophyll a: a comparative study of in vitro, in vivo and in situ methods. *Water Research*, 38(3), 517–522. <https://doi.org/10.1016/j.watres.2003.10.033>
- [26] Giardino, C., Pepe, M., Pietro Alessandro Brivio, Ghezzi, P. M., & Zilioli, E. (2001). Detecting chlorophyll, Secchi disk depth and surface temperature in a sub-alpine lake using Landsat imagery. *Science of the Total Environment*, 268(1-3), 19–29. [https://doi.org/10.1016/s0048-9697\(00\)00692-6](https://doi.org/10.1016/s0048-9697(00)00692-6)
- [27] Karimian, H., Huang, J., Chen, Y., Wang, Z., & Huang, J. (2023). A novel framework to predict chlorophyll-a concentrations in water bodies through multi-source big data and machine learning algorithms. *Environmental Science and Pollution Research*, 30(32), 79402–79422. <https://doi.org/10.1007/s11356-023-27886-2>
- [28] Randolph, K., Wilson, J., Tedesco, L., Li, L., Pascual, D. L., & Soyeux, E. (2008). Hyperspectral remote sensing of cyanobacteria in turbid productive water using

optically active pigments, chlorophyll a and phycocyanin. Remote Sensing of Environment, 112(11), 4009–4019. <https://doi.org/10.1016/j.rse.2008.06.002>

[29] Ansper, A., & Alikas, K. (2018). Retrieval of Chlorophyll a from Sentinel-2 MSI Data for the European Union Water Framework Directive Reporting Purposes. Remote Sensing, 11(1), 64. <https://doi.org/10.3390/rs11010064>

[30] Duan, H., Zhang, Y., Zhang, B., Song, K., & Wang, Z. (2006). Assessment of Chlorophyll-a Concentration and Trophic State for Lake Chagan Using Landsat TM and Field Spectral Data. Environmental Monitoring and Assessment, 129(1-3), 295–308. <https://doi.org/10.1007/s10661-006-9362-y>

## 5.2: REFERENCES LINK

- I. Basak, R., Wahid, K. A., & Dinh, A. (2021). Estimation of the Chlorophyll-A Concentration of Algae Species Using Electrical Impedance Spectroscopy. Water, 13(9), 1223. <https://doi.org/10.3390/w13091223>
- II. <https://www.flinnsci.com/api/library/Download/3ca645e9316b40059f2f73ec24de675b>
- III. Beer-Lambert Law | Transmittance & Absorbance  
Link: <https://www.edinst.com/blog/the-beer-lambert-law/>
- IV. The Beer-Lambert Law  
Link:  
[https://chem.libretexts.org/Bookshelves/Physical\\_and\\_Theoretical\\_Chemistry\\_Textbook\\_Maps/Supplemental\\_Modules\\_\(Physical\\_and\\_Theoretical\\_Chemistry\)/Spectroscopy/Electronic\\_Spectroscopy/Electronic\\_Spectroscopy\\_Basics/The\\_Beer-Lambert\\_Law](https://chem.libretexts.org/Bookshelves/Physical_and_Theoretical_Chemistry_Textbook_Maps/Supplemental_Modules_(Physical_and_Theoretical_Chemistry)/Spectroscopy/Electronic_Spectroscopy/Electronic_Spectroscopy_Basics/The_Beer-Lambert_Law)

### 5.3: APPENDIX

#### Features of Jasco- V-770 Spectrophotometer:

A wide range UV-Visible/Near Infrared Spectrophotometer with a unique optical design featuring a single monochromator and dual detectors for the wavelength range from 190 to 2700nm (3200nm option).

The V-770's single monochromator design provides for maximum light throughput with excellent absorbance linearity. A PMT detector is used for the UV to visible region and a Peltier-cooled PbS detector for the NIR region.

The V-770 UV-Visible/NIR spectrophotometer is operated using Spectra Manager™ Suite. This innovative cross-platform spectroscopy software is compatible with Windows 7 Pro (32- and 64-bit) and Windows 8.1 operating systems.

For simple operation, the handheld iRM has a great look and feel with a colour touch-sensitive screen. Data can also be downloaded to Spectra Analysis on a PC further PC data processing.

The V-700 Series has a growing list of software applications for both Spectra Manager™ and iRM. If you have an application which you don't see listed, please let us know as we may already have it or we can prepare an application designed specifically for your requirements.

<b>Optical System</b>	<b>Czerny-Turner grating mount Single monochromator Fully symmetrical double beam</b>
<b>Light source</b>	Halogen lamp, Deuterium lamp
<b>Wavelength range</b>	190 to 2700 nm (3200 nm option)
<b>Wavelength accuracy</b>	+/-0.3 nm (at 656.1 nm) +/-1.5 nm (at 1312.2 nm)

<b>Wavelength repeatability</b>	+/-0.05 nm (UV-Vis), +/-0.2 nm (NIR)
<b>Spectral bandwidth (SBW)</b>	UV-Visible: 0.1, 0.2, 0.5, 1, 2, 5, 10 nm L2, L5, L10 nm (low stray light mode) M1, M2 nm (micro cell mode)  NIR: 0.4, 0.8, 1, 2, 4, 8, 20, 40 L8, L20, L40 nm (low stray light mode) M4, M8 nm (micro cell mode)
<b>Stray light</b>	1 % (198 nm KCL) 0.0005 % (220 nm NaI) 0.0005 % (340 nm NaNO <sub>2</sub> ) 0.0005 % (370 nm NaNO <sub>2</sub> ) SBW: L2 nm  0.04 % (1420 nm: H <sub>2</sub> O) 0.1 % (1690 nm: CH <sub>2</sub> Br <sub>2</sub> ) SBW: L8 nm
<b>Photometric range</b>	UV-Visible: -4~4 Abs NIR: -3~3 Abs
<b>Photometric accuracy</b>	+/-0.0015 Abs (0 to 0.5 Abs) +/-0.0025 Abs (0.5 to 1 Abs) +/-0.3 %T Tested with NIST SRM 930D
<b>Photometric repeatability</b>	+/-0.0005 Abs (0 to 0.5 Abs) +/-0.0005 Abs (0.5 to 1 Abs) Tested with NIST SRM 930D
<b>Scanning speed</b>	10~4000 nm/min (8000 nm/min in preview mode)
<b>Slew speed</b>	UV-Vis: 12,000 nm/min NIR: 48,000 nm/min
<b>RMS noise</b>	0.00003 Abs (0 Abs, wavelength: 500 nm, measurement time: 60 sec, SBW: 2 nm)
<b>Baseline stability</b>	0.0003 Abs/hour (Wavelength: 250 nm, response: slow and SBW: 2 nm)
<b>Baseline flatness</b>	+/-0.0002 Abs (200 - 2500 nm)
<b>Detector</b>	PMT, Peltier cooled PbS
<b>Standard functions</b>	IQ accessories, Start button, Analog output
<b>Standard programs</b>	Abs/%T meter, Quantitative analysis, Spectrum measurement, Time course measurement, Fixed wavelength measurement, Validation, Daily check, Dual wavelength time course measurement
<b>Dimensions and weight</b>	460(W) x 602(D) x 268(H) mm, 29 kg
<b>Power requirements</b>	150 VA
<b>Installation requirements</b>	Room temperature: 15-30 Celsius, humidity: below 85%

## **CODE:**

### **1. CODE FOR DATA EXTRACTION**

```
import os

import pandas as pd

from google.colab import drive

# Mount Google Drive to access files

drive.mount('/content/drive')

# Input folder containing CSV files

csv_folder_path = '/content/drive/MyDrive/dataset collection csv folder/younger
alage dataset'

# Output CSV file path

csv_output_path = '/content/drive/MyDrive/younger algae.csv'

# Columns and rows to extract

columns_to_extract = ['Unnamed: 1']

rows_to_extract = list(range(18, 1221)) # Generate a range from 18 to 3018

df_list = []

column_offset = 0 # Initialize column offset
```

```
# Loop through all CSV files in the folder
```

```
for file in os.listdir(csv_folder_path):
```

```
    if file.endswith('.csv'):
```

```
        file_path = os.path.join(csv_folder_path, file)
```

```
        df = pd.read_csv(file_path)
```

```
        if all(column in df.columns for column in columns_to_extract):
```

```
            if all(row in df.index for row in rows_to_extract):
```

```
                filtered_df = df.loc[rows_to_extract, columns_to_extract]
```

```
                # Add an offset to the column index
```

```
                filtered_df.columns = [f"{col}_Offset_{column_offset}" for col in
filtered_df.columns]
```

```
                df_list.append(filtered_df)
```

```
                column_offset += 1
```

```
# Check if any dataframes were found
```

```
if df_list:
```

```
    # Concatenate all dataframes in the list along the columns (axis=1)
```

```
    combined_df = pd.concat(df_list, axis=1)
```



```
# Write the combined dataframe to a CSV file
```

```
combined_df.to_csv(csv_output_path, index=False)
```

```
print(f"Filtered data from the folder saved to '{csv_output_path}'.")
```

```
else:
```

```
print("No data found matching the specified columns and rows.")
```

## 2. REGRESSIONS CODES

```
from google.colab import drive
```

```
drive.mount('/content/drive')
```

```
import pandas as pd
```

```
import numpy as np
```

```
import numpy
```

```
import math
```

```
import matplotlib.pyplot as plt
```

```
from matplotlib import pyplot
```

```
from scipy import signal
```

```
from sklearn.svm import SVR
```

```
from sklearn.preprocessing import StandardScaler

from sklearn.model_selection import train_test_split

from sklearn.cross_decomposition import PLSRegression

from sklearn.model_selection import KFold, cross_val_score

from sklearn.metrics import r2_score

from sklearn.metrics import mean_squared_error

from sklearn.metrics import mean_absolute_error

dataset=pd.read_csv('/content/drive/MyDrive/final      dataset      files/machine
learning/Dataset.csv')

dataset

x = dataset

x

x.shape

x = dataset.drop(['per'], axis=1).to_numpy()

wavelength = x[0,0:1001]

wavelength

x= x[1:85,0:1001]
```

```
x
```

```
y= dataset.per.to_numpy()
```

```
y = y[1:85]
```

```
y
```

```
y.shape
```

```
# plotting the signal
```

```
pyplot.plot(wavelength, x.T)
```

```
pyplot.xlabel(' Wavelength')
```

```
pyplot.ylabel('Abs')
```

```
pyplot.title("Spectra")
```

```
pyplot.show()
```

```
#train and test
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,  
random_state=0)
```

```
#standard scalar
```

```
from sklearn.preprocessing import StandardScaler
```

```
scaler = StandardScaler()
```

```
scaler.fit(x_train)
```

```
x_train= scaler.transform(x_train)
```

```
x_test= scaler.transform(x_test)
```

```
k_folds = KFold(n_splits = 10)
```

```
# Calculate first derivative applying a Savitzky-Golay filter
```

```
X = signal.savgol_filter(x, window_length=91, polyorder=3, deriv=1)
```

```
f = pyplot.figure()
```

```
f.set_figwidth(7)
```

```
f.set_figheight(8)
```

```
print("Plot after re-sizing: ")
```

```
pyplot.plot(wavelength, x.T)
```

```
pyplot.xlabel(' Wavelength')
```

```
pyplot.ylabel('Abs')
```

```
pyplot.title("Spectra")
```

```
pyplot.show()
```

SUPPORT VECTOR REGRESSION

```
mod = SVR(C=10, epsilon=0.2)

cv_scores = cross_val_score(mod, x_train, y_train, cv=k_folds)

mod.fit(x_train, y_train)

y_pred = mod.predict(x_test)

r = r2_score(y_test, y_pred)

print("Root Square:")

print(r)

MSE= mean_squared_error(y_test, y_pred)

print("Mean Square Error:")

print(MSE)

y_test = np.array(y_test).astype(float)

y_pred = np.array(y_pred).astype(float)

MSE = np.square(np.subtract(y_test, y_pred)).mean()

RMSE = math.sqrt(MSE)

print("Root Mean Square Error:")

print(RMSE)

y_pred
```

```
import numpy as np

import matplotlib.pyplot as plt

fig,ax = plt.subplots(1)

# plot the data

ax.scatter(y_test,y_pred,color="red", marker="o",)

m, b = np.polyfit(y_test, y_pred, 1)

#add linear regression line to scatterplot

plt.plot(y_test, m*y_test+b)

plt.xlabel('Actual')

plt.ylabel('Predicted concentraton in mg/L')

plt.title("Prediction of chlorophyll A")

plt.show()

from sklearn.datasets import make_regression

make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)

clf = SVR(C=10, epsilon=0.2)

kf = KFold(n_splits=10, shuffle=True, random_state=42)

rmse = 0
```

```
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):

    print(f"Fold {i}:")

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = x_pca[train_index], x_pca[test_index]

    y_train, y_test = y[train_index], y[test_index]

    X_train = X_train.astype(np.float64)

    X_test = X_test.astype(np.float64)

    y_test = y_test.astype(np.float64)

    y_train = y_train.astype(np.float64)

    clf.fit(X_train, y_train)

    ypred = clf.predict(X_test)

    ypred=np.array(ypred).flatten()

    aa=np.array(y_test).flatten()

    mat_plot(aa,ypred)

    rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))

#df = pd.DataFrame(clf.cv_results_)

#df
```

```
rmse = rmse /10
```

```
print("average RMSE",rmse)
```

```
#partial least square regression
```

```
from sklearn.cross_decomposition import PLSRegression
```

```
pls2 = PLSRegression(n_components=5)
```

```
pls2.fit(x_train, y_train)
```

```
PLSRegression()
```

```
Y_pred = pls2.predict(x_test)
```

```
from sklearn.metrics import r2_score
```

```
r = r2_score(y_test, Y_pred)
```

```
print("Root Square:")
```

```
print(r)
```

```
from sklearn.metrics import mean_squared_error
```



```
MSE= mean_squared_error(y_test, Y_pred)

print("Mean Square Error:")

print(MSE)

import math

y_test = np.array(y_test).astype(float)

y_pred = np.array(y_pred).astype(float)


MSE = np.square(np.subtract(y_test, Y_pred)).mean()

RMSE = math.sqrt(MSE)


print("Root Mean Square Error:")

print(RMSE)

fig,ax = plt.subplots(1)

# plot the data

ax.scatter(y_test, Y_pred,color="blue", marker="o",)

m, b = np.polyfit(y_test, Y_pred, 1)
```

```
#add linear regression line to scatterplot
```

```
plt.plot(y_test, m*y_test+b)
```

```
plt.xlabel('Actual')
```

```
plt.ylabel('Predicted concentraton in mg/L')
```

```
plt.title("Prediction of chlorophyll A")
```

```
plt.show()
```

```
from sklearn.datasets import make_regression
```

```
make_regression(n_features=4, n_informative=2, random_state=0, shuffle=False)
```

```
clf = PLSRegression(n_components=5)
```

```
kf = KFold(n_splits=10, shuffle=True, random_state=42)
```

```
rmse = 0
```

```
for i, (train_index, test_index) in enumerate(kf.split(x_pca)):
```

```
    print(f"Fold {i}:")
```

```
    #print("TRAIN:", train_index, "TEST:", test_index)
```

```
    X_train, X_test = x_pca[train_index], x_pca[test_index]
```

```
    y_train, y_test = y[train_index], y[test_index]
```

```
X_train = X_train.astype(np.float64)

X_test = X_test.astype(np.float64)

y_test = y_test.astype(np.float64)

y_train = y_train.astype(np.float64)

clf.fit(X_train, y_train)

ypred = clf.predict(X_test)

ypred=np.array(ypred).flatten()

aa=np.array(y_test).flatten()

mat_plot(aa,ypred)

rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))

#df = pd.DataFrame(clf.cv_results_)

#df

rmse = rmse /10

print("average RMSE",rmse)
```

## DECISION TREE REGRESSION

```
import matplotlib.pyplot as plt

import numpy as np
```

```
from sklearn import tree

from sklearn.tree import DecisionTreeRegressor

clf = tree.DecisionTreeRegressor()

from sklearn.datasets import load_diabetes

from sklearn.model_selection import cross_val_score

from sklearn.tree import DecisionTreeRegressor

regressor = DecisionTreeRegressor(random_state=0)

cross_val_score(regressor,x_test, y_test, cv=3)

reg = DecisionTreeRegressor(max_depth=50)

clf = tree.DecisionTreeRegressor()

clf = clf.fit(x_train, y_train)

D_pred = clf.predict(x_test)

from sklearn.metrics import r2_score

r = r2_score(y_test, D_pred)

print("Root Square:")

print(r)

from sklearn.metrics import mean_squared_error
```

```
MSE= mean_squared_error(y_test, D_pred)

print("Mean Square Error:")

print(MSE)

import math

y_test = np.array(y_test).astype(float)

y_pred = np.array(y_pred).astype(float)

MSE = np.square(np.subtract(y_test, D_pred)).mean()

RMSE = math.sqrt(MSE)

print("Root Mean Square Error:")

print(RMSE)

fig,ax = plt.subplots(1)

# plot the data

ax.scatter(y_test, D_pred,color="blue", marker="o",)

, b = np.polyfit(y_test, D_pred, 1)

#add linear regression line to scatterplot

plt.plot(y_test, m*y_test+b)

plt.xlabel('Actual')
```

```
plt.ylabel('Predicted concentraton in mg/L')

plt.title("Prediction of chlorophyll A")

plt.show()

from sklearn.datasets import make_regression

make_regression(n_features=4, n_informative=2, random_state=0, shuffle=False)

clf = tree.DecisionTreeRegressor()

kf = KFold(n_splits=10, shuffle=True, random_state=42)

rmse = 0

for i, (train_index, test_index) in enumerate(kf.split(x_pca)):

    print(f"Fold {i}:")

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = x_pca[train_index], x_pca[test_index]

    y_train, y_test = y[train_index], y[test_index]

    X_train = X_train.astype(np.float64)

    X_test = X_test.astype(np.float64)

    y_test = y_test.astype(np.float64)

    y_train = y_train.astype(np.float64)
```

```
clf.fit(X_train, y_train)

ypred = clf.predict(X_test)

ypred=np.array(ypred).flatten()

aa=np.array(y_test).flatten()

mat_plot(aa,ypred)

rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))

#df = pd.DataFrame(clf.cv_results_)

#df

rmse = rmse /10

print("average RMSE",rmse)
```

## RANDOM FOREST REGRESSION

```
from sklearn.ensemble import RandomForestRegressor

from sklearn.datasets import make_regression

make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)

regr = RandomForestRegressor(max_depth=2, random_state=0)

regr.fit(x_train, y_train)

R_pred=regr.predict(x_test)
```

```
from sklearn.metrics import r2_score

r = r2_score(y_test, R_pred)

print("Root Square:")

print(r)

from sklearn.metrics import mean_squared_error

MSE= mean_squared_error(y_test, R_pred)

print("Mean Square Error:")

print(MSE)

import math

y_test = np.array(y_test).astype(float)

y_pred = np.array(y_pred).astype(float)

MSE = np.square(np.subtract(y_test, R_pred)).mean()

RMSE = math.sqrt(MSE)

print("Root Mean Square Error:")

print(RMSE)

fig,ax = plt.subplots(1)
```



```
# plot the data

ax.scatter(y_test, R_pred,color="blue", marker="o",)

m, b = np.polyfit(y_test, R_pred, 1)

#add linear regression line to scatterplot

plt.plot(y_test, m*y_test+b)

plt.xlabel('Actual')

plt.ylabel('Predicted concentraton in mg/L')

plt.title("Prediction of chlorophyll A")

plt.show()

from sklearn.datasets import make_regression

make_regression(n_features=4, n_informative=2,random_state=0, shuffle=False)

clf = RandomForestRegressor(max_depth=2, random_state=0)

kf = KFold(n_splits=10, shuffle=True, random_state=42)

rmse = 0

for i, (train_index, test_index) in enumerate(kf.split(x_pca)):

    print(f"Fold {i}:")

    #print("TRAIN:", train_index, "TEST:", test_index)
```

```
X_train, X_test = x_pca[train_index], x_pca[test_index]

y_train, y_test = y[train_index], y[test_index]

X_train = X_train.astype(np.float64)

X_test = X_test.astype(np.float64)

y_test = y_test.astype(np.float64)

y_train = y_train.astype(np.float64)

clf.fit(X_train, y_train)

ypred = clf.predict(X_test)

ypred=np.array(ypred).flatten()

aa=np.array(y_test).flatten()

mat_plot(aa,ypred)

rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))

#df = pd.DataFrame(clf.cv_results_)

#df

rmse = rmse /10

print("average RMSE",rmse)
```

## K-NEAREST NEIGHBOR REGRESSION

```
# Import necessary libraries
```

```
from sklearn.decomposition import PCA
```

```
'''
```

```
from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(x_smooth, y, test_size=0.1,  
random_state=0)'''
```

```
# Apply PCA
```

```
pca = PCA(n_components = 5)
```

```
X_pca = pca.fit_transform(X)
```

```
# Explained variance
```

```
explained_variance = pca.explained_variance_
```

```
total_explained_variance = explained_variance.sum()
```

```
# Print results
```

```
print(f"Explained Variance:\n{explained_variance}")
```

```
print(f"Total Explained Variance: {total_explained_variance:.4f}")
```

```
# Explained variance ratio
```

```
explained_variance_ratio = pca.explained_variance_ratio_
```

```
total_explained_variance_ratio = explained_variance_ratio.sum()
```

```
# Print results
```

```
print(f"\nExplained Variance Ratio:\n{explained_variance_ratio}")
```

```
print(f"Total Explained Variance Ratio: {total_explained_variance_ratio:.4f}")
```

```
# Import necessary libraries
```

```
import numpy as np
```

```
import matplotlib.pyplot as plt
```

```
# Plot explained variance ratio
```

```
cumulative_variance_ratio = np.cumsum(explained_variance_ratio)
```

```
f = plt.figure()
```

```
f.set_figwidth(10)
```

```
f.set_figheight(10)
```

```
plt.plot(cumulative_variance_ratio, marker='o')
```

```
plt.xlabel('Number of Principal Components')
```

```
plt.ylabel('Cumulative Explained Variance Ratio')
```

```
plt.title('Cumulative Explained Variance Ratio by Principal Components')
```

```
plt.show()
```

```
!pip install matplotlib

def mat_plot(a,b): # qq1 is the actual readings andthe b is the predictd

    from sklearn.metrics import mean_absolute_error

    from sklearn.metrics import mean_squared_error

    from sklearn.metrics import r2_score

    import numpy as np

    import matplotlib.pyplot as plt

    print("The R2 ", (r2_score( a, b)))

    print("RMSE:", np.sqrt(mean_squared_error( a, b)))

    #print("MAPE%:", mean_absolute_percentage_error( a, b))

    print("MAE",mean_absolute_error(a,b))

    print("MSE",mean_squared_error(a,b))

    print("RMSE",np.sqrt(mean_squared_error(a,b)))

    fig, ax = plt.subplots()

    ax.plot(b, a, linewidth=0, marker="o", color='C0', markersize=8)

    #plot(x, y, color='green', linestyle='dashed', marker='o', markerfacecolor='blue',
    markersize=12).
```

```
low_x, high_x = ax.get_xlim()

low_y, high_y = ax.get_ylim()

low = max(low_x, low_y)

high = min(high_x, high_y)

ax.plot([low, high], [low, high], ls="-", c=".2", alpha=.4)

#ax.set_title('R2 score')

plt.rcParams.update({'font.size': 20})

ax.set_xlabel("Actual")

ax.set_ylabel("Predicted ")

plt.show()

import matplotlib

import numpy as np

from sklearn.model_selection import KFold

from sklearn.neighbors import KNeighborsRegressor

from sklearn.metrics import mean_squared_error

# Apply PCA

pca = PCA(n_components = 2)
```

```
x_pca = pca.fit_transform(X)

rmse = 0

kf = KFold(n_splits=10, shuffle=True, random_state=42)

for i, (train_index, test_index) in enumerate(kf.split(x_pca)):

    print(f"Fold {i}:")

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = x_pca[train_index], x_pca[test_index]

    y_train, y_test = y[train_index], y[test_index]

    X_train = X_train.astype(np.float64)

    X_test = X_test.astype(np.float64)

    y_test = y_test.astype(np.float64)

    y_train = y_train.astype(np.float64)

    neigh = KNeighborsRegressor( n_neighbors = 1, p = 1, weights = 'distance',
algorithm = 'auto' )

    neigh.fit(X_train, y_train)

    y_pred=neigh.predict(X_test)

    y_pred=np.array(y_pred).flatten()
```

```
qq1=np.array(y_test).flatten()

print( "*****")

#plt.plot(qq1,y_pred)

mat_plot(qq1,y_pred)

rmse = rmse + np.sqrt(mean_squared_error(qq1,y_pred))

print( "Result for each fold " + str(i))

print( "*****")

rmse = rmse /10

print("average RMSE",rmse)

from sklearn.model_selection import cross_val_score

from sklearn.neighbors import KNeighborsRegressor

from sklearn.metrics import mean_squared_error

# Import necessary libraries

from sklearn.decomposition import PCA

# Apply PCA

pca = PCA(n_components = 5)

x_pca = pca.fit_transform(X)
```



```
'''
```

```
from sklearn.model_selection import train_test_split
```

```
x_train, x_test, y_train, y_test = train_test_split(x_pca, y, test_size=0.1,  
random_state=42, shuffle= True)'''
```

```
neigh = KNeighborsRegressor()
```

```
x_pca = x_pca.astype(np.float64)
```

```
y = y.astype(np.float64)
```

```
'''
```

```
y_train = y_train.astype(np.float64)
```

```
y_test = y_test.astype(np.float64)
```

```
'''
```

```
from sklearn.model_selection import KFold
```

```
from sklearn.model_selection import GridSearchCV
```

```
parameters = {'n_neighbors':list(range(1, 5)) , 'p':list(range(1, 5)) ,  
'weights':['distance'], 'algorithm': ['auto']}
```

```
from sklearn.model_selection import GridSearchCV
```

```
kf = KFold(n_splits=10, shuffle=True, random_state=42)
```

```
clf = GridSearchCV(neigh, parameters, cv = kf , return_train_score=False, scoring=
'neg_mean_absolute_error') # scoring='neg_mean_squared_error

clf.fit(x_pca, y)

'''

#clf.cv_results_

ypred = clf.predict(x_test)

ypred=np.array(ypred).flatten()


aa=np.array(y_test).flatten()

mat_plot(aa,ypred)

'''

kf = KFold(n_splits=10, shuffle=True, random_state=42)

rmse = 0

for i, (train_index, test_index) in enumerate(kf.split(x_pca)):

    print(f"Fold {i}:")

    #print("TRAIN:", train_index, "TEST:", test_index)

    X_train, X_test = x_pca[train_index], x_pca[test_index]
```

```
y_train, y_test = y[train_index], y[test_index]

X_train = X_train.astype(np.float64)

X_test = X_test.astype(np.float64)

y_test = y_test.astype(np.float64)

y_train = y_train.astype(np.float64)

clf.fit(X_train, y_train)

ypred = clf.predict(X_test)

ypred=np.array(ypred).flatten()

aa=np.array(y_test).flatten()

#plt.plot(qq1,y_pred)

mat_plot(aa,ypred)

rmse = rmse + np.sqrt(mean_squared_error(aa,ypred))

df = pd.DataFrame(clf.cv_results_)

df

rmse = rmse /10

print("average RMSE",rmse)
```