UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Monitoring Sea Water Pollution from Satellite data

Author:
Yevhen STEPANOV

Supervisor:
Dmytro KARAMSHUK

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2023

Declaration of Authorship

I, Yevhen STEPANOV, declare that this thesis titled, “Monitoring Sea Water Pollution from Satellite data” and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

"The obstacle is the path."

Marcus Aurelius

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Monitoring Sea Water Pollution from Satellite data

by Yevhen STEPANOV

Abstract

This study is motivated by the Ocean Decade declared by the United Nations. We employed machine learning techniques to detect and delineate areas of pollution in the coastal zone of Great Britain, utilizing pollution reports from the Department for Environment Food And Rural Affairs (DEFRA) and the ocean monitoring datasets from the European Space Agency (ESA).

In this study, feature engineering was performed on chlorophyll concentration data. Two datasets were constructed: one with statistical metrics (mean, median, standard deviation, and percentiles) as features, and another with individual cells of the chlorophyll concentration matrix as features, utilizing different matrix n sizes, where n is from 3 to 11, where each element or pixel of the matrix represented a 1km × 1km area. Logistic regression, decision trees, random forest classifier, gradient boosting classifier, and LeNet models were applied. Hyper parameter tuning was conducted to optimize the performance of each model. Among the models, the gradient boosting classifier achieved the highest accuracy of 95.21%. Additionally, the F1 score was determined to be 0.2445, the ROC AUC was 0.7659, and the precision-recall AUC (PR-AUC) was found to be 0.1821.

Detecting and delineating areas of pollution can greatly assist cleaning services in efficiently carrying out their job, resulting in improved remediation and restoration efforts. The identification of pollution areas holds significant implications for the fishing industry, as it enables informed decision-making regarding fishing practices and resource management, ensuring the sustainability and viability of the sector. Moreover, the accurate detection and delineation of pollution areas have the potential to generate substantial economic, social, and environmental benefits by facilitating targeted interventions, protecting ecosystems, preserving marine resources, and fostering a healthier and more resilient environment.

The findings of this study provide valuable insights into the efficacy of classification approaches in identifying and mapping pollution sites in coastal regions using pollution reports from DEFRA.

Acknowledgements

I would like to express my heartfelt gratitude to Dmytro Karamshuk for his exceptional supervision throughout this research project. His invaluable feedback, inspiration, and encouragement have been instrumental in shaping the outcome. I am deeply thankful to the Ukrainian Catholic University and the Faculty of Applied Sciences for organizing the Master's Program in Data Science, which has provided me with an enriching educational experience. I would like to wholeheartedly express my sincere gratitude to Oleksii Molchanovsky and Ruslan Partsey for their exceptional leadership of the program and their remarkable ability to resolve organizational matters. My deepest gratitude goes to my family, whose unwavering motivation has been a constant source of strength and encouragement throughout this endeavor. I would also like to express my profound thanks to my friends and colleagues, whose assistance and inspiration have been invaluable during the two years of study. Furthermore, I am immensely grateful to the Armed Forces of Ukraine for ensuring a secure environment that allowed me to complete this work. Without the collective support and contributions of all those mentioned, this accomplishment would have remained beyond reach.

Contents

List of Figures

List of Tables

This research is dedicated to the world. It is my sincere hope that the findings and outcomes of this study will serve as a contribution to future research endeavors aimed at monitoring water pollution on a global scale. May this work inspire and motivate further studies, leading to timely detection and effective response to contaminants present in water surfaces worldwide. By dedicating our efforts to this cause, we aspire to foster greater awareness and instigate positive change in safeguarding the precious resources of our planet for current and future generations.

Chapter 1

Introduction

The United Nations (U.N.) has designated the period between 2021 and 2030 as the "Decade of Ocean Science for Sustainable Development." (UNESCO, 2021) The primary objective of this initiative is to promote "the science we need for the ocean we want." (UNESCO, 2021). In pursuit of this goal, the U.N. has invited scientists worldwide to propose ideas for Decade Actions. These collaborative projects seek to facilitate the creation of a healthier, more sustainable ocean by 2030 (Minogue and King, 2021).

Seawater pollution poses a substantial threat to the health and sustainability of marine ecosystems, impacting aquatic organisms and human populations reliant on these resources. Traditional methods for monitoring seawater quality rely on in-situ measurements such the one provided by the UK Department for Environment, Food and Rural Affairs (DEFRA) (Merchant, 2017). In-situ measurement refers to the process of collecting data directly from the location or environment of interest, typically through physical sensors or instruments deployed on-site. It provides real-time, accurate, and context-specific measurements, allowing for precise analysis and understanding of the phenomenon being studied. DEFRA (Department for Environment, Food and Rural Affairs) monitors approximately 400 seaside locations across the United Kingdom for various environmental factors, including bathing water quality. These monitoring activities aim to assess and maintain the cleanliness and safety of coastal areas for recreational purposes. The data collected by DEFRA includes measurements and reports related to pollution incidents, water quality parameters, and other relevant environmental indicators. The monitoring covers a wide range of factors, such as bacterial contamination, chemical pollutants, nutrient levels, and other potential sources of pollution.

While in-situ pollution monitoring is limited in spatial and temporal coverage, advancements in satellite remote sensing technology have provided a promising alternative for ocean monitoring on a much larger scale.

According to the (Agency., 2023), one of the symptoms of degraded water quality condition is the increase of algae biomass as measured by the concentration of chlorophyll a. It would validate the application of machine learning to detect pollution: chlorophyll is visible from space and chlorophyll indicates pollution. Accordingly, we can detect pollution by detecting chlorophyll on satellite images.

Remote sensing of chlorophyll concentration has been used as an indicator of water quality and primary productivity affected by water pollution incidents (Lu, 2022). Chlorophyll is a pigment found in phytoplankton, which form the base of the marine food chain. Chlorophyll concentration can be remotely monitored from space using European Space Agency (ESA) satellites, which provide data at a spatial resolution of 1km. This monitoring capability allows for the daily assessment of chlorophyll concentration levels, enabling continuous tracking of changes and patterns in chlorophyll distribution over time.

This master's thesis aims to develop and evaluate a novel approach for monitoring seawater pollution using satellite data from the European Space Agency (ESA). The study will leverage machine learning techniques and the remote sensing data from ESA to detect and map areas of pollution in the coastal zone of Great Britain, utilizing pollution reports from the UK's Department for Environment, Food and Rural Affairs (DEFRA) (Merchant, 2017).

In this study, we pose the following research questions to investigate and address our objectives:

1. What is the optimal shape of data (3×3, 11×11, etc.) and statistical features to detect water pollution based on the concentration of chlorophyll?

2. Which machine learning model performs best at detecting water pollution based on the concentration of chlorophyll?

In this work, various machine learning models including Logistic Regression (regression, 2016), Decision Trees (tree, 2016), Random Forest (forest, 2016), Gradient Boosting Classifier (boosting, 2016), and LeNet (Y. Lecun, 1998) were used detect the cases of water pollution by classifying the segments of an image as polluted or not polluted. These models were chosen for their suitability in classification tasks and their ability to provide insights into the research objectives. The best model we considered achieved an F1 score of 0.24, Receiver Operating Characteristic Area Under the Curve (ROC AUC) of 0.76 and the accuracy of 95.21%.

My contributions can be:

1. Collected a dataset of pollution and chlorophyll concentration in seawater around England.

2. Compared multiple machine learning models across multiple dimensionality of data, to find the approach most suitable for early detection of seawater pollution.

3. Presented first results obtained from a machine learning model trained to detect pollution of seawater based on the concentration of chlorophyll and discussed its practical applicability Furthermore, there has been a notable absence of studies that have conducted a direct comparison between chlorophyll concentration and cases of contamination in the context of seawater pollution. This research seeks to address this gap by examining the relationship between chlorophyll concentration and contamination incidents within England's seawater.

Effective monitoring and management of seawater pollution require a multidisciplinary approach, involving the integration of various data sources and the collaboration of scientists, policymakers, and stakeholders.

1.1 Motivation

The development of a classification model utilizing satellite data to predict pollution incidents based on chlorophyll concentration can generate substantial economic, social, and environmental benefits. Firstly, the model can help identify areas and times where pollution is most likely to occur, enabling proactive measures to prevent incidents. This can mitigate economic losses experienced by local industries, such as fishing or tourism. (Directive, 2020). From an economic perspective, early detection allows for prompt intervention and mitigation measures, minimizing the economic impact of pollution. For example, industries located downstream from a potential pollution source can take immediate actions to protect their water supply, reducing the risk of costly production disruptions or damage to equipment.

Secondly, the model can assist in the development of effective pollution reduction strategies. Policymakers can implement targeted policies by identifying regions and periods most susceptible to pollution, resulting in efficient resource use and cost savings related to cleanup efforts. In terms of environmental benefits, early identification of pollution incidents can prevent or minimize ecological damage. For instance, in a marine ecosystem, the early detection of an oil spill enables the deployment of containment measures, reducing the spread of the pollutants and minimizing harm to marine life, habitats, and coastal ecosystems.

Furthermore, the model's development can enhance environmental monitoring and regulation, promoting sustainable development practices, and attracting environmentally conscious investors to stimulate sustainable industry growth. Socially, early identification enables timely communication and public awareness, promoting public health and safety. For instance, in the case of a chemical spill in a community, early detection allows authorities to issue timely warnings, facilitating evacuation procedures and minimizing potential health risks to residents.

The early identification of pollution incidents can lead to substantial economic benefits by minimizing financial losses, social benefits by safeguarding public health, and environmental benefits by preventing or mitigating ecological damage. These examples illustrate the multi-faceted advantages that arise from timely action and proactive measures in response to pollution incidents.

The future outlook for integrating satellite-based Earth observation into water quality monitoring for the Water Framework Directive (WFD) is promising. Currently, **40%** of surface waters meet the good status, and efforts are being made to improve the status of the remaining waterbodies by 2027. Satellite products offer significant advantages, including monitoring the effectiveness of management measures, providing comprehensive assessments of waterbody structure and function, and extending monitoring to currently unmonitored waterbodies.(Directive, 2020)

The model's predictions can also optimize water quality monitoring staff by freeing up resources to allocate to areas with insufficient staff. This can ensure efficient and comprehensive water quality monitoring, leading to the timely identification and mitigation of pollution incidents.

Moreover, the model can identify the source of pollution by wind direction when pollution has already occurred, allowing for prompt action to contain and mitigate its effects. This can significantly benefit industries reliant on clean water, such as aquaculture and recreational water sports.

The EU Copernicus space program has invested €7.5 billion between 2008 and 2020 to generate terabytes of land and ocean observations daily. The program's constellation of satellites is guaranteed until at least 2030, with plans to replace aging sensors. This ensures that member states can rely on satellite assets for their monitoring requirements. Over the years, national expertise and international collaboration in translating satellite data into water quality metrics have grown, with scientifically rigorous methods published in peer-reviewed papers.(Directive, 2020)

In conclusion, the development of a classification model predicting pollution incidents based on chlorophyll concentration from satellite data can yield significant economic benefits. This includes identifying pollution hotspots, reducing economic impacts, promoting sustainable practices, optimizing water quality monitoring staff, enhancing recreational water activities, and preventing damage to the eco-fauna in the water.

1.2 Thesis Goal

The aim of this study is to construct a classification model that can predict pollution incidents in specific points based on chlorophyll concentration using open data from pollution reports and satellite data from Copernicus Marine Service. The study emphasizes the potential benefits of such a model, including efficient use of resources and cost savings related to cleanup efforts, the ability to identify pollution sources promptly, and the optimization of staff who measure water quality. By accurately predicting pollution incidents, the model can aid in formulating effective pollution reduction strategies and in implementing targeted policies to reduce pollution in vulnerable areas, ultimately promoting sustainable practices and enhancing the quality of life for people who rely on clean water. The objective of this thesis is to devise a methodology for identifying the existence of pollution by utilizing data on chlorophyll concentration. This will involve building classification models to classify instances as either polluted or non-polluted based on their chlorophyll concentrations. The performance of these models will be evaluated using appropriate metrics to identify the best-performing model. Furthermore, this research aims to gain insights into the relationship and interaction between pollution incidents and chlorophyll concentrations to enhance our understanding of this complex phenomenon.

1.3 Thesis Structure

1. Chapter 1 This chapter serves as an introduction to the thesis, providing an overview of the topic, justification for the research's significance, motivation behind the study, and the goals to be achieved and explains the novelty of our study.

2. Chapter 2 provides a summary of the related works in the field, offering an overview of existing papers, studies, and research that are relevant to the subject of the thesis. This chapter serves to situate the current research within the broader context of existing knowledge and helps identify gaps or areas for further exploration.

3. Chapter 3 provides a detailed description of the datasets used in the research, including information on their sources, acquisition methods, and relevant characteristics. This chapter also covers the process of obtaining and merging the datasets, as well as the necessary data preprocessing steps.

4. Chapter 4 presents the methodology employed in the research, outlining the approach to feature engineering, the types of models utilized, the metrics used for evaluation, dataset balancing techniques, cross-validation procedures, and the process of splitting and tuning hyperparameters.

5. Chapter 5 focuses on conducting experiments with different datasets, specifically exploring their varying dimensionalities. This chapter provides a summary and insights gained from these experiments.

6. Chapter 6 serves as the conclusion of the master thesis, summarizing the key findings, contributions, and implications of the research. Additionally, this chapter outlines potential avenues for future research and areas that could benefit from further exploration.

Chapter 2

Related Work

2.1 Monitoring of Water Pollution in Diyala River using High Resolution Satellite Image (A.H.Kadhim, 2012)

The article "Monitoring of Water Pollution in Diyala River using High Resolution Satellite Image" (A.H.Kadhim, 2012) discusses the use of remote sensing technology to monitor water pollution in Diyala River, Iraq. The study aimed to increase the capability of detecting and monitoring the quality of water resources affected by wastewater treatment plant disposal or industrial pollution.

The authors used high-resolution satellite imagery to discriminate several spectrally different classes of water using digital image processing methods, including supervised and unsupervised classification. They also used Landsat 8 imagery and the Water Quality Index (WQI) method to detect changes in water quality between 2013 and 2019. The study found that remote sensing analysis provides several advantages over traditional methods, allowing for effective and quantitative results.The WQI(Water Quality Index) value decreased from 76.29 in 2013 to 63.95 in 2019, indicating a decline in water quality. The authors also identified specific areas of the river where water quality was particularly poor. They suggest that the use of high-resolution satellite imagery can help identify areas where water pollution is most severe, allowing for targeted intervention to improve water quality.

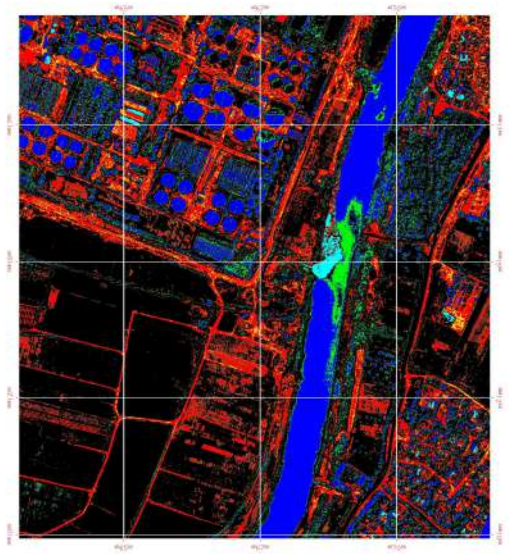


FIGURE 2.1: The Supervised Classification Result

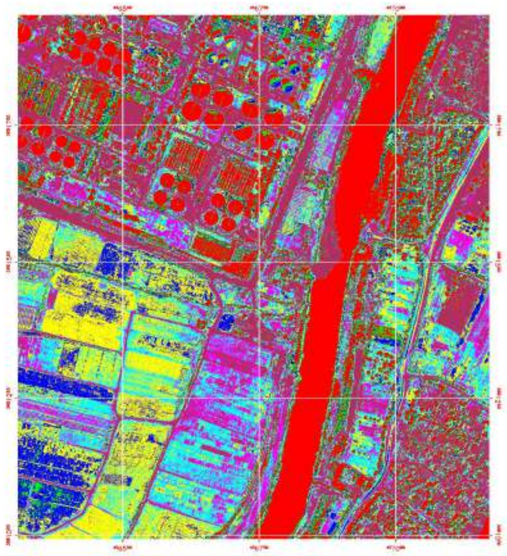


FIGURE 2.2: 7-Class Unsupervised Classification

The Water Quality Index (WQI) uses the four water quality indicators: chlorophyll-a, TN, TP, and turbidity. The WQI was calculated for each sublagoon area by converting annual offsets to a percentage scale, resulting in a final score ranging from 0% to 100% (A.H.Kadhim, 2012).

The study identified areas of the river where water quality was particularly poor and recommended the use of high-resolution satellite images for water pollution detection and monitoring. The findings can be useful for policymakers and environmental agencies to monitor and manage water resources more effectively. Overall, remote sensing analysis methods can play a vital role in investigating global resources, estimating land use, and monitoring environmental quality such as water pollution. The use of high-resolution satellite imagery provides a cost-effective and efficient method for monitoring water quality and identifying areas that require intervention to protect human health and the environment.

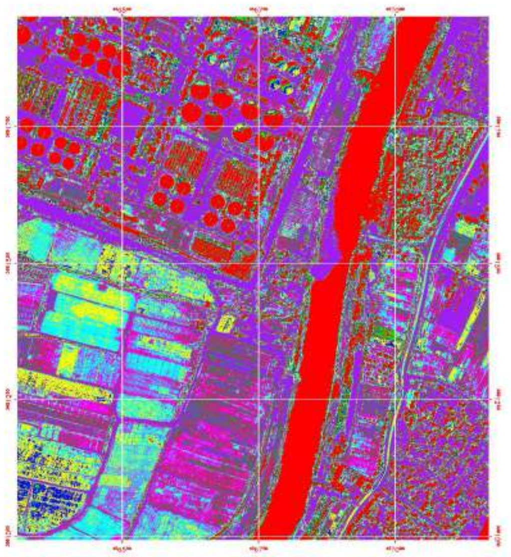


FIGURE 2.3: 9-Class Unsupervised Classification

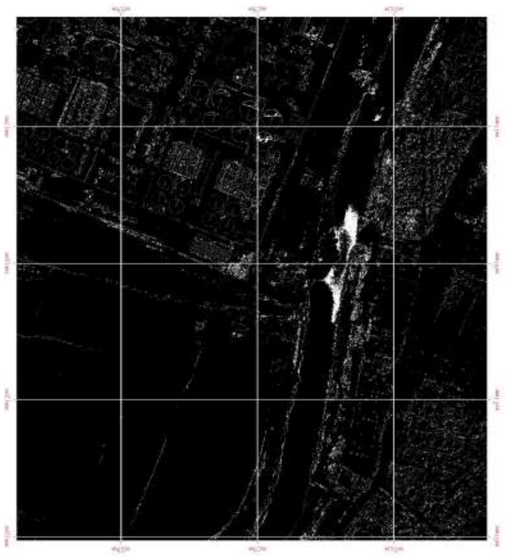


FIGURE 2.4: Threshold Value Pollution Detection

In contrast to the present study, our research centers on a different geographical region and incorporates data from the Copernicus Marine Service alongside DEFRA data, providing a more comprehensive perspective on pollution incidents. Additionally, we use alternative feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

2.2 A new approach to monitor water quality in the Menor sea (Spain) using satellite data and machine learning methods (Casanova, 2021)

In the study (Casanova, 2021) by Casanova, a machine learning approach utilizing Sentinel-2 data was proposed to estimate the concentration of chlorophyll-a in the Menor Sea, a coastal lagoon in Spain. The authors used Random Forest, support vector machine, Artificial Neural Network, and Deep Neural Network algorithms under three feature selection scenarios and several spectral indices in combination with Sentinel 2 bands. The study aimed to provide cost-effective and near-real time information for monitoring the water quality of the Menor Sea, which is declared as a sensitive area to eutrophication due to human activities. The results demonstrated the possibility of estimating chl-a concentration in a cost-effective manner and providing near-real time information for local authorities, tourism, and fishing industry. Remote sensing techniques using satellite data can improve insights about water quality, distribution of toxin-producing algae, and aquatic biogeochemical cycling. The study showed the suitability of Sentinel-2 satellite data for mapping different water quality parameters and the application of machine learning methods to measure marine or lake events such as primary production, harmful algal blooms or algae blooms. The methodology may contribute to improving the predictability of chlorophyll-a concentrations to mitigate the negative effects of high concentrations of phytoplankton and algae over the local populations.

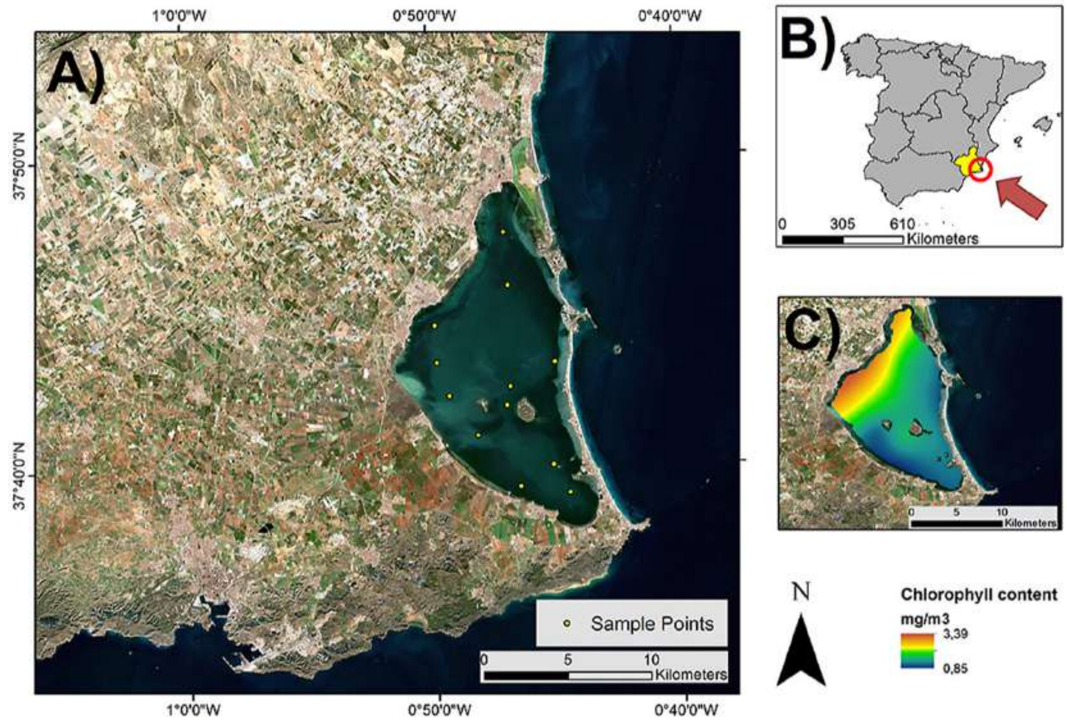


FIGURE 2.5: The Menor sea is an hypersaline coastal lagoon located in the south-east of Spain

The study compared the performance of four machine learning algorithms (random forest, support Vector Machine with Radial Basis Function kernel, ANN, and DNN) in predicting chl-a concentration in the Menor sea, Spain. The best results were achieved by the rf model without feature reduction, with an R2 of 0.92 and RMSE of 0.82 mg/m3. Feature reduction significantly decreased the performance of random forest and support vector machine with radial basis function kernel for one scenario, while the ANN method performed the best under another scenario.

In contrast to the present study, our research centers on a different geographical region and incorporates data from the Copernicus Marine Service alongside DEFRA data, providing a more comprehensive perspective on pollution incidents. Additionally, we use alternative feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

2.3 Detection and Monitoring of Marine Pollution Using Remote Sensing Technologies (Nichol, 2018)

The article (Nichol, 2018) discusses the use of remote sensing technology for monitoring marine pollution, which has become a major concern due to human activities. The article highlights the benefits of using aerial and spaceborne sensors for monitoring oil and chemical spills, sewage, high suspended solids, algal blooms, and solid waste in coastal areas. The article also discusses the technical limitations of the technology, such as the dynamic nature of pollutants and the limited information on the specific spectral response of pollutants. The use of active and hyperspectral airborne sensors is considered superior to spaceborne sensors for monitoring coastal and estuarine pollutants due to their real-time and detailed monitoring capability. Estuarine pollutants are substances or contaminants that are introduced into estuarine environments, leading to adverse effects on water quality and the ecological balance of these semi-enclosed coastal systems. These pollutants can include industrial waste, agricultural runoff, sewage, heavy metals, pesticides, and other harmful substances that have the potential to harm aquatic life and impact the overall health of estuarine ecosystems. In our work, we opt to use spaceborne sensors due to their accessibility in the public domain and their capability to provide coverage over a broader geographical expanse. This choice allows for easier access to data. Marine managers and researchers have expressed significant concern regarding heavy metal pollution, prompting studies to explore the use of airborne hyperspectral data for addressing this issue. Recent advancements in software and computation power have facilitated the wider utilization of remote sensing systems in managing marine resources and pollution. The article concludes that remote sensing technology offers valuable insights into pollution events in environmentally sensitive marine regions. Furthermore, with the continuous progress of remote sensing sensors, more advanced methods are anticipated to emerge for effectively monitoring marine pollution in the future.

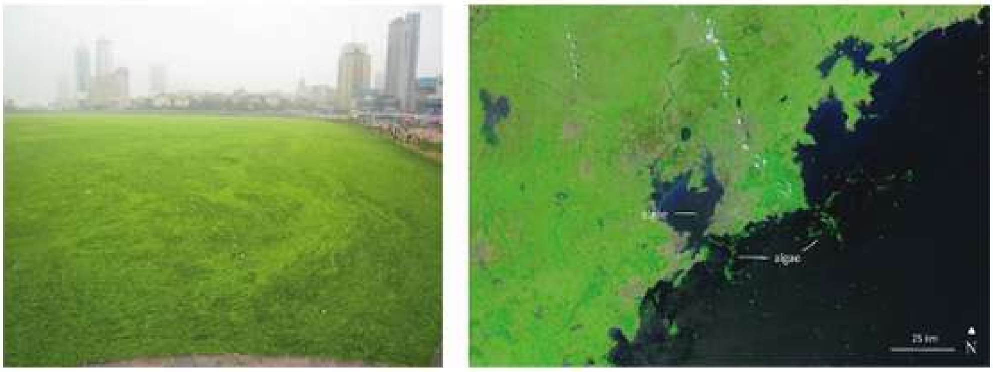


FIGURE 2.6: Spread of green algae along the coast of Qingdao in 2008, when summer Olympics was planned in this coast (source Corey SheranFlickr) (right) and algae visible in MODIS false color image (shortwave, NIR, and Red) (source MODIS rapid response project at NASAGSFC) (left)

Aerial and spaceborne sensors provide comprehensive information about marine pollutants, but there are technical limitations in assessing detailed information about pollutants. Active and hyperspectral airborne sensors are considered superior for monitoring coastal and estuarine pollutants, while spaceborne sensors are reliable for large-scale ocean monitoring. The article highlights the importance of collaboration between the research community and government for using the full potential of remote sensing data in marine pollution management. Recent developments in software and computation power have led to increased use of remote sensing data for managing marine resources and pollution. Based on the findings, the article concludes that the continuous advancement of remote sensing sensors will drive the development of sophisticated methods in the future, specifically tailored for monitoring marine pollution. In contrast to the present study, our research centers on a different geographical region and incorporates data from the Copernicus Marine Service alongside DEFRA data, providing a more comprehensive perspective on pollution incidents. Additionally, we use alternative feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

2.4 A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges (Chen, 2022)

Chen (Chen, 2022) aims to discuss the use of remote sensing data for estimating water quality parameters, particularly Total Suspended Matter (TSM) and Chlorophyll-a (Chl-a) concentration. Different methods have been developed to estimate these parameters using hyperspectral data and multispectral data, and remote sensing has been proven to be an effective tool for monitoring water quality parameters on a regional scale.

Various satellites and sensors are available for remote sensing of water quality, and four different modes of water quality parameter retrieval are commonly used: empirical, analytical, semi-empirical, and artificial intelligence (AI) modes. The empirical mode uses statistical regression formulas to establish a correlation between ground-measured water quality parameter values and the reflectance of specific bands or combinations of bands. Bio-optical models and radiation transmission models are employed in the analytical mode to simulate the propagation of light in water bodies. In contrast, the semi-empirical mode integrates the simplicity of the empirical mode with the accuracy of the analytical mode. The artificial intelligence mode is based on machine learning algorithms that can handle complex and nonlinear relationships between the water quality parameters and remote sensing data.

However, these models face challenges in terms of spatial dependency, temporal limitations, and spectral interaction, which affect their applicability. Possible solutions to these challenges include developing new methods and data, such as regional generalized additive models, machine learning-based models, and environmental variables dependence. Additionally, airborne data and multisensory satellite data can be used to enhance the models' applicability. Combining different methods is also a solution to improve the models' applicability. Overall, remote sensing data combined with these different modes of water quality parameter retrieval can provide a powerful tool for monitoring and managing water quality on a regional scale.

The hybrid spatio-temporal-spectral fusion model for identifying water pollution includes several techniques such as principal component analysis (PCA) (analysis, 2016), non-negative matrix factorization (NMF) (factorization, 2016), and regression analysis. The model integrates spatial, temporal, and spectral data to improve the accuracy of water pollution identification. PCA (analysis, 2016) is used to reduce the dimensionality of the data, while NMF (factorization, 2016) is used for source separation and to identify different pollution sources. Regression analysis is used to predict the concentration of pollutants based on the identified pollution sources. The model also includes a data fusion step that integrates the results from different techniques to provide a more comprehensive understanding of water pollution. Overall, the hybrid spatio-temporal-spectral fusion model is an effective tool for identifying water pollution sources and can aid in the development of targeted pollution control strategies. Overall, the article provides a comprehensive understanding of remote sensing water quality monitoring and sets the groundwork for future research in the field.

In contrast to the present study, our research centers on a different geographical region and incorporates data from the Copernicus Marine Service alongside DEFRA data, providing a more comprehensive perspective on pollution incidents. Additionally, we use alternative feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

2.5 Use of the Sentinel-2 and Landsat-8 Satellites for Water Quality Monitoring: An Early Warning Tool in the Mar Menor Coastal Lagoon (Navarro, 2022)

Navarro (Navarro, 2022) aims to demonstrate the suitability and consistency of using a joint constellation of Landsat-8 and Sentinel-2 satellites for estimating indicators of water quality in the Mar Menor coastal lagoon, which is highly vulnerable to eutrophication. The study used biogeochemical parameters, such as turbidity and chl-a, to validate the satellite imagery's performance in estimating water quality.

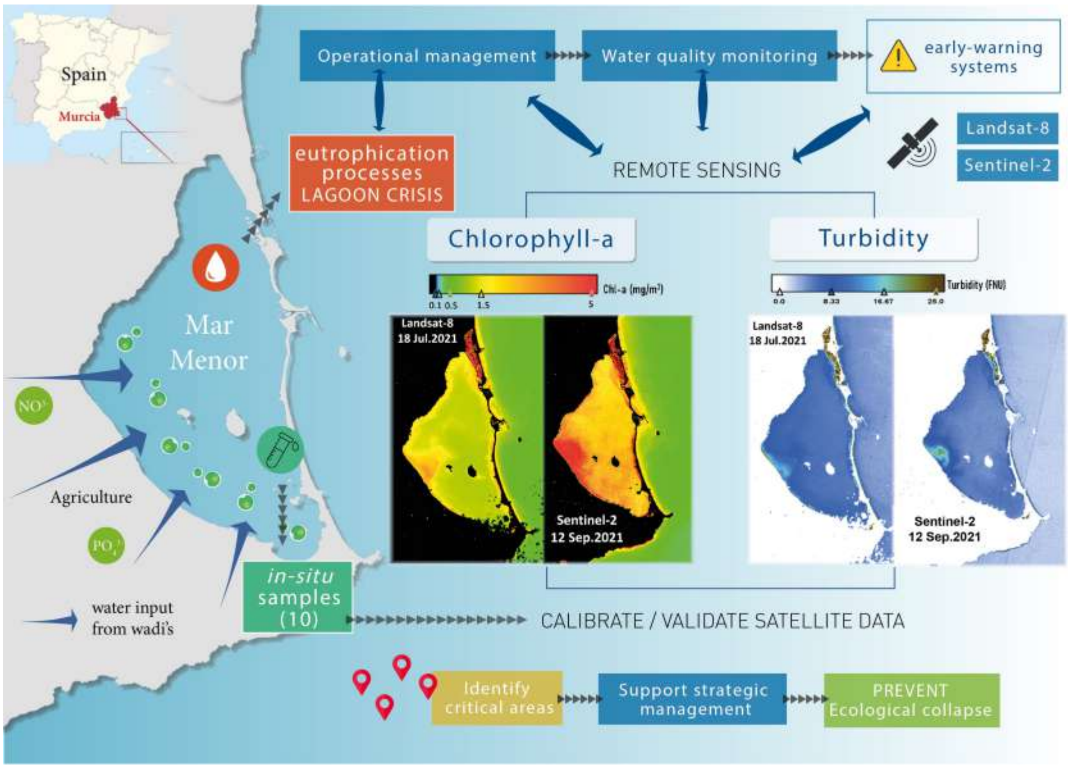


FIGURE 2.7: Satellite Graphical Workflow

The evaluation of the atmospheric and sunglint correction was performed using the ACOLITE (RBINS, 2021) software for both satellites. The study found that using both satellites in tandem can improve mapping strategies and provide appropriate information on a systematic basis and in a cost-effective way. Due to time constraints in our study, we do not incorporate data from multiple satellites. The satellite imagery was capable of early detection of chl-a levels above 3 mg/m3, which generally triggered subsequent blooms during recent years. Multitemporal maps were produced, and the highest turbidity and chl-a levels were consistently located in the western section of the lagoon, particularly at the mouth of the draining Albujon watercourse.

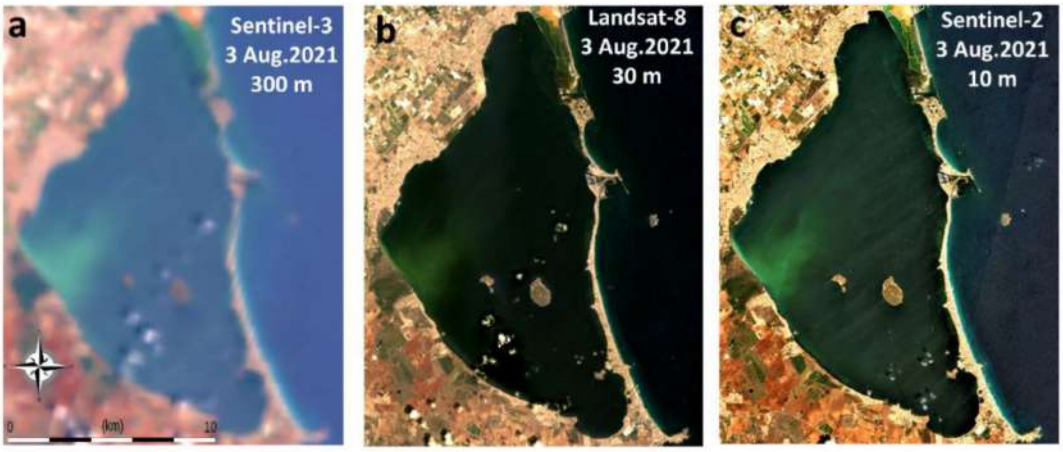


FIGURE 2.8: RGB (Red-Green-Blue) composite image on 3 August 2021 of (a) Sentinel-3 satellite (300 m spatial resolution), (b) Landsat-8 satellite (30 m spatial resolution), and (c) Sentinel-2 satellite (10 m spatial resolution).

The article "Use of the Sentinel-2 and Landsat-8 Satellites for Water Quality Monitoring: An Early Warning Tool in the Mar Menor Coastal Lagoon" (Navarro, 2022) focuses on the use of remote sensing technology for water quality monitoring in the Mar Menor coastal lagoon. In this study, the MODIS Standard (OC3) algorithm was used to estimate the concentration of chlorophyll-a from satellite measurements of ocean color obtained by Sentinel-2 and Landsat-8 satellites. The study demonstrated the potential of the OC3 algorithm in providing early warning of harmful algal blooms and other water quality issues in the Mar Menor lagoon. The results showed a significant correlation between the chlorophyll-a concentration and water quality parameters such as dissolved oxygen, pH, and temperature. This study highlights the usefulness of the OC3 algorithm in conjunction with satellite data for real-time monitoring of water quality in coastal areas, which can help in the timely detection of environmental hazards and the implementation of appropriate management strategies.

The MODIS Standard (OC3) (Gathot Winarso, 2014) algorithm is a method for deriving the concentration of chlorophyll-a in ocean waters from satellite measurements of ocean color obtained by the Moderate Resolution Imaging Spectroradiometer (MODIS). This algorithm uses blue and green spectral bands to estimate the absorption of phytoplankton pigments and the scattering by particles in the water column. The concentration of chlorophyll-a is then calculated based on empirical relationships between the absorption and scattering coefficients and the concentration of chlorophyll-a.

The study highlights the potential of using satellite imagery as an innovative tool to support decision-makers in implementing a joint monitoring strategy and characterizing water quality distribution in the lagoon. The findings of this study could have significant implications for managing water quality in coastal regions vulnerable to eutrophication. In contrast to the present study, our research centers on a different geographical region and incorporates data from the Copernicus Marine Service alongside DEFRA data, providing a more comprehensive perspective on pollution incidents. Additionally, we use alternative feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

2.6 Overview of the Application of Remote Sensing in Effective Monitoring of Water Quality Parameters (Godson Adjovu, 2023)

In the work by Godson Adjovu (Godson Adjovu, 2023), the objective is to provide an overview of remote sensing (RS) applications in accurately retrieving and monitoring various water quality parameters (WQPs) such as chlorophyll-a concentration, turbidity, total suspended solids, colored dissolved organic matter, and total dissolved solids. The study discusses the techniques utilized, as well as the limitations and advantages associated with RS for effectively assessing and monitoring water quality. Remote sensing (RS) allows for the effective retrieval of water quality parameters (WQPs) by categorizing them as either optically active or inactive based on their impact on the optical characteristics measured by RS sensors. RS applications provide decision-makers with the opportunity to accurately quantify and monitor WQPs on a spatiotemporal scale. Numerous studies have explored the use of RS for water quality monitoring, employing empirical, analytical, semi-empirical, and machine-learning algorithms. Optical RS, in particular, has been extensively used for estimating WQPs, while microwave sensors have also been employed in certain cases.

The article discusses how remote sensing (RS) can be used for monitoring water quality parameters (WQPs) through the measurement of optical active parameters. The RS techniques involve measuring changes in empirical or analytical models to WQPs by relating them to remotely sensed reflectance. The optimal wavelength used in WQP measurement depends on the constituent being measured and the sensor characteristics. The article outlines different approaches used to extract WQPs from remotely sensed spectral data, including the empirical, analytical, semi-empirical, and artificial intelligence (AI) methods. AI methods capture both linear and nonlinear relationships and have been applied in water quality retrieval, producing satisfactory results compared to conventional statistical approaches. The article concludes that RS can provide valuable insights into the changes in water quality and aid in the development of management strategies for maintaining healthy ecosystems.

The article explores various approaches to employing remote sensing (RS) for the monitoring of water quality parameters (WQPs) including chlorophyll-a concentration, turbidity, total suspended solids, colored dissolved organic matter, total dissolved solids, and more. The four approaches outlined include the empirical method, analytical method, semi-empirical methods, and artificial intelligence (AI) methods.

The empirical method uses statistical relationships between measured RS spectral values and measured water quality, established using regression techniques. The analytical method involves modeling reflectance using inherent optical properties (IOPs) and apparent optical properties (AOPs), with physical relationships derived between the WQP, underwater light field, and the remotely sensed radiance.

The semi-empirical method combines the empirical and analytical methods by using the spectral characteristics of the parameters, with the appropriate combination of wavebands used as correlates. The spectral radiance is then recalculated to above the surface irradiance reflectance and related to the WQP through regression techniques.

Lastly, the AI method is an implicit algorithm approach that captures both linear and nonlinear relationships using various AI applications such as neural networks (NN) (network, 2016) and support vector machines (SVM) (machine, 2016). AI methods are useful for dealing with complications from various water surfaces, WQP combinations, and sediment deposits. Studies have shown that AI methods, particularly the ANN, outperform regression models in water quality retrieval. (T.Vakili, 2020) (Y.Zhang, 2002)

In contrast to the present study, our research centers on a different geographical region and incorporates data from the Copernicus Marine Service alongside DEFRA data, providing a more comprehensive perspective on pollution incidents. Additionally, we use alternative feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

2.7 Satellite-assisted monitoring of water quality to support the implementation of the Water Framework Directive (Directive, 2020)

The EU Water Framework Directive (WFD) aims to achieve good ecological and chemical status for surface waters and groundwater by 2027.(Directive, 2020) However, there are challenges in monitoring and assessing water quality across Member States. This paper proposes the use of satellite observations to complement conventional monitoring methods and improve the effectiveness of the WFD. Satellite observations can provide better spatial and temporal coverage of waterbodies, quantify elements of environmental status, and enhance the classification of ecological status. The paper recommends the following:

- Recognize satellite observation as an assessment method under the revised WFD and encourage its use to complement national monitoring.
- Establish a satellite observation expert group to harmonize metrics, ensure comparability with existing methods, and advise Member States on best practices.
- Reference the use of satellite-derived metrics, particularly for assessing phytoplankton biomass and bloom frequency, in the Reporting Guidance of the revised WFD.
- Convene a conference to agree on common practices and reporting standards for using satellite-based water quality metrics to support the WFD.

The paper emphasizes the advanced satellite-based instruments available in the EU, such as those within the Copernicus framework, and the significant investments already made. It highlights ongoing research and innovation actions and EU investment in the Copernicus program. The recommendations are supported by experts, stakeholders, and relevant organizations. Implementation of the recommendations requires support from policy makers, national monitoring authorities, and advisory bodies. The formation of an advisory body, potentially within the ECOSTAT working group, is suggested to develop a strategy for the use of satellite-derived water quality products.

To facilitate this integration, agreement at the European level is needed regarding the role of satellite products in supporting the WFD. This would lead to the development of data standards and harmonization criteria. At the national level, investment in the capability to use satellite observation products is necessary. Consistency across member states would require agreement on methodologies, potentially supplied by specialized satellite Earth observation service providers. These methodologies should be transparent and widely accepted to ensure their contribution to reporting environmental status.

It's important to note that satellite observation should not replace existing monitoring practices but complement them. The inclusion of satellite products in statutory reporting may involve additional costs. However, these costs are likely to be small compared to the investment already made in the Copernicus program. Direct cost savings can be expected when satellite data aid in prioritization, more efficient catchment management, and strategic design of in situ sampling programs. Furthermore, satellite-based monitoring of algal blooms can provide early warnings and reduce risks and mitigation costs for various sectors.

In order to enhance the incorporation of satellite-based Earth observation into national and regulatory monitoring for the Water Framework Directive (WFD), several recommendations have been put forward. The European Commission, member states, and water management authorities are urged to acknowledge satellite-based Earth observation as a viable method in the revised Common Implementation Strategy (CIS) guidance for the WFD. This can be achieved by establishing a working group within the CIS ECOSTAT group to deliver guidance on how member states can use satellite observation for robust and cost-effective monitoring while harmonizing metrics across countries.

Furthermore, it is recommended to convene a conference involving the European Commission, member states, and pertinent water authorities to formally acknowledge and endorse the utilization of satellite-based Earth observation metrics for the implementation of the Water Framework Directive (WFD). The goal of the conference would be to design a joint plan of action for harmonized use of satellite data and building institutional capacity.

Overall, integrating satellite-based Earth observation into water quality monitoring can enhance the effectiveness, efficiency, and coverage of the monitoring process. Through the implementation of the suggested measures, the European Commission and member states can effectively leverage the potential of satellite data to enhance water management practices and successfully achieve the objectives outlined in the Water Framework Directive.

2.8 Summary

We summarise this previous literature in Table 2.1 and Table 2.2. Our study differs from these previous works in few important respects.

We leverage a unique fusion of in-situ pollution records and remote sensing data from CMEMS. Unlike other in-situ datasets in the literature, the former constitutes a longitudinal view (multiple years) on the problem across a relatively wide geography (covers all major recreational seaside locations in the country of England) and is conducted with relatively high frequency (once every several days).

Our research focuses on an another geographic area and integrates data from both the Copernicus Marine Service and DEFRA. Furthermore, we use another feature engineering techniques to incorporate and analyze the zone of chlorophyll concentration derived from the Copernicus data.

By focusing on water quality monitoring in the coastal areas of England, our work has practical applications and implications. The North Sea and Celtic Sea are important coastal regions with significant ecological and economic value. Effective monitoring and management of water quality in these areas are crucial for environmental preservation and sustainable resource utilization. Our work can contribute to the development of early warning systems, mitigation strategies, and decision-making tools that aid in the protection and preservation of these coastal environments.

In our study, we focus exclusively on the England region and use data from DEFRA to identify pollution incidents. Additionally, we incorporate chlorophyll concentration data obtained from the Copernicus Marine service, which provides a spatial resolution of 1 km $\times$ 1 km. We also explore various dimensions of the chlorophyll concentration coverage zone to enhance our analysis.

TABLE 2.1: Summary of the Papers

Study Title	**Model Methods**	**Data Features**
(A.H.Kadhim, 2012)	Minimum-Distance Mean Classifier, Parallelepiped Classifier, Gaussian Maximum Likelihood Classifier, K-Mean Classifier	Chlorophyll-a, turbidity
(Casanova, 2021)	Random forest, Support vector machine, Artificial Neural Network, Deep Neural Network	Chlorophyll-a
(Nichol, 2018)	Artificial Neural Network, Support Vector Regression, Random Forest	Chlorophyll-a, diffuse attenuation coefficient, water-leaving radiance spectra
(Chen, 2022)	Support Vector Machines, Artificial Neural Network, Band Ratio Model, First Order Differential Model, Three-band Model, Hybrid Spatio temporal Spectral Fusion Model	Total suspended matter, chlorophyll-a, colored dissolved organic matter, chemical oxygen demand, total nitrogen, total phosphorus
(Navarro, 2022)	OC3 algorithm	chlorophyll-a
(Godson Adjovu, 2023)	Support Vector Machines, Artificial Neural Network	chlorophyll-a, turbidity, total suspended solids, colored dissolved organic matter, total dissolved solids

TABLE 2.2: Summary of the Papers

Study Title	**Geography**	**Surface type**	**Data source**
(A.H.Kadhim, 2012)	Diyala River, Iraq	River	Landsat 8 imagery
(Casanova, 2021)	Menor sea, Spain	Sea	Sentinel-2 data
(Nichol, 2018)	Yellow Sea coast Qingdao, China	Coast	Sentinel-2 Copernicus Open Access Hub
(Chen, 2022)	Zhuhai estuary river China, shallow lakes, lakes in eastern Nebraska	River	USA's Hyperion data, China's HJ-1 satellite HIS data, MERIS imagery data, Sentinel-2 images
(Navarro, 2022)	Mar Menor coastal lagoon, Spain	Lagoon	Landsat-8, Sentinel-2 satellites, Sentinel-3 satellite
(Godson Adjovu, 2023)	Rivers and streams in the Great Plains of Central North America	Rivers	Landsat 9 OLI/TIRS, Landsat 8 OLI/TIRS, Landsat 7 ETM+, Landsat 5 TM, RapidEye images, ASTER, MODIS, Sentinel-2 MSI

Chapter 3

Datasets

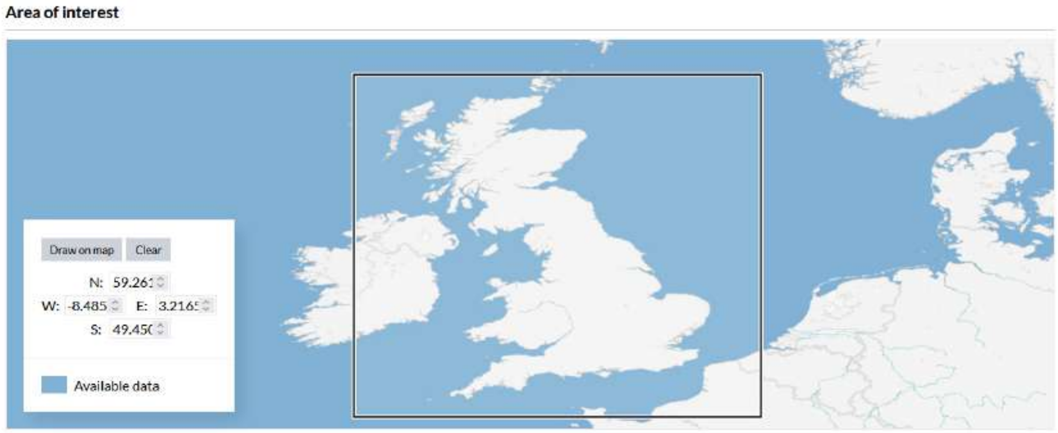
3.1 Geographical region

In this research project, we selected England as the geographical region of interest for monitoring seawater quality. Our choice was dictated by the prominence of public discourse around seawater pollution in the England on the one hand (Laville, 2023), and by the availability of detailed in-situ and remote sensing datasets for english coastal waters on the other hand.

To collect the datasets from the coastal waters of England, we set the bounding box (BBox) with the following coordinates:

Boundary	Value
Northern	59.26106184235473
Western	-8.485977552106107
Eastern	3.216552695395291
Southern	49.45085563487484

It should be noted that while this BBox captures the majority of the waters around England, it may also include some data from the landmass of Ireland.



Area of interest

N: 59.261
W: -8.485 E: 3.216
S: 49.450

FIGURE 3.1: Area of interest

3.2 Remote sensing datasets

To collect remote sensing water quality datasets, we used high-resolution biochemical ocean measurements from the Copernicus Marine Service (CMEMS) (Service, 2014). The Copernicus Marine Service (CMEMS) is a brunch of the European Union Space Programme which provides free global and regional-scale satellite ocean data to enhance and broaden the understanding of the global oceans and to promote the development, protection, and restoration of marine environments for all maritime industries.

This research paper employs three remote sensing datasets from CMEMS.

3.2.1 Global Ocean Colour

The first dataset used is the `cmems_obs-oc_glo_bgc-plankton_my_l4-gapfree-multi-4km` dataset, which is the Global Ocean Colour (Copernicus-GlobColour) Bio-Geo-Chemical, L4 (monthly and interpolated) from Satellite Observations (1997-ongoing) with a spatial resolution of 4 × 4 km.

Table 3.1 displays the variables of the dataset.

Variable	Description
Time	Time of observation
Latitude	Latitude of observation location
Longitude	Longitude of observation location
CHL	Mass concentration of chlorophyll a in sea water
ZSD	Secchi depth of sea water
CDM	Volume absorption coefficient of radiative flux in sea water due
KD	Volume attenuation coefficient of downwelling radiative flux in sea water
BBP	Volume backward scattering coefficient of radiative flux

TABLE 3.1: Variables of the Bio-Geo-Chemical Dataset

3.2.2 Bio-Geo-Chemical indicators

The second dataset used is the `cmems_obs_oc_nws_bgc_geophy_nrt_l4-hr_P1D-m` dataset, which is the North West Shelf Region, Bio-Geo-Chemical, L4, monthly means and interpolated daily observation with a spatial resolution of 0.1 × 0.1 km. The variables of this dataset are depicted in Table 3.2.

Variable	Description
Time	Time of observation
Latitude	Latitude of observation location
Longitude	Longitude of observation location
CHL	Mass concentration of chlorophyll a in sea water
SPM	mass concentration of suspended matter in sea water
TUR	sea water turbidity
BBP	Volume backward scattering coefficient of radiative flux

TABLE 3.2: Variables of the Bio-Geo-Chemical, L4, 0.1 × 0.1 km Dataset

3.2.3 Atlantic Ocean Colour

Finally, the `cmems_obs-oc_atl_bgc-plankton_my_l4-gapfree-multi-1km_P1D` dataset was used, which is the Atlantic Ocean Colour (Copernicus-GlobColour) Bio-Geo-Chemical, L4 (daily interpolated) from Satellite Observations (1997-ongoing) with a spatial resolution of 1×1 km. The variable included in this dataset is mass concentration of chlorophyll a in sea water (CHL).

Variable	Description
Time	Time of observation
Latitude	Latitude of observation location
Longitude	Longitude of observation location
CHL	Mass concentration of chlorophyll a in sea water

TABLE 3.3: Variables of the Bio-Geo-Chemical, L4, 1×1 km Dataset

3.2.4 Temporal and spatial indices

The variables included in each dataset are measured at fixed locations specified by latitude and longitude pairs. The CHL time series is defined for each location and is represented as $\{CHL(s)\}_{t=1}^{T}$, where the location s is defined by a pair of (lat, lon) coordinates, t represents the variable for day, while T represents the observed period of one year.

In this study, we analyzed a one-year time window of CHL measurements, specifically from January 1st, 2021, to December 31st, 2021. We inputted the data length and the coordinates into the Copernicus website for the dataset with `product_id` `OCEANCOLOUR_GLO_BGC_L4_MY_009_104`, which resulted in a total dimension of $365\times253\times327$, referring to day, latitude, and longitude, respectively. The ZSD, CDM, KD, BBP time series is defined for each location and is represented as $\{ZSD(s)\}_{t=1}^{T}$, $\{CDM(s)\}_{t=1}^{T}$, $\{KD(s)\}_{t=1}^{T}$, $\{BBP(s)\}_{t=1}^{T}$, where the location s is defined by a pair of (lat, lon) coordinates, t represents the variable for day, while T represents the observed period of one year.
For the dataset with `product_id` `OCEANCOLOUR_NWS_BGC_HR_L4_NRT_009_209`, which resulted in a total dimension of $365\times8640\times3435$, referring to day, latitude, and longitude, respectively. The CHL, SPM, TUR, BBP time series is defined for each location and is represented as $\{CHL(s)\}_{t=1}^{T}$, $\{SPM(s)\}_{t=1}^{T}$, $\{TUR(s)\}_{t=1}^{T}$, $\{BBP(s)\}_{t=1}^{T}$, where the location s is defined by a pair of (lat, lon) coordinates, t represents the variable for day, while T represents the observed period of one year.

3.2.5 Datasets collection

CMEMS datasets are available through the graphical UI as well as through an API. Despite the advantages of using the API, our experience has revealed some challenges. One of the issues encountered was that the data size for some of the datasets exceeded the maximum import size allowed by the API. Additionally, the API has a timeout for authentication and queries, which means that if the retrieval of data from the Copernicus Marine database takes too long, the dataset may become unusable due to the API timing out.

However, for the subsequent datasets, we encountered problems. Due to the large dataset size of 18.2 Gb for the `product_id` `OCEANCOLOUR_NWS_BGC_HR_L4_NRT_009_209`, the Copernicus UI did not provide the option to download the data, so we used FTP

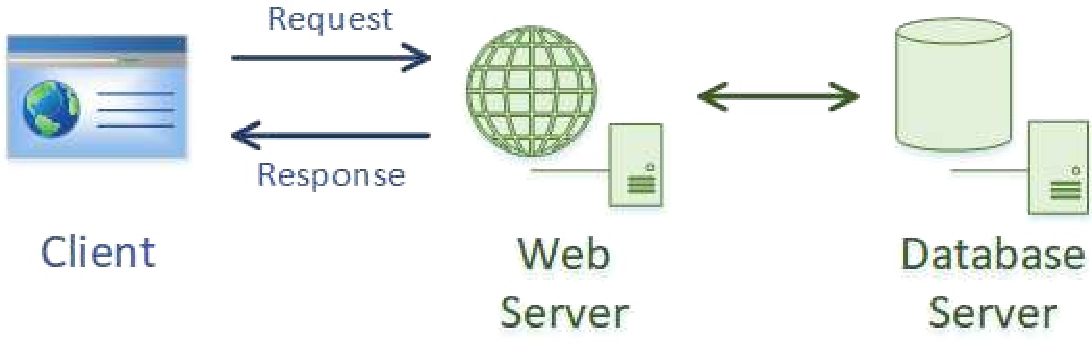


FIGURE 3.2: Copernicus Marine Service API Workflow

to download the entire dataset. Unfortunately, there was no option to select the region of interest using FTP, so we had to download the full dataset and later slice the area of interest for England using the BBox coordinates. Inputting the data length and the coordinates for this dataset resulted in a total dimension of 365×8640×3435, referring to day, latitude, and longitude, respectively.

Similarly, for the dataset with `product_id` `OCEANCOLOUR_ATL_BGC_L4_MY_009_118`, the dataset size was 5.97 Gb. Unfortunately, the Copernicus User Interface (UI) did not offer a direct download option for this particular dataset. Consequently, we resorted to utilizing the FTP (File Transfer Protocol) method to acquire the dataset. Copernicus Marine Service provided us with the necessary credentials to access and download the dataset. We inputted the data length and coordinates, which resulted in a total dimension of 365×4416×5664, referring to day, latitude, and longitude, respectively. The final dimensions of all datasets are presented in Table 3.4.

Product ID	Dimension
OCEANCOLOUR_ATL_BGC_L4_MY_009_118	365X4416X566
OCEANCOLOUR_NWS_BGC_HR_L4_NRT_009_209	365X8640X3435
OCEANCOLOUR_GLO_BGC_L4_MY_009_104	365X253X327

TABLE 3.4: Dataset‘s dimensions

3.3 Pollution label dataset

In addition to remote sensing datasets, we also used the in-situ water quality measurements provided by the England’s Department for Environment Food Rural Affairs (DEFRA) 3.3.

DEFRA monitors water quality indicators across seaside location in England which includes the information on: Bathing water site details, History of in-season sample results, History of annual bathing water classifications, History of abnormal situations, and History of pollution risk forecasts. After careful examination, we determined that we only required the Bathing water site details and History of pollution risk forecasts subsets, which we merged using the unique identifier EUBWID. The full summary of all variables in the dataset are summarised in Table 3.5.

DEFRA provides an API to accesss the datasets.

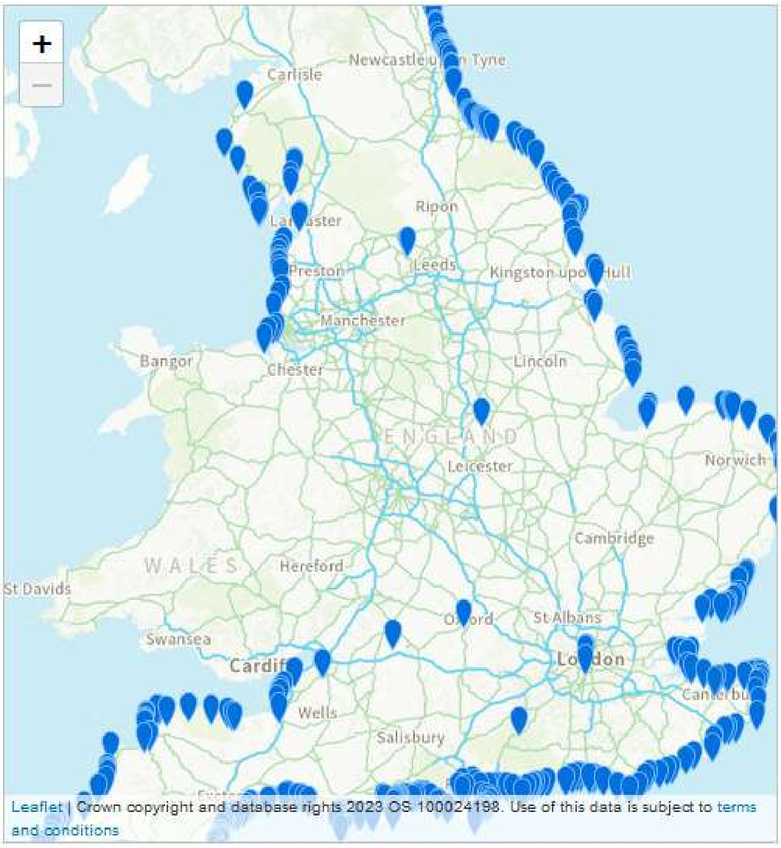


FIGURE 3.3: Bathing Water Quality DEFRA API

3.3.1 Variables

We used variables EUBWID, label, predictedAt, riskLevelLabel, warning, samplePointID, pollutionRiskForecasting indexed by time and location from DEFRA dataset. The samples in the dataset were collected once a day. The variables were indexed based on two dimensions: time and location. The time dimension corresponds to the day of the sample, while the location dimension was indexed by latitude and longitude coordinates.

In addition, we created a new `is_pollution` variable, which takes binary values {`true`, `false`}. This feature is based on the `riskLevelLabel` feature, which can have the values `no-prediction`, `normal`, or `increased`. If the `riskLevelLabel` value is `increased`, it means that there is pollution, otherwise, there is no pollution.

3.3.2 Dataset size

The DEFRA dataset contains 333394 rows and 28 columns spanning from 2013 to 2022. However, there are a few rows in the dataset where several water quality measurements were made during the day, resulting in duplicate values in columns such as "EUBWID," "predictedAt_year," "predictedAt_month," and "predictedAt_day." The total number of rows with several water quality measurements is 3081×28. This indicates that there are several measurements of chlorophyll for a single day. To address this, we applied a rule that if there was contamination within a single day, then there would be only one record with the `is_pollution` label set to `true`. Otherwise,

Subset	Variables
History of pollution risk forecasts	- EUBWID - label - predictedAt - warning - riskLevelLabel
Bathing water site details	- EUBWID - samplePointID - lat - long - pollutionRiskForecasting

TABLE 3.5: Variables in different subsets of the dataset

it would be set to `false` indicating no contamination. We have excluded the "no-prediction" label from our investigation due to its potential to introduce bias into our model. Therefore, our study focuses solely on the prediction of relevant labels within the defined scope of our research.

After applying the aforementioned data preprocessing steps and selecting the measurements from 2021 year alone, the resulting dataset shape is 331735×28.

3.4 Data preprocessing

3.4.1 Data preprocessing pipeline

All the steps involved in the data preprocessing are summarized into a pipeline for preprocessing. The data preprocessing pipeline is displayed in Figure 3.4.

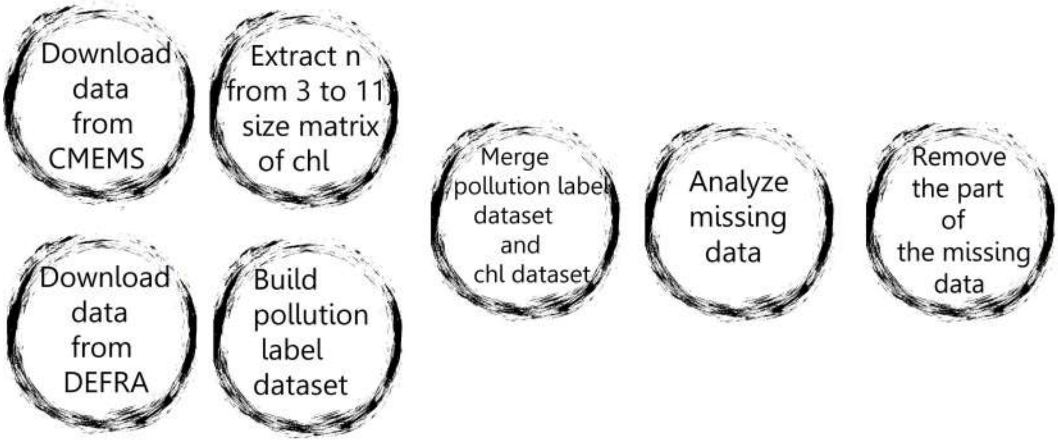

Download data from CMEMS
Extract n from 3 to 11 size matrix of chl
Download data from DEFRA
Build pollution label dataset
Merge pollution label dataset and chl dataset
Analyze missing data
Remove the part of the missing data

FIGURE 3.4: Data preprocessing pipeline

3.4.2 Data merge

In order to combine the pollution label dataset from DEFRA and the chlorophyll concentration dataset from Copernicus marine service, we first extracted the longitude and latitude coordinates from the pollution label dataset. Then, we searched for these coordinates in the chlorophyll concentration dataset. If we did not find an exact match, we selected the nearest coordinates as a replacement. This mapping allowed us to link the coordinates between the two datasets.

We also mapped the time feature using the predicted year, predicted month, and predicted day from the pollution label dataset. If we could not find an exact match for the time in the chlorophyll concentration dataset, we selected the smallest nearest time.

The execution time for obtaining data for different spatial resolutions is displayed in Figure 3.7.

Spatial Resolution	Execution time
1km	22.85 mins - 0.38 hours
4km	707.525 mins - 11.79 hours
0.1km	25.47 mins - 0.42 hours

TABLE 3.6: Summary of processing times for different spatial resolutions.

The missing data percentage for different spatial resolutions is illustrated in Figure 3.7.

Spatial Resolution	Missing data(%))
1km	64.26
4km	78.12
0.1km	87.28

TABLE 3.7: Missing data percentage for different spatial resolutions.

The term "missing data" refers to corrupted data in the context of our study, indicating that certain information was not provided by the Copernicus Marine Service satellite due to weather-related issues or technical problems with the satellite itself. This results in gaps or absence of data in our dataset, which can impact the completeness and reliability of the information available for analysis.

The final dataset consists of 64566 rows and includes the following features: chl_level_0 to chl_level_120, latitude, longitude, and a binary indicator for pollution (is_pollution). These features provide information about chlorophyll levels at various depths and the corresponding geographical coordinates, along with the pollution status.

3.4.3 Processing missing records

To calculate the percentage of missing data in the dataset, we first flatten the dataset, which means we calculate the total number of elements in each pixel. Each pixel represents a 3×3 matrix. Then, we count the number of missing elements in each pixel. The percentage of missing data is then calculated by dividing the total number of missing elements by the total number of elements in the dataset and multiplying by `100%`.

Having observed a high percentage of missing data (`78.12%`) for the spatial resolution of 4km×4km, we have decided to exclude this dataset and focus on datasets with a higher spatial resolution. We have chosen the dataset with a higher spatial resolution primarily because it exhibits a lower percentage of missing data. The missing data refers to information that is not provided from the Copernicus Marine Service satellite due to technical issues or other reasons. By utilizing the dataset with fewer missing data points, we can ensure a more comprehensive and reliable analysis with a richer set of high-resolution information.

For each unique pair of longitude and latitude, we collected all chlorophyll concentration data for all days and computed the total number of chlorophyll elements and total missing data. We then calculated the percentage of missing data for each unique pair of longitude and latitude.

Next, we distributed the pairs into six predefined bins based on the percentage of missing data. These bins include the ranges of (90.0, 100.0), (80.0, 90.0), (70.0, 80.0), (60.0, 70.0), (50.0, 60.0), and (0.0, 50.0). This information is illustrated in Table 3.8.

The size of the values for a specific pair of longitude and latitude indices is equal to 18634.

Percentage Range	**Missing data(%))**
90.0% - 100.0%	50.12%
80.0% - 90.0%	0.48%
70.0% - 80.0%	6.92%
60.0% - 70.0%	20.29%
50.0% - 60.0%	9.79%
0.0% - 50.0%	12.41%

TABLE 3.8: Percentage of unique coordinates pairs with NaN values in different ranges for 0.1km×0.1km spatial resolution

Due to the presence of corrupted data on `45 days (days 1-12 and days 333-365)` in the year 2021, we decided to remove these days from the dataset. The size of the values for a specific pair of longitude and latitude indices is equal to 397. The percentage of unique coordinate pairs with NaN values in different ranges for a spatial resolution of 1km×1km is presented in Table 3.9.

TABLE 3.9: Percentage of unique coordinates pairs with NaN values in different ranges for 1km×1km spatial resolution

Range	**Percentage**
90.0% - 100.0%	22.41%
80.0% - 90.0%	1.51%
70.0% - 80.0%	8.81%
60.0% - 70.0%	24.43%
50.0% - 60.0%	10.57%
0.0% - 50.0%	32.24%

Additionally, we can examine the NaN percentages across unique coordinate pairs, as depicted in Figure 3.5.

In our analysis, we extracted and removed the data points where the percentage of missing values was in the range of `90% - 100.0%` and `80.0% - 90.0%`. We did this because these data points had an excessively high percentage of missing values, which could negatively impact the accuracy of our analysis. By removing these data points, we could ensure that our analysis was focused on more complete and reliable data.

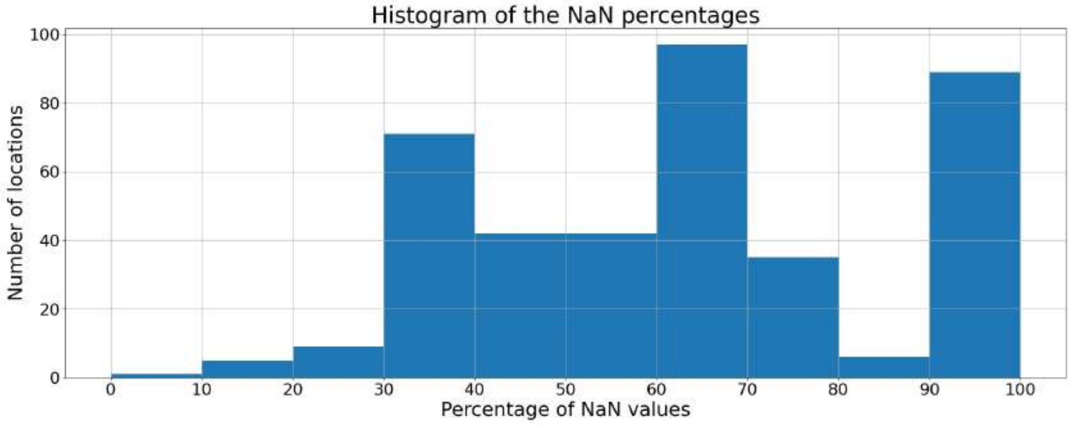


FIGURE 3.5: Histogram of NaN Percentages Across Unique Coordinates Pairs

We have constructed an Empirical Cumulative Distribution Function (ECDF) to visualize the distribution of NaN percentages for unique coordinate pairs. The results are displayed in Figure 3.6.

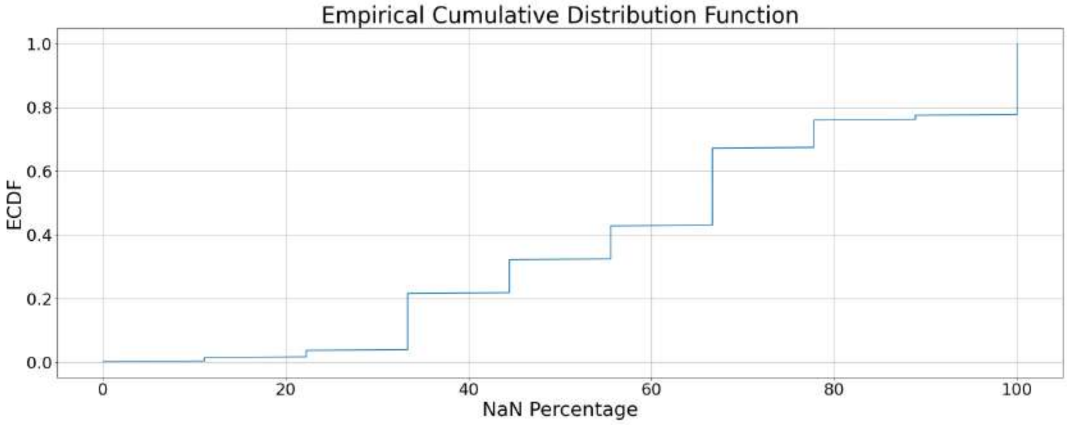


FIGURE 3.6: Empirical Cumulative Distribution Function of NaN Percentage
For Unique Coordinates Pairs

We have created a histogram to analyze the distribution of unique pairs based on the percentage range of missing data. The histogram is depicted in Figure 3.7.

3.4.4 Analysis and Exclusion of Datasets with Different Spatial Resolutions

The concentration level `chl_level` is a measure of chlorophyll-a concentration in the water, which is an indicator of phytoplankton biomass.

After careful consideration, we have decided to exclude the dataset with a spatial resolution of 4km×4km due to a high percentage of missing data. As a result, we have chosen to focus our analysis on datasets with a higher spatial resolution, which are expected to provide more accurate and detailed information for our research.

We have decided to convert the chlorophyll concentration data into the RGB(RGB color model) color format and display the dataset for each day of the year 2021. The process of converting the chlorophyll data to RGB involves normalizing the data

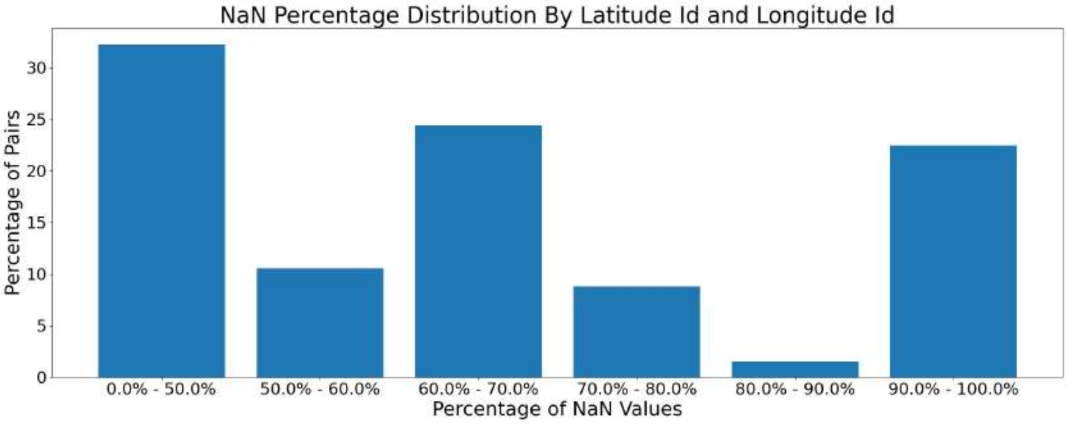


FIGURE 3.7: Histogram of Unique Pair Percentage by NaN Values.

between 0 and 1, scaling it to the 8-bit range of the RGB color space (0-255), and replicating the same grayscale values for each of the three color channels (red, green, and blue). This results in an RGB image where violet color represents land, yellow represents water surface, and different shades of green represent the chlorophyll concentration. Upon thorough examination of the chlorophyll dataset with a spatial resolution of 0.1km×0.1km, we have determined that the dataset is corrupted and inconsistent. This is illustrated in Figure 3.8.

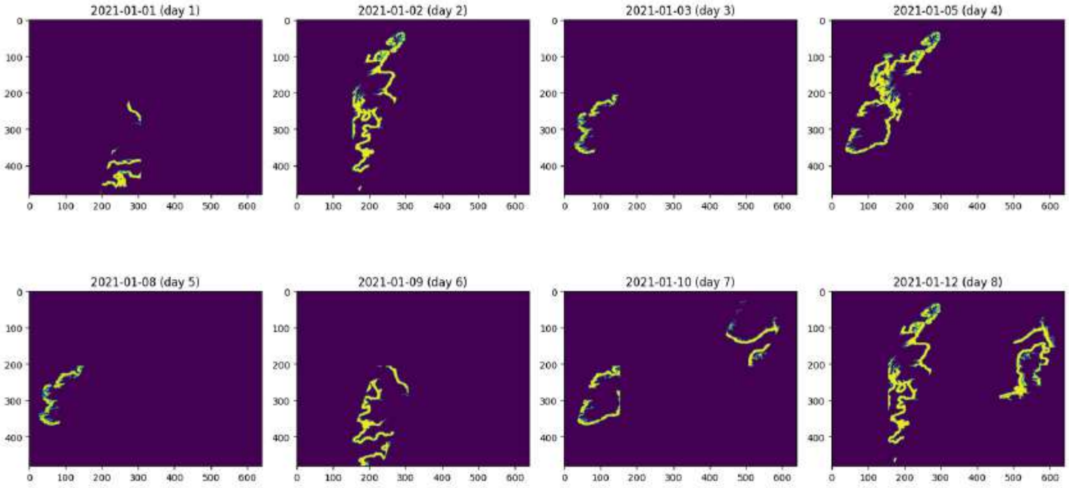


FIGURE 3.8: Chlorophyll Concentration by Day in 2021 for spatial resolution 0.1×0.1km

In the visualization, the color violet represents land, while yellow represents the water surface. The varying shades of green correspond to the different levels of chlorophyll concentration. After careful consideration, we have decided to use the dataset with a spatial resolution of 1km×1km, because, as depicted in Figure 3.9, the data exhibits consistency. We have chosen to visualize the chlorophyll concentration data on a daily basis for the year 2021, using a spatial resolution of 1km×1km.

However, within this dataset, there are instances of distorted data, as observed in Figure 3.10. These distorted data points are specifically associated with the winter months.

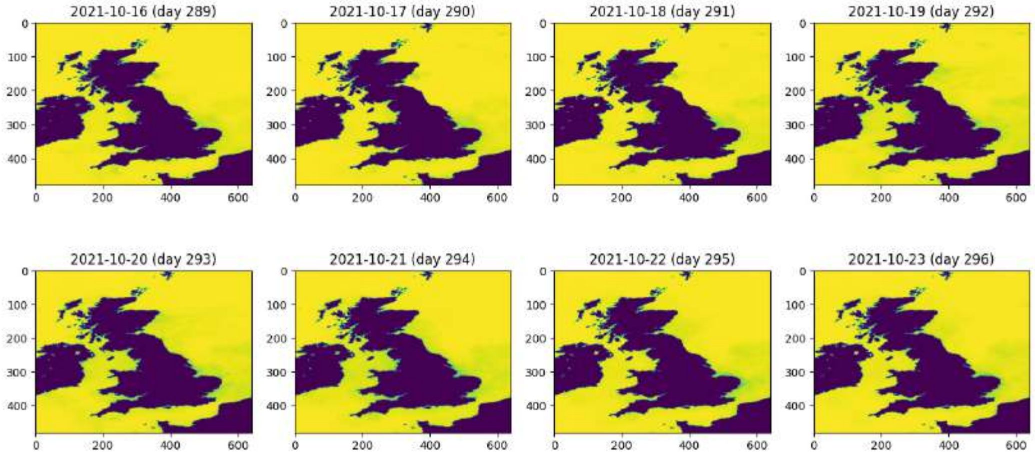


FIGURE 3.9: Chlorophyll Concentration by Day in 2021 for spatial resolution 1×1km

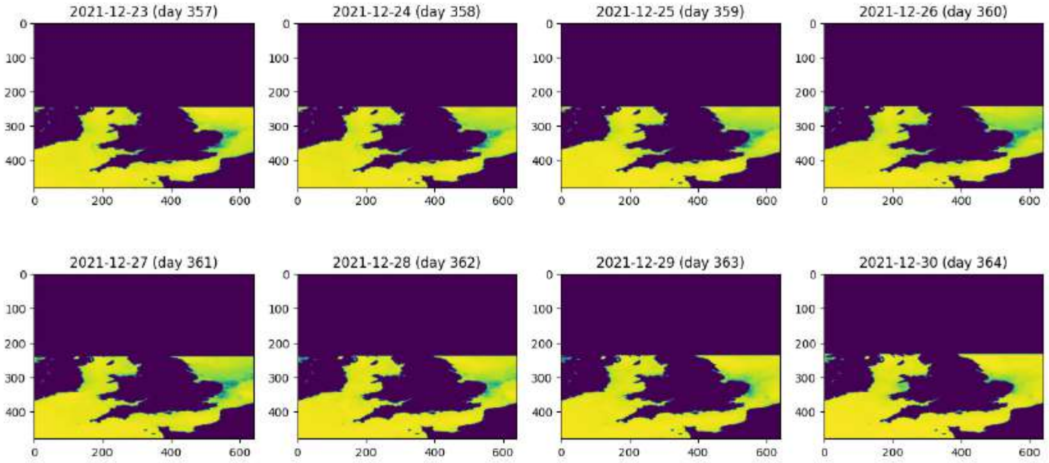


FIGURE 3.10: Distorted Chlorophyll Concentration for spatial resolution 1×1km

In the feature engineering process for the first dataset, we created eight statistical features for each point in a 3×3 matrix. These features include mean, median, standard deviation, minimum, maximum, percentile 25, percentile 50, and percentile 75. In addition to these features, we also included the year, month, and a label indicating whether the pollution level is high or not.

The Sea Chlorophyll Concentration in the England over Time for one particular point in Figure 3.11 As observed on the chart 3.11, there were noticeable bursts in chlorophyll concentration during the months of April and May.

The point is illustrated in Figure 3.12 where we can observe the sea chlorophyll concentration in the England over time for a specific location in Figure 3.11. The corresponding point is marked on the chlorophyll map (see Figure 3.12).

All unique points from the DEFRA pollution dataset are plotted on the chlorophyll map. The corresponding point is marked on the chlorophyll map in Figure 3.13

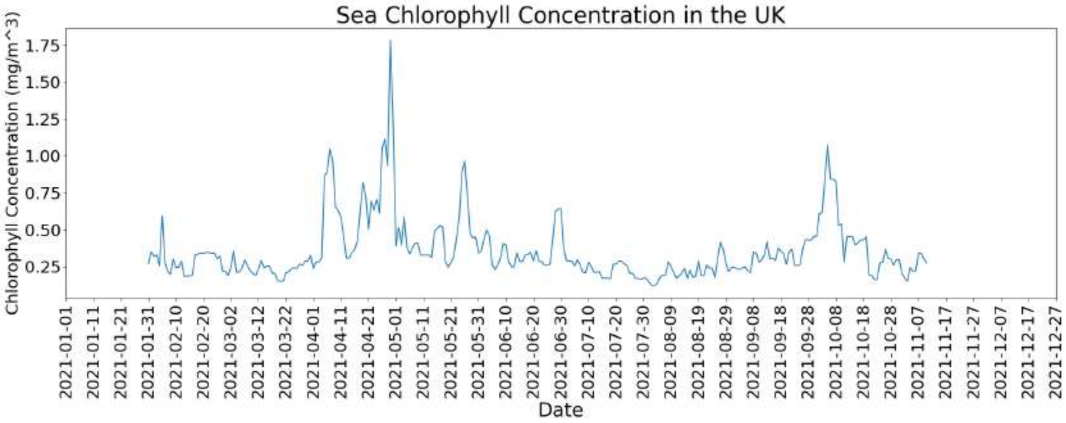


FIGURE 3.11: Sea Chlorophyll Concentration in the England over Time

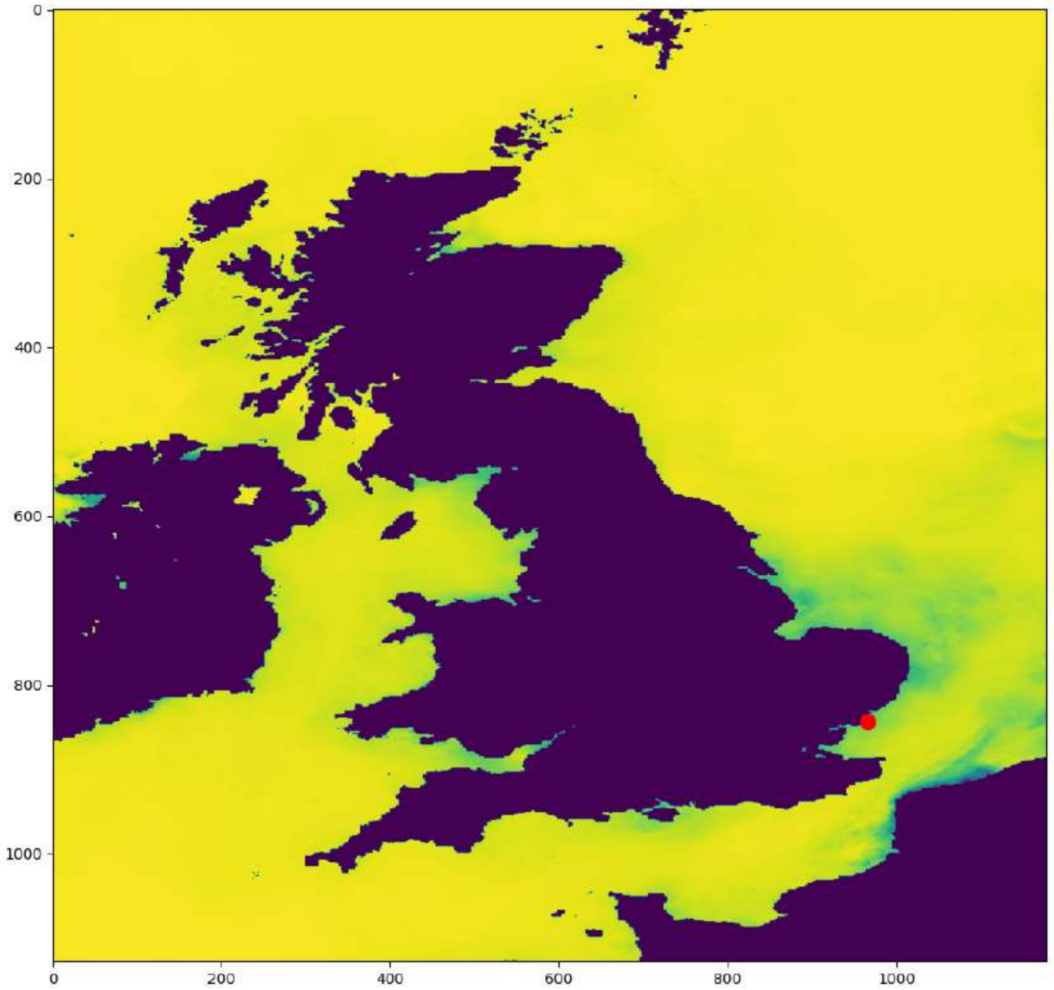


FIGURE 3.12: The red marked point on the Chlorophyll Concentration map

It can be observed that there are 9 points located within landmasses, which have been removed from the dataset.

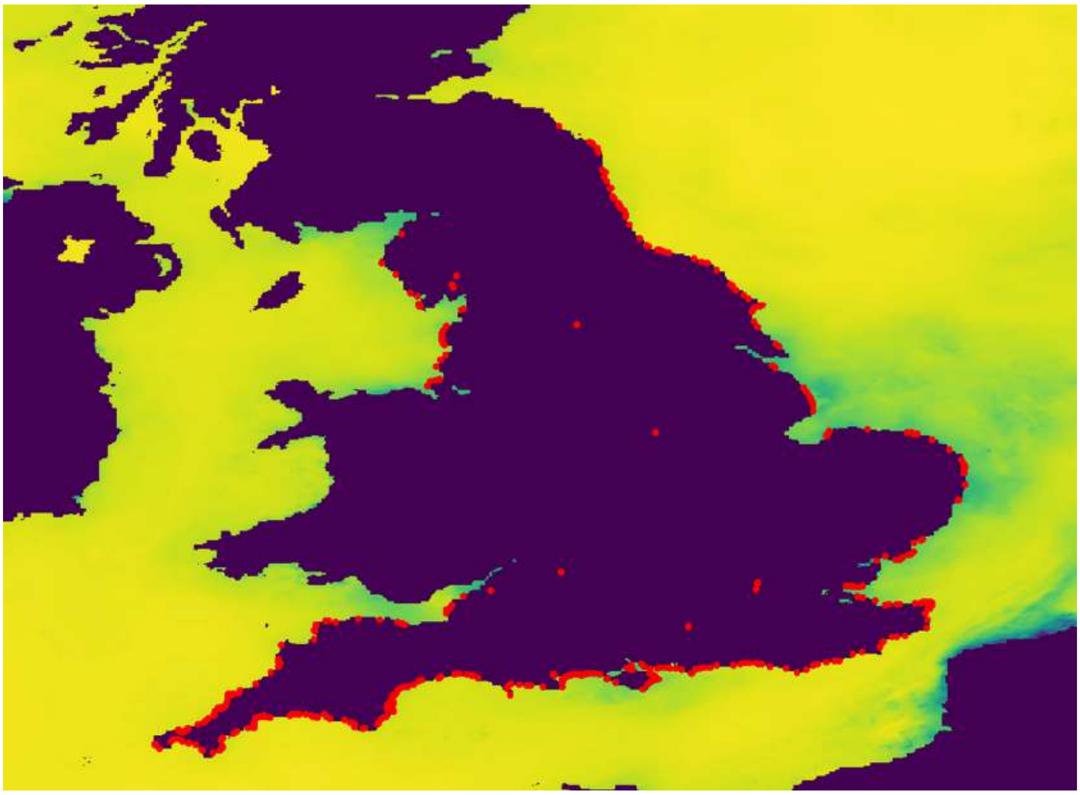


FIGURE 3.13: The red pollution detecting points on the Chlorophyll Concentration map

We have added 15 points to the map of England, each representing a specific location, and the corresponding matrix of chlorophyll concentration for those points. And this is displayed in Figure 3.14.

3.4.5 Result of data preprocessing

To calculate the statistical features, we checked if the middle element of the matrix is a non-NaN value. If it is, we used it as a feature. If it is a NaN value, we calculated the statistical features based on the non-NaN values within the 3×3 matrix. The middle element of the matrix, which corresponds to the point measured by the DEFRA service, is [2] [2].

For the second dataset, we used each element of the 3×3 matrix as a separate feature. This means that for each point with coordinates and time, we have nine features in total. For the remaining missing data, we filled in the values with the mean chlorophyll concentration.

As a result of our feature engineering, we have created two datasets that will be used for further processing and evaluation.

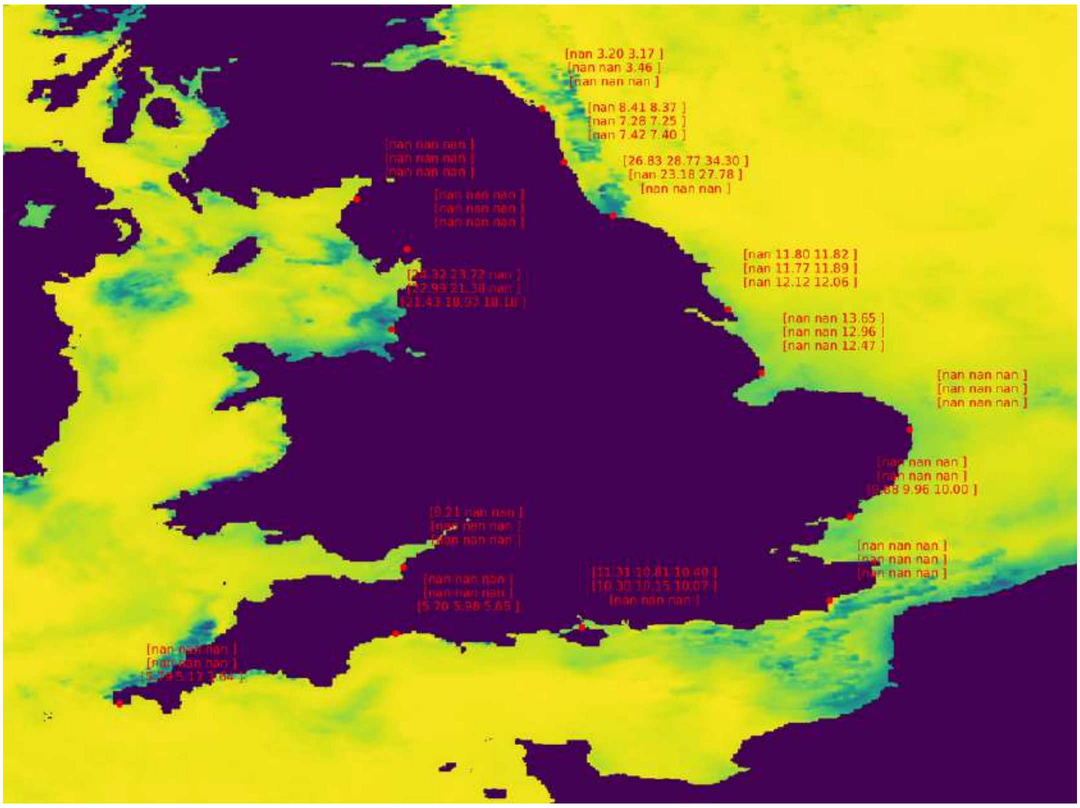


FIGURE 3.14: The 15 points on the Chlorophyll Concentration map with corresponding matrix

Chapter 4

Methodology

4.1 Problem formulation

The focus of this current work is on the problem of "nowcasting" pollution incidents: We aim to provide real-time or near-real-time predictions for the current pollution levels using freely available remote sensing data from CMEMS.

To this end, we formulate the problem of pollution prediction as a binary classification problem: $\hat{Y}_t^s = f(x_t^s)$, where Y_t^s is a binary variable indicating a record of a pollution incident in location s on day t as defined by DEFRA. x_t^s is a feature vector describing the state of the marine environment in location s recorded by CMEMS on day t. Our goal is to train a function f to map a feature vector X_t^s to a prediction $\hat{Y}_t^s$ of a pollution incident Y_t^s.

In our study, we formulated several research questions to guide our investigation and analysis. These questions serve as the foundation for our study's objectives and aim to address specific aspects related to our research topic:

1. What is the optimal shape of data (3×3, 11×11, etc.) and statistical features to detect water pollution based on the concentration of chlorophyll?
2. Which machine learning model performs best at detecting water pollution based on the concentration of chlorophyll?

4.2 Feature engineering

To define vectors X_t^s we use the following approach. Since, MSES data is provided in a grid with a fixed step (e.g. 0.1km x 0.1km), we first identify the cell c^s in the grid which contains location s. We then consider a window $s_{n\times n}$ of $n \times n$ cells centered around c^s.

For each MSES variable $x_t^s \in X_t^s$ (e.g. CHL, SPM, TUR, etc.) recorded on day t, we construct a vector $x_t^{s_{n\times n}}$ resulting from querying and flattening all measurements in the window $s_{n\times n}$ on day t.

In addition, we compute a number of statistics for each vector $x_t^{s_{n\times n}}$, namely: Mean, Median, Standard Deviation, Minimum Value, Maximum Value, 25th Percentile, 50th Percentile and 75th Percentile. We will discuss the efficacy of this feature engineering step in comparison to using raw features in Chapter 5.

4.3 Normalization

The purpose of applying StandardScaler (sklearn, 2023a) to the chlorophyll concentration is to standardize the feature and bring it to a common scale. This can be beneficial for machine learning algorithms that assume the input data follows a normal distribution or have sensitivity to feature scales.

The process of normalization involves transforming the data such that it follows a standard normal distribution with a mean of 0 and a standard deviation of 1. This is achieved by subtracting the mean value of each feature and dividing it by its standard deviation. The formula for standardization is:

$$z = \frac{\text{chlorophyll concentration} - \mu}{\sigma}$$

where z is the standardized value, chlorophyll concentration represents the original chlorophyll concentration, μ is the mean of the chlorophyll concentration, and σ is the standard deviation of the chlorophyll concentration.

4.4 Train test split

We used the train test split (learn.org, 2014) function from the sklearn.model selection module to implement the train-test split technique. The testing set consisted of 20% of the total data, while the remaining 80% was allocated to the training set.

The main objective of the train-test split technique is to evaluate the model's performance and its ability to generalize to unseen data. By segregating the data into distinct sets, we could train the model on the training set to learn patterns and relationships within the data. The independent testing set was then used to assess how accurately the model could predict unseen instances.

4.5 Dataset balancing

The dataset used for the analysis is imbalanced, meaning that there is a significant disparity in the number of instances between the two classes. Specifically, there are 62,560 instances classified as not pollution incidents, while only 2,006 instances are classified as pollution incidents.

This class imbalance can pose challenges when training a machine learning model. Since the majority class (not pollution incidents) dominates the dataset, the model may become biased and tend to favor predicting the majority class more accurately. As a result, it may struggle to effectively identify and predict instances of pollution incidents, which are relatively rare compared to the not pollution incidents.

In the case of the Random Forest Classifier model (forest, 2016), We employed a balancing approach by setting the parameter 'class weight' to 'balanced'. This technique automatically adjusts the weights of the classes based on their frequencies in the training data. It helps address the issue of imbalanced classes by giving more weight to the minority class during the training process. This way, the model can learn to better handle the imbalanced nature of the data and make more accurate predictions for both classes.

For the other models, We used the RandomOverSampler (sklearn, 2023b) technique. RandomOverSampler is a method for addressing class imbalance by oversampling the minority class. It randomly duplicates instances from the minority class until it reaches a similar number of instances as the majority class. The advantage of this approach is that it allows the model to be exposed to more instances of the minority class, thereby reducing the bias towards the majority class and potentially improving the model's ability to generalize and make accurate predictions for both classes.

However, it's important to note that RandomOverSampler has its limitations. Applying oversampling techniques to the minority class can result in overfitting, causing the model to become overly specialized to the training data and perform inadequately on unseen data. Moreover, if the dataset is already substantial, oversampling may substantially prolong the training time and demand additional memory resources.

It's worth mentioning that We applied the balancing technique only to the training dataset and not the entire dataset. This ensures that the model's evaluation and performance metrics are based on unbiased data, as the test dataset remains in its original distribution.

4.6 Models

4.6.1 Baseline

To establish the baseline for our model, we determine the probability of non-pollution by dividing the number of samples labeled as non-pollution by the total number of samples present in the dataset.

4.6.2 Logistic Regression

Logistic Regression models the conditional probability of the pollution indicator $(y = 1)$ given the chlorophyll concentration features $(x_1, x_2, \ldots, x_n)$. The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ represent the coefficients of the logistic regression model, and the sigmoid function is used to transform the linear combination of the features and coefficients into a probability value between 0 and 1, i.e.: (regression, 2016)

$$P(y = 1 \mid x_1, x_2, \ldots, x_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n)}}$$

One of the main reasons for selecting Logistic Regression is its simplicity and interpretability. It is a linear model that assumes a linear relationship between the predictors (chlorophyll concentration in this case) and the log-odds of the binary outcome (pollution indicator). The model estimates the coefficients that represent the influence of the predictors on the probability of the binary outcome.

4.6.3 Decision Trees classifier

The Decision Trees algorithm is presented in Figure 4.1.

```
if Chlorophyll Concentration > threshold_1:
    if Chlorophyll Concentration > threshold_2:
        Prediction: Pollution Indicator = 1
    else:
        Prediction: Pollution Indicator = 0
else:
    Prediction: Pollution Indicator = 0
```

FIGURE 4.1: Decision Trees Algorithm

One of the primary reasons for selecting a Decision Tree Classifier is its simplicity and interpretability. Similar to decision trees, Decision Tree Classifiers also represent a hierarchical structure where each internal node corresponds to a decision based on a particular feature and threshold. Each leaf node in the tree represents the final prediction or classification.

The hierarchical structure of Decision Tree Classifiers allows for straightforward interpretation and comprehension of the decision-making process. By traversing the tree from the root to a specific leaf node, one can easily understand the sequence of decisions made by the classifier to arrive at a particular classification outcome.

Moreover, Decision Tree Classifiers excel at capturing non-linear relationships between the features, such as chlorophyll concentration, and the target variable, such as a pollution indicator. They can identify complex interactions and patterns within the data that may not be apparent through simple linear relationships. This capability makes Decision Tree Classifiers particularly effective in scenarios where the relationships between the input features and the target variable are non-linear in nature.

In summary, Decision Tree Classifiers offer simplicity, interpretability, and the ability to capture non-linear relationships, making them a valuable choice for tasks that require transparent decision-making processes and the analysis of complex data.

4.6.4 Random Forest Classifier

The Random Forest Classifier (forest, 2016) is selected for several reasons. Firstly, it is an ensemble learning method that combines multiple decision trees to make predictions. This ensemble approach brings numerous benefits, such as improved accuracy and increased reliability compared to using a single decision tree. By aggregating the predictions from multiple trees, the Random Forest Classifier can mitigate the risk of overfitting and reduce the impact of individual trees' errors or biases.

Furthermore, Random Forests have the ability to capture more complex relationships within the data. Due to the randomness introduced during the training process, each decision tree in the forest is exposed to a different subset of the data. This variability helps the ensemble to capture diverse patterns and consider different perspectives, which can enhance the model's overall predictive performance. By combining the predictions from multiple trees, the Random Forest Classifier can effectively capture non-linear relationships, interactions, and higher-order dependencies that may exist in the data.

Another advantage of the Random Forest Classifier is its relative ease of implementation. While it is a sophisticated ensemble model, it does not require extensive hyperparameter tuning or fine-tuning compared to some other complex models. The Random Forest algorithm has default parameter settings that often yield good results across various datasets. This characteristic makes it a practical choice, especially when time and computational resources are limited.

In summary, the Random Forest Classifier is selected for its ensemble approach, which improves accuracy and reliability. It excels at capturing complex relationships in the data and offers simplicity in implementation with minimal hyperparameter tuning requirements. These advantages make the Random Forest Classifier a popular and effective choice for various classification tasks.

4.6.5 Gradient Boosting Classifier

The Gradient Boosting Classifier is an ensemble learning method that combines multiple weak learners, typically decision trees, to make predictions. It sequentially trains the weak learners on the residuals of the previous models, focusing on improving the model's performance with each iteration.

The training algorithm for Gradient Boosting Classifier proceeds as follows:

Initialization: Initially, all training instances are given equal weights. The first weak learner is trained on the entire dataset.

Training Iterations: For each iteration:

Prediction: The current ensemble of weak learners makes predictions on the training data.

Residual Calculation: The difference between the actual target values and the predictions is calculated. These differences, called residuals, represent the errors of the current ensemble.

Training of Weak Learner: A new weak learner is trained to predict the residuals. It focuses on capturing the patterns in the residuals that the current ensemble fails to capture.

Update Ensemble: The new weak learner is added to the ensemble, and its contribution is determined by a learning rate parameter. The learning rate controls the weight given to each weak learner in the ensemble.

Weight Update: The weights of the training instances are updated based on the residuals. Instances with larger residuals are given higher weights, so that the next weak learner can focus on correcting these instances.

Final Prediction: The final prediction is obtained by combining the predictions of all the weak learners in the ensemble. The contribution of each weak learner is weighted based on the learning rate.

The iterative process continues until a predefined number of iterations is reached or a stopping criterion is met. The Gradient Boosting Classifier effectively combines the strengths of multiple weak learners to create a strong predictive model. It excels in capturing complex non-linear relationships and can provide accurate predictions for the binary pollution indicator based on the chlorophyll concentration features.

4.6.6 Convolutional Neural Networks (LeNet)

LeNet **??** is a convolutional neural network (CNN) model designed for image classification tasks (Y. Lecun, 1998). We chose to use LeNet for our study:

Transformation of Chlorophyll Concentration to Images: We transform the matrix of chlorophyll concentration, which is 11x11 in size, into images. By considering each value in the matrix as a pixel intensity, we can create images for input into the LeNet model.

Convolutional Networks for MNIST: The MNIST database is widely used for digit recognition, containing 60,000 training images and 10,000 testing images. Each image in MNIST is 28x28 pixels (database, 2010). MNIST classification tasks commonly involve the use of convolutional networks, and LeNet is a classic CNN architecture specifically designed for such tasks. LeNet's architecture (database, 2010) consists of convolutional layers, pooling layers, and fully connected layers.

Generalization to Chlorophyll Concentration: While originally designed for the MNIST dataset, LeNet's convolutional layers can learn and extract relevant features from various types of images, including our transformed chlorophyll concentration images. Although the image size in our case is 11x11, LeNet can still be applied effectively since it is designed to handle various image sizes.

And also the LeNet model was selected for our study for several reasons:

1. Proven Performance: LeNet is a well-established and widely used CNN architecture that has shown excellent performance in image classification tasks. It

was one of the pioneering models in the field of deep learning and has been successfully applied to various image recognition tasks, including the MNIST dataset.

2. Suitability for Small-Sized Images: The LeNet architecture is particularly suitable for small-sized images, such as the 11x11 transformed chlorophyll concentration images in this scenario. LeNet's design includes convolutional layers with small filter sizes, which enables the model to effectively capture local patterns and features in the images.

3. Effective Feature Extraction: LeNet's convolutional layers are designed to learn and extract meaningful features from images. This is crucial for image classification tasks as it allows the model to automatically identify relevant patterns and structures that can discriminate between different classes. The hierarchical structure of convolutional layers followed by pooling layers in LeNet helps in progressively abstracting features from the input images.

4. Efficient Training: LeNet's architecture strikes a good balance between model complexity and training efficiency. It contains a moderate number of trainable parameters, making it easier and faster to train compared to more complex CNN architectures. This is especially beneficial when dealing with limited computational resources or when working with smaller datasets.

5. Availability of Precedent: LeNet has been widely used and studied in the deep learning community, which means there are numerous resources, implementations, and insights available for reference. This makes it easier to find guidance, troubleshoot issues, and leverage existing knowledge when working with LeNet.

Overall, the selection of LeNet for this scenario was based on its proven performance, suitability for small-sized images, effective feature extraction capabilities, efficient training characteristics, and the availability of precedent in the deep learning community. These factors contribute to making LeNet a reliable and appropriate choice for image classification tasks, including the classification of transformed chlorophyll concentration images.

4.7 Metrics

4.7.1 ROC AUC

True positive rate (sensitivity) is a measure that quantifies the proportion of actual positive cases correctly identified by a classification model, indicating its ability to detect positive instances.

True negative rate (specificity) is a metric that represents the proportion of actual negative cases accurately identified by a classification model, indicating its capability to correctly identify negative instances.

To calculate the True Positive Rate (Sensitivity), divide the number of true positive predictions by the sum of true positive predictions and false negative predictions.

To calculate the True Negative Rate (Specificity), divide the number of true negative predictions by the sum of true negative predictions and false positive predictions.

The ROC curve is a graphical representation of the model's performance as the classification threshold varies. The plot illustrates the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) for various threshold values. The area under this curve, referred to as the ROC AUC score, quantifies the model's capability to differentiate between positive and negative cases.

A higher ROC AUC score indicates better performance, as it implies a higher ability to correctly classify instances. A score of 0.5 represents a random classifier, while a score of 1 indicates a perfect classifier.

4.7.2 Average precision

To understand the Average Precision Score, we need to consider precision and recall. Precision measures the proportion of correctly predicted positive instances (true positives) out of all instances predicted as positive. Recall measures the ratio of accurately predicted positive instances to the total number of actual positive instances.

The Average Precision Score combines precision and recall by calculating the area under the precision-recall curve. This curve plots the precision values against different levels of recall, typically achieved by varying the classification threshold. The score is computed by calculating the average precision at different recall levels and then taking the mean of those values.

A higher Average Precision Score indicates better performance, as it implies a higher proportion of correctly predicted positive instances with respect to both precision and recall. It provides an overall measure of the model's ability to rank and retrieve positive instances accurately based on the chlorophyll concentration and the binary pollution indicator.

Precision (Positive Predictive Value) is a measure that quantifies the proportion of correctly predicted positive instances out of all instances predicted as positive by a classification model, indicating the model's accuracy in identifying true positives.

Recall (Sensitivity) is a metric that represents the proportion of actual positive cases correctly identified by a classification model, indicating its ability to detect positive instances.

PR AUC stands for Precision-Recall Area Under the Curve. It is a performance metric used to evaluate the quality of a binary classification model, focusing on the trade-off between precision (positive predictive value) and recall (sensitivity). PR AUC represents the area under the precision-recall curve, which plots precision against recall at various classification thresholds. A higher PR AUC value indicates better model performance in achieving high precision and recall simultaneously.

4.7.3 Accuracy

When evaluating the accuracy of a model, we can use the following metrics:

True Positive (TP): The number of the instances where the model correctly predicts pollution when it is present.

True Negative (TN): The number of the instances where the model accurately predicts the absence of pollution.

False Positive (FP): The number of the instances when the model incorrectly predicts pollution when there is none.

False Negative (FN): The number of the instances when the model fails to predict pollution when it is actually present.

We calculate accuracy as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

4.7.4 F1 score

The F1 Score integrates precision and recall into a unified metric, providing an overall evaluation of the model's performance. Precision quantifies the ratio of true positives to all instances predicted as positive, while recall measures the ratio of true positives to all actual positive instances.

The F1 Score is determined by computing the harmonic mean of precision and recall. It provides a balanced measure that considers both precision and recall simultaneously. The harmonic mean emphasizes situations where both precision and recall are high, giving equal importance to both metrics.

A higher F1 Score indicates better performance, as it implies a higher balance between precision and recall. It shows the model's ability to achieve accurate positive predictions (high precision) while capturing most of the actual positive instances (high recall) based on the chlorophyll concentration and the binary pollution indicator.

4.8 Cross validation

Cross-validation is a method employed to evaluate the performance and generalizability of a model.

Cross-validation involves dividing the training dataset into multiple subsets or folds. In this case, the dataset was split into 5 equal parts or folds. The model is then trained and evaluated multiple times, with each fold serving as the validation set while the remaining folds are used for training. This process is repeated for each fold, ensuring that every data point is used for both training and validation.

4.9 Tuning hyper parameters

4.9.1 DecisionTreeClassifier model

For the DecisionTreeClassifier model, We employed the following hyperparameters to optimize its performance:

1. Criterion: This hyperparameter specifies the criterion used for node splitting in the decision tree. By selecting 'entropy', the model uses information gain based on the entropy of the target variable to make decisions about feature splits.

2. Min samples leaf: This hyperparameter determines the minimum number of samples required to be present at a leaf node. By setting it to 2, the model ensures that each leaf node has at least 2 samples, which helps prevent overfitting and promotes generalization.

3. Max depth: None indicates that there is no limit on the maximum depth of the decision tree, allowing it to grow until all leaves are pure or until the minimum number of samples required for a split is reached.

4. Min samples split: 2 specifies the minimum number of samples required to perform a split at an internal node. With a value of 2, the model will only split a node if it contains at least 2 samples.

4.9.2 LogisticRegression model

For the LogisticRegression model, We used the following hyperparameters to optimize its performance:

1. C: The hyperparameter C represents the inverse of the regularization strength. A higher value of C indicates weaker regularization. In this case, C was set to 100, implying a relatively low regularization strength, which allows the model to focus more on fitting the training data.

2. Penalty: The penalty hyperparameter determines the type of regularization used in the logistic regression model. Setting it to 'l1' indicates the adoption of L1 regularization, also known as Lasso regularization. L1 regularization encourages sparsity in the model's coefficients, favoring a subset of important features.

3. Solver: The solver hyperparameter determines the algorithm used for optimization during model training. 'liblinear' is a solver specifically designed for logistic regression problems. It is efficient for small-to-medium-sized datasets and supports both L1 and L2 regularization.

4.9.3 RandomForestClassifier model

For the RandomForestClassifier model, the following hyperparameters were found to yield the best performance:

1. Max depth: This hyperparameter specifies the maximum depth of each decision tree in the random forest. A depth of 10 means that each tree in the forest is allowed to have a maximum of 10 levels or splits.

2. Min samples leaf: This hyperparameter determines the minimum number of samples required to be at a leaf node in each decision tree. With a value of 4, each leaf node in the trees must have at least 4 samples.

3. Min samples split: This hyperparameter sets the minimum number of samples required to perform a split at an internal node in each decision tree. By setting it to 2, the model will only split a node if it contains at least 2 samples.

4. N estimators: This hyperparameter denotes the number of decision trees to be included in the random forest ensemble. A higher number of estimators generally leads to better model performance, but it also increases computational complexity.

4.9.4 GradientBoostingClassifier model

The selected model is the GradientBoostingClassifier, which has been optimized with the following hyperparameters:

1. Learning rate: The learning rate determines the contribution of each tree in the gradient boosting ensemble. A learning rate of 0.1 means that each tree's prediction is scaled down by a factor of 0.1, controlling the impact of each individual tree on the final model.

2. Max depth: The max depth hyperparameter defines the maximum depth of each tree in the gradient boosting ensemble. With a value of 7, each tree is allowed to have a maximum of 7 levels or splits, enabling the model to capture more complex relationships in the data.

3. N estimators: The n estimators hyperparameter specifies the number of boosting stages or iterations in the gradient boosting process. By setting it to 300, the ensemble will consist of 300 trees, which increases model complexity and improves its ability to learn from the data.

4. Subsample: The subsample hyperparameter determines the fraction of samples used for training each tree in the ensemble. In this case, a value of 0.8 indicates

that 80% of the data is randomly sampled with replacement for each tree, introducing randomness and reducing the likelihood of overfitting.

4.9.5 LeNet model

The chosen model is LeNet, and it has been configured with the following hyperparameters:

1. Batch size: The batch size hyperparameter determines the number of training examples processed in a single iteration or mini-batch during training. A batch size of 64 means that 64 samples will be processed at a time before updating the model's parameters.

2. Dropout rate: The dropout rate hyperparameter controls the regularization technique known as dropout. A dropout rate of 0.1 means that during training, 10% of the model's input units will be randomly set to 0 at each update, reducing the model's reliance on specific features and improving generalization.

3. Epochs: The epochs hyperparameter specifies the number of times the entire training dataset is passed through the model during training. With 200 epochs, the model will be trained on the complete dataset 200 times, allowing it to learn from the data and improve its performance over iterations.

4. Learning rate: The learning rate hyperparameter determines the step size at each iteration during the optimization process. It controls the rate at which the model's parameters are adjusted. A learning rate of 0.001 implies small updates to the model's parameters, which can prevent overshooting and help the model converge to the optimal solution.

Chapter 5

Experiments

We conducted an experiment where we varied the value of n between 3 and 11 to investigate its impact on the matrix chlorophyll concentration.

5.1 What is the optimal shape of data (3×3, 11×11, etc.) and statistical features to detect water pollution based on the concentration of chlorophyll?

5.1.1 Statistical features of chlorophyll concentration dataset

In this experiment, we focused on extracting the chlorophyll concentration from a dataset represented as a $n \times n$ matrix, where n from 3 to 11. The dataset consists of statistical features related to chlorophyll concentration, but it is imbalanced. These models were used with their default hyperparameters. The results can be observed in Table 5.1.

Based on the analysis of the Statistical features of chlorophyll concentration dataset, it has been determined that the optimal shape of the data is represented by an 11x11 matrix. This specific matrix configuration provides the best results in terms of performance metrics for the models evaluated.

Among the evaluated models, the random forest classifier emerged as the top-performing model, demonstrating superior accuracy, ROC AUC, and PR AUC scores within the context of the 11x11 matrix.

1. Accuracy: The baseline model achieved the highest accuracy of 0.0587.
2. ROC AUC: The random forest classifier attained a ROC AUC score of 0.5509.
3. PR AUC: Similarly, the random forest classifier obtained the highest PR AUC score of 0.0480.

Based on these results, we can conclude that among all the models considered for the 11x11 matrix, the random forest classifier performs the best for the Statistical features of chlorophyll concentration dataset.

5.1.2 Each pixel of chlorophyll concentration as a features dataset

In this experiment, we aimed to extract chlorophyll concentration from a dataset represented as as a $n \times n$ matrix, where n from 3 to 11. Each pixel in the matrix represents a feature for the chlorophyll concentration. Nevertheless, it is crucial to acknowledge that the dataset exhibits class imbalance, indicating an uneven distribution of chlorophyll concentration values. All models were used with their default hyperparameters. The results can be observed in Table 5.2.

Matrix Size	Model	F1 Score	ROC AUC	PR-AUC
3x3	Baseline	0.0605	0.5048	0.0314
3x3	Logistic Regression	0.0000	0.5000	0.0306
3x3	Random Forest	0.0179	0.5757	0.0497
3x3	Decision Trees	0.0253	0.5469	0.0346
3x3	Gradient Boosting	0.0000	0.5681	0.0382
5x5	Baseline	0.0604	0.5038	0.0313
5x5	Logistic Regression	0.0000	0.5000	0.0306
5x5	Random Forest	0.0147	0.5343	0.0349
5x5	Decision Trees	0.0233	0.5345	0.0330
5x5	Gradient Boosting	0.0000	0.5411	0.0339
7x7	Baseline	0.0597	0.4988	0.0310
7x7	Logistic Regression	0.0000	0.5000	0.0306
7x7	Random Forest	0.0197	0.5550	0.0426
7x7	Decision Trees	0.0366	0.5417	0.0340
7x7	Gradient Boosting	0.0050	0.5442	0.0367
9x9	Baseline	0.0598	0.4996	0.0310
9x9	Logistic Regression	0.0000	0.5000	0.0306
9x9	Random Forest	0.0147	0.5488	0.0401
9x9	Decision Trees	0.0353	0.5282	0.0328
9x9	Gradient Boosting	0.0049	0.5360	0.0354
11x11	Baseline	0.0587	0.4919	0.0306
11x11	Logistic Regression	0.0000	0.5000	0.0306
11x11	Random Forest	0.0339	0.5509	0.0480
11x11	Decision Trees	0.0374	0.5248	0.0329
11x11	Gradient Boosting	0.0098	0.5280	0.0375

TABLE 5.1: Models metrics for different matrix sizes of statistical features dataset

Upon analyzing the models using the 11x11 matrix for the Each pixel of chlorophyll concentration as a features dataset, we observe the following best performance metrics:

1. Accuracy: The baseline model achieved the highest accuracy of 0.1569.
2. Random Forest Classifier: The random forest classifier closely followed the baseline model, with an accuracy of 0.1564. It also obtained a relatively high ROC AUC score of 0.7188 and a PR AUC score of 0.1255.

Based on these results, we can conclude that among all the models considered for the 11x11 matrix with Each pixel of chlorophyll concentration as a features dataset, the random forest classifier performs the best. Although it has a slightly lower accuracy than the baseline model, it compensates with significantly higher ROC AUC and PR AUC scores, indicating better overall performance.

Furthermore, the Each pixel of chlorophyll concentration as a features dataset has demonstrated the best results compared to Statistical features of chlorophyll concentration dataset. Therefore, for further study and analysis, I use the Each pixel of chlorophyll concentration as a features dataset for improved insights and outcomes.

Matrix Size	Model	F1 Score	ROC AUC	PR-AUC
3x3	Baseline	0.0605	0.5048	0.0314
3x3	Logistic Regression	0.0000	0.5000	0.0306
3x3	Random Forest	0.0339	0.5931	0.0609
3x3	Decision Trees	0.0773	0.5531	0.0393
3x3	Gradient Boosting	0.0050	0.5704	0.0397
5x5	Baseline	0.060382	0.503775	0.031298
5x5	Logistic Regression	0.000000	0.500000	0.030587
5x5	Random Forest	0.099323	0.636162	0.115877
5x5	Decision Trees	0.116743	0.568111	0.045211
5x5	Gradient Boosting	0.009828	0.600200	0.067640
7x7	Baseline	0.059678	0.498836	0.030999
7x7	Logistic Regression	0.000000	0.499920	0.030587
7x7	Random Forest	0.139130	0.709112	0.154011
7x7	Decision Trees	0.148241	0.580216	0.052423
7x7	Gradient Boosting	0.019851	0.671053	0.088856
9x9	Baseline	0.059783	0.499568	0.031043
9x9	Logistic Regression	0.000000	0.499920	0.030587
9x9	Random Forest	0.154839	0.727279	0.173810
9x9	Decision Trees	0.156716	0.570948	0.053088
9x9	Gradient Boosting	0.034483	0.699704	0.104475
11x11	Baseline	0.0604	0.5038	0.0313
11x11	Logistic Regression	0.0000	0.5000	0.0306
11x11	Random Forest	0.1564	0.7188	0.1255
11x11	Decision Trees	0.1569	0.5666	0.0530
11x11	Gradient Boosting	0.0443	0.7138	0.1185

TABLE 5.2: Models metrics for different matrix sizes of Each pixel as features dataset

Based on our investigation, it is found that utilizing a dataset where each pixel represents an individual feature yields more favorable results for further analysis, as supported by the metrics presented in 5.2. A comparison with the metrics from 5.1 further reinforces the superior performance of this approach. This higher level of granularity and detailed information captured within the chlorophyll concentration dataset enables a more comprehensive understanding of the underlying patterns and dynamics of water pollution detection.

In response to the research question posed in Chapter 4 regarding the optimal shape of data and statistical features for detecting water pollution based on chlorophyll concentration, our analysis indicates that an 11×11 data shape yields the most favorable results. This conclusion is supported by the metrics presented in 5.2, which were compared with the metrics from 5.2. The evaluation of these metrics highlights the superior performance and effectiveness of the 11×11 data shape in detecting water pollution accurately and efficiently.

5.2 Which machine learning model performs best at detecting water pollution based on the concentration of chlorophyll?

5.2.1 11x11 dataset with the best hyper parameters

After evaluating the performance metrics of different datasets using models with default hyperparameters, we have chosen to focus on further research using the 11x11 matrix dataset. This dataset comprises chlorophyll concentration data, with each pixel representing a feature. Our next set of experiments were conducted using this dataset to explore its potential.

We have performed hyperparameter tuning for several models, including LogisticRegression, RandomForestClassifier, DecisionTreeClassifier, and GradientBoostingClassifier. Additionally, we have used the LeNet network for analyzing an 11×11 dataset.

In the hyperparameter tuning process, we adjusted the parameters of each model to optimize their performance. This involved finding the best combination of hyperparameters that would improve the accuracy or other relevant metrics of the models.

Furthermore, for the 11×11 dataset, we employed the LeNet network, which is a convolutional neural network architecture commonly used for image recognition tasks. By utilizing this network, we aimed to leverage its capabilities in analyzing the chlorophyll concentration data represented by each pixel in the 11×11 matrix.

5.2.2 DecisionTreeClassifier model

During the hyperparameter tuning process, the best parameter configuration for the DecisionTreeClassifier model was found to be:

$criterion : entropy, max\_depth : None,' min\_samples\_leaf' : 2,' min\_samples\_split' : 2$

In summary, these findings highlight the specific hyperparameters used, their purpose in optimizing the DecisionTreeClassifier model, and the best parameter configuration obtained through the hyperparameter tuning process.

The model with the best hyperparameters for the DecisionTreeClassifier outperformed the model with default parameters in terms of accuracy, F1 score, ROC AUC, and PR-AUC, achieving higher values in all metrics. It exhibited a significantly improved ability to correctly classify instances and discriminate between classes compared to the default parameter model.

5.2.3 LogisticRegression model

After conducting hyperparameter tuning, the best parameter configuration for the LogisticRegression model was found to be:

$C : 100, penalty : l1, solver : liblinear$

The model with the best hyperparameters for the Logistic Regression classifier significantly improved the performance compared to the model with default parameters. It achieved an accuracy of 0.6398, indicating a higher proportion of correctly classified instances. The F1 score improved from 0.0000 to 0.0839, suggesting a better balance between precision and recall. The ROC AUC also increased from 0.5000 to 0.5911, indicating an improved ability to discriminate between classes. Additionally, the PR-AUC improved from 0.0306 to 0.0386, indicating a better trade-off between precision and recall.

5.2.4 RandomForestClassifier model

The best parameter configuration for the RandomForestClassifier model consists of a maximum tree depth of 10, a minimum of 4 samples per leaf node, a minimum of 2 samples for splitting internal nodes, and an ensemble of 200 decision trees. These hyperparameters were determined through hyperparameter tuning to optimize the performance of the random forest model.

The model with the best hyperparameters for the RandomForestClassifier showed improvements in performance compared to the model with default parameters. It achieved an accuracy of 0.8523, indicating a higher percentage of correctly classified instances. The F1 score improved from 0.1564 to 0.1580, suggesting a slight enhancement in the balance between precision and recall. The ROC AUC increased from 0.7188 to 0.7404, indicating an improved ability to discriminate between classes. Additionally, the PR-AUC improved from 0.1255 to 0.1272, suggesting a better trade-off between precision and recall.

5.2.5 GradientBoostingClassifier model

In summary, the optimized GradientBoostingClassifier model uses a learning rate of 0.1, a maximum tree depth of 7, 300 boosting iterations, and a subsampling rate of 0.8. These hyperparameters have been selected through the tuning process to achieve the best performance for the model.

The model with the best hyperparameters for the GradientBoostingClassifier exhibited significant improvements in performance compared to the model with default parameters. It achieved an accuracy of 0.9521, indicating a high proportion of correctly classified instances. The F1 score improved from 0.0443 to 0.2445, suggesting a substantial enhancement in the balance between precision and recall. The ROC AUC increased from 0.7138 to 0.7659, indicating a significantly improved ability to discriminate between classes. Additionally, the PR-AUC improved from 0.1185 to 0.1821, indicating a better trade-off between precision and recall.

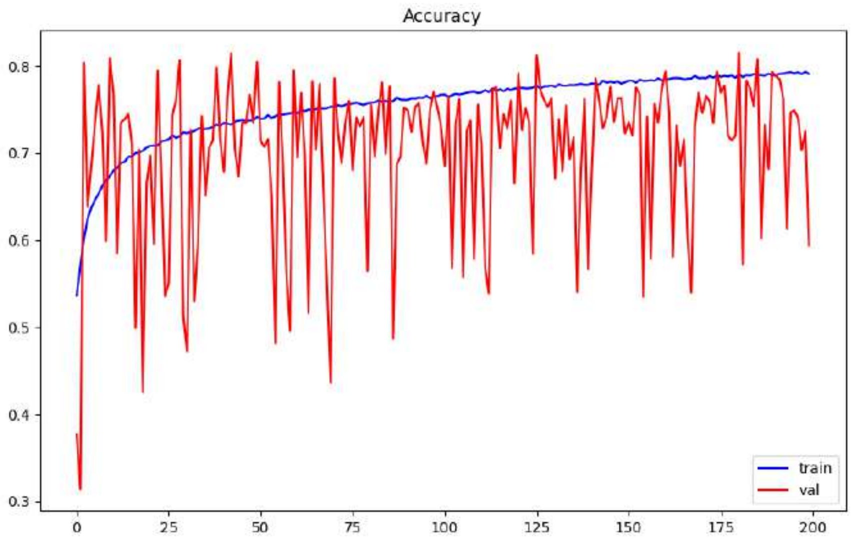


FIGURE 5.1: LeNet accuracy

5.2.6 LeNet model

In summary, the LeNet model has been configured with a batch size of 64, a dropout rate of 0.1, trained for 200 epochs, and a learning rate of 0.001. These hyperparameters were chosen to optimize the performance and training of the LeNet network for the specific task or dataset at hand. The accuracy of the LeNet model is depicted in Figure 5.1.

In our specific scenario, we used a separate validation set to fine-tune the hyperparameters for LeNet, a convolutional neural network architecture. To ensure optimal parameter selection, we utilized a cross-validation parameter, denoted as 'cv', set to a value of 5. This setting facilitated the division of the dataset into five separate folds, enabling comprehensive cross-validation throughout the training process.

5.3 Summary

As a result, we have compiled the performance metrics for all models on the balanced dataset, which are presented in the following table 5.3.

Model	Accuracy	F1 Score	ROC AUC	PR AUC
Baseline	0.2083	0.0605	0.5048	0.0314
DecisionTreeClassifier	0.9521	0.2013	0.5890	0.0692
LogisticRegression	0.6398	0.0839	0.5911	0.0386
RandomForestClassifier	0.8523	0.1580	0.7404	0.1272
GradientBoostingClassifier	**0.9521**	**0.2445**	**0.7659**	**0.1821**
LeNet	0.5935	0.0981	0.7530	0.2092

TABLE 5.3: Final Performance Metrics

Based on the final performance metrics, the DecisionTreeClassifier and GradientBoostingClassifier models achieved the highest accuracy of 0.9521. The DecisionTreeClassifier model performed better in terms of the F1 score with a value of 0.2013, while the GradientBoostingClassifier model outperformed the others in terms of F1 score with a value of 0.2445. DecisionTreeClassifier, RandomForestClassifier, and

GradientBoostingClassifier, generally performed better than the baseline and simpler model (LogisticRegression). They showed improvements across multiple metrics, including accuracy, F1 score, ROC AUC, and PR AUC. These models were able to capture more intricate patterns in the data, leading to enhanced performance in classification tasks.

In response to the research question posed in Chapter 4 regarding the machine learning model that performs best in detecting water pollution based on chlorophyll concentration, our analysis reveals that the GradientBoostingClassifier models achieved the highest accuracy of 0.9521. This result is based on the metrics presented in Table 5.3. The evaluation of these metrics demonstrates the superior performance of the GradientBoostingClassifier models in accurately detecting water pollution in relation to chlorophyll concentration.

The following charts are presented for visual analysis of the metrics:
The F1 scores for different models are displayed in Table 5.2.

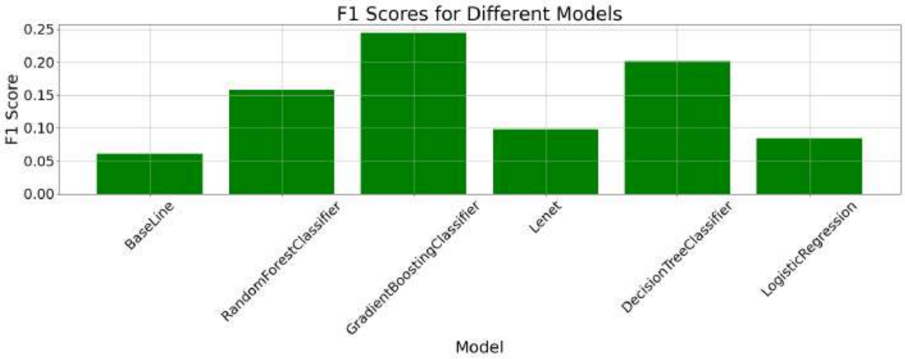


FIGURE 5.2: F1 score for different models

The ROC AUC scores for different models are displayed in Table 5.3.

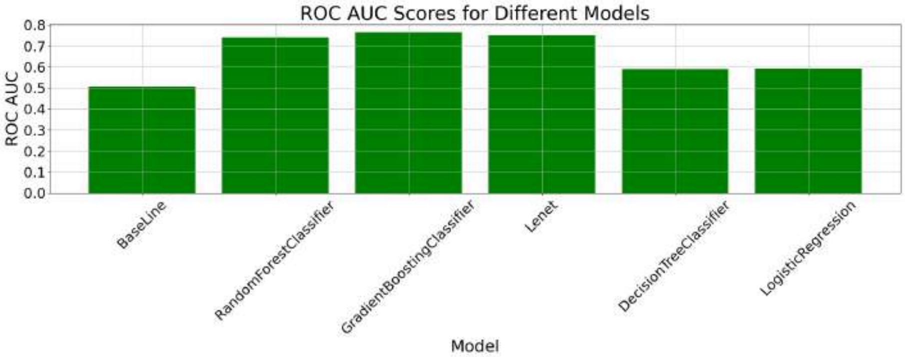


FIGURE 5.3: ROC AUC score for different models

The PR AUC scores for different models are displayed in Figure 5.4.
The Accuracy scores for different models are displayed in Table 5.5.

5.3.1 Comparative Analysis of F1 Score, Pollution Incidents, and Average Chlorophyll Level by Month

Next, we split the test dataset into separate datasets based on the months. Each month's dataset was evaluated individually to obtain the F1 score for that specific month.

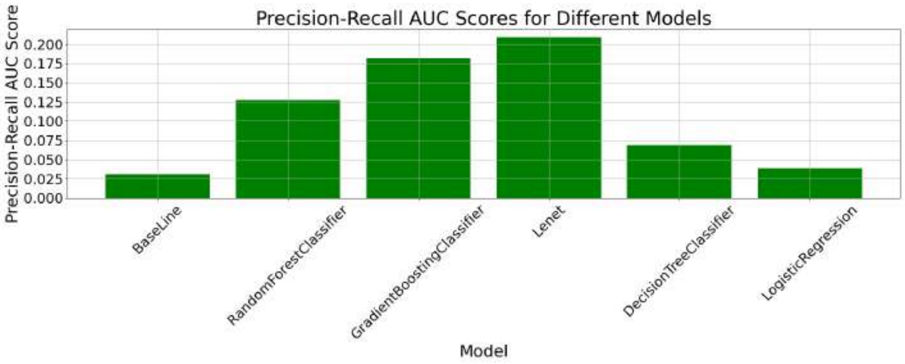


FIGURE 5.4: PR AUC score for different models

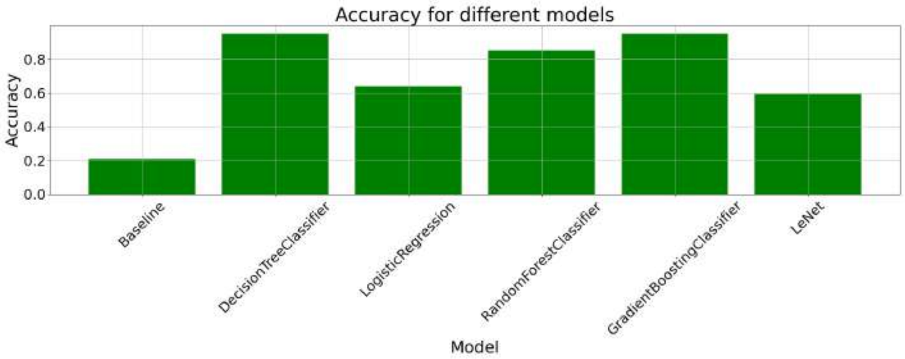


FIGURE 5.5: Accuracy score for different models

Winter months were excluded from the table because a significant number of rows had to be removed due to corruption issues from the Copernicus marine service. Therefore, the F1 score metric for the winter months is not available in the table. The F1 score metric for months is displayed in Table 5.4.

Month	F1 Score
March	0
April	0.526
May	0.608
June	0.729
July	0.494
August	0.655
September	0.666
October	0.5

TABLE 5.4: F1 Score for Each Month

Visually, we can examine the F1 score metric for each month in Figure 5.6.

We have constructed a histogram that represents the number of pollution incidents by month, which is displayed in Figure 5.7. Additionally, we compared this histogram with the F1 score metric for each month.

Upon analysis, it is evident that the number of pollution incidents increased in July, while the corresponding F1 score decreased. Conversely, in September, although the number of pollution incidents remained high, the F1 score was also high. These findings suggest that there may not be a clear pattern between the increase in

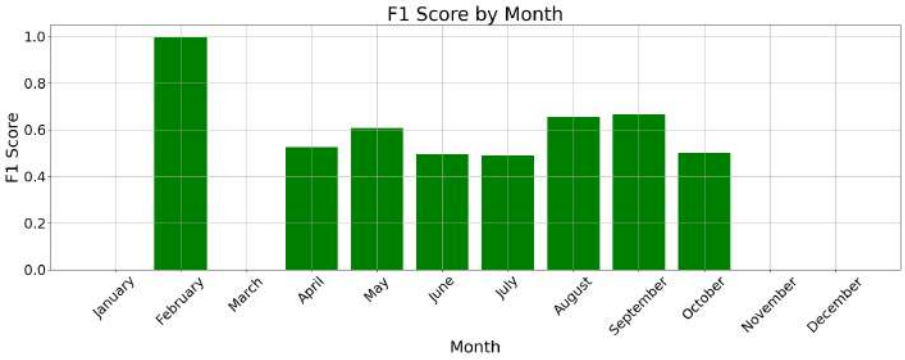


FIGURE 5.6: The F1 score metric for each month

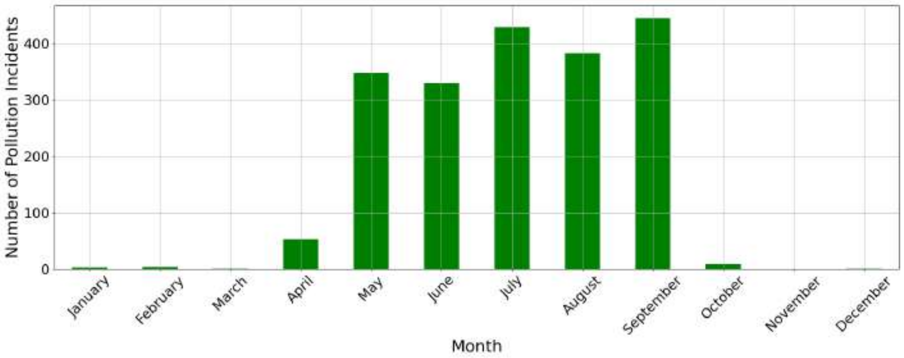


FIGURE 5.7: Pollution incidents by month

the number of pollution incidents and the F1 score.

We have constructed a histogram that represents the number of pollution incidents by month, which is displayed in Figure 5.8. Additionally, we compared this histogram with the F1 score metric for each month.

The purpose of this experiment was to establish a correlation between the F1 score and pollution incidents by month or average chlorophyll levels by month. However, no significant relationship was observed between them. Specifically, when the F1 score decreased in July, there was no noticeable impact on the occurrence of pollution incidents by month or the average chlorophyll levels by month.

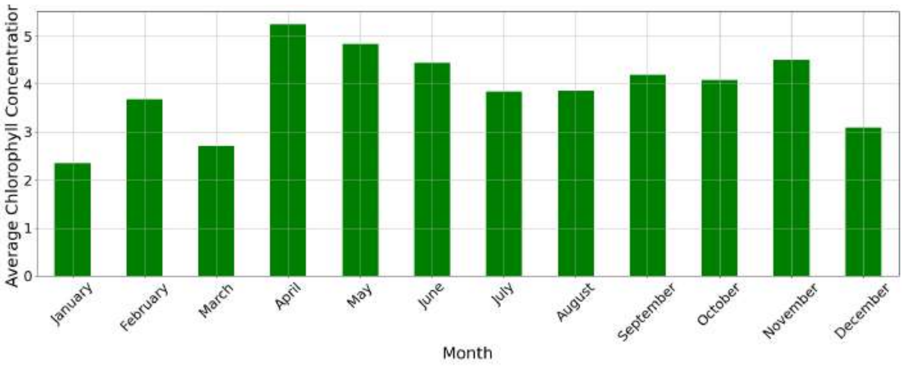


FIGURE 5.8: Average chlorophyll by month

By comparing the number of pollution incidents and the average chlorophyll concentration by month, it becomes evident that there is an inverse relationship between the two variables. Specifically, when the number of pollution incidents is lower, the average chlorophyll concentration tends to be higher. This suggests that a decrease in pollution incidents may contribute to an increase in chlorophyll levels in the studied environment. For the months of April and May, there are fewer pollution incidents recorded, but the chlorophyll level is relatively high. In contrast, the months of July and September exhibit a higher number of pollution incidents, accompanied by lower chlorophyll levels compared to other months. These observations highlight the potential inverse relationship between pollution incidents and chlorophyll levels, where a decrease in pollution incidents is associated with higher chlorophyll levels, while an increase in pollution incidents is linked to lower chlorophyll levels.

According to the visual analysis of the F1 score and average chlorophyll level by month, no clear patterns or correlations can be observed. This suggests that a more comprehensive and detailed study is required to investigate the relationship between these two variables. Further analysis, such as statistical tests or advanced modeling techniques, may be necessary to uncover any potential underlying patterns or dependencies between the F1 score and average chlorophyll level.

5.3.2 Comparative Analysis of F1 Score, Pollution Incidents, and Average Chlorophyll Level by Region

We split the test dataset into regions around England and aggregated the data for each region.

Pseudocode for the `divide test dataset into squares` **algorithm:**

1. First, the algorithm identifies the minimum and maximum values of latitude and longitude from the given pairs.
2. It then calculates the number of rows and columns required in the grid based on the specified square size.
3. An empty grid is initialized to hold the squares.
4. The algorithm iterates over each pair and assigns it to the corresponding square in the grid.
5. For each latitude and longitude pair, the algorithm calculates the row and column indices within the grid based on the minimum latitude and longitude values and the specified square size.
6. If a pair falls within a specific square defined by its row and column indices, it is added to that square's set of pairs.
7. Each non-empty square, along with its set of pairs, is appended to a row.
8. If a row contains at least one non-empty square, it is added to the grid.
9. Finally, the algorithm returns the resulting grid, which represents the partitioned data.

Following the F1 score calculation for each region, the results are presented in Figure 5.9.

We have generated the number of pollution incidents by region 5.10 and the average chlorophyll level by region 5.11.

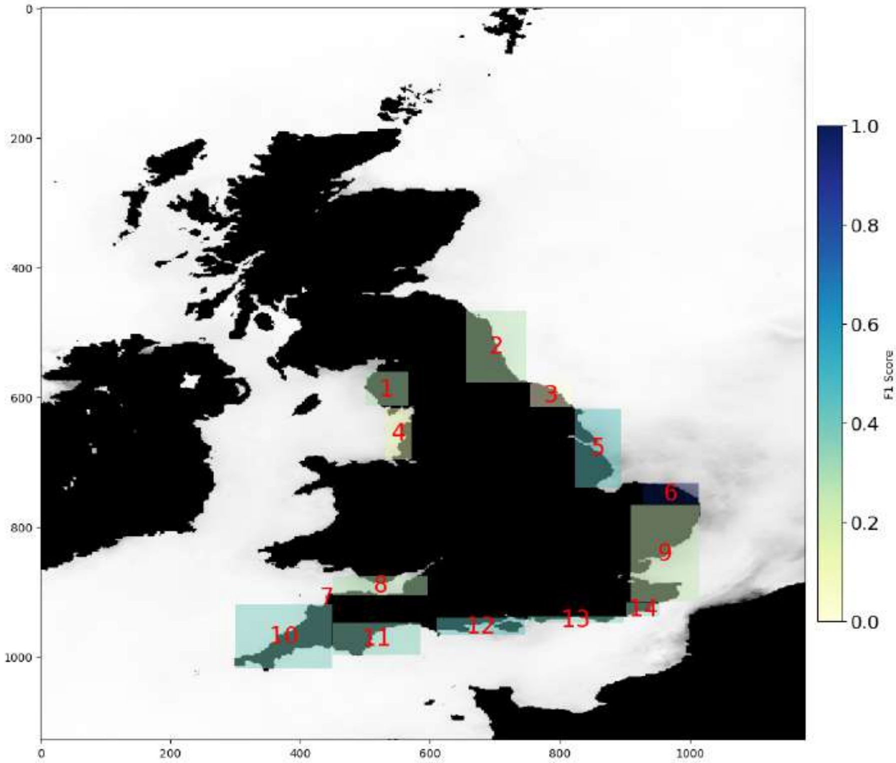


FIGURE 5.9: The F1 score metric for each region

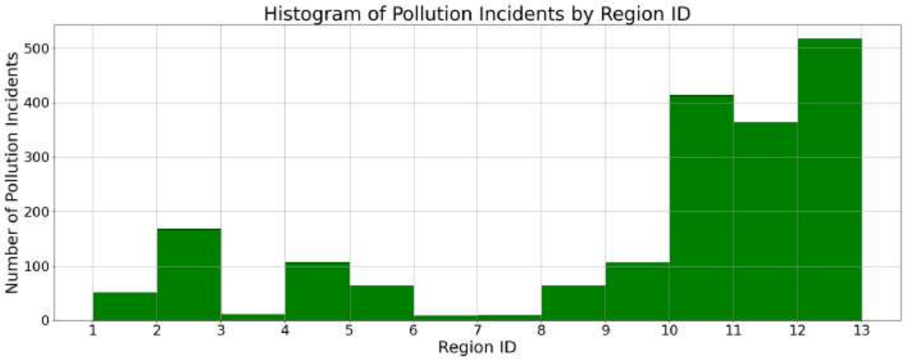


FIGURE 5.10: The number of pollution incidents by region

By comparing the F1 score by region chart with the number of pollution incidents, we observe that regions 10, 11, 12, and 13 exhibit high F1 scores along with the highest levels of chlorophyll. On the other hand, region 5 shows a high F1 score but a low number of pollution incidents. This finding suggests that further investigation is warranted for future studies.

By comparing the F1 score by region chart with the average chlorophyll level, we do not observe any discernible patterns. There is no observed correlation between the average chlorophyll concentration categorized by region and the F1 score.

5.3.3 Confusion matrix and precision-recall curve

The best model is gradient boosting classifier. Among the evaluated models, the gradient boosting classifier achieved the highest accuracy with a value of 0.9521, indicating its superior performance in correctly classifying instances. In terms of F1

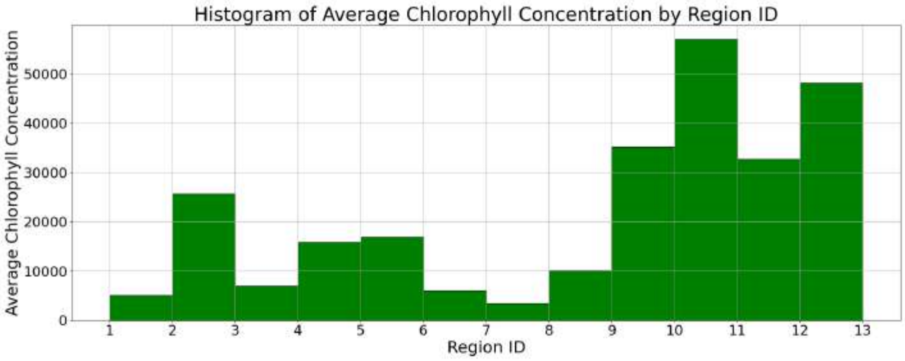


FIGURE 5.11: The average chlorophyll by region

score, the model attained a score of 0.2445, reflecting a balanced precision and recall trade-off. The ROC AUC (Receiver Operating Characteristic Area Under the Curve) value for the gradient boosting classifier was 0.7659, indicating its effectiveness in distinguishing between positive and negative instances. Additionally, the PR AUC (Precision-Recall Area Under the Curve) of 0.1821 suggests the model's ability to maintain precision while correctly identifying positive instances.

Based on the evaluation results, the best model Gradient boosting classifier is determined using the confusion matrix and precision-recall curves. These metrics provide valuable insights into the performance and predictive capabilities of the model.

The confusion matrix allows us to analyze the classification accuracy of the model by presenting the counts of true positive, true negative, false positive, and false negative predictions for each class. It provides a comprehensive view of the model's performance across different classes and helps identify any misclassifications or imbalances. The confusion matrix is presented in Figure 5.12.

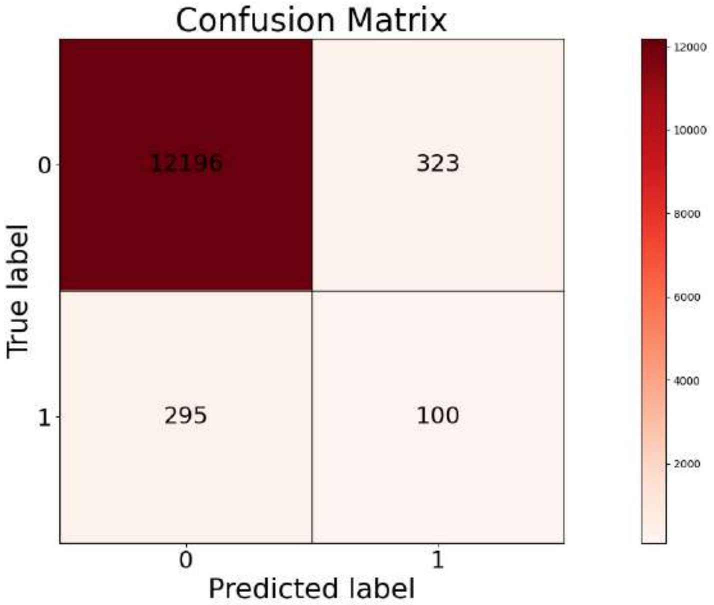


FIGURE 5.12: Confusion Matrix

The model achieved a high number of true positives (TP), with 12,196 instances correctly predicted as positive. This indicates that the model effectively identified and classified a significant portion of the positive class accurately. However, the model had a relatively low number of true negatives (TN), with only 100 instances correctly predicted as negative. This suggests that the model had some difficulty correctly identifying and classifying instances belonging to the negative class. Additionally, there were a moderate number of false positives (FP), with 323 instances incorrectly predicted as positive. These instances were classified as positive by the model, but in reality, they belonged to the negative class. Similarly, the model had a moderate number of false negatives (FN), with 295 instances incorrectly predicted as negative. These instances were classified as negative by the model, but in reality, they belonged to the positive class. In summary, the model exhibited a strong performance in correctly classifying positive instances (TP) but struggled to accurately identify negative instances (TN). The presence of false positives (FP) and false negatives (FN) suggests areas where the model's predictions may have been less accurate. Further analysis and improvements may be necessary to enhance the model's performance, particularly in correctly classifying negative instances.

On the other hand, precision-recall curves offer a graphical representation of the model's precision and recall values at different classification thresholds. Precision measures the proportion of correctly classified positive instances out of all instances predicted as positive, while recall calculates the proportion of correctly classified positive instances out of all actual positive instances. The curve illustrates the tradeoff between precision and recall, allowing us to choose an appropriate threshold for our specific needs.

Based on the given confusion matrix:

1. True positive (TP): 12196
2. False positive (FP): 323
3. False negative (FN): 295
4. True negative (TN): 100

The confusion matrix represents the performance of a binary classification model. In this case, the model correctly predicted 12,196 instances of class 1 (positive) as positive (TP) and incorrectly classified 323 instances of class 0 (negative) as positive (FP). Moreover, the model wrongly classified 295 instances of class 1 as negative (FN) and correctly predicted 100 instances of class 0 as negative (TN). The model makes much more true positive predictions than true negative ones, but only because there are much more positive cases in the data.

Recall (also known as sensitivity or true positive rate) is the proportion of actual positive instances correctly identified by the model. Recall = TP / (TP + FN) = 12196 / (12196 + 295) = 0.9766 = 97.66

Precision is the proportion of positive predictions that are actually correct. Precision = TP / (TP + FP) = 12196 / (12196 + 323) = 0.9747 = 97.47

Therefore, the recall is 97.66% and the precision is 97.47%.

The recall of 97.66% indicates that the model correctly identified a high percentage of actual positive instances (12196 out of 12491). This suggests that the model is effective in capturing the true positives and has a low false negative rate (295 out of 12491). In other words, it successfully detected the majority of positive instances in the dataset.

The precision of 97.47% signifies that among the instances predicted as positive by the model, the majority (12196 out of 12519) were indeed true positives. This indicates that the model has a low false positive rate (323 out of 12519). It demonstrates that when the model predicts an instance as positive, there is a high likelihood that it is indeed a positive instance.

Overall, the high recall and precision values indicate that the model performs well in accurately classifying positive instances, achieving a good balance between capturing positive instances and minimizing false positive and false negative predictions.

The precision-recall curve is depicted in Figure 5.13.

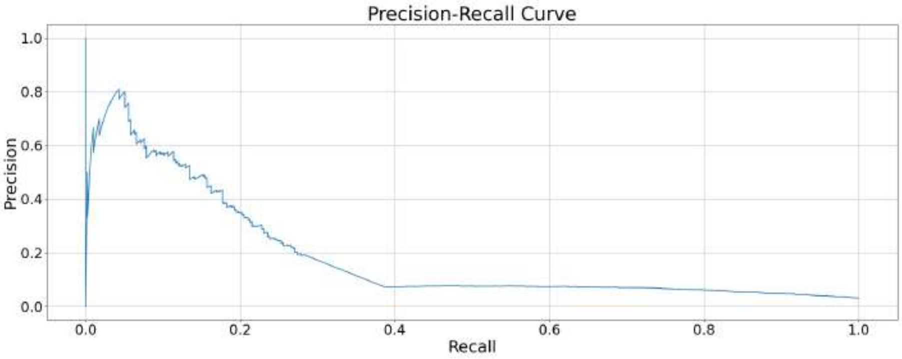


FIGURE 5.13: Precision Recall Curve

By analyzing the confusion matrix and precision-recall curves, we can determine the best model based on its ability to accurately classify instances, handle class imbalances, and strike an optimal balance between precision and recall. These evaluation metrics provide valuable insights into the model's performance and can guide decision-making in selecting the most suitable model for the task at hand.

Before, we would need to perform 12914 manual measurements to detect the cases of water pollution in your test set. Those measurements are expensive (in terms of money and human-hours), and each one would have a probability of

$$P = \frac{12914}{395} \approx 33\%$$

of detecting the actual case of pollution.

With your model acting as a tool for early detection, we can detect 97.66% of all pollution cases (recall) by only checking the 423 cases that were reported by your model as positive. In this scenario, each measurement will have a P = 97.47% chance of detecting actual pollution (precision).

Chapter 6

Conclusions

Based on the research conducted, which involved data on chlorophyll concentration from the Copernicus Marine Service and pollution reports from DEFRA, the goal was to classify the presence of pollution based on chlorophyll concentration.

The study included experiments using different matrix sizes (3x3, 5x5, 7x7, 9x9, and 11x11) of chlorophyll concentration as features. Feature engineering was performed, resulting in the creation of two datasets: one with statistical metrics (mean, median, standard deviation, and percentiles) as features, and another with each cell of the chlorophyll concentration matrix as a feature.

Evaluation metrics such as F1 score, ROC AUC, and PR AUC were used to assess the performance of the models. The results indicated that the dataset with each cell of the chlorophyll concentration matrix as a feature, combined with a gradient boosting classifier, yielded the best metrics.

- Accuracy: The model achieved an accuracy of 0.9521, indicating a high overall correct classification rate.

- F1 Score: The F1 score, which combines precision and recall, was determined to be 0.2445. This metric considers both the model's ability to correctly identify positive instances (precision) and its ability to capture all positive instances (recall).

- ROC AUC: The receiver operating characteristic area under the curve (ROC AUC) was calculated as 0.7659. This metric assesses the model's ability to distinguish between positive and negative instances across different classification thresholds.

- PR-AUC: The precision-recall area under the curve (PR-AUC) was found to be 0.1821. This metric evaluates the trade-off between precision and recall and provides insights into the model's performance when dealing with imbalanced datasets.

Specific hyperparameters were tuned for the gradient boosting classifier, including a learning rate of 0.1, a maximum tree depth of 7, 300 boosting iterations, and a subsampling rate of 0.8.

In conclusion, the research demonstrated that using each cell of the chlorophyll concentration matrix as a feature, along with a gradient boosting classifier, produced the best performance in classifying pollution based on chlorophyll concentration. These findings provide valuable insights for future studies in the field of pollution classification using remote sensing data.

6.1 Future research directions

1. **Comparative Analysis with Other Satellite Services:** In order to enhance the understanding of chlorophyll concentration and its relationship with pollution incidents, it is recommended to acquire chlorophyll concentration datasets from other satellite services such as Aqua MODIS, Landsat, GeoEye-1, and WorldView-2/3/4. By comparing the results obtained from these datasets with the current research findings, a more comprehensive understanding of the spatial distribution and variability of chlorophyll concentration can be achieved.

2. **Comprehensive Study of Optical Properties in Sea Water:** Conduct an extensive investigation into the optical properties of sea water, including sea water turbidity, mass concentration of suspended matter, volume backward scattering coefficient of radiative flux, volume attenuation coefficient of downwelling radiative flux, Secchi depth, and volume absorption coefficient of radiative flux. Explore the interrelationships between these properties, their spatial and temporal variations, and their impact on light transmission and water quality in marine ecosystems. This research will provide a deeper understanding of the optical characteristics of sea water and their ecological significance, contributing to improved monitoring and management strategies for coastal and marine environments.

3. **Time Series and Advanced Time Series Models:** To improve the accuracy of pollution incident predictions, it is suggested to explore a range of time series models. Traditional models like Autoregressive Integrated Moving Average (ARIMA), Seasonal ARIMA (SARIMA), and Exponential Smoothing (ES) can be employed. Additionally, advanced models such as Vector Autoregression (VAR), Long Short-Term Memory (LSTM), Gaussian Process Regression (GPR), Seasonal-Trend Decomposition using LOESS (STL), Bayesian Structural Time Series (BSTS), Recurrent Neural Networks (RNNs), and other architectures of Convolutional Neural Networks (CNNs) can be investigated. These models offer the potential to capture complex patterns and relationships within the time series data.

4. **Contamination Direction Prediction:** Once I have identified the cases of contamination, figure out the direction in which the contamination will move.

5. **Exploring Additional Dataset Balancing Approaches:** In addition to the currently applied dataset balancing approaches, further investigation can be conducted on other methods such as Random Undersampling, Synthetic Minority Over-sampling Technique (SMOTE), Adaptive Synthetic Sampling (ADASYN), and Data Augmentation. These techniques can help mitigate the effects of imbalanced datasets and potentially improve the performance of the classification models.

6. **Hypothesis Testing on Mission Date and Tides:** It is worth exploring the hypothesis that the mission date of the satellite and the occurrence of tides may be related. By analyzing the data and investigating potential correlations or patterns between mission dates and tidal variations, new insights into the dynamics of chlorophyll concentration and pollution incidents can be gained.

7. **Relationship between Pollution Incidents and Chlorophyll Concentration:** Further investigation can be conducted to identify any potential patterns or relationships between the number of pollution incidents and the average chlorophyll concentration. This analysis can provide valuable insights into the ecological dynamics of the marine environment and help understand the impact of chlorophyll concentration on pollution incidents.

By undertaking these future advancements, the research can contribute to the advancement of knowledge in the field of marine chlorophyll concentration analysis, pollution incident prediction, and environmental monitoring.

Appendix A

Links to materials

A.1 Code

The public repository with code is available by the link:
`https://github.com/yest89/PollutionProject`

A.2 Dataset

Dataset is available by the link:
`https://drive.google.com/drive/folders/1YXlfv5yWUEHngXq0b52FgBp3LsD9Rzf1`

Bibliography

Agency., United States Environmental Protection (2023). *National aquatic resource surveys. indicators: Chlorophyll a.*

A.H.Kadhim, S.M.Ali A.S.Mahdi (2012). “Monitoring of Water Pollution in Diyala River using High Resolution Satellite Image”. In: *College of Science / Baghdad University*. https://www.researchgate.net/publication/236576836_Monitoring_of_Water_Pollution_in_Diyala_River_using_High_Resolution_Satellite_Image.

analysis, Principal component (2016). *Principal component analysis*.

boosting, Gradient (2016). *Gradient boosting*.

Casanova, Diego Gomez Pablo Salvador Julia Sanz José Luis (2021). “A new approach to monitor water quality in the Menor sea (Spain) using satellite data and machine learning methods”. In: *Environmental Pollution* Volume 286. DOI: \url{https://doi.org/10.1016/j.envpol.2021.117489}.

Chen, Haibo Yang Jialin Kong Huihui Hu Yao Du Meiyan Gao Fei (2022). “A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges”. In: *Remote sensing* 1770. https://www.mdpi.com/2072-4292/14/8/1770.

database, MNIST (2010). *MNIST database*.

Directive, The EU Water Framework (2020). “Satellite-assisted monitoring of water quality to support the implementation of the Water Framework Directive”. In: *European Commission*. https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5cc8cbc7d&appId=PPGMS.

factorization, Non negative matrix (2016). *Non-negative matrix factorization*.

forest, Random (2016). *Random forest*.

Gathot Winarso, Yennie Marini (2014). “MODIS STANDARD (OC3) CHLOROPHYLL-A ALGORITHM EVALUATION IN INDONESIAN SEAS”. In: *Winarso* 11. https://jurnal.lapan.go.id/index.php/ijreses/article/view/2597.

Godson Adjovu Haroon Stephen, David James Sajjad Ahmad (2023). “Overview of the Application of Remote Sensing in Effective Monitoring of Water Quality Parameters”. In: *Remote Sens.* 1938. https://doi.org/10.3390/rs15071938.

Laville, Sandra (2023). *30 water treatment works released 11bn litres of raw sewage in a year, study suggests*.

learn.org, imbalanced (2014). *RandomOverSampler*.

Lu, Liufeige (2022). “Remote sensing quantitative monitoring of chlorophyll a concentration in Taihu Lake based on measured spectrum of water surface”. In: *BIO Web Conf.* dx.doi.org/10.1051/bioconf/20225501015.

machine, Support vector (2016). *Support vector machine*.

Merchant, Nathan (2017). “A Review of the Tools Used for Marine Monitoring in the UK: Combining Historic and Contemporary Methods with Modeling and Socioeconomics to Fulfill Legislative Needs and Scientific Ambitions”. In: *Marine Ecosystem Ecology*. https://doi.org/10.3389/fmars.2017.00263.

Navarro, Isabel Caballero Mar Roca Juan Santos-Echeandía Patricia Bernárdez Gabriel (2022). "Use of the Sentinel-2 and Landsat-8 Satellites for Water Quality Monitoring: An Early Warning Tool in the Mar Menor Coastal Lagoon". In: *Remote Sens.* 2744. https://doi.org/10.3390/rs14122744.

network, Neural (2016). *Neural network*.

Nichol, Sidrah Hafeez Man Sing Wong Sawaid Abbas Janet (2018). "Detection and Monitoring of Marine Pollution Using Remote Sensing Technologies". In: *Monitoring of Marine Pollution*. http://dx.doi.org/10.5772/intechopen.81657.

RBINS (2021). *ACOLITE*.

regression, Logistic (2016). *Logistic regression*.

Service, Copernicus Marine (2014). "Copernicus Programme". In: *Mercator Ocean International*. https://marine.copernicus.eu/.

sklearn (2023a). *Standard Scaler*.

— (2023b). *Train test split*.

tree, Decision (2016). *Decision tree*.

T.Vakili, J.Amanollahi (2020). "Determination of optically inactive water quality variables using Landsat 8 data: A case study in Geshlagh reservoir affected by agricultural land use." In: *J. Clean. Prod.* 119134. https://doi.org/10.1016/j.jclepro.2019.119134.

UNESCO (2021). "The ocean decade. The Science We Need for the Ocean We Want." In: *UNESCO*. https://oceandecade.org/.

Y. Lecun L. Bottou, Y. Bengio P. Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE*. https://ieeexplore.ieee.org/document/726791.

Y.Zhang J.Pulliainen, S.Koponen M.Hallikainen (2002). "Application of an empirical neural network to surface water quality estimation in the Gulf of Finland using combined optical data and microwave data." In: *Remote Sens.* 81. https://doi.org/10.1016/S0034-4257(02)00009-3.