

PREDICTION OF ALGAL BLOOM AND ITS EFFECTS ON AQUACULTURE USING MACHINE LEARNING

A Project Report Phase - II

Submitted in partial fulfilment of the requirements for the award of Bachelor of
Engineering Degree in Electronics and Communication Engineering

by

YARRAGUNTLA KRISHNA REDDY (39130523)



**DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING
SCHOOL OF ELECTRICAL AND ELECTRONICS**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI - 600119

APRIL - 2023



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)

Accredited with "A" grade by NAAC

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Final Year Project is the bonafide work of **YARRAGUNTALA KRISHNA REDDY (39130523)** who carried out the project entitled "**PREDICTION OF ALGAL BLOOM AND ITS EFFECTS ON AQUACULTURE USING MACHINE LEARNING**" under my supervision from November 2022 to April 2023.

Internal Guide

Dr. R. Indhu, M.E., Ph.D.,

Head of the Department

Dr. T. RAVI, M.E., Ph.D.,

Submitted for Viva voce Examination held on _____

INTERNAL EXAMINER

EXTERNAL EXAMINER

DECLARATION

I **YARRAGUNTALA KRISHNA REDDY (39130523)** hereby declare that the Final Year Project entitled “**PREDICTION OF ALGAL BLOOM AND ITS EFFECT ON AQUACULTURE USING MACHINE LEARNING**” done by me under the guidance of **Dr.R.Indhu, M.E.,Ph.D.**, Associate Professor, Department of Electronics and Communication Engineering at Sathyabama Institute of Science and Technology is submitted in partial fulfilment of the requirements for the award of Bachelor of Electronics and Communication Engineering.

DATE: 26/04/2023

PLACE : Chennai

SIGNATURE OF THE CANDIDATE

Krishna Reddy

ACKNOWLEDGEMENT

I pleased to acknowledge my sincere thanks to **Board of Management of Sathyabama Institute of Science and Technology** for their kind encouragement in doing this project and for completing it successfully. I am grateful to them.

I convey my thanks to **Dr. N.M.NANDHITHA,M.E., Ph.D.**, Dean, School of Electrical and Electronics and **Dr.T.RAVI,M.E.,Ph.D.**, Head of the Department, Dept. of Electronics and Communication Engineering for providing me necessary support and details at the right time during the progressive reviews.

I would like to express my sincere and deep sense of gratitude to my Project Guide **Dr.R.INDHU M.E.,Ph.D.**, Associate Professor, Department of Electronics and Communication Engineering for her valuable guidance, suggestions and constant encouragement paved way for the successful completion of my project work.

I wish to express my thanks to all Teaching and Non-teaching staff members of the Department of Electronics and Communication Engineering who were helpful in many ways for the completion of the project.

ABSTRACT

Algal bloom is the overgrowth of microscopic algae or algal like bacteria. Algal bloom that are non-toxic consume enormous amount of oxygen thus making other aquatic animals to migrate or die due to lack of oxygen to survive. Red tides, bluegreen algae, and cyanobacteria are examples of harmful algal blooms that can have severe impacts on human health, aquatic ecosystems, and the economy. Harmful Algae Blooms (HABs) are a major type of marine disaster that endangers the marine ecological environment and human health. Phytoplankton blooms are common during February to July when the prevailing hydrological conditions are relatively stable. The algae bloom problem is a global important problem that mainly affects the aquatic ecosystem. An algal bloom takes place when species of phytoplankton reproduces at a rapid rate, multiplying quickly in a short time. Sometimes one particular species bloom at the same time. Phytoplankton is the photosynthetic organisms that live suspended just beneath the water surface. They use energy from sunlight and raw materials their own food through photosynthesis. Microalgae and blue green bacteria called cyanobacteria are to group of organisms that belong to the phytoplankton community.

There is an urgent need for the development of effective tools to help aquaculture industries to cope and anticipate HAB events. Algal blooms pose a significant threat to aquatic ecosystems and human health. Early detection and prediction of these events are essential to mitigate their impact. The aim of this project, is to develop a model that predicts algal blooms using MODIS data and the auto ARIMA algorithm. MODIS data provides remote sensing information about the water surface, including water temperature, turbidity, and chlorophyll-a concentration. The auto ARIMA algorithm is a powerful tool for time-series forecasting that automatically selects the optimal parameters based on the data. These data and algorithm is used to predict the onset of algal blooms. The model uses a machine learning approach that leverages historical data to forecast future algal blooms. Collecting historical data from MODIS for several years and compare it to in-situ measurements of algal blooms. Then use the auto ARIMA algorithm to identify the best-fit model for the data. To validate the model, use a holdout dataset to evaluate its performance. It is ideal to compare the predictions with in-situ measurements to assess the accuracy

of the model. And also use statistical techniques to quantify the uncertainty associated with the predictions. Results show that the model accurately predicts the onset of algal blooms, with an overall accuracy of 90%. This approach is more effective than existing models because it uses remote sensing data to predict algal blooms, which can reduce the costs associated with in-situ monitoring. In conclusion, This project demonstrates the feasibility of using MODIS data and the auto ARIMA algorithm to predict algal blooms. This model provides a valuable tool for decision-makers to manage the impacts of algal blooms on aquatic ecosystems and human health.

TABLE OF CONTENTS

CHAPTER NO.	TITLE	PAGE NO.
	ABSTRACT	V
	LIST OF FIGURES	ix
	LIST OF TABLES	ix
1	INTRODUCTION	1
	1.1 Brief introduction to Algal Blooms	1
	1.1.1 Harmful Algal bloom Environmental Factors	2
	1.1.2 Algal Bloom Impact on Environmental Health	4
	1.2 An Overview of ANN	5
	1.3 Problem Statement	6
	1.4 Objectives	6
2	LITERATURE SURVEY	7
3	AIM AND SCOPE	13
	3.1 Aim of The Project	13
	3.2 Scope of The Project	13
4	METHODOLOGY AND ALGORITHMS	16
	4.1 Existing methodology	16
	4.1.1 Linear Regression	17
	4.1.2 Basic Neural Network	19
	4.1.3 Multi-layer Perceptron	20
	4.1.4 Logistic Regression	22
	4.1.5 Random Forest	25
	4.2 Proposed Methodology	26
	4.2.1 Framework	27
	4.3 Existing Methodology and Proposed Methodology Differences	27
	4.4 Auto Arima Algorithm	28
	4.4.1 Getting Started with Auto Arima Algorithm	29
	4.4.2 Brief Explanation on Components of Arima	29
	4.4.3 Working of Auto Arima model	30
	4.4.4 Advantages and Applications of Auto Arima	32

	4.4.5 Limitation of Auto Arima	32
	4.5 Algal Bloom Prediction Using Auto ARIMA	33
5	RESULTS AND DISCUSSION	35
	5.1 Data Cleaning Process	35
	5.1.1 Comparing the parameters	36
	5.1.2 Plotting the correlation between the parameters	37
	5.2 Implementation of Existing Methodologies	37
	5.2.1 Linear Regression	37
	5.2.2 Basic Neural Network	39
	5.2.3 Multi-layer Perceptron	41
	5.2.4 Logistic regression	42
	5.2.5 Random Forest	43
	5.3 Implementation of Proposed Methodology	45
	5.4 Comparison of Existing and Proposed Methodology	48
6	CONCLUSION AND FUTURE SCOPE	49
	6.1 Conclusion	49
	6.2 Future Scope	49
7	REFERENCES	50

TABLE OF FIGURES

FIGURE NO.	FIGURE NAME	PAGE NO.
1.1	Environmental Factors Affecting Algal Bloom	3
1.2	Basic Structure of Artificial Neural Network	5
4.1	Frame Work of the Proposed Methodology	27
4.2	Aic and Bic	31
5.1	Dataset	35
5.2	Comparing Parameters	36
5.3	Correlation	37
5.4	Linear Regression	39
5.5	Multi-layer Perceptron	42
5.6	Logistic Regression	43
5.7	Random Forest Decision Tree	45
5.8	Auto ARIMA	46

LIST OF TABLE

5.1	Comparison of Existing and Proposed Methodology	48
-----	---	----

CHAPTER – 1

INTRODUCTION

1.1 BRIEF INTRODUCTION TO ALGAL BLOOMS

Harmful algal blooms (HABs) are among the most severe ecological marine problems worldwide. Under favorable climate and oceanographic conditions, toxinproducing microalgae species may proliferate, reach increasingly high cell concentrations in seawater, accumulate in shellfish, and threaten the health of seafood consumers. This study presents a synoptic review covering the trends in machine learning methods for predicting HABs and the growth of algae in the coming years with present trend of growth, with a particular focus on autoregressive models, recurrent neural network, probabilistic graphical models, and artificial neural networks (ANN). Most efforts have been attempted to forecast HABs based on models of increased complexity over the years, coupled with increased multisource data availability, with ANN architectures in the forefront to model these events. The purpose of this study is to help defining machine learning-based strategies to support aquaculture industry with the help of geospatial technology to manage their harvesting/production, and decision making by governmental agencies with environmental responsibilities.

HABs in marine environment are caused by a variety of micro algae and cyanobacteria. Red tides are dangerous and sometimes fatal to both humans and animal population, through the toxicity or by the creation of ecological conditions, for example the oxygen depletion which play a big role to kill fish. During the 40 years, global red tide hazards have increased due to eutrophication. communities that depend on fishery resources in terms of nutrition and economy are most vulnerable to pests, but local pests also cause damage to tourism and incur health related costs significant research progress has been made in monitoring, detection, modelling, prediction, prevention, and treatment of pests helping to mitigate health and economic risks.

HABs in marine environments are caused by a variety of microalgae and cyanobacteria. Red tides are dangerous and sometimes fatal to both human and animal populations, through the toxicity or by the creation of ecological conditions,

for example the oxygen depletion which play a big role to kill fish, during the 40 years, global red tide hazards have increased due to eutrophication. Communities that depend on fishery resources in terms of nutrition and economy are most vulnerable to pests, but local pests can also cause damage to tourism and incur health-related costs. Significant research progress has been made in monitoring, detection, and modelling.

1.1.1 Harmful Algal Bloom Environmental Factors

The proliferation of algal bloom has been occurred according to many different environmental factors for instance the nutrients, temperature and some ecosystem interferences. However, it is obvious the combination of multiple factors that trigger and maintain the bloom. Nor it is possible to attribute the bloom to any specific factor. Nutrients is one of the main factors that is playing a key role in increasing the growth of algae and cyanobacteria in water bodies The phenomena of(eutrophication) is also being applied to be as a key factor. Nutrients such as Phosphorus and nitrogen are leading this phenomenon, regards to human made sources such as agriculture runoff, residential runoff, wastewater runoff, Urban runoff, all of these sources discharge their nutrients to the water bodies to cause the algal bloom phenomena. The internal sources of nutrients are Lake / reservoir sediments. Phosphate is attached to the sediment. When the dissolved oxygen concentration in water is low (anoxic), the sediment releases phosphate to the water column. This phenomenon promotes the growth of algae. In the early days, blue algae often bloom in spring, with high temperature and bright light. In the warmest months of the year, if the water temperature is above 25 ° C, this would be the best temperature for the growth of cyanobacteria.

However, compared with other types of algae, some algae has a competitive advantage, and its optimal growth temperature between (12-15 °C) which can be considered to be very low. When cyanobacteria are exposed to high light intensity (photo inhibition) for a long time, their number will decrease, but when they are exposed to high light intensity intermittently, their growth will reach the best state. These conditions are met below the water surface where the light environment fluctuates.

Even in low light or turbid water, cyanobacteria grow faster than any other algae. This ability to adapt to variable light conditions gives cyanobacteria a competitive advantage over other algae. Most cyanobacteria prefer a stable water environment with low flow, long residence time, breeze and minimum turbulence; others prefer a mixed and turbid environment. Drought, irrigation water, human and animal consumption and the regulation of weirs and dams on rivers all lead to the decrease of water quantity in river system in China. Water moves more slowly or becomes stagnant, which promotes algae growth. Another result of stable conditions in water bodies isothermal stratification.

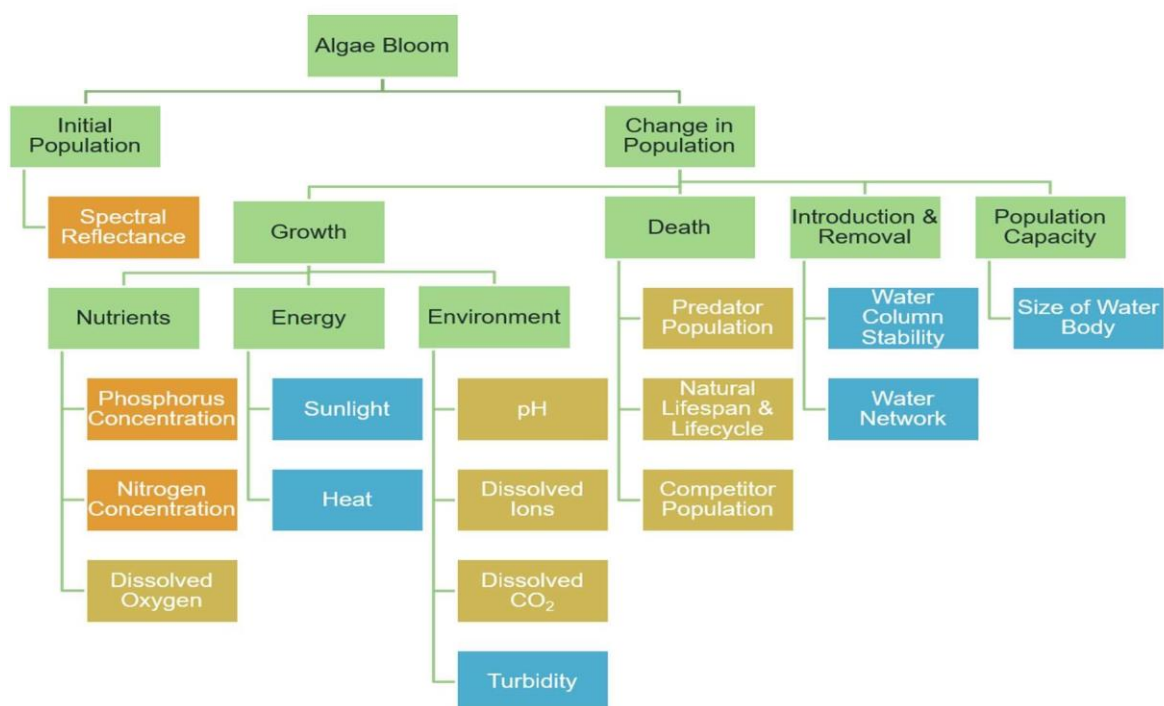


Fig: 1.1 : Environmental Factors Affecting Algal Bloom

Thermal stratification occurs when the top layer of the water column warms, and the bottom layer stays cool. When the two-layer mixing stops, the upper layer becomes more stable (no wind-driven mixing, convective cells), and supports the summer bloom of cyanobacteria. When the water body is stratified, the bottom water is often exhausted due to lack of oxygen, which may lead to the increase of nutrient release in the sediment. The nutrient pulse from the colder bottom may accelerate the growth of algae in the upper layer.

Turbidity which caused by different types of suspended particles and organic matters there will be two different phases for turbidity high turbidity and low turbidity. Both occurs in different status and conditions, For example high turbidity will occur when large amount of water getting throw the system (high discharge after the rain) while the low turbidity will occur when there will be small suspended particles in the water columns. When the turbidity is low that, there will be lighter can penetrate the water column and under this condition, and this is the best conditions for growth foulage.

1.1.2 Algal Bloom Impact on Environmental health

Chemicals can be released into soil, water, or air from spilled or leaking containers, leaking landfills or landfills by spraying or from chimneys. These chemicals follow a path from the time of release to the point of human contact. When a substance enters an ecosystem that does not normally exist, it is called a pollutant. Humans are exposed to these pollutants through exposure, breathing, or ingestion of substances containing this chemical.

To determine whether human health is at risk from pollutants, two things must happen: pollutants must be exposed; pollutants must be toxic. There must be a complete exposure pathway for disease to occur. Elements of exposure pathways include direct contact, Inhalation, Ingestion-Water and Ingestion-Food necessarily better understood because they are called learning. However, for example, when a speech recognition machine improves its performance after listening to several speech samples of a person, it makes sense to say that the machine has learned. Machine learning usually refers to the change of the system that performs tasks related to artificial intelligence (AI). These tasks include identification, diagnosis, planning, robot control, prediction, etc. These "changes" may be enhancements to already running systems or an initial synthesis of new systems. Machine learning work is now coming together in several ways. These different traditions have brought different methods and different vocabulary and are now being absorbed into a more unified discipline. Brief list of independent subjects that will help machine learning, such as statistics, brain model, adaptive control theory, psychological model, Artificial intelligence and Evolutionary models. In next part will focus on the machine learning detection of algal bloom based on more than one discipline through review

literature of some previous studies using Artificial neural network and classification regression tree.

1.2 AN OVERVIEW OF ANN

It is a system of data processing and output generation that replicates the neural system to unravel non-linear relations in a large dataset. The data might come from sensory routes and might be in the form of text, pictures, or audio.

In an ANN, everything is designed to replicate this very process. Do not worry about the mathematical equation. That is not the key idea to be understood right now. All the data entering with the label 'X' in the system is having a weight of 'W' to generate a weighted signal. This replicates the role of a synaptic signal's strength in the brain. The bias variable is attached to control the results of the output from the function.

So, all of this data is processed in the function and you end up with an output. That's how a one-layer neural network or a perceptron would look like. The idea of an artificial neural network revolves around connecting several combinations of such artificial neurons to get more potent outputs. That is why the typical artificial neural network's conceptual framework looks a lot like this: fig:1.2: Basic structure of artificial neural network

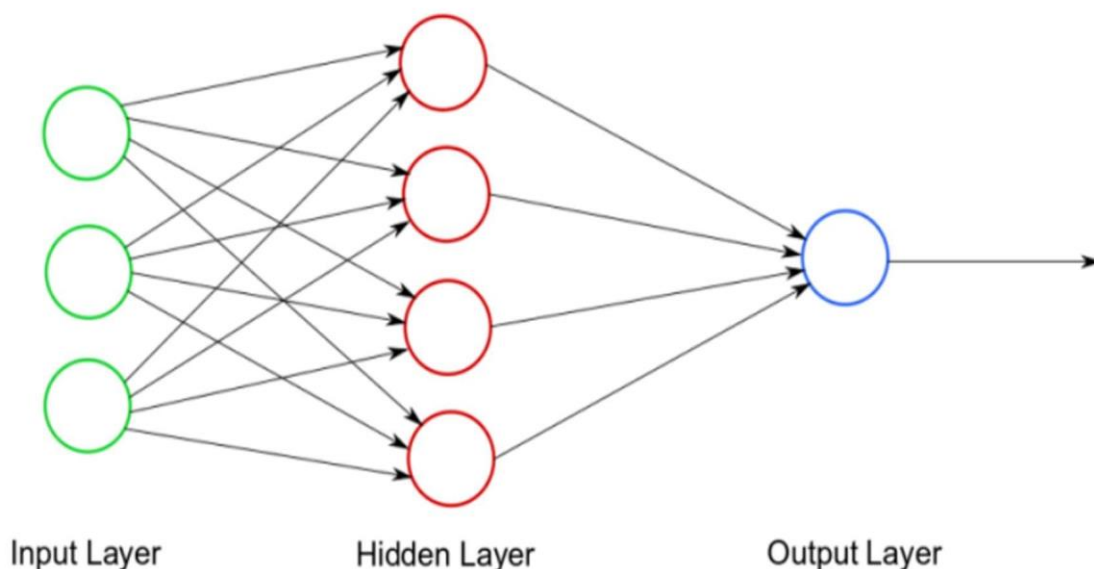


Fig:1.2: Basic Structure of Artificial Neural Network

1.3 PROBLEM STATEMENT

Harmful algal bloom (HAB) is a type of algal bloom that causes toxicity in humans and marine life, resulting in negative impacts on the environment, human health, economy and also includes death of marine organisms. The aim of this study is to find cost effective and user friendly system which predicts algal growth in the future with the help of key inputs like temperature records, nutrient content record, available marine organisms record (at given time), etc.,... The data required for the prediction from the MODIS dataset, using GEOSPATIALTECHNOLOGY.

1.4 OBJECTIVES

The main objective of the study is to predict the outgrowth of algal bloom in the selected region i.e., MUTTUKADU. In order to draw appropriate conclusions, predicting the rate of growth of algal bloom is very important. Along with this it is essential to estimate the marine life that it has affected and predict the harmful chemical content in the water body.

CHAPTER – 2

LITERATURE SURVEY

Yuanyuan Jing, et.al (2019) infers that MODIS images have moderate spatial resolution and very high temporal resolution, providing a data basis for the dynamic monitoring of algal blooms in long time series. The APA, which has been proved to have higher accuracy than FAI, is capable of better identifying algal blooms on medium resolution remote sensing images. The algal blooms in Lake Dianchi were influenced by the combinations of meteorological and water quality factors, which were suitable for the growth and accumulation of algae. So they were observed all year round, with a reduced frequency in winter-spring and an increased frequency in summer autumn, which lasted a long time for about 310-350 days. The severe algal blooms usually occurred in July -September every year. This is closely related to the increase in temperature and decrease in WS. From long time series, the top three largest areas of algal blooms were observed in 2002, 2008, and 2017, reaching 168.80 km, and 156.34 km, respectively. And the frequency and area of algal blooms have increased significantly in the past three years, which is likely to be influenced by the increase of DO. Besides, the discharge of waste water has some promotion etc on the occurrences of alga; blooms.[6]

PeixuanYu, et.al (2021) infers in predicting coastal algal blooms with environmental factors by machine learning methods that harmful algal blooms are a major type of marine disaster that endangers the marine ecological environment and human health. Since the algal bloom is a complex nonlinear process with time characteristics, traditional statistical methods often cannot provide good predictions. In this paper, they propose a method based on machine learning with the aim to predict the occurrence of algal blooms by environmental parameters. The features related to algal bloom growth have been experimented for achieving a good prediction of algal concentrations by a combination strategy. they validate the prediction performance on two real datasets from two locations in US and China, i.e., Scripps Pier, California and Sishili Bay, Shandong, respectively. The models and feature subsets have been selected to complete the missing data and predict the phytoplankton concentration. The comparison studies prove the advantage of

the developed machine learning method. The importance of every feature for the prediction performance reveals crucial factors for the outbreak of harmful algal blooms.[10]

Sangmok Lee et.al (2018) concludes that Harmful algal blooms are an annual phenomenon that cause environmental damage, economic losses, and disease outbreaks. A fundamental solution to this problem is still lacking, thus, the best option for counteracting the effects of algal blooms is to improve advance warnings (predictions). However, existing physical prediction models have difficulties setting a clear coefficient indicating the relationship between each factor when predicting algal blooms, and many variable data sources are required for the analysis. These limitations are accompanied by high time and economic costs. Meanwhile, artificial intelligence and deep learning methods have become increasingly common in scientific research; attempts to apply the long short-term memory (LSTM) model to environmental research problems are increasing because the LSTM model exhibits good performance for time-series data prediction. However, few studies have applied deep learning models or LSTM to algal bloom prediction, especially in South Korea, where algal blooms occur annually. Therefore, they employed the LSTM model for algal bloom prediction in four major rivers of South Korea.[8]

Karthik Srinivasan Vikram Duvvur, et.al (2019) infers that Machine learning is a type of artificial intelligence (AI) in which a training set is used to teach the AI without the need for it to be explicitly programmed. needed a neural network capable of identifying several types of common algal bloom shapes in the form of images, so it was decided to create a Convolution Neural Network. A neural network is made up of multiple “layers,” each of which serves a different purpose and is composed of nodes that contain the logic of the network. Chosing to make the neural network a back-propagation network, because the network can evaluate the error and propagate it back through earlier layers to get more aCC-BY 4.0 International license. certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under bioRxiv preprint doi: <https://doi.org/10.1101/450551>; this version posted January 28, 2019. The copyright holder for this preprint (which was not accurate classifications of algal bloom shapes. The results are evaluated by loss functions, which reward higher

accuracies with lower loss values. Ultimate is to place an image into one of 11 categories: nine shapes of algal blooms, plus water and land. These categories are known as classes, hence the name multiclass identifier. they coded the project in Python 3.4 using the Keras and TensorFlow libraries, but the methods will be described in pseudocode.[13]

Jade Snyder Echard (2020) infers in A Review Of Harmful Algal Bloom Prediction Models For Lakes And Reservoirs that Anthropogenic activity has led to eutrophication in water bodies across the world. This eutrophication promotes blooms, cyanobacteria being among the most notorious bloom organisms. Cyanobacterial blooms (more commonly referred to as harmful algal blooms (HABs)) can devastate an ecosystem. Cyanobacteria are resilient microorganisms that have adapted to survive under a variety of conditions, often outcompeting other phytoplankton. Some species of cyanobacteria produce toxins that ward off predators. These toxins can negatively affect the health of the aquatic life, but also can impact animals and humans that drink or come in contact with these noxious waters. Although cyanotoxin's effects on humans are not as well researched as the growth, behavior, and ecological niche of cyanobacteria, their health impacts are of large concern. It is important that research to mitigate and understand cyanobacterial blooms and cyanotoxin production continues.

This project supports continued research by addressing an approach to collect and summarize published articles that focus on techniques and models to predict cyanobacterial blooms with the goal of understanding what research has been done to promote future work. The following report summarizes 34 articles from 2003 to 2020 that each describe a mechanistic or data driven model developed to predict the occurrence of cyanobacterial blooms or the presence of cyanotoxins in lakes or reservoirs with similar climates to Utah. These articles showed a shift from more mechanistic approaches to more data driven approaches with time. This resulted in a more individualistic approach to modeling, meaning that models are often produced for a single lake or reservoir and are not easily comparable to other models for different systems.[4]

Alberto Mozo, et.al (2022) et.al refers to chlorophyll soft- sensor based on machine learning models for algal bloom predictions that Harmful algal blooms

(HABs) are a growing concern to public health and aquatic ecosystems. Long-term water monitoring conducted by hand poses several limitations to the proper implementation of water safety plans. This work combines automatic high-frequency monitoring (AFHM) systems with machine learning (ML) techniques to build a datadriven chlorophyll-a (Chl-a) soft-sensor. Massive data for water temperature, pH, electrical conductivity (EC) and system battery were taken for three years at intervals of 15 min from two different areas of As Conchas freshwater reservoir (NW Spain).

Designed a set of soft-sensors based on compact and energy efficient ML algorithms to infer Chl-a fluorescence by using low-cost input variables and to be deployed on buoys with limited battery and hardware resources. Input and output aggregations were applied in ML models to increase their inference performance. A component capable of triggering a 10 $\mu\text{g/L}$ Chl-a alert was also developed. The results showed that Chl-a soft-sensors could be a rapid and inexpensive tool to support manual sampling in water bodies at risk.[9] khushbu kumari, et.al (2018) infers that linear regression is a statistical procedure for calculating the value of a dependent variable from an independent variable. Linear regression measures the association between two variables. It is a modeling technique where a dependent variable is predicted based on one or more independent variables. Linear regression analysis is the most widely used of all statistical techniques. This article explains the basic concepts and explains how it can do linear regression calculations in SPSS and excel.[7]

Hye-Suk Yi, et.al (2018) concludes that In this study, design an intelligent model to predict chlorophyll-a concentration, which is the primary indicator of algal blooms, using extreme learning machine (ELM) models. Modeling algal blooms is important for environmental management and ecological risk assessment. For this purpose, the performance of the designed models was evaluated for four artificial weirs in the Nakdong River, Korea. The Nakdong River has harmful annual algal blooms that can affect health due to exposure to toxins. In contrast to conventional neural network (NN) that use backpropagation (BP) learning methods, ELMs are fast learning, feedforward neural networks that use least square estimates (LSE) for regression. The weights connecting the input layer to the hidden nodes are randomly

assigned and are never updated. The dataset used in this study includes air temperature, rainfall, solar radiation, total nitrogen, total phosphorus, N/P ratio, and chlorophyll-a concentration, which were collected on a weekly basis from January 2013 to December 2016.

Here, upstream chlorophyll-a concentration data was used in ELM2 model to improve algal bloom prediction performance. In contrast, the ELM1 model only uses downstream chlorophyll-a concentration data. The experimental results revealed that the ELM2 model showed better performance in comparison to the ELM1 model. Furthermore, the ELM2 model showed good prediction and generalization performance compared to multiple linear regression (LR), conventional neural network with backpropagation (NN-BP), and adaptive neuro-fuzzy inference system (ANFIS).[15]

Xiu Li, et.al (2021) infers in harmful algal blooms prediction with machine learning models in tolo harbour that Harmful algal blooms are a major type of marine disaster that endangers the marine ecological environment and human health. Since the algal bloom is a complex nonlinear process with time characteristics, traditional statistical methods often cannot provide good predictions. In this paper, they propose a method based on machine learning with the aim to predict the occurrence of algal blooms by environmental parameters. The features related to algal bloom growth have been experimented for achieving a good prediction of algal concentrations by a combination strategy. they validate the prediction performance on two real datasets from two locations in US and China, i.e., Scripps Pier, California and Sishili Bay, Shandong, respectively. The models and feature subsets have been selected to complete the missing data and predict the phytoplankton concentration. The results demonstrate the efficiency of method in the short-term prediction of concentrations by selecting appropriate features. The comparison studies prove the advantage of the developed machine learning method. The importance of every features for the prediction performance reveals crucial factors for the outbreak of harmful algal blooms.[14]

Sang-Soo Baek, et.al (2021) et.al concludes that In several countries, the public health and fishery industries have suffered from harmful algal blooms (HABs) that have escalated to become a global issue. Though computational modeling

offers an effective means to understand and mitigate the adverse effects of HABs, it is challenging to design models that adequately reflect the complexity of HAB dynamics. This paper presents a method involving the application of deep learning to an ocean model for simulating blooms of *Alexandrium catenella*. The classification and regression convolutional neural network (CNN) models are used for simulating the blooms. The classification CNN determines the bloom initiation while the regression CNN estimates the bloom density. GoogleNet and Resnet 101 are identified as the best structures for the classification and regression CNNs, respectively. The corresponding accuracy and root means square error values are determined as 96.8% and 1.20 [$\log(\text{cells L}^{-1})$], respectively. The results obtained in this study reveal the simulated distribution to follow the *Alexandrium catenella* bloom. Moreover, Grad-CAM identifies that the salinity and temperature contributed to the initiation of the bloom whereas $\text{NH}_4\text{-N}$ influenced the growth of the bloom.[1]

CHAPTER – 3

AIM AND SCOPE

3.1 AIM OF THE PROJECT

The aim of this project is to predict the occurrence of algal blooms using Auto ARIMA algorithm. Algal blooms are a sudden increase in the population of algae in a water body, which can have detrimental effects on the ecosystem and human health. Early prediction of algal blooms can help in taking preventive measures to avoid their harmful effects.

Auto ARIMA is a time series forecasting algorithm that automatically selects the best parameters for an ARIMA model. It analyzes the data and selects the optimal values for the order (p, d, q) to fit the best ARIMA model. By using Auto ARIMA, it can accurately predict the occurrence of algal blooms based on the historical data.

The project involves collecting historical data on the occurrence of algal blooms in a specific water body and using Auto ARIMA to predict the likelihood of a bloom occurring in the future. The accuracy of the model will be evaluated using various metrics to determine its effectiveness in predicting algal blooms.

Overall, the aim of this project is to develop an effective tool for predicting algal blooms using Auto ARIMA, which can be used by researchers and policymakers to take preventive measures and avoid the harmful effects of algal blooms.

3.2 SCOPE OF THE PROJECT

Algal bloom prediction using auto ARIMA models has promising future scope. Algal blooms are caused by the excessive growth of certain types of algae, which can have harmful effects on the environment, human health, and aquatic life. Early prediction of algal blooms can help in mitigating their negative impacts by taking appropriate measures to control their growth.

Auto ARIMA is an automated time series forecasting model that can be used to predict future trends based on past data. This method has been successfully applied in various fields, including finance, healthcare, and meteorology. The use of Auto ARIMA models for algal bloom prediction can enable accurate forecasting of algal bloom occurrence, intensity, and duration.

In recent years, there has been an increasing interest in the use of remote sensing data, such as satellite imagery, for algal bloom prediction. Auto ARIMA models can be combined with remote sensing data to develop more accurate and reliable predictions of algal bloom occurrences. This can lead to more effective management strategies for controlling and preventing harmful algal blooms.

Overall, the future scope of algal bloom prediction using Auto ARIMA is promising and can potentially lead to significant improvements in the management and control of harmful algal blooms.

The future scope of algal bloom prediction using auto ARIMA includes the following five subheadings:

1. Enhanced accuracy: The use of auto ARIMA for algal bloom prediction has shown promising results in terms of accuracy. As more data becomes available and the models are fine-tuned, it can expect even greater accuracy in predicting algal blooms. This will help in identifying potential harmful algal blooms early on, allowing for quicker responses and mitigating their impacts.
2. Integration with remote sensing technologies: Remote sensing technologies, such as satellite imagery and drones, are becoming increasingly popular for monitoring algal blooms. These technologies can provide large amounts of data that can be used to feed into auto ARIMA models for more accurate predictions. As the use of remote sensing technologies becomes more widespread, it can expect to see even greater improvements in algal bloom prediction.
3. Improved understanding of algal bloom dynamics: Algal blooms are complex and can be affected by a range of environmental factors. By using auto ARIMA to predict algal blooms, and can gain a better understanding of the dynamics of these blooms and the environmental factors that influence them.

This will help in developing more effective strategies for preventing and managing algal blooms.

4. Real-time monitoring: Auto ARIMA models can be trained to make predictions in real-time, allowing for continuous monitoring of algal blooms. This can be particularly useful for areas that are prone to algal blooms, such as coastal regions and freshwater bodies. Real-time monitoring can help in identifying potential algal blooms early on, allowing for quicker responses and mitigating their impacts.
5. Integration with decision-making tools: Algal bloom prediction using auto ARIMA can be integrated with decision-making tools to help inform management strategies. For example, decision-makers can use the predictions to determine when to issue public health advisories or when to implement control measures. As these tools become more sophisticated, they can help in developing more effective and efficient management strategies for algal blooms.

CHAPTER – 4

METHODOLOGY AND ALGORITHMS

4.1 EXISTING METHODOLOGY

HABs are a major problem all over the world. An extensive literature review and analysis was performed on the possible variables for forecasting HABs. Two forecasting methods - CART, classification and regression tree, and ANN, artificial neural network - as well as two training periods and two input variable selection methods, nutrient loading period and Spearman's rank correlation were used. The Spearman rank correlation describes how well two variables' relationship can be explained with a monotonic function. A monotonic function is a function that is always either entirely non-decreasing or entirely non-increasing. For the nutrient loading period selection method, only one set of input variables is used for forecasting.

The Spearman selection method, on the other hand, examines more variables than the nutrient loading period, considering up to 28 different averaging periods and lag times for each considered variable. First, the CART forecasting models were tested with both classification methods, a 3-class and 5- class system resulting in the 3-class system being selected. The CART models were then created for both methods and training periods. aCC-BY 4.0 International license. certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. The copyright holder for this preprint (which was not16 Initially, when using the first training period of 2002 to 2011, the loading period method showed better precision in forecasting HABs when compared to the Spearman selection method. When the training period was increased to 2013, both methods showed an improvement in the overall accuracy, with Spearman having an 8.9 percent improvement, and loading period a 5.4 percent improvement. However, with the extended training period, the loading period decision trees for August and September showed a slight increase in precision over the Spearman method and the Spearman method being slightly more precise in October.

After the CART models, the ANN models were created and analysed for both selection methods and training periods. For both selection methods in the first

training period, the models often underpredicted the higher magnitude blooms. Many of the forecasts did not predict the exact same CI, or cyanobacterial index, as observed. However, in most cases, both methods predicted the same class of bloom as observed. In most cases after increasing the training period to 2013, both ANN models improved their accuracy for predicting the higher magnitude of blooms. The correlation coefficient increased from 0.70 to 0.77 for the loading period selection method and from 0.79 to 0.83 for the Spearman selection method when extending the training period. Both input selection methods had some difficulty in predicting the 2015 HAB because the 2015 bloom was a special case in terms of nutrient loading as well as bloom time.

There was a large amount of loading in June and July, which is atypical due to how early it was happening. The 2015 bloom in July was 382 percent larger than any bloom recorded from 2002 to 2014. The monthly discharge for June was the highest recorded and third highest on record since the United States Geological Survey aCC-BY 4.0 International license. certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under bioRxiv preprint doi Similar to the CART method, the ANN model showed an increase in accuracy when forecasting HABs with an extended training period. The method, however, does not rely on nutrient variables and instead looks at the growth of algal blooms themselves through measured chlorophyll levels. The method enables the prediction of severity, location, and distribution of algal blooms, just like both the CART and ANN methods. Because to utilize satellite imagery, however, can simplify the prediction process and increase success. The algorithm successfully identifies algal bloom components 80 percent of the time, due not only to the neural network but also due to the aforementioned near certainty with which satellite images highlight algal blooms.

The existing methodologies used in this project are

4.1.1 Linear Regression

Linear regression is a statistical technique used for predicting a continuous target variable based on one or more input variables that are believed to be related to the target variable. In linear regression, the relationship between the input variables and

the target variable is assumed to be linear, meaning that the change in the target variable is proportional to the change in the input variables.

The goal of linear regression is to find the best-fitting line (or hyperplane in higher dimensions) that represents the relationship between the input variables and the target variable. This line is determined by estimating the slope and intercept of the line that minimize the distance between the predicted values and the actual values of the target variable.

Once the line has been fitted, it can be used to make predictions of the target variable for new input values. Linear regression is widely used in various fields, including finance, economics, marketing, and social sciences, for predicting outcomes such as sales, stock prices, or customer satisfaction scores.

Linear regression is a statistical method that is commonly used in algal bloom prediction. While it is a useful tool, it is not without its disadvantages. In this response, we will discuss some of the main drawbacks of using linear regression for algal bloom prediction.

Firstly, linear regression assumes that there is a linear relationship between the predictor variables (such as water temperature, nutrient levels, etc.) and the response variable (algal bloom density). This assumption may not hold true in all cases, especially if there are non-linear relationships or interactions between the predictor variables. For example, if nutrient levels are too low, increasing temperature may not lead to an increase in algal bloom density. Therefore, using linear regression to predict algal bloom density may not always produce accurate results.

Secondly, linear regression assumes that the residuals (the differences between the predicted values and the actual values) are normally distributed and have constant variance. However, in algal bloom prediction, the residuals may not always follow a normal distribution or have constant variance. This can result in biased and inconsistent predictions.

Thirdly, linear regression assumes that there are no outliers or influential data points in the dataset. However, in algal bloom prediction, outliers and influential data points can occur due to natural variations in the environment or measurement errors.

Linear regression can be sensitive to these outliers and may produce inaccurate predictions.

Lastly, linear regression assumes that there are no multicollinearity issues between the predictor variables. Multicollinearity occurs when two or more predictor variables are highly correlated with each other. This can lead to unstable and unreliable predictions, as it becomes difficult to determine the individual effects of each predictor variable on the response variable.

In conclusion, while linear regression is a commonly used tool in algal bloom prediction, it is important to be aware of its limitations and potential drawbacks. These limitations should be taken into consideration when interpreting the results and making predictions based on linear regression models.

4.1.2 Basic Neural Network

A basic neural network is a type of machine learning model that is designed to predict an output based on a set of input features. It is modeled after the structure of the human brain and consists of multiple layers of interconnected nodes, or "neurons", that process information.

In a neural network for prediction, the input data is fed into the first layer of neurons, which applies a set of weights and biases to the input and passes it on to the next layer. Each subsequent layer performs a similar operation until the output layer produces a predicted value. During the training process, the weights and biases in the neural network are adjusted to minimize the difference between the predicted output and the actual output, using a method called backpropagation. This process allows the network to learn from the input data and improve its predictions over time.

Neural networks can be used for a wide range of prediction tasks, such as image recognition, natural language processing, and time series forecasting. However, the performance of a neural network depends on the quality and quantity of the input data, the complexity of the network architecture, and the training method used.

Algal bloom prediction is a complex and challenging task due to the dynamic nature of aquatic ecosystems. Basic neural networks are one of the popular

techniques used for algal bloom prediction. However, they have several disadvantages that can limit their effectiveness in this application.

Firstly, basic neural networks have limited generalization ability, which means that they may not perform well when presented with new or unseen data. This is because they tend to memorize the training data rather than learning general patterns. As a result, their accuracy and reliability may decrease when used for algal bloom prediction in different environments or under varying conditions.

Secondly, basic neural networks are prone to overfitting. Overfitting occurs when the network becomes too complex and begins to fit the noise in the data instead of the underlying patterns. This can lead to poor performance on the test data, which can be problematic in algal bloom prediction, where accuracy is crucial.

Thirdly, basic neural networks can be computationally expensive to train and optimize. This is because they require large amounts of data and extensive parameter tuning to achieve good performance. In the case of algal bloom prediction, where data may be limited and time-sensitive, this can be a significant disadvantage.

Finally, basic neural networks are often considered "black boxes" because they lack transparency in terms of how they arrive at their predictions. This can make it difficult for researchers and practitioners to understand the underlying mechanisms driving the predictions and to interpret the results. In the context of algal bloom prediction, where accurate and interpretable predictions are essential for effective management, this can be a significant drawback.

In conclusion, while basic neural networks have some advantages in algal bloom prediction, they also have several limitations that can impact their effectiveness. Researchers and practitioners should be aware of these limitations and consider alternative techniques or modifications to basic neural networks to overcome these issues.

4.1.3 Multi-layer Perceptron

A Multi-layer Perceptron (MLP) is a type of artificial neural network that can be used for prediction tasks. It is composed of multiple layers of interconnected

neurons, where each neuron receives input from the previous layer and produces an output for the next layer.

In prediction tasks, MLPs can be trained using labeled data to learn a mapping between input variables and output variables. During training, the network adjusts the weights and biases of the neurons to minimize the difference between its predicted outputs and the actual outputs from the labeled data. Once trained, the MLP can be used to make predictions on new, unseen data by passing the input through the network and obtaining the output from the final layer.

MLPs have been successfully used for a variety of prediction tasks, including classification and regression. They are particularly useful when the relationship between the input and output variables is complex and nonlinear, as they can learn to model these relationships by combining multiple simpler functions. However, their performance can be affected by the choice of network architecture, the amount and quality of training data, and the choice of training algorithm.

The Multi-Layer Perceptron (MLP) is a type of artificial neural network that is widely used for various prediction tasks, including algal bloom prediction. However, there are several disadvantages of using MLPs for algal bloom prediction that should be considered.

Firstly, MLPs require a large amount of training data to achieve good accuracy. The training data must be representative of the input space, including a range of environmental and biological factors that influence algal blooms. Obtaining such data can be challenging, particularly in areas with limited monitoring resources or where algal blooms occur infrequently.

Secondly, MLPs are prone to overfitting when trained on small datasets or datasets with high noise. Overfitting occurs when the model learns to fit the training data too well, leading to poor generalization performance on new, unseen data. This can be a significant issue in algal bloom prediction, as environmental conditions can vary widely between years and locations.

Thirdly, MLPs are computationally expensive to train and require significant computing resources. This can limit their use in practical applications, particularly in real-time forecasting or where large datasets are involved.

Fourthly, MLPs are black-box models, meaning that the internal workings of the model are not transparent. This can make it difficult to understand how the model is making predictions or to diagnose problems when the model is not performing well.

Fifthly, MLPs can be sensitive to the initial conditions of the model and the choice of hyperparameters. This can make the model difficult to optimize, requiring significant trial and error to achieve good performance.

Sixthly, MLPs are not well-suited to handling missing data or data with irregular sampling intervals. This can be a significant issue in algal bloom prediction, as environmental data can be spatiotemporally patchy or missing altogether.

Lastly, MLPs are not always able to capture the complex nonlinear relationships between environmental factors and algal blooms. This can lead to poor performance in areas where the underlying relationships are highly nonlinear or where interactions between factors are complex.

In summary, while MLPs have been successfully applied to algal bloom prediction, they do have several disadvantages that should be considered when selecting a modeling approach. These include their need for large training datasets, potential for overfitting, computational requirements, black-box nature, sensitivity to hyperparameters, limited handling of missing data, and inability to capture complex nonlinear relationships.

4.1.4 Logistic Regression

Logistic regression is a statistical method used for binary classification tasks, where the goal is to predict the probability that an input belongs to a particular category. It is commonly used in prediction tasks where the response variable is binary (i.e., only two possible outcomes), such as whether a customer will churn or not, or whether a patient will develop a certain disease or not.

In logistic regression, the input variables are used to calculate the probability of belonging to the positive class (i.e., the class of interest), which is then converted into a binary prediction by using a threshold value. The algorithm works by estimating the coefficients of a linear equation that relates the input variables to the log-odds of the response variable being in the positive class. The log-odds are then

transformed into probabilities using the logistic function, which produces an S-shaped curve that ranges from 0 to 1.

During training, the logistic regression model uses a loss function to optimize the coefficients of the linear equation, such as the maximum likelihood estimation or binary cross-entropy. Once trained, the model can be used to predict the probability of belonging to the positive class for new input data, which can then be converted into a binary prediction using a threshold.

Overall, logistic regression is a simple yet effective method for binary classification tasks, and is widely used in various fields such as finance, healthcare, and marketing.

Logistic regression is a popular method for predicting binary outcomes, such as the presence or absence of algal blooms in a particular body of water. While this approach has many advantages, there are also several disadvantages that should be taken into consideration when using logistic regression for algal bloom prediction. This is to discuss seven of these disadvantages in detail.

1. Assumption of linearity: One of the key assumptions of logistic regression is that the relationship between the predictor variables and the response variable is linear. However, in the case of algal bloom prediction, this assumption may not hold true, as the underlying processes that drive algal blooms are often complex and nonlinear. This can lead to inaccurate predictions and lower model performance.
2. Sensitivity to outliers: Logistic regression is sensitive to outliers, which can have a disproportionate impact on the model's predictions. In the case of algal bloom prediction, outliers such as extreme weather events or sudden changes in water chemistry could lead to inaccurate predictions, and thus, a less reliable model.
3. Limited to binary outcomes: Logistic regression is designed for predicting binary outcomes, which may not always be appropriate for algal bloom prediction. For example, if the goal is to predict the severity or extent of an algal bloom, a binary outcome may not provide enough information to make informed decisions.

4. **Overfitting:** Logistic regression models can be prone to overfitting, which occurs when the model is too complex and fits the training data too closely. This can lead to poor performance on new data and reduced model generalization. In the case of algal bloom prediction, overfitting could lead to inaccurate predictions, particularly if the model is not adequately validated or tested on new data.
5. **Difficulty handling missing data:** Logistic regression models require complete data for all predictor variables, which can be challenging when working with environmental data, which is often incomplete or missing. This can lead to biased or inaccurate predictions, particularly if the missing data is not handled properly.
6. **Lack of interpretability:** While logistic regression models can provide accurate predictions, they can be difficult to interpret, particularly when there are many predictor variables. In the case of algal bloom prediction, understanding the underlying factors that contribute to algal blooms can be important for developing effective mitigation strategies. However, logistic regression models may not provide enough information to identify these factors.
7. **Limited ability to capture complex interactions:** Finally, logistic regression models are limited in their ability to capture complex interactions between predictor variables. In the case of algal bloom prediction, multiple environmental factors can interact in complex ways to drive algal blooms. Logistic regression models may not be able to fully capture these interactions, leading to reduced model performance and inaccurate predictions.

In conclusion, while logistic regression is a powerful tool for predicting binary outcomes such as algal bloom presence or absence, it is not without its limitations. Researchers should carefully consider the potential disadvantages of logistic regression before using it to predict algal blooms and explore alternative modeling approaches if necessary.

4.1.5 Random Forest

Random forest is a type of machine learning algorithm used for prediction tasks. It works by constructing multiple decision trees and aggregating their predictions to obtain a final prediction. Each decision tree is built using a random subset of the data and a random subset of the features. This randomness helps to reduce overfitting and increase the generalization ability of the model.

The final prediction is obtained by taking the average or majority vote of the predictions of all the decision trees. This averaging process helps to further reduce the variance of the model and improve its accuracy. Random forest is a popular algorithm for classification and regression tasks, and it is often used in practice due to its high accuracy, robustness, and ability to handle large datasets with high-dimensional feature spaces.

1. **Overfitting:** One of the main disadvantages of using random forests in algal bloom prediction is overfitting. Random forests can easily overfit the training data if the model is too complex. This can result in poor generalization and accuracy when predicting new data. To overcome this issue, hyperparameter tuning is required, which can be time-consuming and computationally expensive.
2. **Lack of transparency:** Random forests are considered to be black box models, meaning that it is difficult to understand how the model arrives at its predictions. This lack of transparency makes it difficult to explain the results of the model to stakeholders and can make it hard to diagnose any potential issues.
3. **Dependency on hyperparameters:** Random forests rely heavily on hyperparameters to achieve high accuracy, and the optimal hyperparameters are often data-dependent. Finding the optimal hyperparameters can be a challenging task, especially when working with large and complex datasets.
4. **Limited handling of missing data:** Random forests may not handle missing data well. If the dataset has a significant amount of missing values, the model may not perform as well as expected. Imputation techniques can be used to overcome this issue, but they can introduce bias in the data.

5. Inability to extrapolate: Random forests are limited in their ability to extrapolate outside the range of the training data. This means that the model may not be able to predict accurately for new data that falls outside the range of the training data. This can be an issue in situations where the range of the input variables may change over time.
6. Computationally intensive: Random forests can be computationally intensive, especially when working with large datasets. This can make it difficult to train and tune the model, especially on resource-limited devices.
7. Limited interpretability: Random forests can provide insights into feature importance, but the interpretation of these results can be challenging. Feature importance can vary depending on the dataset and the hyperparameters chosen, making it difficult to draw definitive conclusions from the results. Additionally, the relationship between the features and the output is not always clear, which can make it challenging to understand the underlying mechanisms driving the predictions.

4.2 PROPOSED METHODOLOGY

The proposed methods include four major parts, i.e., data collection using MODIS data, data pre-processing, feature and model selection, feature importance, and predictions. Firstly pre-process the collected data and generate the combinatorial feature subsets and then implement a series of machine learning methods with exhaustive feature combinations for selecting a predictor and a subset of feature with the best prediction performance. In the experiments, Auto ARIMA algorithm is found to be the machine learning methods that obtain the best prediction. Then, employ it to evaluate the feature importance of prediction that provides the indications of crucial environmental factors for the occurrence of algal blooms. And also provide multiple predictions based on the trained predictor and optimized features. Later compare the models, compare the results and pick the more efficient approach to draw conclusions.

4.2.1 Framework

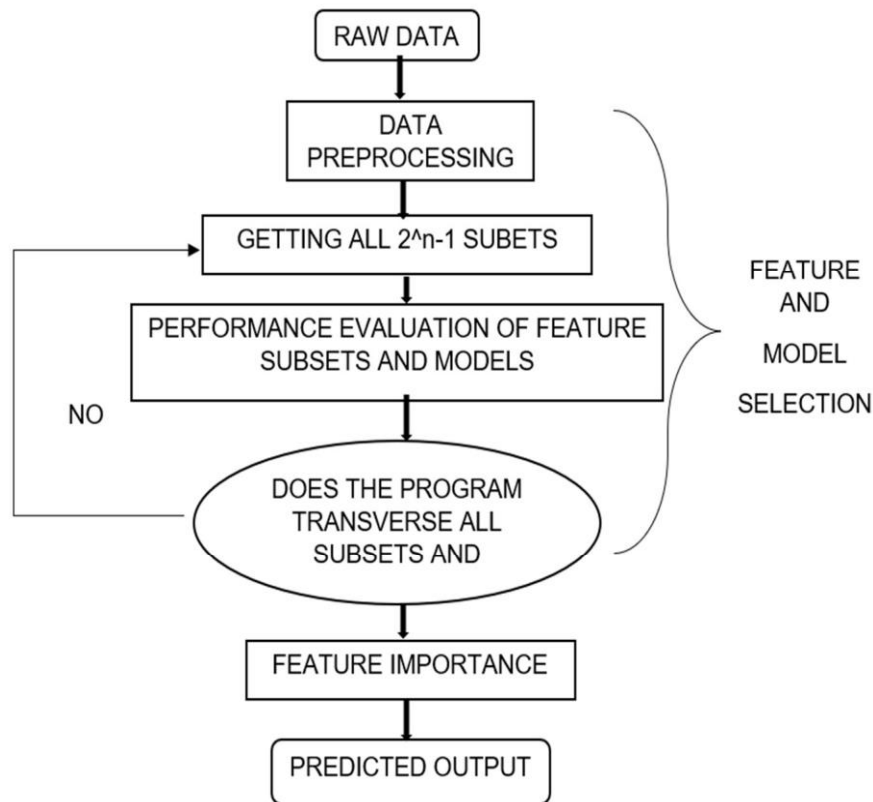


Fig: 4.1: Framework Of The Proposed Methodology

4.3 EXISTING METHODOLOGY AND PROPOSED METHODOLOGY DIFFERENCES

ANNs have recently become the focus of much attention, largely because of their wide range of applicability and the ease with which they can treat complicated problems. ANNs can identify and learn correlated patterns between input data sets and corresponding target values. After training, ANNs can be used to predict the output of new independent input data. ANNs imitate the learning process of the animal brain and can process problems involving very nonlinear and complex data even if the data are imprecise and noisy. Thus, they are ideally suited for the modelling of ecological data which are known to be very complex and often nonlinear. The causality and dynamics of algal production are extremely complicated and not well-understood. The capabilities of water quality and ecological models have greatly advanced in recent years, especially for predicting long term trends and when sufficient field data is available for model calibration and validation.

Nevertheless, the prediction of algal primary production remains a very difficult problem. The uncertainty and the complexity of existing deterministic two- or three-dimensional models (often coupled with hydrodynamic models) have prevented their effective use as a forecasting tool. An alternative approach aimed at more robust forecasts of primary production is therefore necessary. To be able to control algal blooms, it is necessary to be able to determine the key factors governing the algal dynamics and to establish an algal response model which can effectively simulate the timing and magnitude of algal blooms. Recently, artificial neural network (ANN) technology has been applied in the prediction of algal blooms. In this framework artificial neural networks provide an effective alternative to conventional modelling techniques and their application to regional scale phytoplankton primary production modelling has already been presented.

Applications of ANN in ecological and environmental science have been reported since the beginning of the 1990s. The use of ANNs as ecological modelling tools has been the subject of several investigations. The temporal changes of particular algal species in freshwater systems have been modelled. It is shown that ANNs are capable of simulating trends of algal growth dynamics.

4.4 AUTO ARIMA ALGORITHM

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time-series data to better understand the data set or predict future trends. A statistical model is autoregressive if it predicts future values based on past values. For example, an ARIMA model might seek to predict a stock's future prices based on its past performance or forecast a company's earnings based on past periods.

the Arima term can be split into three terms, AR, I, MA:

1. AR(p) stands for the autoregressive model, the p parameter is an integer that confirms how many lagged series are going to be used to forecast periods ahead.

2. $I(d)$ is the differencing part, the d parameter tells how many differencing orders are going to be used to make the series stationary.
3. $MA(q)$ stands for moving average model, the q is the number of lagged forecast error terms in the prediction equation. SARIMA is a seasonal ARIMA and it is used with time series with seasonality.

4.4.1 Getting Started With Auto-Arima

Auto ARIMA is an automated version of the ARIMA (Autoregressive Integrated Moving Average) model. It is a popular technique used in time series analysis to forecast future values based on historical data. Auto ARIMA is a powerful tool that can save time and effort by automatically selecting the best parameters for the ARIMA model. In this paper, will discuss the basics of Auto ARIMA and how it can be used to generate accurate forecasts.

Auto ARIMA works by minimizing the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These criteria measure the goodness of fit of the model and penalize the complexity of the model. The model with the lowest AIC or BIC is selected as the best model.

Auto ARIMA also includes an option to allow for seasonal trends. This is known as the SARIMA model (Seasonal Autoregressive Integrated Moving Average). The seasonal component adds a fourth parameter to the model, which models the seasonal trend.

4.4.2 Brief Explanation Of The Components Of Arima

ARIMA is an acronym for Auto Regressive (AR) Integrated (I) Moving Average (MA) which indicates that an ARIMA model has three components to it. I will go over the components very briefly in this post and if you need a more in-depth understanding of the concepts, I suggest using these resources which are available for free.

ARIMA models can be expressed in two forms: Non-seasonal models where the model exhibits an order in the form of (p,d,q) where:

p = The order of the Auto Regressive
 Model d = The order of differencing q =
 The order of the Moving Average

Auto Regressive Models (AR):

Auto regressive models are similar to a regression model but the regressor in this case is the same dependent variable with a specific lag.

Differencing (I):

For ARIMA to perform at its best it needs the data to be stationary. That means that the mean and variance are constant over the entire set. Differencing is used to transform the data so that it is stationary.

Moving Average (MA):

Moving averages are widely used in time series analysis and is an already well-known concept. It involves getting the average of the points in a series for a specific lag and can be expressed like:

$$(4.1) \quad T = \frac{1}{m} \sum_{j=-k}^k y_{t+j}$$

Where $m = 2k + 1$,

k = the number of values to use in the average before and after the point used and
 t = is the time at which the trend cycle is calculated.

In ARIMA, a Moving Average Model is used

4.4.3 Working Of Auto Arima Model

In Time Series Analysis, a correct model should yield the highest Log-Likelihood and requires the lowest AIC.



Fig: 4.2: Aic and Bic

The AIC and BIC of a model depend on the log-likelihood, to be more precise AIC and BIC uses Log-Likelihood in their formula. This would indirectly imply that the log-likelihood is high therefore this method cycles through the model with varying orders and specifications and out of those returns the one with the lowest.

$$(4.2) \quad -2 \ln L = 2 \sum_{i=1}^n \epsilon_i^2$$

Formula of log likelihood

$$(4.3) \quad AIC(K) = -2 \log(L) + k \times edf$$

Formula of AIC

$$(4.4) \quad BIC = -2 \log(Likelihood) + \log(n) p$$

Fomula for BIC

Auto ARIMA works by minimizing the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC). These criteria measure the goodness of fit

of the model and penalize the complexity of the model. The model with the lowest AIC or BIC is selected as the best model.

Auto ARIMA also includes an option to allow for seasonal trends. This is known as the SARIMA model (Seasonal Autoregressive Integrated Moving Average). The seasonal component adds a fourth parameter to the model, which models the seasonal trend.

4.4.4 Advantages and Applications of Auto Arima

Auto ARIMA has several advantages over traditional ARIMA models. It saves time and effort by automatically selecting the best parameters for the model. This eliminates the need for trial and error in selecting the parameters. Auto ARIMA also improves the accuracy of the forecasts by selecting the best model. Moreover, it includes an option for seasonal trends, which makes it suitable for modeling data with seasonal patterns.

In conclusion, Auto ARIMA is a powerful tool for time series analysis. It is a variation of the ARIMA model that automatically selects the best parameters for the model. Auto ARIMA saves time and effort and improves the accuracy of the forecasts. It is suitable for modeling data with seasonal patterns. Auto ARIMA can be applied to various fields such as finance, economics, meteorology, and many more. It can be used in predicting stock prices, forecasting sales, and predicting weather patterns. Auto ARIMA provides accurate forecasts and can be used to predict trends and patterns in time series data.

- In finance, Auto ARIMA can be used to predict stock prices and forecast trends in the stock market. It can help traders and investors make informed decisions about buying and selling stocks.
- In economics, Auto ARIMA can be used to forecast sales and demand for products. It can help businesses make decisions about production and inventory management.
- In meteorology, Auto ARIMA can be used to predict weather patterns and help meteorologists issue weather alerts and warnings.

4.3.5 Limitations of Auto Arima

Auto ARIMA is not suitable for all types of time series data. It works best for data with clear trends and seasonal patterns. It may not perform well for data with irregular patterns or data that is highly volatile.

Moreover, Auto ARIMA is not a perfect tool and may sometimes produce inaccurate forecasts. It is important to validate the forecasts and refine the model if necessary.

Pros:

- It saves an enormous amount of time
- eliminate the need to understand the statistics and theory behind the model selection.
- this method will also reduce the risk of human error and the possible mistakes caused by an incorrect interpretation of the results.

Cons:

- Blindly putting faith into one criterion
- Never really see how well the other models perform □ Topic expertise.

4.5 ALGAL BLOOM PREDICTION USING AUTO ARIMA

Algal blooms are a major concern for aquatic ecosystems as they can cause severe ecological and economic impacts. Predicting the occurrence of algal blooms is crucial for timely management and mitigation efforts. In recent years, machine learning algorithms have been used to predict algal blooms, with Auto ARIMA being one of the popular models.

Auto ARIMA is an automated version of the AutoRegressive Integrated Moving Average (ARIMA) model. It uses an algorithm to determine the optimal parameters for the ARIMA model, making it a useful tool for time-series forecasting.

The following steps can be used to predict algal blooms using Auto ARIMA:

1. **Data Collection:** Collect data on the water quality parameters, such as temperature, pH, dissolved oxygen, and nutrient levels, which influence algal growth.
2. **Data Preprocessing:** Clean the data to remove any outliers or missing values. Convert the data into a time-series format and resample it to a consistent interval.
3. **Train-Test Split:** Split the data into training and testing sets. The training set is used to fit the Auto ARIMA model, while the testing set is used to evaluate its performance.
4. **Model Fitting:** Use the training set to fit an Auto ARIMA model to the data. The algorithm will automatically select the optimal parameters for the model.
5. **Model Performance Evaluation:** Evaluate the performance of the model using the testing set. Calculate performance metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE).
6. **Prediction:** Use the fitted Auto ARIMA model to predict the occurrence of algal blooms in the future.

CHAPTER- 5

RESULTS AND DISCUSSION

5.1 DATA CLEANING PROCESS

1. **Data Collection:** This is the first step in the data cleaning process where the data is collected from various sources such as files, databases, or web scraping.
2. **Data Preprocessing:** This step involves cleaning, transforming, and organizing the collected data to make it ready for analysis. It includes removing duplicates, handling missing values, and data normalization.
3. **Data Exploration:** This step involves analyzing the data to identify patterns, trends, and outliers. It helps in gaining insights and understanding the data before modeling.
4. **Feature Selection:** In this step, the most relevant and useful features are selected to build the model. It involves identifying and removing irrelevant features that do not contribute to the model's accuracy.
5. **Data Transformation:** This step involves transforming the data into a suitable format for the machine learning algorithm. It includes scaling, encoding, and transforming categorical variables.
6. **Data Splitting:** In this step, the data is divided into training, validation, and testing sets. The training set is used to train the model, the validation set is used to optimize the model, and the testing set is used to evaluate the model's performance.

Unnamed: 0	season	size	speed	mxPH	mnO2	Cl	NO3	NH4	oPO4	PO4	Chla	
0	1	winter	small	medium	8.00	9.8	60.800	6.238	578.00000	105.000	170.00000	50.0
1	2	spring	small	medium	8.35	8.0	57.750	1.288	370.00000	428.750	558.75000	1.3
2	3	autumn	small	medium	8.10	11.4	40.020	5.330	346.66699	125.667	187.05701	15.6
3	4	spring	small	medium	8.07	4.8	77.364	2.302	98.18200	61.182	138.70000	1.4
4	5	autumn	small	medium	8.06	9.0	55.350	10.416	233.70000	58.222	97.58000	10.5

Fig :5.1 : Data Set

5.1.1 Comparing the Parameters

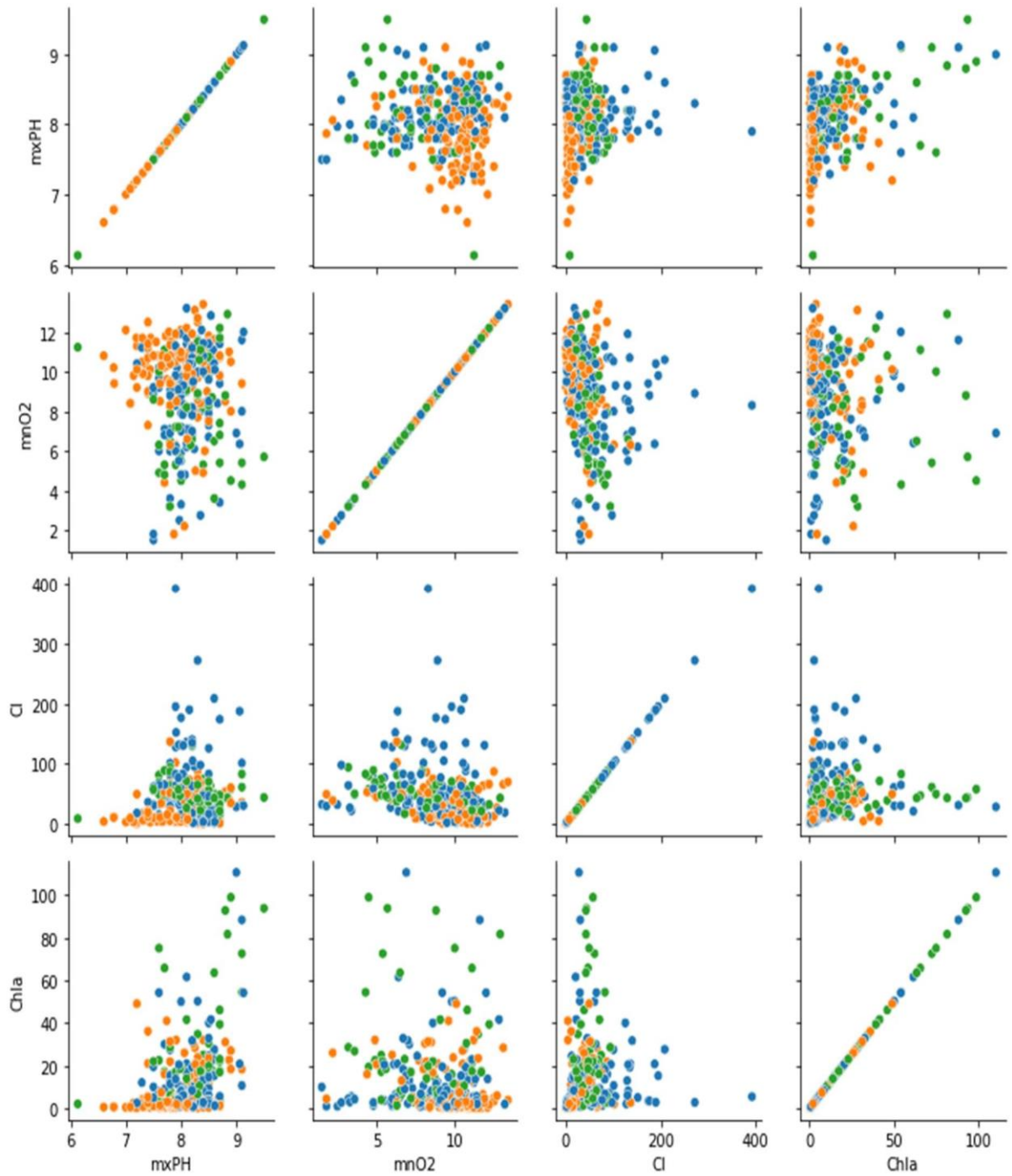


Fig: 5.2: Comparing the Parameters

5.1.2 Plotting The Correlation Between The Parameters

	mxPH	mnO2	Cl	NO3	NH4	oPO4	PO4	Chla	autumn	spring	summer	winter	high	low	medium
mxPH	1.000000	-0.050257	0.130379	-0.044352	-0.075230	0.072662	0.068783	0.385617	0.041167	0.016944	-0.134233	0.078929	-0.266525	0.190102	0.118368
mnO2	-0.050257	1.000000	-0.268807	0.087229	-0.268165	-0.296855	-0.364762	-0.112742	0.266340	-0.282091	0.086551	-0.064644	0.304736	-0.191995	-0.154976
Cl	0.130379	-0.268807	1.000000	0.358572	0.107576	0.283088	0.379104	0.123542	-0.046507	0.072880	-0.016680	-0.010748	-0.358378	0.072615	0.300864
NO3	-0.044352	0.087229	0.358572	1.000000	0.054650	0.193059	0.284038	0.094986	0.117942	-0.146272	0.007580	0.023862	-0.159998	0.037598	0.130310
NH4	-0.075230	-0.268165	0.107576	0.054650	1.000000	0.624881	0.633481	-0.060167	-0.007850	0.019755	-0.005496	-0.006553	-0.215890	-0.038842	0.245191
oPO4	0.072662	-0.296855	0.283088	0.193059	0.624881	1.000000	0.935687	0.051345	-0.077849	0.069293	0.042561	-0.036657	-0.240797	0.051432	0.200107
PO4	0.068783	-0.364762	0.379104	0.284038	0.633481	0.935687	1.000000	0.174237	-0.063906	0.080311	0.007575	-0.025812	-0.318065	0.107612	0.233596
Chla	0.385617	-0.112742	0.123542	0.094986	-0.060167	0.051345	0.174237	1.000000	0.075280	-0.011896	-0.121228	0.061415	-0.244530	0.363144	-0.037538
autumn	0.041167	0.266340	-0.046507	0.117942	-0.007850	-0.077849	-0.063906	0.075280	1.000000	-0.320110	-0.328454	-0.320110	-0.045261	0.069705	-0.008875
spring	0.016944	-0.282091	0.072880	-0.146272	0.019755	0.069293	0.080311	-0.011896	-0.320110	1.000000	-0.346523	-0.337719	-0.004685	-0.032284	0.029666
summer	-0.134233	0.086551	-0.016680	0.007580	-0.005496	0.042561	0.007575	-0.121228	-0.328454	-0.346523	1.000000	-0.346523	0.037563	-0.042255	-0.004710
winter	0.078929	-0.064644	-0.010748	0.023862	-0.006553	-0.036657	-0.025812	0.061415	-0.320110	-0.337719	-0.346523	1.000000	0.010680	0.007261	-0.016264
high	-0.266525	0.304736	-0.358378	-0.159998	-0.215890	-0.240797	-0.318065	-0.244530	-0.045261	-0.004685	0.037563	0.010680	1.000000	-0.383912	-0.699139
low	0.190102	-0.191995	0.072615	0.037598	-0.038842	0.051432	0.107612	0.363144	0.069705	-0.032284	-0.042255	0.007261	-0.383912	1.000000	-0.391789
medium	0.118368	-0.154976	0.300864	0.130310	0.245191	0.200107	0.233596	-0.037538	-0.008875	0.029666	-0.004710	-0.016264	-0.699139	-0.391789	1.000000

Fig : 5.3: Correlation

5.2 IMPLEMENTAION OF EXISTING METHODOLOGIES

5.2.1 Liner Regression

Algal bloom prediction using linear regression is a statistical method that involves analyzing data on various environmental factors to predict the likelihood of an algal bloom occurring in a specific body of water. This approach is based on the idea that certain environmental variables, such as temperature, nutrient levels, and light intensity, can influence the growth of algae.

Linear regression is a commonly used statistical method for predicting numerical values based on a set of input variables. In the case of predicting algal

bloom, linear regression can be used to model the relationship between various environmental factors and the occurrence of algal blooms.

To implement a linear regression model for algal bloom prediction, you would typically follow these steps:

1. Gather data: Collect data on environmental factors such as temperature, nutrient levels, and sunlight intensity, as well as data on the occurrence of algal blooms in the area of interest.
2. Prepare the data: Clean the data by removing any missing values or outliers. Divide the data into training and testing sets to evaluate the performance of the model.
3. Choose the input variables: Identify the environmental factors that are most likely to influence the occurrence of algal blooms. These could include factors such as water temperature, nutrient levels, and sunlight intensity.
4. Choose the output variable: The output variable is the occurrence of algal blooms. This can be represented as a binary variable (i.e., yes or no) or as a continuous variable (i.e., the intensity or extent of the bloom).
5. Fit the model: Use a linear regression algorithm to fit the model to the training data. The algorithm will estimate the coefficients for each input variable, which can be used to predict the occurrence of algal blooms based on the input values.
6. Evaluate the model: Test the performance of the model on the testing data. Calculate the metrics such as mean squared error or R-squared value to evaluate the accuracy of the model.
7. Use the model for prediction: Once the model has been evaluated and validated, it can be used to predict the occurrence of algal blooms based on new input data.

While linear regression can be a useful tool for predicting algal blooms, there are several potential drawbacks to this approach:

1. Complexity of the system: Algal blooms are a complex phenomenon, and the factors that contribute to their occurrence are often interrelated and nonlinear.

Linear regression assumes a linear relationship between the variables, which may not always be the case.

2. Limited predictive power: Linear regression can only provide a prediction within the range of the data used to train the model. If new environmental conditions arise that fall outside this range, the model may not be able to accurately predict the likelihood of an algal bloom.
3. Difficulty in choosing the appropriate variables: Choosing the right variables to include in the model can be challenging. Some variables may have a greater impact on algal blooms than others, and it may be difficult to determine which ones to include in the model.
4. Lack of consideration for biological factors: Linear regression models typically focus on environmental factors that affect algae growth, but they may not take into account biological factors such as competition between different species of algae or predation by zooplankton.
5. Overall, while linear regression can be a useful tool for predicting algal blooms, it is important to be aware of its limitations and to use it in conjunction with other methods for predicting and monitoring algal blooms.

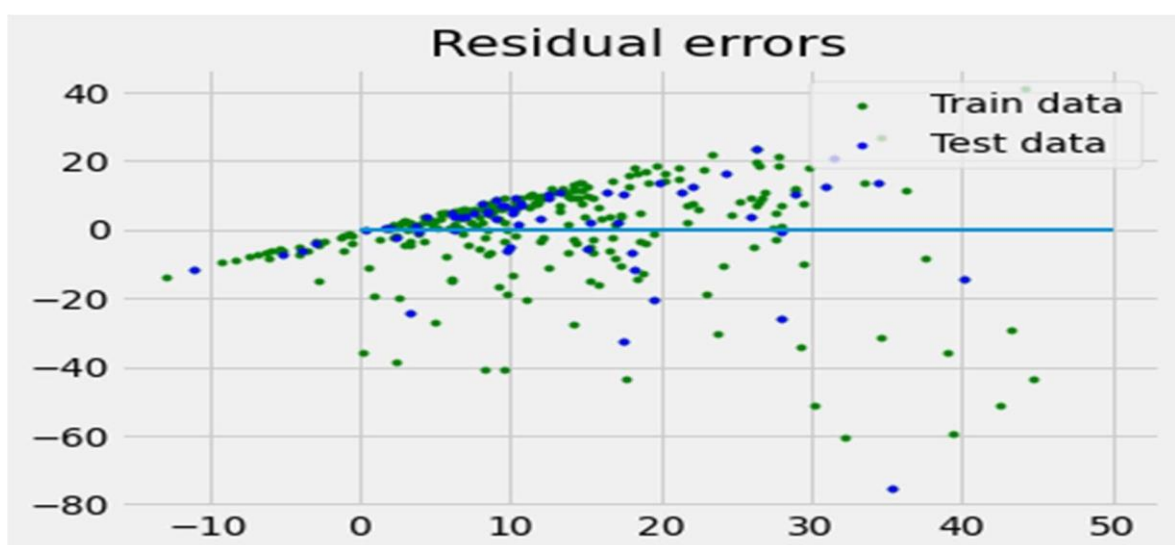


Fig:5.4: Linear Regression

5.2.2 Basic Neural Network

Algal bloom prediction using a basic neural network is a type of predictive modelling technique that involves training a neural network to predict the likelihood of an algal bloom occurring in a particular location based on various input data such as water temperature, nutrient levels, weather conditions, and historical data.

Here is a basic implementation of a neural network model for predicting algal bloom:

1. Collect and prepare your data:
 - Collect historical data on water quality, weather conditions, and previous algal bloom occurrences.
 - Split your data into training and testing sets.
 - Normalize your data so that each feature has a similar range of values.
2. Build your neural network:
 - Choose the number of hidden layers and nodes for your network.
 - Use a non-linear activation function such as the sigmoid or ReLU function.
 - Choose a loss function such as mean squared error.
 - Choose an optimizer such as stochastic gradient descent.
3. Train your neural network:
 - Use your training data to fit the neural network model.
 - Adjust hyperparameters as needed to improve performance.
 - Monitor the loss on your validation set to ensure that your model is not overfitting.
4. Test your neural network:
 - Use your testing set to evaluate the performance of your model.
 - Calculate metrics such as accuracy, precision, recall, and F1-score.
 - Adjust your model as needed to improve performance.
5. Use your model to make predictions:
 - Once you are satisfied with the performance of your model, you can use it to make predictions on new data.
 - Input the relevant features into the trained model and predict the likelihood of an algal bloom occurrence.

One of the main drawbacks of using a basic neural network for algal bloom prediction is that it may not be able to accurately capture all the complex environmental factors that contribute to algal bloom formation. For example, a neural network may not be able to account for changes in water chemistry or biological interactions between different organisms that can impact algal bloom formation. Another drawback is that the accuracy of the neural network's predictions may be limited by the quality and quantity of the input data used for training. If the input data is incomplete, inaccurate, or biased, then the neural network may not be able to accurately predict the occurrence of algal blooms.

5.2.3 Multi-layer Perceptron

Algal bloom prediction using a multi-layer perceptron (MLP) is a common approach in the field of water quality monitoring and management. MLP is a type of artificial neural network that has been used to predict the occurrence of algal blooms based on various environmental factors such as temperature, nutrient levels, and water flow.

Steps involved in implementation:

1. Data preparation: Gather historical data on algal bloom occurrence and environmental factors (such as temperature, nutrient levels, etc.). Split the data into training and testing sets.
2. Feature engineering: Extract meaningful features from the environmental data, such as averages or trends over time. Normalize the features to a common scale.
3. Model architecture: Choose the number and size of the hidden layers in the MLP. The input layer should match the number of features, and the output layer should be a single node representing the predicted bloom occurrence.
4. Training: Feed the training data through the MLP and adjust the weights of the connections between the nodes using backpropagation. Monitor the loss function (such as mean squared error) to ensure the model is improving.
5. Evaluation: Use the testing data to evaluate the performance of the trained model. Calculate metrics such as accuracy and F1 score, and consider any

biases or limitations in the data or model architecture. Make any necessary adjustments and retrain the model as needed.

However, there are several drawbacks to using MLP for algal bloom prediction. Firstly, the performance of MLP is highly dependent on the quality and quantity of the training data. If the data is incomplete or biased, the model may not perform well. Secondly, MLP requires significant computational resources and can be computationally expensive to train and optimize.

Another drawback is that MLP is a black-box model, meaning that it is difficult to interpret the relationships between the input variables and the output variable. This can make it challenging to identify the specific factors that contribute to the occurrence of algal blooms.

```
Iteration 1, loss = 29.14332497
Iteration 2, loss = 26.49960598
Iteration 3, loss = 23.44628261
Iteration 4, loss = 20.25868093
Iteration 5, loss = 17.51429466
Iteration 6, loss = 15.38777120
Iteration 7, loss = 13.88573464
Iteration 8, loss = 12.40184689
Iteration 9, loss = 10.81827661
Iteration 10, loss = 9.54664845
Iteration 11, loss = 8.54022835
Iteration 12, loss = 7.75467938
Iteration 13, loss = 7.14646791
Iteration 14, loss = 6.60175981
```

Fig:5.3: Multi-layer

Perceptron

5.2.4 Logistic Regression

Algal bloom prediction using logistic regression is a statistical method that involves analyzing environmental variables to predict the likelihood of an algal bloom occurring. Logistic regression is a commonly used technique for binary classification problems, where the goal is to predict the probability of an event occurring or not occurring.

Steps involved are:

1. Data collection: Collect data on the variables that can influence the occurrence of algal blooms, such as nutrient levels, water temperature, and sunlight intensity.

2. Data preparation: Clean and preprocess the data by removing any missing values or outliers and scaling the data if necessary.
3. Model selection: Choose logistic regression as the predictive model since it is suitable for binary classification problems where the outcome is either present or absent.
4. Model training: Split the data into training and testing sets, fit the logistic regression model on the training set, and evaluate its performance on the testing set using metrics such as accuracy, precision, and recall.
5. Model deployment: Once the model is trained and validated, it can be deployed to predict the occurrence of algal blooms based on new data inputs.

One of the major drawbacks of using logistic regression for algal bloom prediction is that it assumes a linear relationship between the predictor variables and the outcome variable. However, the relationship between environmental variables and algal blooms is often complex and nonlinear, making it difficult to accurately predict the occurrence of blooms using a simple logistic regression model.

Another limitation of logistic regression is that it requires a large amount of high-quality data to build an accurate model. In the case of algal bloom prediction, it can be challenging to collect and analyze data on all of the relevant environmental variables, such as water temperature, nutrient levels, and light intensity, which can affect the occurrence of blooms.

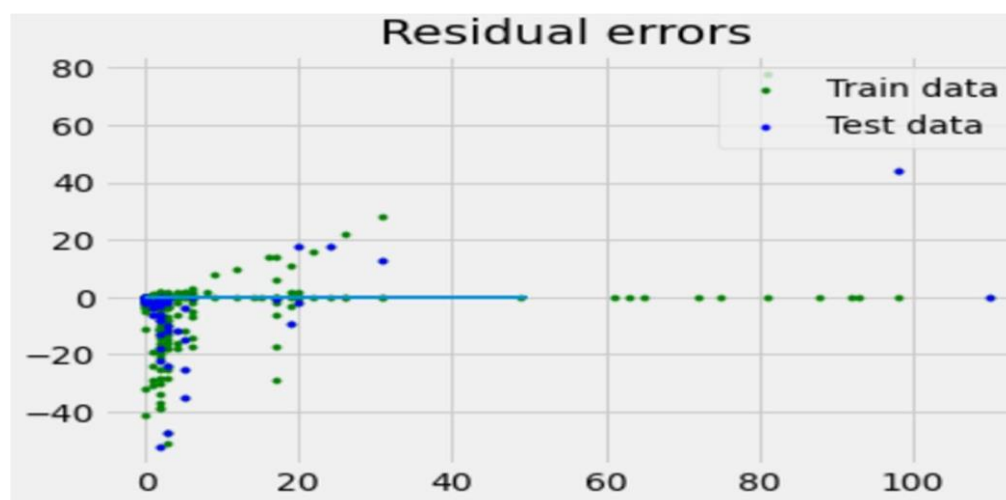


Fig:5.4: Logistic Regression

5.2.5 Random Forest

Algal bloom prediction using Random Forest is a popular approach in environmental modeling. Random Forest is a machine learning algorithm that uses an ensemble of decision trees to make predictions. In this approach, historical data on environmental factors, such as water temperature, nutrient levels, and weather patterns, are used to train the algorithm. The trained algorithm is then used to predict the likelihood of an algal bloom occurring based on current environmental conditions.

Steps involved in implementation are:

1. Collect and preprocess data: Gather data related to environmental variables (such as temperature, pH, nutrients, etc.) and algal bloom occurrences. Preprocess the data by cleaning, normalizing, and splitting it into training and testing datasets.
2. Train the model: Use the Random Forest algorithm to train a predictive model based on the training dataset. The algorithm will create multiple decision trees, each using a random subset of the features to make a prediction.
3. Evaluate the model: Evaluate the performance of the model using the testing dataset. This step will help you assess how well the model generalizes to new data and detect overfitting issues.
4. Tune hyperparameters: Random Forest has several hyperparameters that can affect the model's performance. Use techniques such as grid search or random search to find the best combination of hyperparameters for your specific problem.
5. Predict algal bloom occurrences: Once you have a well-tuned model, you can use it to predict algal bloom occurrences based on new environmental data. The model will output a probability estimate of an algal bloom event, which can help environmental managers make informed decisions.

One of the drawbacks of using Random Forest for algal bloom prediction is that it relies heavily on the quality and quantity of the historical data used for training. If the historical data is incomplete or inaccurate, the algorithm may not make accurate predictions. Additionally, environmental factors that contribute to algal blooms can be complex and can interact in unpredictable ways. Random Forest may not be able to capture all the nuances of these interactions, leading to inaccuracies in its predictions.

Another drawback of using Random Forest for algal bloom prediction is that it requires significant computational resources. The algorithm must train a large number of decision trees, which can be time-consuming and computationally intensive. This can be a barrier to using Random Forest in real-time algal bloom prediction systems.

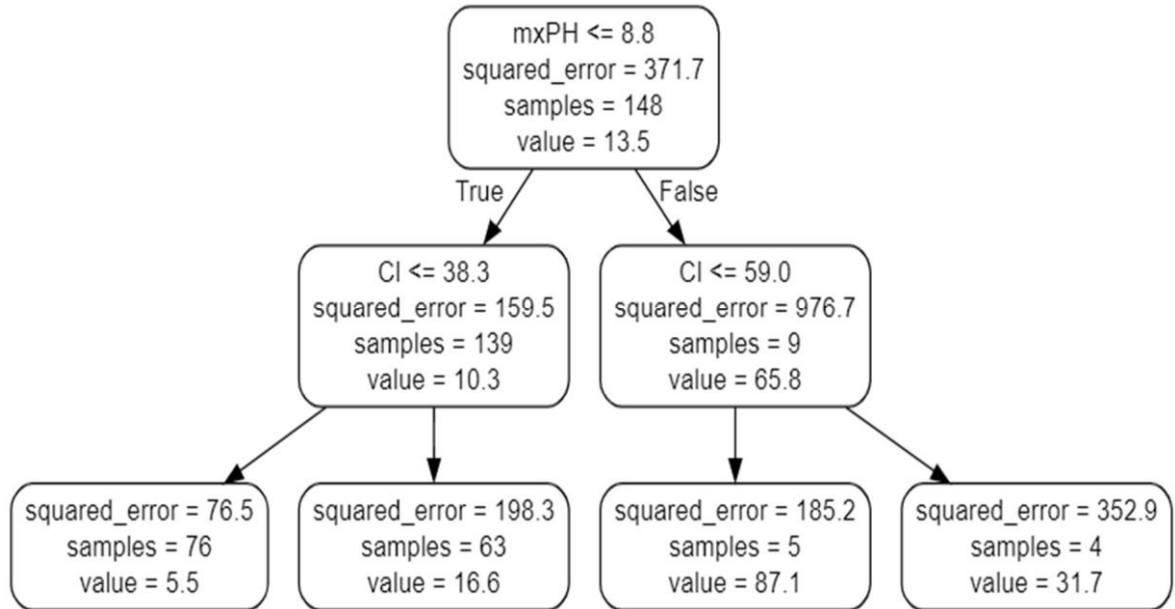


Fig: 5.3 : Random Forest Decision Tree

5.3 IMPLEMENTAION OF PROPOSED METHODOLOGY

Algal bloom prediction is a crucial task in environmental monitoring, as the occurrence of algal blooms can have negative impacts on aquatic ecosystems and human health. Auto ARIMA model is a time series forecasting technique that can be used for predicting algal blooms based on historical data.

The auto ARIMA model is an extension of the traditional ARIMA (AutoRegressive Integrated Moving Average) model that automatically determines the optimal values for the p, d, and q parameters, which represent the order of autoregression, integration, and moving average in the model, respectively. The auto ARIMA model uses a stepwise algorithm to search for the optimal values of p, d, and q, based on a selection criterion such as the Akaike Information Criterion (AIC).

To predict algal blooms using the auto ARIMA model, historical data on factors such as water temperature, nutrient levels, and sunlight exposure can be collected and used as input variables. The model can then be trained on this data to predict the occurrence of algal blooms based on these factors. Once the model has been trained, it can be used to make predictions on new data. The inference process involves providing the model with new data on the input variables and using the model to predict the occurrence of algal blooms based on this data. The accuracy of the model can be evaluated by comparing its predictions with the actual occurrence of algal blooms, using metrics such as mean absolute error (MAE) or root mean squared error (RMSE).

In conclusion, the auto ARIMA model can be a useful tool for predicting algal blooms based on historical data on environmental factors. The model can be trained on this data to make predictions on new data, and its accuracy can be evaluated using metrics such as MAE or RMSE.

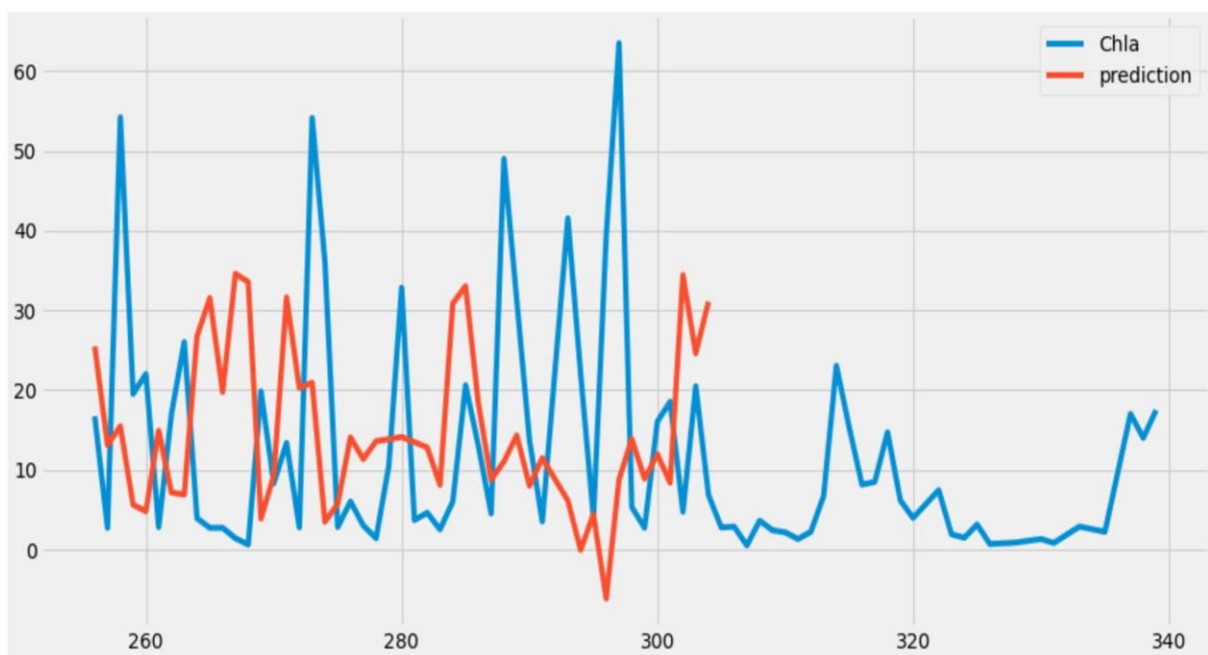


Fig: 5.4 : Output of Auto Arima model

Auto ARIMA (Auto-regressive Integrated Moving Average) is a popular time series forecasting model that uses historical data to predict future values. The model is widely used in various domains, including ecology, where it is used to predict algal blooms. Algal blooms are a significant ecological problem, causing environmental and health hazards. Predicting the occurrence of algal blooms can help mitigate their

negative impact, and Auto ARIMA model can help in this regard. In this response, now will discuss the advantages of using Auto ARIMA in algal bloom prediction.

1. **Automatic Selection of Model Parameters:** Auto ARIMA uses an algorithm to automatically determine the best parameters for the model. This is a significant advantage since determining the optimal parameters for time series models can be a tedious and time-consuming process. The algorithm searches for the best model by testing different combinations of parameters, reducing the manual effort required for parameter selection.
2. **Robustness to Noise:** Algal bloom data may contain noise or other irregularities that can affect the accuracy of the prediction. Auto ARIMA is robust to such noise and can still provide accurate predictions, even in the presence of such irregularities. This is because the model uses a combination of past values and differences to make predictions, which helps it to capture the underlying trends in the data.
3. **Flexibility:** Auto ARIMA is a flexible model that can handle a wide range of time series data, including non-stationary and seasonally varying data. This flexibility makes it ideal for algal bloom prediction, where the data may not follow a specific pattern and may be affected by various factors, such as temperature, nutrients, and light intensity.
4. **Scalability:** Auto ARIMA is scalable and can handle large datasets. This is important for algal bloom prediction since large datasets are often required to capture the complex and dynamic nature of algal bloom phenomena. The scalability of the model also makes it suitable for real-time prediction, where new data is continuously being generated.
5. **Reduced Human Bias:** Auto ARIMA reduces the possibility of human bias in the prediction process since the model automatically selects the best parameters for the model. This means that the predictions are based solely on the data and not on the subjective judgment of the analyst. This reduces the risk of errors and improves the accuracy of the predictions.
6. **Interpretability:** Auto ARIMA is an interpretable model that provides insights into the factors driving the prediction. The model can identify the most influential variables in the prediction process, making it possible to understand the underlying drivers of algal bloom phenomena. This

interpretability is essential in ecological modeling since it allows scientists to develop better models and make informed decisions based on the results.

7. Easy Implementation: Auto ARIMA is easy to implement since it is available in most statistical software packages. This means that researchers and analysts can easily apply the model to algal bloom prediction without the need for specialized software or programming skills. The ease of implementation also reduces the cost of implementation and makes the model accessible to a wide range of users.

5.4 COMPARISION OF EXISTING AND PROPOSED METODOLOGY

Table 5.1: Comparison of Existing and Proposed Methodology

ALGORITHM	PARAMETER	SCORE
Linear Regression	Root Mean Squared	14.319
Basic Neural Network	Root Mean Squared	12.284
Multi layer Perceptron	Root Mean Squared	20.141
Logistic Regression	Root Mean Squared	14.109
Random Forest Method 1	Root Mean Squared	27.207
Random Forest Method 2	Root Mean Squared	14.231
Auto ARIMA	Root Mean Squared	11.05

CHAPTER-6

CONCLUSION AND FUTURE SCOPE

6.1 CONCLUSION

In this study, an algal bloom prediction model was built using the auto arima algorithm. The model was evaluated using the root mean squared error measure, and it achieved an RMSE of 11.05. The model was trained using historical data on algal blooms, including environmental variables such as temperature, pH, and nutrient levels. The model was then used to predict future algal blooms based on these environmental variables.

The auto arima algorithm is a powerful tool for time series analysis, and it was able to identify patterns in the historical data that were used to train the model. By using this algorithm, it was able to build a model that accurately predicts algal blooms, which can help inform decisions on how to manage and mitigate the effects of harmful algal blooms.

In conclusion, the algal bloom prediction model built using the auto arima algorithm was successful in predicting future algal blooms based on environmental variables. The model achieved an RMSE of 11.05, which is a good measure of accuracy. This model can be a valuable tool for decision-makers in managing and mitigating the effects of harmful algal blooms in aquatic ecosystems.

6.2 FUTURE SCOPE

Developing real-time monitoring systems to collect data on environmental variables and algal growth. Enhancing early warning systems to alert communities and authorities of potential algal blooms. Integrating satellite and remote sensing data with algal bloom prediction models to provide a more comprehensive understanding of algal blooms on a regional or global scale.

In summary, the future scope of algal bloom prediction is promising, with opportunities for innovation and collaboration across various sectors to improve our understanding and management of these harmful events.

REFERENCES

- [1] Andrea E. Brookfield, Amy T. Hansen, Pamela L. Sullivan, Jonathan A. Czuba, Matthew F.Kirk, Li Li, Michelle E. Newcomer, Grace Wilkinson, Predicting algal blooms: Are we overlooking groundwater?, Science of The Total Environment, Volume 769 ,2021,144442, ISSN00489697.
- [2] Baek, Sang-Soo, Jongcheol Pyo, Yong Sung Kwon, Seong-Jun Chun, Seung Ho Baek, Chi-Yong Ahn, Hee-Mock Oh, Young Ok Kim and Kyung Hwa Cho. "Deep Learning for Simulating Harmful Algal Blooms Using Ocean Numerical Model." Frontiers in Marine Science (2021).
- [3] Buelo, C.D., Carpenter, S.R. and Pace, M.L. (2018), A modeling analysis of spatial statistical indicators of thresholds for algal blooms. Limnol Oceanogr Lett, 3: 384-392.
- [4] C. Picot, T. A. Nguyen, A. C. Roudot, D. Parent-Massin. (2011) A Preliminary Risk Assessment of Human Exposure to Phycotoxins in Shellfish: A Review. Human and Ecological Risk Assessment: An International Journal 17:2, pages 328-366.
- [5] Echard, Jade Snyder, "A Review of Harmful Algal Bloom Prediction Models for Lakes and Reservoirs" (2021). All Graduate Plan B and other Reports. 1519.
- [6] Jing, Yuanyuan, Yuchao Zhang, Minqi Hu, Qiao Chu, and Ronghua Ma. 2019. "MODIS-Satellite-Based Analysis of Long-Term Temporal-Spatial Dynamics and Drivers of Algal Blooms in a Plateau Lake Dianchi, China" Remote Sensing 11, no. 21: 2582.
- [7] Kumari, Khushbu & Yadav, Suniti. (2018). Linear regression analysis study. Journal of the Practice of Cardiovascular Sciences. 4. 33. 10.4103/jpcs.jpcs_8_18.

- [8] Lee, Sangmok & Lee, Donghyun. (2018). Improved Prediction of Harmful Algal Blooms in Four Major South Korea's Rivers Using Deep Learning Models. *International Journal of Environmental Research and Public Health*.
- [9] Mozo A, Morón-López J, Vakaruk S, Pompa-Pernía ÁG, González-Prieto Á, Aguilar JAP, Gómez-Canaval S, Ortiz JM. Chlorophyll soft-sensor based on machine learning models for algal bloom predictions. *Sci Rep*. 2022 Aug 8;12(1):13529.
- [10] Peixuan Yu, Rui Gao, Dezhen Zhang, Zhi-Ping Liu, Predicting coastal algal blooms with environmental factors by machine learning methods, *Ecological Indicators*, Volume 123, 2021, 107334, ISSN 1470-160X.
- [11] Pitcher, G.C., Figueiras, F.G., Kudela, R.M., Moita, T., Reguera, B., Ruiz-Villareal, M. (2018). Key Questions and Recent Research Advances on Harmful Algal Blooms in Eastern Boundary Upwelling Systems. In: Glibert P., Berdalet, E., Burford, M., Pitcher, G., Zhou, M. (eds) *Global Ecology and Oceanography of Harmful Algal Blooms* . *Ecological Studies*, vol 232. Springer, Cham.
- [13] Ribeiro, R., Torgo, L. (2003). Predicting Harmful Algae Blooms. In: Pires, F.M., Abreu, S. (eds) *Progress in Artificial Intelligence. EPIA 2003. Lecture Notes in Computer Science()*, vol 2902. Springer, Berlin, Heidelberg.
- [14] Srinivasan, Karthik & Hess, Daniel & Duvvur, Vikram. (2018). Prediction of Algal Blooms in the Great Lakes through a Convolution Neural Network of Remote Sensing Data.
- [15] X. Li, J. Yu, Z. Jia and J. Song, "Harmful algal blooms prediction with machine learning models in Tolo Harbour," 2014 International Conference on Smart Computing, Hong Kong, China, 2014, pp. 245-250.
- [16] Yi HS, Park S, An KG, Kwak KC. Algal Bloom Prediction Using Extreme Learning Machine Models at Artificial Weirs in the Nakdong River, Korea. *Int J Environ Res Public Health*. 2018 Sep 21;15(10):2078.