INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA INFORMÁTICA




Integração de Várias Fontes de Dados para a Previsão de Florações de Algas Nocivas Cianobactérias Usando Aprendizagem Automática

MARCO ANDRÉ MORGADO QUERIDO
Junho de 2025

POLITÉCNICO
DO PORTO

ISEP INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO

Integration of Multiple Data Sources on the Prediction of Harmful Algal Blooms Using Machine Learning

Marco André Morgado Querido

A dissertation submitted in partial fulfillment of the requirements for the degree of Master of Science, Specialisation Area of Software Engineering

Supervisor: Dr. Ivo Pereira
Co-Supervisors: Dr. Bruno Cunha, Dr.ª Ivone Amorim, Dr. Hugo Barbosa

Porto, June 28, 2025

Statement of Integrity

I hereby declare that I have conducted this academic work with integrity.

I have not plagiarised or applied any form of undue use of information or falsification of results along the process leading to its elaboration.

Therefore, the work presented in this document is original and authored by me, having not previously been used for any other end. The exceptions are explicitly recognised in the section "Ethical considerations" of the first chapter. This section also states how AI tools were used and for what purpose.

I further declare that I have fully acknowledged the Code of Ethical Conduct of P.PORTO.

ISEP, Porto, June 28, 2025

Dedicatory

This dissertation is dedicated to all who, through their presence, support, and faith in me, enabled me to take this journey. It is a personal and academic triumph as much as it is a collective effort instigated by the generosity, patience, and support of so many gifted individuals.

To professors Bruno Cunha, Hugo Barbosa, Ivo Pereira and Ivone Amorim, thank you for your direction, for walking with me step by step throughout this dissertation, and for your experience disseminated so richly over the last year. Your direction gave me light when I wandered, and power when I was exhausted.

To my family, the biggest support of my life. Your love, encouragement and support gave me strenght through every moment. To my mom and dad especially, your constant support, motivation and encouragement have guided me through the darkest times and lifted me up in the brightest times. This achievement is yours as much as mine. Thank you for always believing in me.

To my girlfriend, thank you for your patience, your calm, and your encouragement. You were with me every step of the way, through stress, fatigue, and through every triumph. Your encouragement was everything to me.

To my friends, who were with me along this path, and told me that this was not a path to be traveled alone. Your presence and sense of humor made the difference.

A special and heartfelt thank you to Professor Cristina Soares. Your knowledge in chemistry and your guidance were beyond description. Your kindness was a shining light at crucial moments, and your joy will never be forgotten.

To all those who, in one way or another, contributed to making this achievement a reality, I extend my heartfelt thank you. This dissertation is proof of your belief in me, and to you, then, I dedicate this dissertation, in gratitude.

Abstract

This dissertation addresses the issue of forecasting cyanobacterial harmful algal blooms (cyanoHABs), a source of harm to aquatic ecosystems and human health via toxin production and water quality degradation. Motivated by the limitations of traditional forecast methods based on single-source data, this work explores multi-source integration to further optimize forecast precision.

The conceptual model is based on the assumption that a combination of several biogeochemical, physical, and meteorological parameters better characterizes cyanobacteria bloom variability's multi-faceted drivers. Based on Copernicus Marine Service data, the model combines chlorophyll-a concentration, as a factor for the magnitude of the bloom, with parameters such as sea surface temperature (SST), rising significant wave height (SSWH), nutrient concentration (e.g., phosphate, ammonium), net primary production (nppv), phytoplankton biomass (phyc), and euphotic depth (zeu).

The approach used was based on the CRISP-DM methodology. The importance of predicting cyanoHAB was realized, and data understanding and preparation involved collecting, cleaning, and preprocessing multi-source time-series data. Ensemble classifiers (Random Forest, Bagging, XGBoost) were used for chlorophyll-a classification and regression models (Random Forest Regressor, ARIMA, SARIMA, LSTM, GRU, CNN) for forecasting trends of chlorophyll-a in the modeling phase. Performance comparison employed ROC AUC, precision, and recall for classification tasks and R^2 and RMSE for regression.

Results show ensemble classifiers labeled chlorophyll-a with almost perfect accuracy and ROC AUC values close to 1.00, and they noted biogeochemical features nppv, phyc, and zeu as the most predictive. Random Forest Regressor was best for regression on time-series ($R^2 = 0.594$), simulating short-term chlorophyll-a patterns accurately. Though, under oversmoothing or instability with noise in the data, the traditional models (ARIMA, SARIMA) and deep learning models (LSTM, GRU, CNN) were not as good.

These findings confirm that multi-source data integration evidently enhances cyanoHAB forecasting and that the use of ensemble machine learning models to make accurate and interpretable predictions is confirmed. The dissertation ends by observing that environmental factors need to be enhanced in prediction models and explainable AI approaches incorporated to build confidence and improve decision-making for water quality management.

Keywords: CyanoHABs, Machine Learning, Artificial Intelligence, Multi-source Data Merging, cyanoHABs Prediction, Water Security

Resumo

Florações de algas nocivas cianobactérias (cyanoHABs) são fenómenos naturais caracterizados pelo crescimento anormal de cianobactérias em volumes de água, frequentemente associados a grandes perdas económicas, riscos graves à saúde animal e humana devido à produção de toxinas pelas cyanoHABs e destruição de ecossistemas aquáticos. Embora já se tenham identificado mecanismos de controlo físico-químico que influenciam os processos que originam e mantêm as cyanoHABs, existem, no entanto, lacunas significativas de conhecimento relativamente aos fatores que desencadeiam e mantêm estes fenómenos.

Nas últimas décadas, muitos estudos têm demonstrado como as abordagens de aprendizagem automática e de inteligência artificial podem ser bastante eficazes na criação de modelos preditivos para eventos de cyanoHABs, utilizando dados ambientais. No entanto, a revisão da literatura demonstra que a maioria dos estudos até agora concentra-se em dados univariáveis, o que limita a capacidade dos modelos de capturar a complexidade multivariada que influencia a dinâmica das cyanoHABs. Perante isso, esta dissertação pretende preencher essa lacuna, explorando a integração de diversas fontes de dados, incluindo variáveis meteorológicas, físicas e biogeoquímicas, para melhorar a previsibilidade dos eventos de cyanoHABs.

O principal objetivo desta dissertação é avaliar a eficácia de alguns modelos de aprendizagem automática que utilizam dados de diferentes fontes para prever florações de cyanoHABs. Para isso, são seguidas três etapas essenciais: realizar uma revisão bibliográfica abrangente sobre como a aprendizagem automática tem sido aplicada na previsão de cyanoHABs; analisar e preparar conjuntos de dados históricos do Copernicus Marine Service; selecionar, treinar e validar diversos modelos preditivos para tarefas de classificação e previsão temporal de clorofila-a. Todo o processo metodológico foi feito ao longo do ciclo CRISP-DM, que abrange desde a compreensão do negócio e dos dados, passando pela preparação e engenharia de características, até à modelação e avaliação, com a exceção da implementação dos resultados.

Os resultados demonstram que modelos como o Random Forest, Bagging e XGBoost obteram uma precisão preditiva excecional na classificação dos níveis de clorofila-a, tendo valores próximos de 1,00 de ROC AUC, com a indicação de altos níveis de capacidade de discriminação entre diferentes classes de intensidade de florações. As análises de importância de características indicaram que variáveis biogeoquímicas, como a produtividade primária líquida de biomassa (nppv), concentração de fitoplâncton como carbono (phyc) e profundidade da zona eufótica (zeu) possuem um peso significativamente maior do que variáveis físicas orientadas ás ondas e vento (como analysed_sst, VCMX e VHM0_WW). Relativamente a previsões temporais, o Random Forest Regressor teve bons resultados, alcançando o melhor desempenho na previsão de tendências sazonais de clorofila-a, com valores de R^2 de até 0,594 com erros muito reduzidos ($MSE \approx 0$, $RMSE \approx 0.02$ e $MAE \approx 0.01$), enquanto que modelos tradicionais como ARIMA e SARIMA, bem como redes neuronais como LSTM, GRU e CNN, apresentaram resultados insatisfatórios, com previsões excessivamente suaves ou instáveis.

As conclusões mostram claramente como a combinação de dados de diferentes fontes pode melhorar a capacidade preditiva dos modelos de cyanoHABs. Destacam também que a integração das variáveis biogeoquímicas e físicas fornece uma compreensão mais completa da dinâmica das cyanoHABs. Além disso, os resultados indicam a importância de investir em abordagens de inteligência artificial explicável (XAI), assegurando que os modelos sejam compreensíveis e utilizáveis pelas autoridades responsáveis pela qualidade da água.

Trabalhos futuros podem incluir a adição de dados meteorológicos locais, como os que estão disponíveis no portal do Instituto Português do Mar e da Atmosfera (IPMA). Além disso, outras variáveis ambientais importantes, como a turbidez e a direção do vento, podem ser consideradas para fortalecer e tornar os modelos mais práticos. Estas descobertas destacam o potencial da aplicação da aprendizagem automática com dados de diversas fontes como uma estratégia fundamental para apoiar a gestão da segurança hídrica e reduzir os impactos de florações nocivas.

Palavras-chave: Florações de algas nocivas, Aprendizagem automática, Inteligência artificial, Fusão de dados multiorigem, Previsão de floração, Segurança hídrica

Contents

List of Figures

List of Tables

List of Abbreviations

ACL3D-Pix2Pix	Attention-Controlled Lightweight 3D architecture
ADP	Adaptive Dynamic Programming
AGM	Algal Growth Model
AI	Artificial Intelligence
AMDNN	Algal Morphology Deep Neural Network
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
aOIF	Artificial Ocean Iron Fertilization
ARIMA	Autoregressive Integrated Moving Average
AUROC	Area Under Receiver Operating Characteristic
B1D-CNN	Blending 1 Dimensional - Convolutional Neural Network
CANet	Convolutional Attention Network
CMA-ES	Covariance Matrix Adaptation Evolutionary Strategies
CNN	Convolutional Neural Network
ConvLSTM	Convolutional Long Short-Term Memory
CRISP-DM	Cross-Industry Standard Process for Data Mining
CSTD	Combined States Temporal Differential
CyanoHAB	Cyanobacterial Harmful Algal Bloom
DL	Deep Learning
DNN	Deep Neural Network
DNN-SA	Deep Neural Network - Sensitivity Analysis
ECMWF	European Centre for Medium-Range Weather Forecasts
EDA	Exploratory Data Analysis
EFDC	Environmental Fluid Dynamics Code
ESN	Echo State Network
ERA5	ECMWF Reanalysis v5
FAI	Floating Algae Index
GBDT	Gradient Boosting Decision Trees
GBM	Gradient Boosting Machines
GNN	Graph Neural Network
GRU	Gated Recurrent Unit
HAB	Harmful Algal Bloom
HNLC	High-Nutrient, Low-Chlorophyll
HSSD	Hidden States Spatial Differential
HTC	Hybrid Task Cascade
IronEx	Iron Enrichment Experiment
KNN	K-Nearest Neighbors
LO	Last Observation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MARE	Mean Absolute Relative Error

MFWAM	Météo-France WAve Model
MIM	Memory In Memory
ML	Machine Learning
MNDWI	Modified Normalized Difference Water Index
MODIS	Moderate-Resolution Imaging Spectroradiometer
MSE	Mean Squared Error
NARNET	Non-linear Autoregressive Neural Network
N-BEATS	Neural Basis Expansion Analysis for Interpretable Time Series Forecasting
NDCI	Normalized Difference Chlorophyll Index
NDVI	Normalized Difference Vegetation Index
OIF	Ocean Iron Fertilization
OLCI	Ocean and Land Colour Instrument
PCA	Principal Component Analysis
Pix2Pix	Image-to-Image Translation Network
PORTIC	Porto Research, Technology & Innovation Center
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
RL	Reinforcement Learning
RMSE	Root Mean Squared Error
ROC AUC	Receiver Operating Characteristic Area Under the Curve
SABI	Surface Algal Blooms Index
SAR	Synthetic Aperture Radar
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	SARIMA with Exogenous Variables
SeaWiFS	Sea-viewing Wide Field-of-view Sensor
SEEDS	Subarctic Pacific iron Experiment for Ecosystem Dynamics Study
SHAP	SHapley Additive exPlanations
SOFeX	Southern Ocean Iron Experiment
SST	Sea Surface Temperature
SSWH	Sea Surface Wave Height
StyleGAN2 ADA	Style Generative Adversarial Network 2 with Adaptive Discriminator Augmentation
SVM	Support Vector Machine
TE	Transfer Entropy
TL	Transfer Learning
TN	Total Nitrogen
TOC	Total Organic Carbon
TP	Total Phosphorus
VCMX	Sea Surface Wave Maximum Height
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
YOLO	You Only Look Once

Chapter 1

Introduction

This chapter aims to outline the primary objective of the dissertation, which was proposed by the stakeholders from Porto Research, Technology & Innovation Center (PORTIC) emphasizing the significance of integrating multiple data sources for predicting Harmful Algal Blooms (HABs) through Machine Learning (ML). To accomplish this, it was suggested to test various ML models that utilize merged data sources. Additionally, this introductory chapter details the methodology employed for the literature review and provides an overview of the document's structure.

1.1 Contextualization

Cyanobacterial Harmful Algal Blooms (cyanoHABs) refer to instances of unusual blooms of cyanobacteria in freshwater, marine, or brackish water bodies that are usually caused by nutrient loading, temperature, light, and rest regimes (McKinley et al., (2024)). Although cyanobacteria are native and pivotal components of primary productivity, their uncontrolled growth can result in deadly toxins production, such as microcystins, saxitoxins, and cylindrospermopsins. These substances can cause adverse health effects, including death, through ingestion, seafood poisoning, or accidental exposure (Weir, Kourantidou, and Jin, (2022)). CyanoHABs lower oxygen levels and create dead zones that destabilize aquatic foodwebs. Their frequency and size are also increasing globally due to climate change, agricultural runoff, and poor treatment of wastewater and thus constituting a novel environmental and public health issue (Martino, Gianella, and Davidson, (2020)).

1.1.1 Economical Impact of CyanoHABs

Economically, cyanoHABs impact aquaculture, fisheries, tourism, and public health. In Scotland, *Dinophysis spp.* have annually reduced shellfish production by 15%, equivalent to £1.37 million, due to contaminating and triggering harvesting closure enforced by regulation or legislation (Martino, Gianella, and Davidson, (2020)). The blooms also entail costs related to monitoring and mitigation system, which further reduce industry profit margins (McKinley et al., (2024)). Communities that depend on fishing, as in like Long Beach, Washington, have lost up to $843,675 from closures triggered by blooms, according to (Weir, Kourantidou, and Jin, (2022)).

Tourism and recreation also are adversely affected by water quality since it drives away tourists and undermines local economies. in Washington, the closing of clam harvesting due to cyanoHABs results in annual income losses of approximately $40 million for tourism-related companies (Weir, Kourantidou, and Jin, (2022)). Health issues caused by toxins that are produced during cyanoHABs, such as gastrointestinal and neurological diseases, are the cause of high healthcare expenses and loss of productivity (Martino, Gianella, and Davidson, (2020)).

CyanoHABs are also supply chain interruptors and impact already vulnerable groups such as women, marginalized communities, and migrant laborers who rely on these jobs (McKinley et al., (2024)). Successful response strategies are also linked with early warning systems, such as the Pacific Northwest HAB Bulletin, and participatory public engagement in monitoring and management (Weir, Kourantidou, and Jin, (2022)).

Following up on attempts to model the complex behavior of cyanoHABs, Artificial Intelligence (AI) and ML are proving to be reliable prediction platforms. ML models use nutrients, temperature, and salinity to predict bloom occurrences with good precision. These platforms, enabled by satellites, IoT sensors, and archival information, provide low false alarm rates with real-time detection. Through early intervention, e.g., adjusting the timing of harvest or warning the public, AI-driven solutions reduce environmental harm and economic losses due to cyanoHABs.

1.1.2 Data Accessibility and Regional Focus

In spite of the increasing interest worldwide in employing ML and AI for cyanoHAB prediction, there is an impressive lack of public high-quality datasets available in Portugal exclusively dedicated to this topic. While other research investigates Portuguese water quality and bloom-condition scenarios, they utilize small lake or sporadic point samplings rather than drawing on long-term, large-scale datasets covering Portugal's extensive coast, where cyanoHAB risk and impact are particularly high. Global research efforts, however, benefits from large datasets and continuous monitoring schemes allowing the development of robust predictive models. This unawareness limits the potential for developing and training advanced ML models for Portuguese aquatic environments and shortens possibilities of explainable and transferrable AI application in national aquaculture and environmental monitoring. In broader scientific literature, three general trends can be discerned in the application of data and AI to cyanoHAB research.

First, there are numerous studies wherein ML algorithms are used to forecast water quality and algal blooms using observations from outside Portuguese regions. These studies would mainly use large-scale observation of Chinese, North American, Indian, or Northern European reservoirs and lakes where government and public funding have provided access to uninterrupted satellite imagery, in-situ data, and past records. ML algorithms like Support Vector Machines (SVM), Random Forests (RF), Artificial Neural Networks (ANN), and Long Short-Term Memory (LSTM) networks are being widely used for the classification of bloom events, chlorophyll-a prediction or eutrophication trend analysis. Systematic review, on the other hand, utilizes such reservoir water quality monitoring methodologies without focusing specifically on data from Portugal, as in Nikoo et al., (2025). Ramaraj et al., (2023) applied Deep Learning (DL) models, including Adaptive Neuro-Fuzzy Inference System (ANFIS), Long Short-Term Memory (LSTM), and Non-linear Autoregressive Neural Network (NARNET), to predict water quality at Kolavai Lake in India, using satellite data and empirical models. However, the study is geographically limited to South Asia. While such observations demonstrate the ease with which AI deployment can be executed for environmental observation, they are of limited utility towards deciphering the cyanoHAB processes in Portugal.

On the other hand, there are few articles published combining Portuguese data and ML techniques precisely for the prediction of cyanoHAB events. One study utilizes a seven-year time series with data from the Portuguese Institute for Sea and Atmosphere (IPMA) to develop a model predicting shellfish farm closure in the region of Sagres, a major aquaculture hub in Southwest Portugal. With cutting-edge ML methods such as RF, Extreme Gradient Boosting (XGBoost), Multi-Layer Perceptrons (MLPs), AutoML, and explainable AI methods such as Shapley values, the paper demonstrates the effectiveness and practicability of using AI to forecast bloom-induced closure (O'Donncha et al., (2024)). This research is unique in being the first known research to integrate directly Portuguese environmental data into AI tools with the aim of confronting cyanoHAB prediction, offering a gold standard approach for future research and supporting the distinct need for further region-specific research.

Lastly, some of the literature makes reference to other works that, although developed in cooperation with Portuguese institutions, apply their models and experiments in non-Portuguese areas. An excellent example is a study on the use of unmanned surface vehicles in observing algal bloom fronts through adaptive ML algorithms and recursive least squares fitting, with data obtained from the Baltic Sea (Fonseca et al., (2021)). Even though the study involves involvement of the University of Porto and utilization of software platforms designed in Portugal, it does not utilize data based on Portuguese waters. Instead, it focuses on improving bloom monitoring capabilities using robotics and AI, rather than predicting toxic cyanobacterial blooms. Still, it represents good progress in environmental monitoring but does not directly contribute to cyanoHABs forecasting in Portuguese marine ecosystems.

The following chapters will explore in greater detail the methodological framework adopted for this dissertation, including data collection, processing, and modeling strategies. Special focus will be placed on the role of environmental variables and the performance of ML techniques in predicting cyanoHAB occurrences.

1.2 Problem Statement

CyanoHABs are natural phenomena that are characterized by a rapid grow of cyanobacteria in water and can result in substantial economic losses, due to harm to drinking water supplies, commercial fishing, wildlife, property values, recreation, and tourism. Toxins produced by some of these events threaten human and animal health. Some of the physicochemical factors behind cyanoHABs dynamics are well known, but there are still major gaps in our understanding of the conditions that trigger and sustain them over time. In a recent study by Marrone et al., (2024), ML and AI are seen as tools with huge potential to create holistic models to understand HABs dynamics by integrating data from multiple sources and ultimately increase the security of our water sources. Even though there are already a few studies in the literature that apply ML in the context of forecasting cyanoHABs, they are mostly limited to one data source. In this context, the main objective of this dissertation is to explore the use of multiple data sources for predicting cyanoHABs using ML.

1.3 Objectives

This dissertation focuses on improving the prediction of cyanoHABs by leveraging the capabilities of ML to integrate multiple data sources effectively. Accordingly, the next sections of the dissertation are structured into the following 3 stages:

- The first stage is a detailed literature review, with the intention of studying and analyzing the state-of-the-art methods used in cyanoHABs prediction. This review aims to reveal the strengths, limitations, and gaps in the existing ML approaches to ground innovation.
- It is then followed by a close analysis of the various available datasets, considering quality and completeness and their possible utility for cyanoHABs prediction. It includes water quality measurements, climatic variables, and other environmental parameters indicating their possible integration. Issues like heterogeneity of data and temporal incompatibilities also represent serious issues.
- Finally, with them in mind, the aim of this dissertation is to choose and test several different ML models that use such disparate datasets. The models attempt to leverage the strengths of each data type to enhance predictive accuracy and reliability, while addressing the inherent challenges of data integration.

With these well-defined steps, this dissertation will allow a thorough and systematic development of a robust predictive model that advances current knowledge on cyanoHABs prediction.

1.4 Methodology

To ensure robust and transparent identification, selection, and synthesis of literature about ML based predictions of cyanoHABs, a systematic review of the current scientific literature, using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology has been performed.

According to Sohrabi et al., (2021), the framework presents the logical steps of formulating a protocol of review, systematically searching databases, selecting studies in line with predetermined eligibility criteria, assessing the quality of the included studies, and synthesizing findings in a structured manner. The PRISMA methodology, combined with Zotero [1], a free research tool, allowed to retrieve the most relevant research articles for the dissertation, to later be carefully analyzed one by one. This approach made it possible to understand the current state of the art for the prediction of cyanoHABs, data sources used and detected flaws in the research area. These factors are important for the creation and refinement of the ML model developed in the dissertation.

1.5 Ethical Considerations

The dissertation's primary aim is to advance the prediction of cyanoHABs using ML by integrating multiple data sources. Any scientific work must follow ethical standards in its purpose, methodology, and execution, so considering this, the dissertation follows the Instituto Politécnico do Porto, (n.d.) and the Ordem dos Engenheiros, (n.d.) to ensure integrity,

[1] https://www.zotero.org

transparency, and respect for intellectual property during the literature review and the predictive model development stages. For the literature review phase, all articles retrieved using the PRISMA methodology are open-access and come from public scientific repositories. At the end of the dissertation, every article referenced has its corresponding citation well defined, according to the predetermined and established standard. All software tools used in the research are free to use and require no subscription plans.

For the predictive model development stage, all datasets utilized come from publicly available databases, ensuring that the data is open-access and free to use. They purely contain information related to the area of domain of the dissertation, without any sensitive data. Publicly available code from other projects or repositories was utilized to test the predictive models, and is open-source.

Additionally, some of the diagrams were generated using *napkin.ai*, an AI-based tool used for converting conceptual ideas and text information into visual representations and graphics, which contributed to the clarity of the methodology and data processing steps. In a few segments, *humanize.ai* was employed to improve the readability and reduce AI-generated text, ensuring a more natural, human-like expression without altering the content of the text. The use of these tools is acknowledged in full transparency and aligns with the ethical standards adopted throughout this dissertation.

1.6 Structure

This document has 5 main chapters: Introduction, Related Work, Methodology, Results and Discussion, and Conclusions.

The Introduction shows the initial problem statement that mentions the complexities of cyanoHAB prediction and the necessity for an innovative ML based solution, along with the objectives proposed in order to solve the problem. This dissertation also defines the methodology used for the literature review and the ethical considerations involved. It provides context on the current economic impact of cyanoHABs, along with the available regional data.

The State-of-the-Art section describes the research methodology used, including the research questions defined, the scientific repositories assessed, the search terms defined, and the final research query used. It also presents the inclusion and exclusion criteria applied to refine the search, the number of publications extracted at each PRISMA stage, and the results of the literature review, presenting answers to each of the predefined research questions.

The Methodology section explains the rationale behind the data sources' selection, their merging into a singular dataset, and its analysis (including data cleaning, exploratory analysis, model selection and evaluation, and time series forecasting techniques).

The Results and Discussion section shows the results obtained from the integration of the multiple data sources and the subsequent analysis. It also discusses the performance of the ML model, interprets the outcomes, compares them with findings from the literature, and highlights both the contributions and limitations of the work.

Finally, the Conclusion section summarizes the main findings of the dissertation, reflects on the effectiveness of the proposed approach, and provides insights into potential future work, including improvements to the current methodology and the exploration of additional data sources or modeling techniques to further enhance cyanoHABs prediction.

Chapter 2

State of the Art

This chapter examines the current state of the art and contextualizes the subject matter addressed in this proposal. To this end it presents the research methodology used to investigate the prediction of HABs using ML, with the integration of multiple data sources. Subsequently, the results are described, followed by an in-depth look on each research question and its answers.

2.1 Research Methodology

The methodology employed in the dissertation is the PRISMA methodology for conducting a systematic review, which aims to synthetize recent advances within the field of cyanoHABs forecasting specially emphasizing the integration of multidata sources. For each approach, the applications and the data types used, the predictive models analyzed or developed, their respective predictive behavior, and identified limitations will also be discussed.

The PRISMA framework ensures that the process is methodical and transparent; thus, the findings can be well-documented, reproducible, and suitable for comparative analysis. As has been observed in Sohrabi et al., (2021), PRISMA highlights the creation of appropriately worded research questions for identifying the most essential aspects of the subject matter in question. It also allows for systematic exploration of the relationships between fundamental concepts, study focus and its broader context.

An starting list of research articles is recovered from selected scientific databases and then narrowed down by intensive selection, with tight inclusion and exclusion criteria in order to ensure the relevance of each article. After iterative screening, non-essential articles are progressively discarded until the final selection represents the most relevant contributions to the field of cyanoHAB prediction using ML.

2.1.1 Research Questions

In line with the objectives of the dissertation, different integration methods of various data sources are proposed for exploration to further improve the prediction of HABs using ML. This allows the possibility to combine different data streams to improve the accuracy and reliability of the predictive model. A set of research questions was formulated to guide the application of the PRISMA methodology in evaluating the effectiveness of each domain within the proposed model development. Table 2.1 shows the research questions.

Table 2.1: Research Questions

Identifier	Research Question
RQ1	What are the current state-of-the-art ML approaches used for predicting cyanoHABs?
RQ2	What types of data sources are typically used in the prediction of cyanoHABs?
RQ3	How can the integration of multiple data sources improve the predictive accuracy of ML models for cyanoHABs?
RQ4	What ML techniques are best suited for integrating heterogeneous data sources in the context of cyanoHABs prediction, and how can they influence the key factors to correctly identify cyanoHABs?
RQ5	How does the performance of a multi-source integrated ML model compare to its corresponding single-source models in forecasting cyanoHABs under different environmental conditions?

First, the review of the state of the art on ML approaches used for forecasting cyanoHABs, sets a background understanding of the present status of development in the use of ML models in environmental prediction. This investigation will determine how those models are used in predicting cyanoHABs and determine the evolution regarding their use to handle complex environmental factors.

Next, the advantages of data integration from multiple sources will be explained in developing better predictive accuracy, which is very much required for improving the performance of cyanoHABs models as opposed to using data from single sources. This dissertation analyzed the general type of data used in cyanoHAB prediction and highlighted the limitations of relying solely on individual data sources. Therefore, the integration of multiple data sources will be shown as necessary for overcoming such limitations.

After that, a comprehensive investigation will be done on the best-fitting ML techniques in the integration of heterogeneous sources and their roles in pinpointing important factors for predicting cyanoHABs. Moreover, the performance of multi-source integrated models versus single-source models is discussed by assessing their performances under various environmental conditions.

Finally, the challenges that affect the application of ML models for the prediction of cyanoHAB are discussed, considering opportunities and limitations when designing innovative, multi-source integrated solutions. Table 2.1 depicts the sequence of the research questions which will be answered based on the findings from the retrieved publications.

2.1.2 Scientific Repositories

The selection of scientific repositories is mandatory while conducting a systematic review. As highlighted in Sohrabi et al., (2021), the 2020 update of the PRISMA methodology places emphasis on the requirement for transparency at the first stage, known as Identification, which entails identifying each data source used to collect the records. For this review, three data repositories were selected: Web of Science, IEEEXplore, and Science Direct, as outlined in Table 2.2. The parameters for each database were carefully defined to address key differences between them, and these will be further discussed in the following section.

Table 2.2: Scientific Repositories

Identifier	Repository	URL
SR1	ACM Digital Library	`https://dl.acm.org/`
SR2	IEEE Xplore	`https://ieeexplore.ieee.org/`
SR3	ScienceDirect	`https://www.sciencedirect.com/`

2.1.3 Search Terms

Defining search terms is an important step to establish a clear connection between the project's objectives and the planned research. By formulating a combination of keywords aligned with the targeted domains, the process ensures the retrieval of articles that closely match the dissertation's focus, resulting in a more comprehensive and relevant collection of scientific contributions (Sohrabi et al. (2021)).

Table 2.3: Search Terms

Identifier	Keywords
Chemical Engineering	("Cyanobacterial Harmful Algal Blooms" OR "Harmful Algal Blooms" OR "Algal Blooms") AND ("Seaweeds" OR "macroalgae" OR "algae")
Computer Science	("Artificial Intelligence" OR "Machine Learning")

As presented in Table 2.3, two domains were identified as being the major areas to explore. The application of terms such as algal blooms or seaweeds scopes specifications to terms related to the field of Environmental Chemistry, and the keywords AI and ML guarantee that we look into knowledge on these subjects. To look into solutions for the problem of the dissertation, a combination of keywords for Chemical Engineering and Computer Science was done so that the concatenation of all terms would be included in a single string query, as portrayed in Table 2.4.

Table 2.4: Research Query

("Cyanobacterial Harmful Algal Blooms" OR "Harmful Algal Blooms" OR "Algal Blooms") AND ("Artificial Intelligence" OR "Machine Learning") AND ("Seaweeds" OR "macroalgae" OR "algae")

2.1.4 Inclusion and Exclusion Requirements

Inclusion and exclusion criteria play a critical role in determining the final set of studies to be included in the review. These defined restrictions streamline the decision-making process during the article screening phase throughout the research. Some of these criteria are even implemented as advanced search rules to filter the data effectively. Tables 2.5 and 2.6 outline, with identifiers, all the requirements used to justify why certain articles were excluded or nearly excluded, ensuring that only the most relevant studies contribute to the systematic review. This approach reflects one of the recent updates to the PRISMA checklist, as documented in Sohrabi et al., (2021).

Table 2.5: Inclusion Requirements

Identifier	**Inclusion Requirement**
IR1	The article is part of a collection of peer-reviewed publications
IR2	The article belongs to the field of Computer Science and Chemical Engineering
IR3	The article focuses on contributing with relevance to the study domains
IR4	The article describes a system, a framework, or an application scenario with both theoretical and practical knowledge bases
IR5	The article has evidence of the evaluation and validation of the proposed conclusions

Table 2.6: Exclusion Requirements

Identifier	**Exclusion Requirement**
ER1	The article is not open access
ER2	The article is not written in English
ER3	The article was published in 2021 or more recently
ER4	The article is not from a journal or a conference proceeding
ER5	The article is not focused on the use of AI or ML for the prediction of cyanoHABs or cyanobacteria
ER6	The article does not mention any type of used datasets with raw data

2.1.5 Publications Extraction

The publication extraction process adhered to the PRISMA methodology, encompassing three key phases: Identification, Screening, and Inclusion. Figure 2.1 illustrates the updated 2020 PRISMA framework, detailing the steps undertaken to evaluate and select articles for inclusion.

During the Identification phase, articles were retrieved from specified repositories using advanced search criteria, focusing on publications from 2021 to 2024, in English, and in peer-reviewed journal or conference formats. These criteria aligned with inclusion requirements IR1, IR2, and IR3, as specified in Table 2.5. Repository-specific commands facilitated the search; for instance, the "ALL" command was applied in SR1 to broaden the query scope, yielding 24 articles, while SR2 and SR3 retrieved 89 and 960 articles, respectively, without additional commands. Zotero was employed to identify and remove 15 duplicate entries, resulting in 1058 unique articles for evaluation.

The Screening phase, the most critical stage, involved three rounds of rigorous evaluation. In the initial round, abstracts were screened to rule out irrelevant articles, bringing the number down to 272. The second round consisted of reading through the introductions of the articles to check if they met the inclusion criteria, removing another 90 articles. The third round involved a detailed full-text review, using the inclusion and exclusion criteria with accuracy. This procedure resulted in deletion of articles for certain reasons: 1 for ER1, 12 for ER4, 114 for ER5, and 14 for ER6.

In the Inclusion phase, a total of 46 articles were selected as the final corpus to address the research questions. This structured and systematic process ensured that there was only a selection of the most appropriate and highest-quality articles, a secure basis upon which to conduct the systematic review.

Figure 2.2 illustrates the distribution of these articles by year and scientific repository. The most referenced year was 2023, followed by 2024 and 2022.

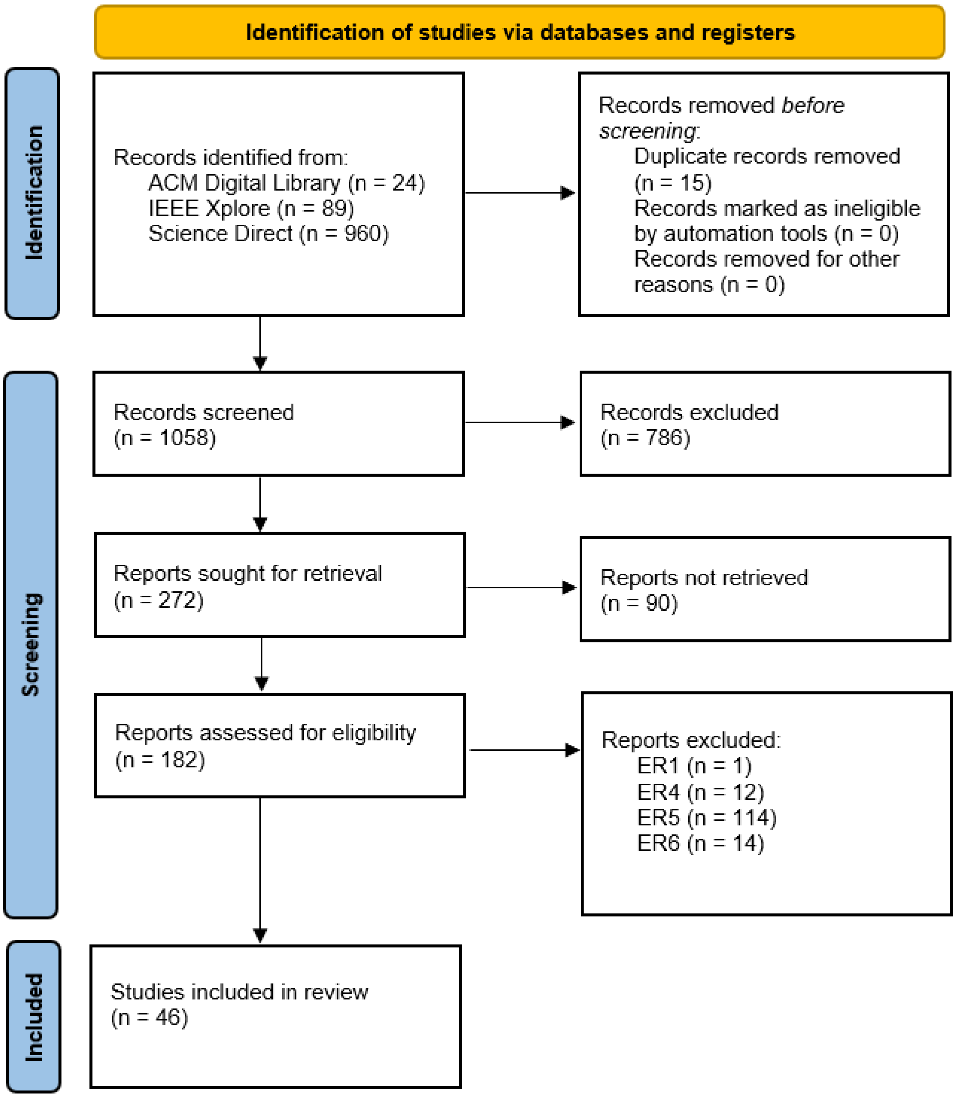

Figure 2.1: PRISMA Diagram with the Retrieved Articles, According to Sohrabi et al. (2021)

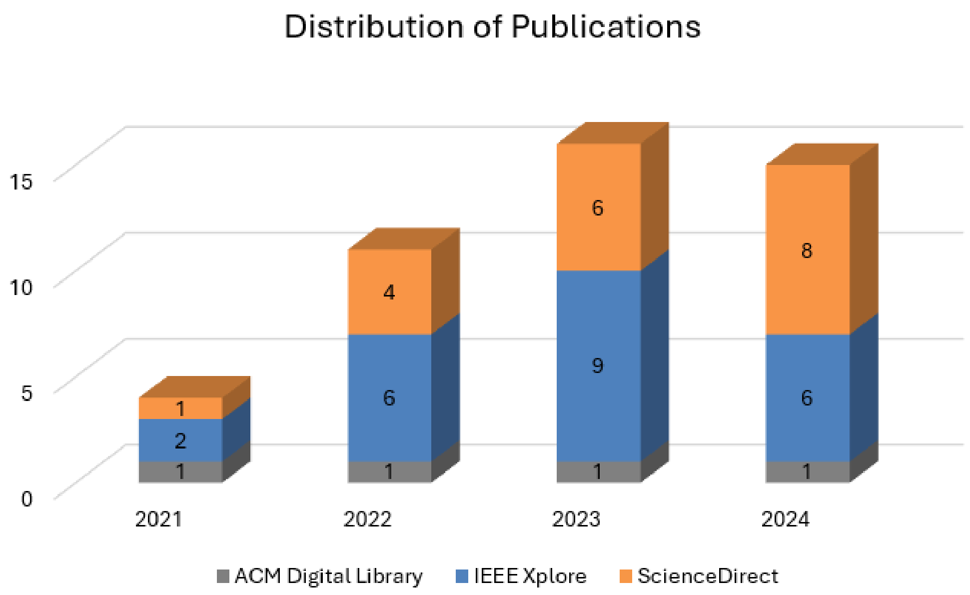

Figure 2.2: Distribution of Publications

2.2 Results

This section outlines the articles collected through the methodology described. Its purpose is to document all the retrieved publications and explain the relevance of their selection in addressing the research questions effectively.

2.2.1 Current state-of-the-art ML approaches used for predicting cyanoHABs

Most of the articles that populate the scientific repositories describe state-of-the-art ML approaches for the prediction of cyanoHABs, highlighting great enhancements in techniques suitable to deal with complex environmental systems. Yang et al., (2023) for example, identified a number of ML methods, one of which is the RF.

According to Sujatha et al., (2023), RF was highly effective in the classification of microcystin toxicity levels and prediction of cyanoHABs. This model presents higher F1-scores and Area Under Receiver Operating Characteristic (AUROC) values compared to the rest of the ML models applied for these tasks. Another study by Molares et al., (2022) presented a comparison between six ML models in managing precautionary closures of mussel farming caused by HABs. Among these, the best overall results were given by the k-Nearest Neighbor (KNN).

Lanuza et al., (2024) applied SVM for the classification of algal bloom severity into three levels: low, moderate, and high. In this paper, SVM was chosen because it is reliable and highly accurate for handling multi-class classification tasks.

The LSTM algorithm is considered among the most advanced methods of modeling temporal sequences, therefore treating nonlinear dynamics within the prediction of cyanoHABs. In fact, the LSTM model outperforms state-of-the-art models in time-series data prediction such as SVM and ANN, for its dynamic and nonlinear behavior. Indeed, this was proven by

Rostam et al., (2021). A study by Luo et al., (2024) introduced the HabLSTM model, a new LSTM-type model for spatiotemporal nonstationary HABs forecasting. The model has two blocks: the Hidden States Spatial Differential (HSSD) block for identifying short-term spatiotemporal patterns and the Combined States Temporal Differential (CSTD) block for identifying long-term spatiotemporal patterns. HabLSTM outperformed existing methods like Convolutional Long Short-Term Memory (ConvLSTM), Memory In Memory (MIM), and Predictive RNN (PredRNN), especially in capturing nonstationary dynamics. Few-shot learning and Transfer Learning (TL) were mentioned by Wang et al., (2024) and Molares et al., (2023), and applied for the classification of algae species, especially in cases with limited data. The framework Algal Morphology Deep Neural Network (AMDNN) with TL outperformed the others in accuracy.

Convolutional Neural Networks (CNNs) are dominant in recent works. In a contribution by Yang et al., (2021) CNNs and TL were used to detect the harmful algae species and chlorophyll-a concentration. Further, Blending 1 Dimensional - Convolutional Neural Network (B1D-CNN) was suggested for retrieving chlorophyll-a from Sentinel-2 satellite data in Salah et al., (2023). Besides, this study dealt with identifying cyanobacteria for HABs research by using the You Only Look Once (YOLO) framework to propose an improved version of CNN called YOLOv5, mentioned in Li et al., (2023), allowing real-time recognition of HAB-producing cyanobacteria taxa. Chen et al., (2022) presented a new technique called Convolutional Attention Network (CANet), to apply for multivariate time series forecast of cyanobacteria bloom combining CNN and the mechanism of attention. Particularly, the model was effective in applying global and local feature extraction, being remarkably important in order to capture dependency and the relationship between the time series variables. Fang et al., (2022) proposed Cascade R-CNN and Hybrid Task Cascade (HTC) models with other best-performing models. Inokuchi et al., (2024) utilized deep CNNs VGG11 and ResNet to identify red tides from GCOM-C satellite imagery. ResNet was as high as 89% accurate compared to VGG11 since it possesses a residual learning process. This is evidence of the effectiveness of CNNs in extracting spatial information from remote sensing data for cyanoHAB prediction.

Reinforcement Learning (RL) was integrated with models such as Environmental Fluid Dynamics Code (EFDC) for the autonomous calibration of water quality parameters in view of chlorophyll-a predictions, being one of the most important indicators of cyanoHABs according to Hong et al., (2023).

The architectures like Neural Basis Expansion Analysis for Interpretable Time Series Forecasting (N-BEATS), CNN, and LSTM ware some of the DL used in multi-horizon forecasting of chlorophyll-a levels and cyanoHAB prediction mentioned by Martin-Suazo et al., (2024) and Aranay et al., (2022).

Penne et al., (2023) presented Independent Component Analysis (ICA) as a supervised-independent model to detect algal blooms through hyperspectral imaging. Unlike supervised models, ICA untangles the mixed signals into statistically independent components with no requirement for labeled ground truth, making it extremely well-suited to dynamic water environments with minimal available ground truth information.

Other hybrid and advanced models were also mentioned in the literature by Ramgirkar et al., (2024), Wen et al., (2024), and Wang et al., (2023), and include the integration of multiple models, such as Gradient Boosting Decision Trees (GBDT), LSTM, and Transfer Entropy (TE) network inference combined with Graph Neural Networks (GNN), enhancing model performance by capturing spatiotemporal interactions and improving accuracy in predicting cyanoHABs. Zhu et al., (2024) applied sophisticated explainable ML models (i.e., XGBoost, RF, and GBDT) in predicting such parameters of the main eutrophication drivers as dissolved inorganic nitrogen (DIN), soluble reactive phosphorus (SRP), and chemical oxygen demand (COD). XGBoost achieved high prediction R^2 scores of 0.92 for SRP, and when used with SHAP (SHapley Additive exPlanations) values, the interpretability of feature importance was unlocked, with demonstration of improvements in prediction and explainable tools within the sector.

2.2.2 Types of data sources used in the prediction of cyanoHABs

The types of data sources employed in cyanoHABs prediction are quite vast in recent projects, emphasizing the importance of satellite imagery, in-situ measurements, and climate data.

Satellite data, such as that derived from Sentinel-2, provide multispectral data. Challenges include environmental interference like clouds, high concentrations of suspended particulate matter, and mixed pixels (e.g., algae-water or algae-cloud mixtures). These limit the standalone accuracy of index-based methods, such as the Floating Algae Index (FAI), that are sensitive to environmental conditions and very often misclassify algal blooms of low concentration, according to Yang et al., (2023).

Historical water quality measurements used by Rostam et al., (2021) analyzed the level of chlorophyll-a, dissolved oxygen, and nutrients. The complex interaction of physical, chemical, and biological processes cannot be described by a single quantification. Temporal and spatial outliers also occurred in the process, which further complicate generalization.

Wang et al., (2024) worked with microscopic algae images from local ecosystems and external datasets (e.g., Hong Kong IFCB dataset, Macau freshwater), but these small local datasets were insufficient for training DL models, while variability in algae morphology and environmental conditions can limit model generalization. The Moderate-Resolution Imaging Spectroradiometer (MODIS) aqua satellite data, includes measurements of chlorophyll-a, particulate organic carbon, nitrate, phosphate, and water temperature, and its use is present in Karmakar et al., (2024) and Srilatha et al., (2023). Its deficiencies arised due to satellite data being susceptible to atmospheric disturbance, cloud cover, and low spatial resolution, and in-situ measurements for validation are costly and time-consuming. The CNN model used in Yang et al., (2021) had used algae images of toxic and non-toxic species of cyanoHABs, including data obtained from research centers and hobbyists. But the single CNN model struggled with misclassifications due to morphological similarities between species. Multiple models may be required for reliable identification.

Data from pontoon monitoring systems for in-situ measurements of dissolved oxygen, chlorophyll-a, water temperature, and hyperspectral images from drones were used in Hong et al., (2023). A disadvantage of this dataset is that point-specific data may not represent spatial variability, and hyperspectral images can be noisy and limited to specific survey times.

The ABF model combines MODIS satellite data and a range of spectral indices like Normalized Difference Vegetation Index (NDVI), Floating Algal Index (FAI), Modified Normalized Difference Water Index (MNDWI) and Surface Algal Blooms Index (SABI), with environmental data like temperature and rainfall derived from Google Earth Engine. All of the data sources also possess constraints like interference by cloud, non-uniform temporal resolution, and inability to create early warning in the absence of predictive factors according to Ananias et al., (2023).

CYGNSS GNSS-R data has high temporal resolution (approximately 4 hours) and cloud-penetrating capability, which overcomes the limitations of optical sensors. However, its low spatial resolution (approximately 25 km) and lack of inherent spectral content require bloom indicators like chlorophyll-a to be modeled indirectly, lowering its independent accuracy said by Jia et al., (2024).

Other data sources used are for example high-resolution water quality parameters from in-situ sensors or historical data on chlorophyll-a concentrations mentioned in Zhang et al., (2021), but the limited data availability or imbalanced datasets can make predictions unreliable, especially when there are fewer samples of high-toxicity levels.

2.2.3 How the integration of multiple data sources improves the predictive accuracy of ML models for cyanoHABs

It becomes important to integrate several data sources from the available base of knowledge in order to improve the predictive accuracy of ML models in cyanoHABs prediction. The combination of heterogeneous sources of data such as satellite imagery, in-situ measurements, and climate variables helps ML models acquire fine-grained and dynamic patterns that control bloom occurrences.

For instance, the integration of Sentinel-2 with other spectral indices such as FAI enhances environmental issues such as cloud cover and suspended particulate matter and provides higher reliability of index-based methods by a better model performance (Yang et al., (2023)).

Algal bloom forecasting was improved through multivariate time-series data capturing temporal and environmental fluctuations (Zhang et al., (2022)). MODIS satellite imagery, along with in-situ data and strong ML methods, offers a more trustworthy and firm means for forecasting biogeochemical dynamics (Karmakar et al., (2024)). Integration with Sentinel-2 imagery and global in-situ data assisted in attaining superior forecast through higher spatial and temporal diversity, particularly under dynamic water conditions (Salah et al., (2023)). Indeed, using datasets from several lakes and integrating nutrient data with other environmental parameters increased the generalizability and robustness of cyanoHAB prediction models, as seen in the work of Sujatha et al., (2023).

Combining MODIS with Synthetic Aperture Radar (SAR) data enhanced model robustness under conditions such as cloudy environment conditions and improves the overall general prediction accuracy (Lyu et al., (2023)). By incorporating multivariate water quality parameters such as pH, temperature, and turbidity into their data preprocessing technique, such as normalization, they can attain improved performances by making the data treated consistent (Wu et al., (2022)). By integrating the ground measurements with the Sentinel-2 space-based data, models utilized the strength of high-resolution spatial information coupled with calibrated in-situ data in making their predictions (Khan et al., (2022)).

Combining MODIS-extracted spectral indices with meteorological data allowed Du et al., (2024) to make accurate, pixel-scale predictions across multiple spatial scales. The model can provide early warning of up to seven days, replicating intricate bloom patterns in large lakes such as Taihu and Chaohu.

By incorporating realistic Style Generative Adversarial Network 2 with Adaptive Discriminator Augmentation (StyleGAN2-ADA)-generated algae patterns and neural style transfer in real lake/reservoir images, the model upgraded 24–50% IoU from training sets of actual images. This is a confirmation of the demand for mixed, combined image data to extend the detection of blooms to real scene scenarios (Barrientos et al., (2023)). Incorporation of OLCI (Ocean and Land Colour Instrument) reflectance into in situ phycocyanin values and auxiliary spectral indices (e.g., Normalized Difference Chlorophyll Index (NDCI)) led to enhanced and stable model performance. The addition helped in the better discrimination of bloom susceptibility conditions at different trophic states (Pacilio et al., (2024)).

Similarly, other various datasets from different locations increased the generalization of the model by including environmental variability, hence increasing the precision of the prediction of the model in Tiwary et al., (2023).

2.2.4 ML techniques best suited for integrating heterogeneous data sources in the context of cyanoHABs prediction, and how they influence the key factors to correctly identify cyanoHABs

All articles evaluated effective ML techniques to integrate heterogeneous data sources in the prediction of cyanoHABs, but some of those techniques outperformed others and stand out.

Hybrid and ensemble approaches are more predominant in the findings due to their flexibility and capacity to handle multiple data modalities. These advanced ML techniques will resolve the challenges of heterogeneous data source integration and improve the identification of key drivers of cyanoHAB events. Similarly, the application of such methods will move toward more accurate, actionable predictions based on the complementary strengths of multiple datasets in support of proactive HABs monitoring and management.

N-BEATS supported univariate and multivariate inputs, decomposing the data into trend and seasonality components, which enhances interpretability and precision. It separated trend and seasonality to enable concentration on long-term patterns of cyanoHABs and improve accuracy in the forecast, according to Martin-Suazo et al., (2024).

Deep active learning enables selecting informative samples during the training process, whereas the genetic algorithm determines the most relevant predictors out of a large set of features. The algorithms improved the model's performance by ensuring only influential climatic and environmental variables like wind speed and average temperature, respectively, will be used, according to Aranay et al., (2022).

Wen et al., (2022) used Principal Component Analysis (PCA) for dimensionality reduction and Autoregressive Integrated Moving Average (ARIMA) for modeling periodicity in environmental variables with respect to the prediction of algal blooms. These approaches therefore captured significant drivers like phosphate and nitrate and reduced noise, hence improving the predictiveness. RF had better performance in handling heterogeneous data, ranking the predictor's importance.

Deep Neural Networks (DNN) offer flexible modeling in complex relationships. An example is the interaction between land use and meteorological conditions. The most important predictors in this relation involve sample location, shore samples, woody wetlands, wind speed, and drought conditions. These features were the major determinants of the specificity and accuracy of the predicted presence of HABs (Choi et al., (2021)).

DNN-based Sensitivity Analysis (DNN-SA) identifies the most informative water quality parameters, improving model performance and interpretability. Hyperparameter tuning was applied to DNN, Gradient Boosting Machines (GBM), and RF models in order to improve learning and reduce errors by Rahaman et al., (2024).

ACL3D-Pix2Pix (Attention-Controlled Lightweight 3D architecture - Image-to-Image Translation Network) demonstrated the strength of combining GANs, ConvLSTM layers, attention, and residual networks. Each module addresses different aspects of multi-source fusion, from temporal consistency to spatial feature saliency and gradient continuity, making the model very competent at capturing subtle bloom variations (Wang et al., (2023)).

Attention-augmented DNNs with custom spectral feature engineering (e.g., band ratios) and Huber loss achieved success in the prediction of cyanoHABs in optically complex lakes. The models were robust to varying chlorophyll-a concentrations and offered stable performance under temporal and spatial conditions (Grendaite et al., (2024)).

The research of Abdullah et al., (2022) illustrates how TL can allow pretrained CNNs to generalize to new datasets, and GANs to imitate underrepresented algae classes. This setup is especially suitable to scenarios in which diverse sources of data must be combined for improving the robustness of classification, e.g., where field sampling is sparse or unbalanced.

2.2.5 Performance of a multi-source integrated ML model compared with single-source models in forecasting cyanoHABs under different environmental conditions

These multimodal models combine an enormous array of diverse data, from satellite imagery, in-situ water quality measurements, to climate data, thereby giving a wide view into the environmental dynamics that shape cyanoHABs.

The proposed Time-varying Adversarial Generative Model, integrated with an Echo State Network (ESN) and Adaptive Dynamic Programming (ADP), outperformed the traditional models such as LSTM, Covariance Matrix Adaptation Evolutionary Strategies (CMA-ES) and Algal Growth Model (AGM) in performance, mentioned by Zhang et al., (2022). The multi-source model with an ensemble of MODIS data, GBDT, and LSTM performed better than a single-source model, whereas LSTM performed better for the prediction in temporal trends, according to Srilatha et al., (2023).

Self-Sampling and Semicorrelated Co-Training outperformed traditional and deep models such as RF, SVM, and U-Net, with an accuracy of 90.52% and great robustness against environmental factors such as cloud cover seen in Lyu et al., (2023).

ANN produced an accuracy of 87.3%, GBDT achieved an accuracy of 86.8%, and the SVM achieved an accuracy of 83.9% with feature selection. This combination yielded the least Mean Squared Error (MSE) and highest R^2 scores among its comparators across both datasets. The integration of the many datasets allowed GBDT to achieve an R^2 of 0.972 in the Sassafras dataset and 0.728 in Lake Okeechobee, giving the model great generalizability and reliability, as seen in Tiwary et al., (2023).

Recurrent models like LSTM and Gated Recurrent Unit (GRU) outperformed CNNs and RF, particularly for extreme values of chlorophyll-a that may be indicative of eutrophication, according to Aslan et al., (2022).

For multi-day forecasting, the multisource ML models, GRU and Multi-Layer Perceptron (MLP), outperformed the single-source baselines. Xia et al., (2024) realized that GRU performed the best for 5-day and 7-day ahead predictions with better Nash-Sutcliffe efficiency, Mean Absolute Error (MAE), and Mean Absolute Relative Error (MARE) values compared to other, simpler baselines like ARIMA and Last Observation (LO).

As compared to single-source, thresholding-based approaches, the ABD model realized over 20% accuracy and F1-score improvement. Atmospheric interference as well as blooming pattern fine detection invariance characterize multi-source integration of ML superiority, according to Ananias et al., (2023).

The multi-source model presented by Balcacer et al., (2023) performed better than visual-only models (e.g., YOLO), at 79% accuracy using the combination of drone, field, and contextual information. Visual-only models alone would classify algae as objects like boats or umbrellas, illustrating the strength of the multi-source, transformer-based approach.

Locally trained models (SVR, LMDN) compared equally to semi-empirical models and global ML approaches and decreased errors by 20% and bias by 50%. Global models overestimated chlorophyll-a, especially in less eutrophic waters. The results from Chegoonian et al., (2022) support localized, multi-source models as an improved approach to accurate, real-time detection of cyanoHAB.

Chapter 3

Methodology

Understanding and predicting toxic cyanobacteria-dominated HABs necessitates high resolution, multivariate environmental data sets that cover a wide range of spatial and temporal measurements. Due to the multifaceted interaction of physical, chemical, and biological oceanographic processes, the first step in this research was a comprehensive literature search for publicly available data sets that are able to include these dynamics. In general, the aim was to locate local datasets with data from as close as the Portuguese coast since this region is typically poorly represented in the literature. No available Portuguese datasets already contain the necessary environmental and biological parameters, such as chlorophyll-a content, Sea Surface Temperature (SST), and Sea Surface Wave Height (SSWH), at the temporal frequency necessary for successful HAB forecasting, therefore this dissertation relied on Iberian Copernicus Marine Service data, high-quality freely accessible oceanographic data having sufficient spatial and temporal resolution. This chapter outlines the methodology steps undertaken to collect, process, and analyze this data, which constitute the foundation for constructing predictive models capable of detecting both seasonal cycles as well as instantaneous bloom events in this geographically limited environment.

3.1 Methodology Summary

The research methodology employed in this dissertation follows the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach, which structures the process of data mining into six phases that are iterative, but distinct and related: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment. All of them were customized towards the predictive goal of forecasting cyanoHABs in Portugal's marine ecosystem, more precisely the Atlantic coast.

As illustrated in Figure 3.1, the CRISP-DM framework provided a structured foundation to guide the entire workflow, from data acquisition to model evaluation.

Business Understanding was committed to defining the overall goal of the research: testing predictive models that are able to predict cyanoHAB events based on environmental variables with appropriate spatial and temporal resolution. The lack of national datasets for the most important biological and physical variables made it necessary to select to work with freely available Copernicus Marine Service data.

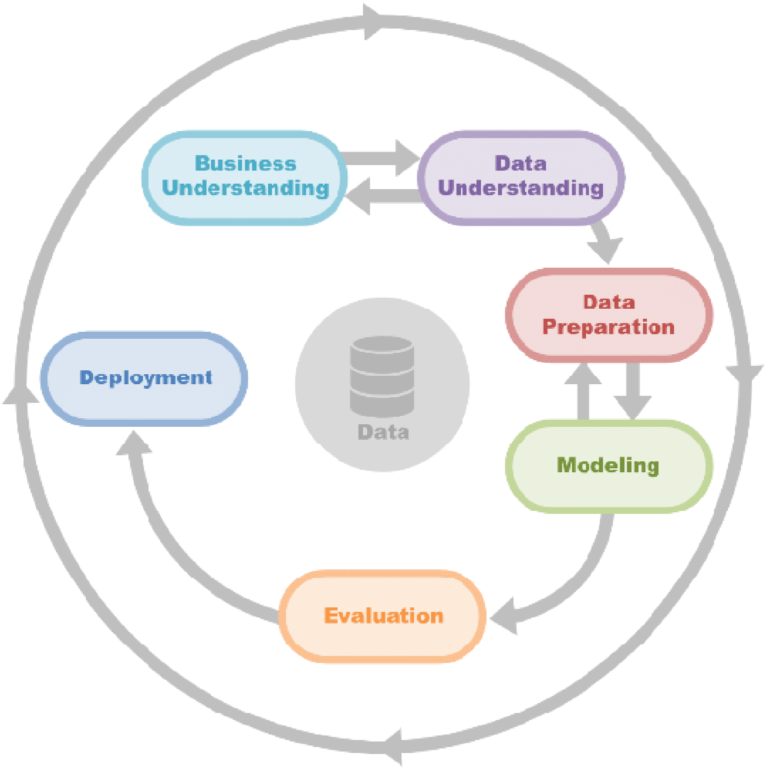

Figure 3.1: CRISP-DM Framework, According to Schröer et al., (2021)

Data Understanding involved identifying which data sets would be capable of capturing the dynamics of bloom formation. Three data sets were selected: one for biogeochemistry (nutrients and chlorophyll-a) and one for thermal patterns (sea surface overall temperature), and one for oceanographic conditions (wave height and direction). These data sets were checked for homogeneous temporal coverage.

Data Preparation was the most extensive stage, wherein thousands of daily '.nc' files for years were fetched. Using Python libraries such as 'xarray' and 'pandas', the daily files were filtered, processed, and grouped to form monthly and yearly data sets. These were manually joined and merged into three general domain-specific data sets. Temporal synchronization was performed in all three domains before they were merged into a common dataset of over two million records.

Modeling was done by applying machine learning algorithms to the final dataset to predict concentrations of chlorophyll-a and identify bloom-prone conditions. Several models were employed for comparing accuracy and ability to handle seasonality and short-term fluctuation in bloom behavior.

Evaluation was conducted via classification metrics such as precision, recall, F1-score, and ROC AUC to measure model performance. Temporal stratification was also considered so that models would not be skewed by specific seasons or anomalies within the dataset.

Finally, though Deployment was not carried out as part of the dissertation, all the workflows and scripts were made reproducible and deployable with ease into real-time systems, with a clear foundation for future operational integration into cyanoHAB early warning systems.

3.2 Data Collection and Preparation

The data used in this dissertation was collected from the Copernicus Marine Service, which provides high-quality, periodically updated environmental data for ocean monitoring and analysis.

The article by Yoon et al., (2018) presents the important role played by satellite data in ocean iron fertilization (OIF) experiments with special focus on observation and interpretation of biological response over high-nutrient, low-chlorophyll (HNLC) waters. Ocean-color satellite records, such as Sea-viewing Wide Field-of-view Sensor (SeaWiFS), have been used to track chlorophyll-a concentration response to fertilization events with extended bloom duration and scales not seen by shipboard measurements alone, e.g., a 150 km ribbon, forming bloom in the course of the SOIREE experiment. Satellite observations also enabled patches to be followed prior to experiments in such a manner as to enable investigators to detect and follow natural chlorophyll-a hotspots where fertilization would be applied. Satellite altimetry was also employed to investigate mesoscale ocean processes and surface fronts, like the Antarctic Polar Front, that control bloom expansion. Satellite retrospective analysis also provided the opportunity for large-scale, long-term assessment of effects of artificial and natural fertilizations. All these uses demonstrate the vitally important function of satellite remote sensing in guiding, testing, and scaling OIF research.

Three datasets were used in this dissertation, each relating to a specific physical or biogeochemical oceanographic domain:

- The Atlantic - Iberian Biscay Irish - Ocean Wave Reanalysis dataset provides hourly high-resolution wave parameters including wave height, period, direction, and wind wave/swell properties from the Météo-France WAve Model (MFWAM) model and European Centre for Medium-Range Weather Forecasts (ECMWF) Reanalysis v5 (ERA5) reanalysis wind forcing. The dataset is important to characterize ocean surface dynamics within the study region.
- The European North West Shelf / Iberia Biscay Irish Seas – High Resolution ODYSSEA L4 SST Analysis delivers daily gap-free SST maps at 0.02° resolution based on satellite observations made by both infrared and microwave radiometers. SST is a key element in the study of ocean-atmosphere interaction and seasonal variability.
- Finally, the Atlantic - Iberian Biscay Irish - Ocean Biogeochemistry NON-ASSIMILATIVE Hindcast dataset, with the PISCES biogeochemical model, provides necessary variables such as chlorophyll-a, oxygen, nitrate, and primary productivity at daily, monthly, and yearly resolution, making ecological and biological assessments.

Figure 3.2 is a diagrammatic representation of the integration of the datasets. In it, the three separate datasets are each processed separately and then merged into a single integrated dataset. This integration was necessary in order to obtain temporal and spatial consistency across physical and biogeochemical domains for harmonized analysis and modeling workflows in the subsequent stages of the dissertation.

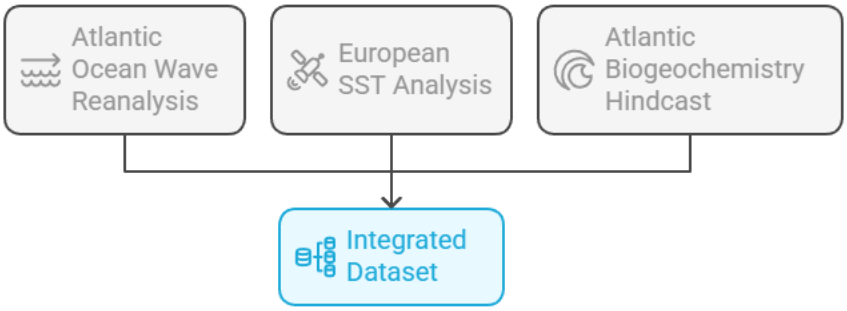

Figure 3.2: Datasets Merge View

3.2.1 Sea Surface Wave Height Dataset

The wave data set includes a good variety of parameters describing the physical motion of ocean surface waves, either wind-generated or swell-generated. Wave parameters such as direction, period, and significant height are recorded for different kinds of waves: primary swell, secondary swell, and wind waves, whose frequency is given in Table 3.1:

Table 3.1: Selected Wave Parameters from the Copernicus Ocean Wave Dataset

Parameter	**Code [Unit]**
Sea surface primary swell wave from direction	VMDR_SW1 [°]
Sea surface primary swell wave mean period	VTM01_SW1 [s]
Sea surface primary swell wave significant height	VHM0_SW1 [m]
Sea surface secondary swell wave from direction	VMDR_SW2 [°]
Sea surface secondary swell wave mean period	VTM01_SW2 [s]
Sea surface secondary swell wave significant height	VHM0_SW2 [m]
Sea surface wave from direction	VMDR [°]
Sea surface wave from direction at variance spectral density maximum	VPED [°]
Sea surface wave maximum crest height	VMXL [m]
Sea surface wave maximum height	VCMX [m]
Sea surface wave mean period (inverse frequency moment)	VTM10 [s]
Sea surface wave mean period (second frequency moment)	VTM02 [s]
Sea surface wave period at spectral density maximum	VTPK [s]
Sea surface wave significant height	VHM0 [m]
Sea surface wave stokes drift x velocity	VSDX [m/s]
Sea surface wave stokes drift y velocity	VSDY [m/s]
Sea surface wind wave from direction	VMDR_WW [°]
Sea surface wind wave mean period	VTM01_WW [s]
Sea surface wind wave significant height	VHM0_WW [m]

Some other properties are crest height maximum, wave drift velocity of stokes, and spectral density-derived values, which play crucial roles in determining energy propagation and sea state conditions. These properties are significant to depict surface turbulence and transport processes, which may directly influence surface phytoplankton distribution and concentration, such as cyanobacteria.

```
time,longitude,latitude,VHM0,VTM02,VTM10,VMDR,VPED,VTPK,VCMX,VMXL,VHM0_WW,VHM0_SW1,VHM0_SW2,VMDR_WW,VMDR_SW1,VMDR_SW2,VTM01_WW,VTM01_SW1,VTM01_SW2,VSDX,VSDY
2018-10-01,-19.0,26.0,1.27,6.56,9.429999,341.01,335.0,12.33,2.37,1.4499999,0.049999997,0.72999996,0.64,21.309998,340.33002,297.19,1.4599999,11.55,9.94,-0.01,-0.02
2018-10-01,-19.0,26.027779,1.27,6.5699997,9.44,340.72998,335.0,12.32,2.36,1.4499999,0.03,0.72999996,0.64,21.089996,340.47,297.12,1.4,11.53,9.94,-0.01,-0.02
2018-10-01,-19.0,26.055557,1.26,6.58,9.45,340.44,335.0,12.309999,2.36,1.4499999,0.03,0.72999996,0.64,21.080002,340.47998,297.07,1.4,11.53,9.929999,-0.01,-0.02
2018-10-01,-19.0,26.083336,1.26,6.5899997,9.46,340.16,335.0,12.3,2.36,1.4399999,0.03,0.74,0.64,21.059998,341.19,296.99,1.4,11.4,9.929999,-0.01,-0.02
2018-10-01,-19.0,26.111115,1.26,6.6,9.469999,339.88,335.0,12.3,2.35,1.4399999,0.03,0.72999996,0.65,21.410004,341.21,296.91,1.4,11.4,9.92,-0.01,-0.02
2018-10-01,-19.0,26.138893,1.25,6.6099997,9.49,339.33002,335.0,12.28,2.34,1.4399999,0.02,0.74,0.65,20.230011,341.47998,296.76,1.4,11.37,9.92,-0.01,-0.02
2018-10-01,-19.0,26.166672,1.25,6.62,9.5,339.07,335.0,12.2699995,2.34,1.43,0.0,0.72999996,0.65,20.830002,341.51,296.68,0.0,11.36,9.92,-0.01,-0.02
2018-10-01,-19.0,26.19445,1.25,6.6299996,9.51,338.81,335.0,12.259999,2.34,1.43,0.0,0.72999996,0.65,20.87001,341.55,296.6,0.0,11.349999,9.92,-0.01,-0.02
2018-10-01,-19.0,26.222229,1.25,6.64,9.5199995,338.56,335.0,12.25,2.33,1.43,0.0,0.72999996,0.65,20.910004,341.59,296.53,0.0,11.349999,9.92,-0.01,-0.02
```

Figure 3.3: Copernicus Ocean Wave Data Sample

Figure 3.3 shows a sample from the Atlantic - Iberian Biscay Irish - Ocean Wave Reanalysis dataset.

3.2.2 Sea Surface Temperature Dataset

The SST data set, shown in Table 3.2, provides a focused collection of variables describing surface temperature conditions, ice presence, and associated metadata:

Table 3.2: Selected SST Parameters from the Copernicus SST Dataset

Parameter	**Code [Unit]**
Estimated error standard deviation of analysed SST	`analysis_error` [K]
Land sea ice lake bit mask	`mask`
Sea ice area fraction	`sea_ice_fraction`
Sea surface foundation temperature	`analysed_sst` [K]

The sea surface base temperature (analysed_sst) is a primary input for the evaluation of the thermal regime that supports or constrains the growth of algae. Other variables such as the SST analysis standard deviation (analysis_error) indicate data quality and uncertainty. The existence of a land/sea/ice/lake mask helps guarantee spatial filtering during preprocessing, and sea ice fraction gives seasonal context that can indirectly influence bloom patterns. Such temperature-based variables are crucial to understanding both long-term seasonal cycles and short-term thermal anomalies that can initiate cyanoHAB development.

```
time,lat,lon,analysed_sst,analysis_error,mask,sea_ice_fraction
2018-10-01,9.01,-20.99,301.54,0.13,1.0,0.01
2018-10-01,9.01,-20.97,301.54,0.13,1.0,0.01
2018-10-01,9.01,-20.95,301.56,0.13,1.0,0.01
2018-10-01,9.01,-20.93,301.57,0.13,1.0,0.01
2018-10-01,9.01,-20.91,301.56,0.13,1.0,0.01
2018-10-01,9.01,-20.89,301.56,0.13,1.0,0.01
2018-10-01,9.01,-20.87,301.54999999999995,0.13,1.0,0.01
2018-10-01,9.01,-20.85,301.53,0.13,1.0,0.01
2018-10-01,9.01,-20.83,301.51,0.13,1.0,0.01
```

Figure 3.4: Copernicus Sea Surface Temperature Data Sample

Figure 3.4 shows a sample from the European North West Shelf / Iberia Biscay Irish Seas – High Resolution ODYSSEA L4 SST Analysis dataset.

3.2.3 Biogeochemical Dataset

Table 3.3 presents the selected biogeochemical parameters based on the PISCES model, a critical factor in detailing marine ecosystems' ecological and nutrient status:

Table 3.3: Selected Biogeochemical Parameters from the Copernicus Ocean Biogeochemistry Dataset

Parameter	**Code [Unit]**
Euphotic zone depth	`zeu` [m]
Mass concentration of chlorophyll-a in sea water	`chl` [mg/m^3]
Mole concentration of ammonium in sea water	`nh4` [mmol/m^3]
Mole concentration of dissolved inorganic carbon in sea water	`dissic` [mol/m^3]
Mole concentration of dissolved iron in sea water	`fe` [mmol/m^3]
Mole concentration of dissolved molecular oxygen in sea water	`o2` [mmol/m^3]
Mole concentration of nitrate in sea water	`no3` [mmol/m^3]
Mole concentration of phosphate in sea water	`po4` [mmol/m^3]
Mole concentration of phytoplankton as carbon in sea water	`phyc` [mmol/m^3]
Mole concentration of silicate in sea water	`si` [mmol/m^3]
Net primary production of biomass as carbon per unit volume	`nppv` [mg/m^3/day]
Sea water pH reported on total scale	`ph`
Surface partial pressure of carbon dioxide in sea water	`spco2` [Pa]

Parameters encompass major nutrient concentrations such as nitrate, ammonium, phosphate, silicate, and iron, all of which impact algal production. Concentration of chlorophyll-a is used as a surrogate for phytoplankton Knbmass, while net primary production (NPPV) and phytoplankton carbon content (PHYC) are other indices of biological activity in the photic zone. Supporting parameters such as pH, oxygen concentration, and partial pressure of carbon dioxide supply information on chemical status of the ocean and potential bloom cause or inhibitor. These traits are highly useful for the simulation of cyanobacterial bloom phenomena because they retain both biological potential and constraint.

```
1  time,longitude,latitude,depth,po4,phyc,si,nh4,chl,fe,no3,nppv,zeu,o2,spco2,dissic,ph
2  2018-10-01 12:00:00,-19.0,26.166666,0.50576,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,214.0,43.98001,2.1600003,8.0156
3  2018-10-01 12:00:00,-19.0,26.166666,1.5558553,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,214.0,43.98001,2.1600003,8.0116005
4  2018-10-01 12:00:00,-19.0,26.166666,2.6676817,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,214.0,43.98001,2.1600003,8.013599
5  2018-10-01 12:00:00,-19.0,26.166666,3.8562799,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,214.0,43.98001,2.1600003,8.0116005
6  2018-10-01 12:00:00,-19.0,26.166666,5.1403613,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,214.0,43.98001,2.1600003,8.0116005
7  2018-10-01 12:00:00,-19.0,26.166666,6.5430336,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,214.0,43.98001,2.1600003,8.0116005
8  2018-10-01 12:00:00,-19.0,26.166666,8.092519,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,215.0,43.98001,2.1600003,8.0156
9  2018-10-01 12:00:00,-19.0,26.166666,9.82275,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,215.0,43.98001,2.1600003,8.0156
10 2018-10-01 12:00:00,-19.0,26.166666,11.77368,0.060001373,1.1999817,1.1,0.0,0.099999994,0.0008,0.0,1.0,100.6,215.0,43.98001,2.1600003,8.0116005
```

Figure 3.5: Copernicus Ocean Biogeochemistry Data Sample

Figure 3.5 shows a sample from the Atlantic - Iberian Biscay Irish - Ocean Biogeochemistry NON-ASSIMILATIVE Hindcast dataset.

3.2.4 Final Dataset

Each of the individual datasets has huge amounts of up to +20 Megabytes of data for a singular day in a year, so the goal was to retrieve all this daily information from each dataset with a threshold in mind (1000 lines from each singular day), so that the final merged dataset is not too extensive and resource consuming.

It's important to mention that some datasets have information organized by month, merged by year, but each month only has data from a specific day of that month. This detail is what led to the recovery of daily data.

Figure 3.6 presents the data recovery approach employed by this research. The process ends with the concatenation of three independent datasets, one for each unique data domain (wave reanalysis, SST, and biogeochemistry). The datasets are concatenated to produce the `final_dataset`, comprising data from 2018 to 2023.

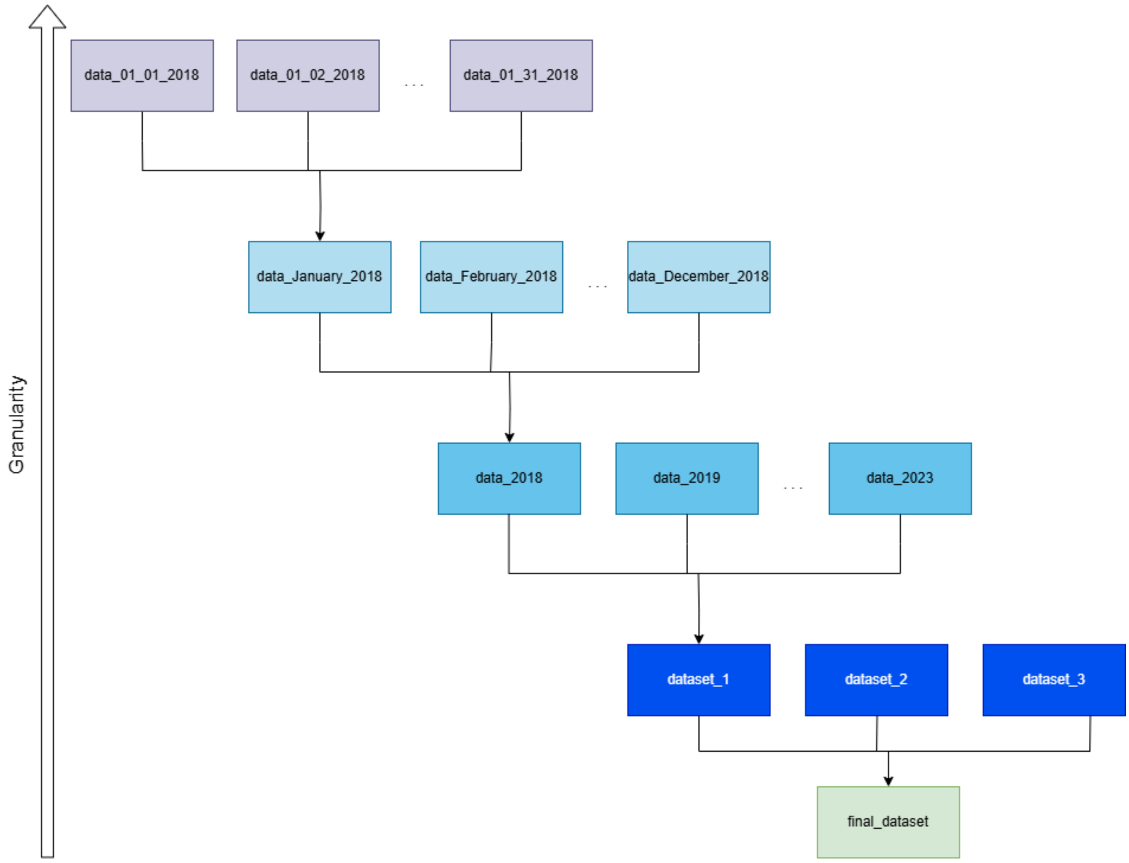

Figure 3.6: Data Recovery Strategy View

For all of these, a hierarchical organization based on temporal granularity is employed. The process begins with daily `.nc` data files such as `data_01_01_2018`, `data_01_02_2018`, and `data_01_31_2018`, which are aggregated into monthly files (`data_January_2018`, `data_February_2018`, `data_December_2018`).

These monthly datasets are subsequently aggregated into yearly collections (e.g., `data_2018`, `data_2019`, `data_2023`), and then grouped by domain into consolidated datasets (`dataset_1`, `dataset_2`, and `dataset_3`).

Finally, the three domain-specific datasets are trimmed to contain an equal number of rows and merged line-by-line into the unified `final_dataset`. This structured approach ensures temporal coherence and consistency across all sources, facilitating subsequent steps such as data analysis, time series forecasting, and predictive model design.

To pull, filter, and aggregate the massive daily volumes of data, Python was used as the primary programming language. Libraries such as xarray, pandas, and fileinput made parsing and converting numerous .nc files a reality, allowing scalable management of the data. Python scripts were developed to routinely extract daily environmental parameters and aggregate them into temporally-aligned monthly and yearly aggregates. The procedure left room for the final aggregated dataset that is both temporally aligned and computationally feasible to be constructed. The aggregated dataset is the ground upon which the remaining phases of data analysis and modeling in this chapter are built.

Figure 3.7 illustrates the step-by-step process utilized during data retrieval and cleaning in this dissertation:

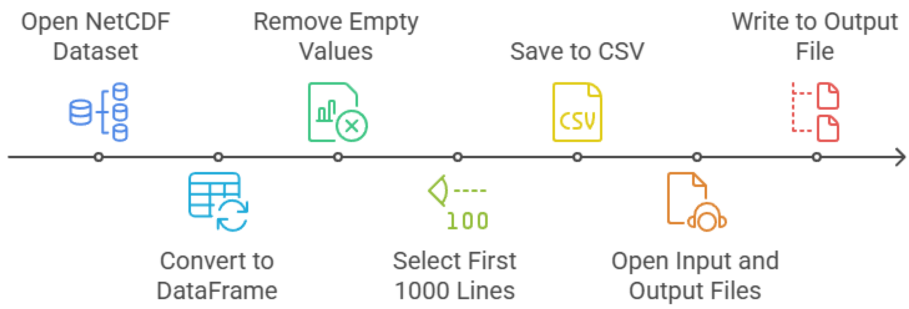

Figure 3.7: Step-by-step Process of Data Retrieval and Cleaning

It begins with opening the NetCDF dataset containing multidimensional environmental data. The dataset is then converted into tabular format using a DataFrame to enable manipulation. Missing or blank values are removed to ensure data integrity. From the cleaned data, just the first 1000 rows are pulled in to decelerate processing in early stages. The resultant subset is saved as CSV for compatibility and ease. Finally, the CSV is read into a fresh output file from which the cleaned data is returned, completing the process of extraction and formatting for further processing. All Copernicus data is stored as .nc files, so it is essential to use the xarray library to retrieve and open files with that extension.

Figure 3.8 illustrates the procedure adopted for gathering different output files into a single consolidated dataset:

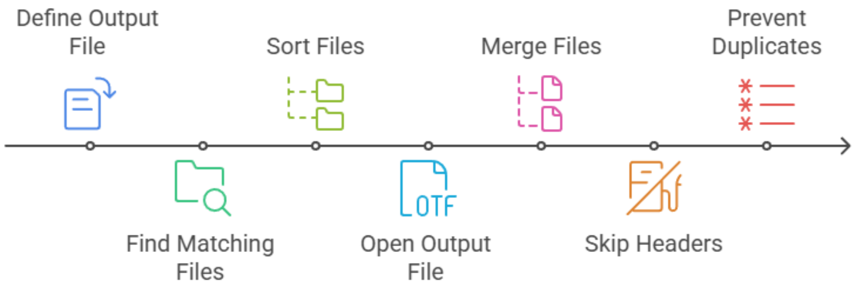

Figure 3.8: Data Merging Strategy View

The procedure starts by declaring the final output file where all gathered material will be stored. The system then goes on to scan directories to find all appropriate text files that meet predetermined naming conventions. Having been found, the files are then sorted in terms of keeping chronological order, which plays an essential role in maintaining time-series consistency. The output file is opened and all the input files are processed, skipping the headers on all except the first to avoid redundancy. A duplicate checking and prevention mechanism is used in merging so that there would be no duplicate lines written and data would remain unique and integral in the final merge output. After getting the data separated by years, each separate yearly data was manually joined in a separate file using Notepad, forming each of the 3 compiled datasets that will be later merged into the final dataset.

Figure 3.9: Final Dataset Merge View

Figure 3.9 illustrates the final data preprocessing stage where domain-specific datasets are cleaned and merged into one dataset for model construction.

It begins with the loading of the cleaned CSV files for the three environmental domains, SSWH, SST, and chlorophyll-a concentration. The latitude and longitude columns, which are unnecessary for modeling, are dropped to reduce the load. First, the number of rows in all data sets is harmonized by truncating each to the smallest data set, maintaining temporal consistency. Third, the data sets are merged line-for-line to produce a single structured table with harmonized attributes, ready for use in machine learning. We end up with one final dataset in csv format, that will be used for preprocessing and analysis. The dataset has 2,188,000 lines.

Chapter 4

Data Preprocessing and Exploratory Analysis

This chapter records data preprocessing and Exploratory Data Analysis (EDA) prior to model construction. Following data collection and joining in Chapter 3, the cleansed dataset of wave, thermal, and biogeochemical variables was cleaned and examined further. This guarantees structurally consistent, analytically reasonable, and tractable data to temporal modeling as well as machine learning-based classification.

To enable a flexible and iterative research process, a set of Jupyter Notebooks was established. These notebooks are the foundation for all subsequent data-driven processes, including filtering of datasets, EDA, testing of different models, and time series forecasting.

4.1 Data Recovery and Modeling

The data cleaning phase consisted of a series of simple preprocessing operations. These included the removal of duplicate records, handling missing values, conversion of time formats for proper indexing, and value conversion to ensure we are working with metric units. Several datasets were also merged into a single common format based on common keys, and the corresponding temporal columns were extracted in order to enable time-aware modeling.

No outliers were removed during data preparation in this research because outliers in environmental data typically reflect true, significant events, like occasional extreme high spikes of chlorophyll-a for toxic HABs, rather than data error. Those occasional but valuable events are essential to capturing the full range of variation of oceanographic systems and especially useful for developing forecast-focused predictive models for bloom detection and forecasting. Removing these values can result in the underestimation of bloom intensity or frequency, ultimately biasing the analysis and weakening the power of the models to simulate extreme, though ecologically and economically significant, events. Retaining outliers ensures that the models will remain sensitive to typical patterns as well as to rare, high-impact events.

Figure 4.1 illustrates the main steps followed in the data cleaning process on the final dataset, which highlights the transformations performed to encourage consistency, temporal coordination, and unit standardization across all sources. Some of these steps were made for each individual dataset, but following good practises, were repeated to ensure good data.

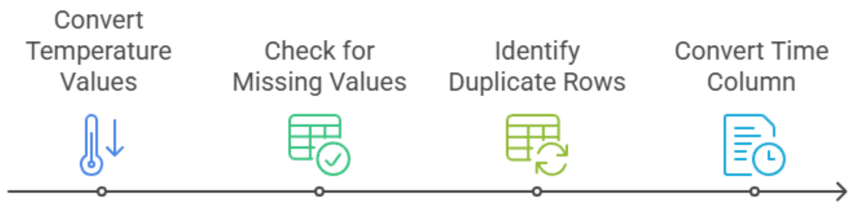

Figure 4.1: Data Cleaning Steps for Final Dataset

4.2 Exploratory Data Analysis

Following data cleaning, EDA focused on identifying the structure, quality, and trends of the data. Descriptive statistics of key variables were calculated, distributions were shown in histograms and boxplots, and temporal and seasonal trends were explored using line plots. Correlation matrices and heatmaps were also used to identify patterns between environmental variables such as chlorophyll-a concentrations, SST, and significant wave height (VHM0).

These notebooks also form the foundation for subsequent modeling stages when various ML models were tested and compared. Lastly, the time series analysis module investigated temporal patterns of chlorophyll-a concentration and related variables and laid the ground for predictive modeling in subsequent phases of the dissertation.

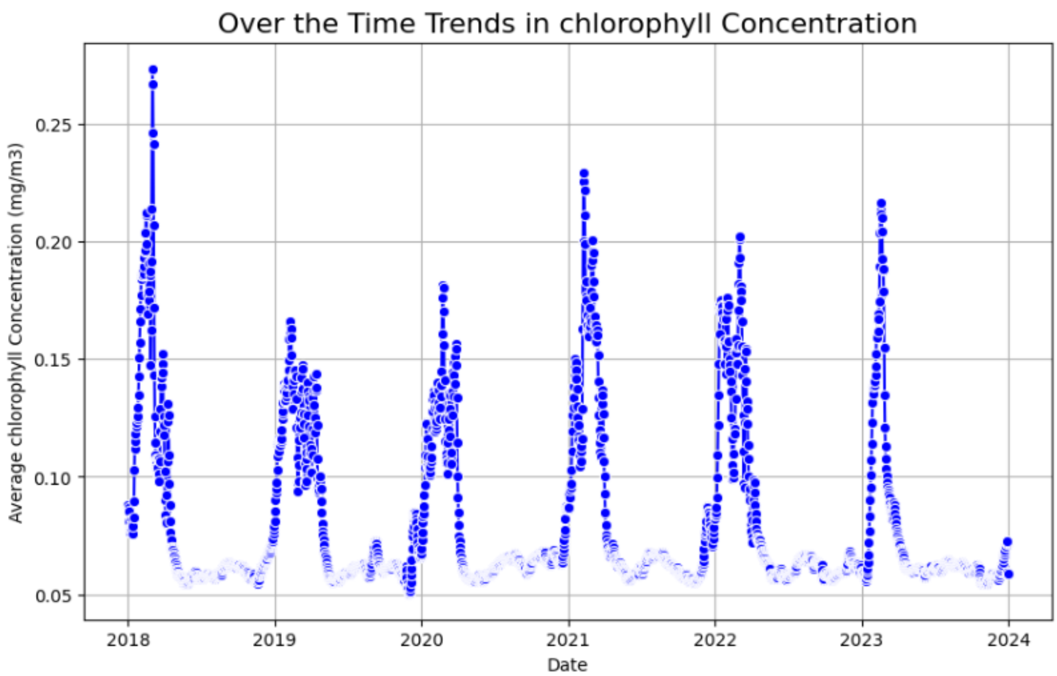

Figure 4.2: Trends In Chlorophyll-a Concentration

The seasonality in the time series plot represented in Figure 4.2 has some yearly-like highs in chlorophyll-a concentration between 2018 and 2024:

The patterns are in line with phytoplankton growth following environmental condition changes such as enhanced SST, levels of available nutrients, or wave activity regime. The highest was in early 2018, at over 0.27, and other years are slightly lower but repeating yearly highs. At intervals between blooms, chlorophyll-a levels fall to relatively stable minima of around 0.05–0.07, a characteristic which suggests the periodic return of these blooms. In general, this illustrate the temporal consistency of the algal blooms and supports the hypothesis of an important role of seasonal environmental conditions in the variability of chlorophyll-a in the area.

Figures 4.3 and 4.4 show temporal trends in SST and Sea Surface Wave Maximum Height (VCMX) between the years 2018-2024, which may be associated with the dynamics of the previously shown chlorophyll-a concentration.

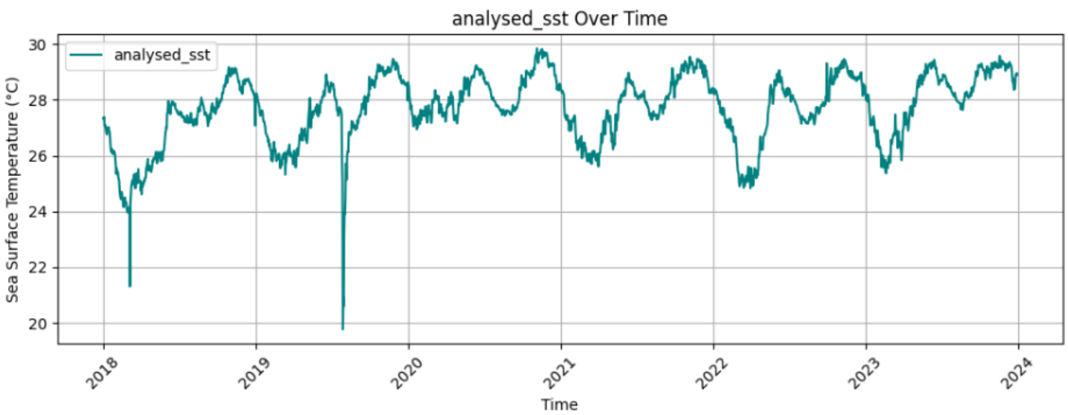

Figure 4.3: Trends in Sea Surface Temperature Values

Figure 4.3 shows that SST shows a clear seasonality, its values oscillating around 24°C to 29°C, reaching their annual maximum each time, predominantly during warmer months. These summer SST maxima fall in line closely with maxima in chlorophyll-a concentration, consistent with the interpretation that increased temperatures enhance phytoplankton production and concomitant increases in chlorophyll-a. The association suggests that SST is a key factor for inducing seasonal algal blooms, potentially by enhancing metabolic rates and stratification favorable to bloom development.

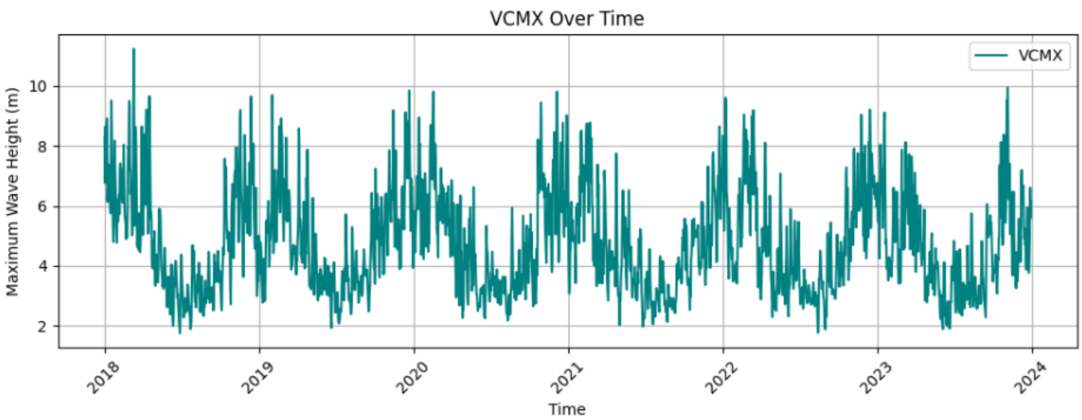

Figure 4.4: Trends In Sea Surface Wave Maximum Height Values

Figure 4.4 indicates VCMX variability that is seasonally occurring with more irregular maxima. Although VCMX is not in absolute concurrence with chlorophyll-a maxima, its fluctuations can conceivably influence the dispersion or suppression of blooms. For instance, diminished wave heights might be in agreement with tranquil, stratified water conditions that would favor bloom establishment, whereas elevated wave action might well entrain the water column and prevent bloom augmentation.

In Yoon et al., (2018), iron is described as a limiting micronutrient in HNLC waters. The article addresses results from artificial ocean iron fertilization (aOIF) experiments such as iron enrichment experiment (IronEx), Southern Ocean iron experiment(SOFeX), and Subarctic Pacific iron Experiment for Ecosystem Dynamics Study (SEEDS) that featured spectacular increases in chlorophyll-a levels, sometimes up to 27 times, following iron enrichment. These experiments confirm that iron availability can increase phytoplankton productivity and cause community structural changes, particularly towards diatom-dominated blooms when silicate is also abundant.

Similarly, Moore et al., (2013) characterizes iron limitation as a common characteristic of HNLC regions such as the Southern Ocean or even in the Pacific. It points out that the productivity of phytoplankton in these regions often is elevated only following iron enrichment, confirming iron as a key limiting nutrient in these oceans. The article also refers to mesoscale IronExs that demonstrated recurring patterns of iron-stimulated biological response in both bottle and in situ experiments.

Figure 4.5 shows the temporal patterns in iron (Fe) concentration from 2018 to 2024, which exhibit annual recurring peaks in agreement with those of chlorophyll-a concentration and SST:

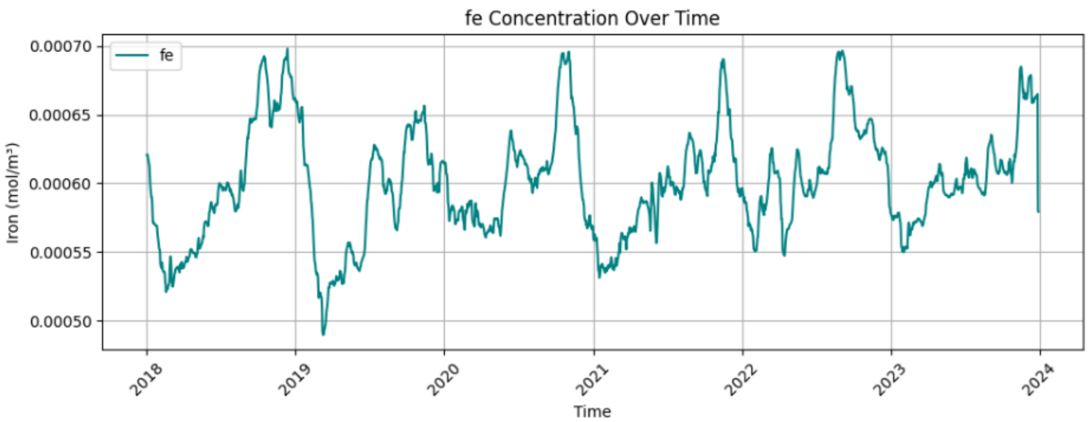

Figure 4.5: Trends In Fe Values

The Fe concentrations range mostly between 0.00055 mol/m³ and 0.00070 mol/m³, with typical peak values occurring during or slightly before the times of high chlorophyll-a concentration. This temporal coincidence suggests that Fe may have a facilitating role in the stimulation of phytoplankton growth, since iron is known as a micronutrient required for photosynthesis and for chlorophyll-a formation. When viewed in the light of the previous analyses, there is no doubt that chlorophyll-a concentration peak rates are most likely to be compelled by a synergistic effect of elevated SST and Fe availability, supporting optimal phytoplankton bloom conditions. While SST seems to act as a main driver by warming surface waters, sporadic bursts in Fe concentration likely enhance the biological response, amplifying bloom events.

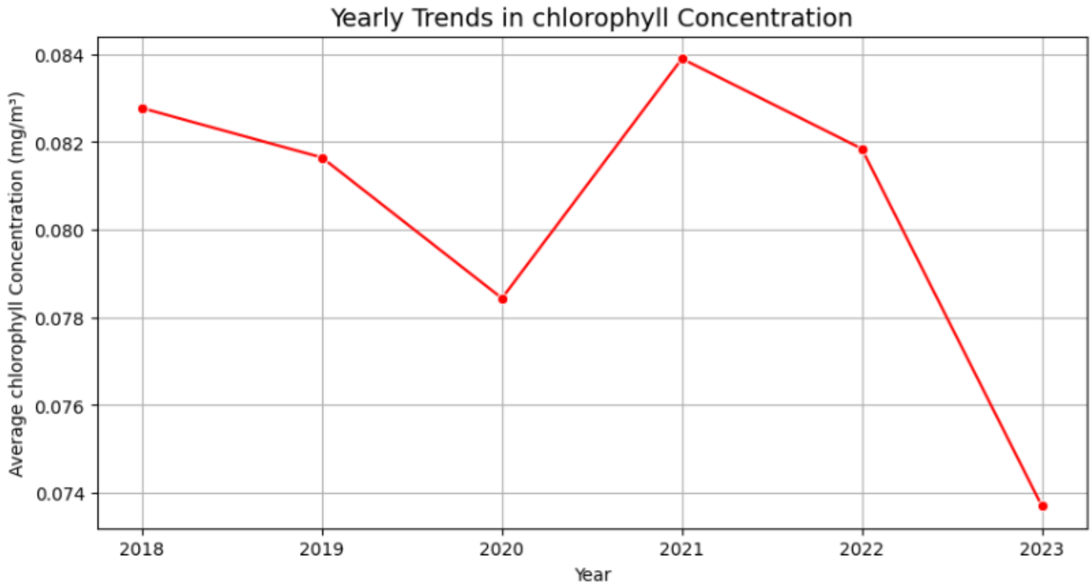

Figure 4.6: Yearly Trends In Chlorophyll-a Concentration

Figure 4.6 shows the yearly average trends of chlorophyll-a concentration from 2018 to 2023, with an oscillating but generally declining trend. The highest average is achieved in 2021, after which there is a steep drop throughout 2022 to its lowest in 2023. This general decreasing trend suggests that although seasonal peaks still occur each year (as previously), the intensity and/or duration of phytoplankton blooms may be decreasing year by year.

If contrasted with previous exams, this trend can mirror environment or nutrient availability change. For example, although there is no noticeable seasonal shift in SST and Fe concentration, subtle decreasing in peak levels of Fe or VCMX may affect the magnitude or duration of bloom. Also, climate change or anthropogenic activity can subtly alter the balance of variables underlying high chlorophyll-a persistence. So, while short-term variables like SST and Fe still control annual blooms, long-term variables could be damping overall productivity, as shown by this downward trend.

The plot shown in Figure 4.7 is a PCA plot of the features in the data, projecting the data points on the first two principal components. This indicates the overall structure and variance of the multidimensional dataset on a two-dimensional map with minimal loss of information.

The point spread is close and compact on the right side of the plot, especially between Principal Component 1 values of -10 to 0, with greater spread towards the left side between 0 to 4, which means that most of the variance lies along PC1, which likely captures the dominant patterns in environmental and biological characteristics (e.g., Fe, chlorophyll-a, SST). Spread vertically along PC2 indicates additional variation, but less so. This plot implies that while there is some variance in the data, most of the variance can be accounted for by a couple of controlling factors, lending credence to the hypothesis that key traits such as SST, Fe concentration, and VHM0 might be primary drivers of chlorophyll-a dynamics.

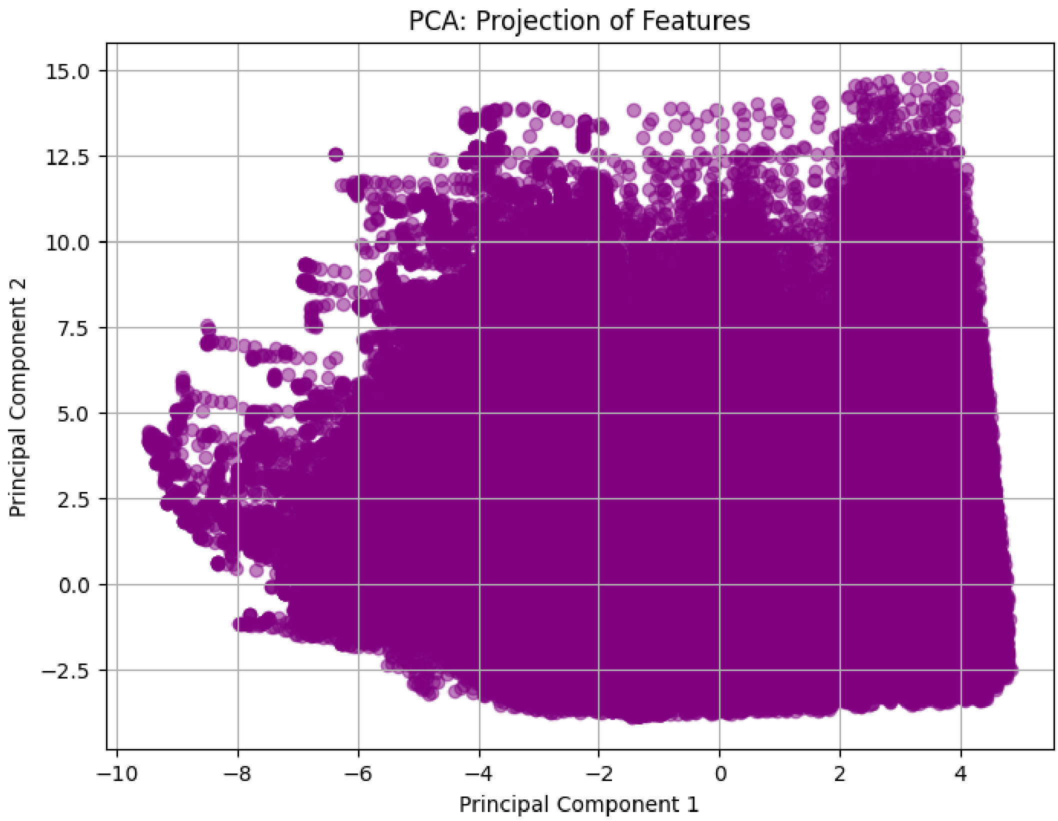

Figure 4.7: PCA Projection Features

Some studies reveal notable reported correlations between physico-chemical parameters and concentration levels of chlorophyll-a, an algal biomass index. Notable correlations with chlorophyll-a and nutrients, total nitrogen (TN) and total phosphorus (TP), with increased correlation coefficients (R^2 values of up to 0.66) were reported in the article by Ly et al., (2021). This study also found that temperature and light induce algae growth whereas greater flow rates suppress it. Similarly, in Cho et al., (2014), the most crucial factors controlling Chl-a were Total Organic Carbon (TOC) and pH, and these bettered TN and TP, which were low in correlation, surprisingly so, most likely because the hyper-eutrophic status of the lake is such that saturation of nutrients makes them less predictable. Furthermore, Stow et al., (2003) determined that nitrogen and phosphorus loads, solar irradiance, and water level were first-order variables in explaining Chl-a variability, with nitrogen-related variables being the first to emerge as predictive variables in ML models. In general, these results highlight the multi-modal and site-specific interaction of nutrients, physical conditions, and meteorological conditions in regulating algal bloom dynamics.

To more rigorously identify drivers of chlorophyll-a variation, correlation matrices were used to examine chlorophyll-a correlations with other environmental and nutrient-derived characteristics. Strength and direction of linear relationship quantification is enabled by these correlation heatmaps, providing insight into which characteristics could be driving or co-varying with phytoplankton process.

In Figure 4.8, which illustrates chlorophyll-a correlations with other nutrients, we can observe high positive correlations with phyc (r = 0.77) and nppv (r = 0.82), indicating that phytoplankton biomass and net primary productivity are very closely related to chlorophyll-a concentration.

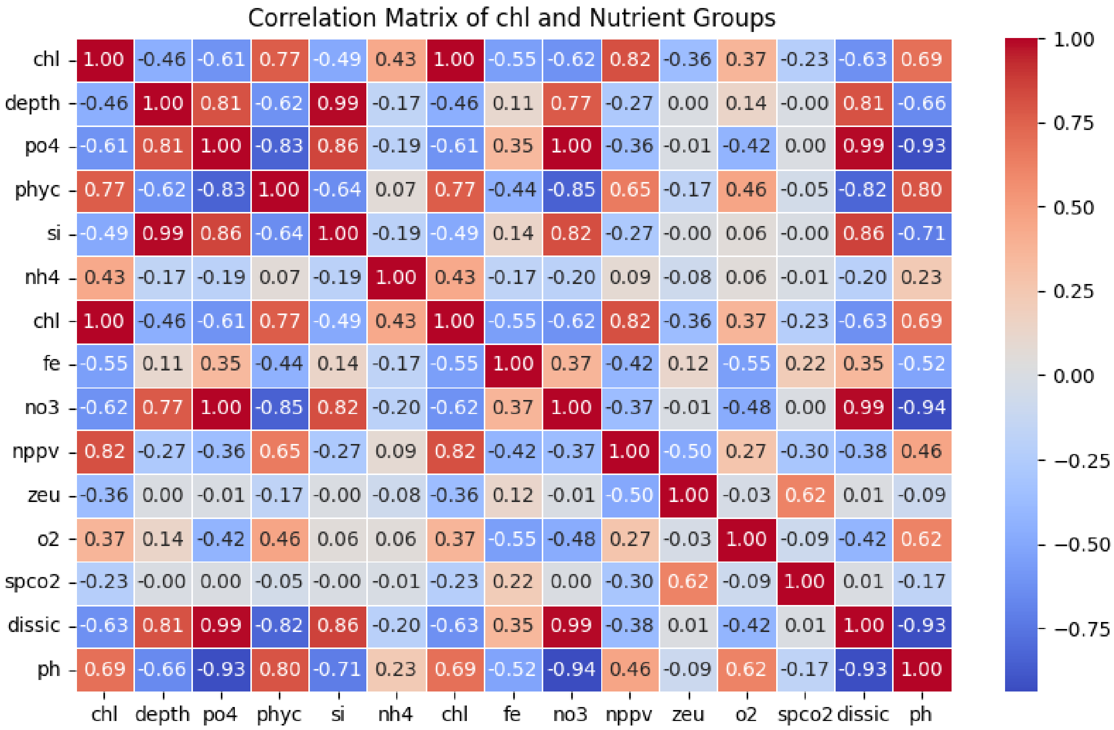

Figure 4.8: Correlation Matrix of chl and Nutrient Groups

po4 (r = -0.61) and si (r = -0.49) on the other hand have negative correlations. Studies have reported that chlorophyll-a may decline when phosphate levels increase, particularly after bloom events or in eutrophic waters during internal loading phases (Painter, (2024)). This suggests that phosphate accumulation often follows a bloom, rather than precedes it, possibly due to internal loading, organic matter decomposition, or reduced uptake by declining phytoplankton communities. In coastal regimes like the Bay of Bengal, silicate has been shown to correlate negatively with chlorophyll-a during certain phases, indicating shifts in species composition or nutrient limitation dynamics (Paul et al., (2008)).

Interestingly, iron also exhibits a moderate negative correlation with chlorophyll (r = -0.55). In a study conducted in the coastal area of Zhanjiang Bay (Zhang et al., (2024)), higher Fe concentrations were similarly associated with reduced chlorophyll-a ($r \approx -0.28$ to -0.51). This indicates that oxidative stress or nutrient deficiencies could be the cause, or potentially because excess iron can inhibit photosynthesis, or it can indicate non-productive seasons (Zhang et al., (2024), Bonnet et al., (2018), Moore et al., (2023)).

Together, these trends suggest that, in this environment, Po4, Si, and Fe are not the primary limiting nutrients and that their increased concentrations may follow bloom events rather than drive them.

In Figure 4.9, the SST correlation matrix shows weak but negative SST correlation with chlorophyll-a (r = -0.19), which is normal since algae tend to grow on cold water. This is to say that while SST seasonality trends appear visually in sync with chlorophyll-a maxima, the direct linear relationship is weak, likely due to complex, non-linear dynamics or the fact that many interacting variables tend to co-occur when blooms happen.

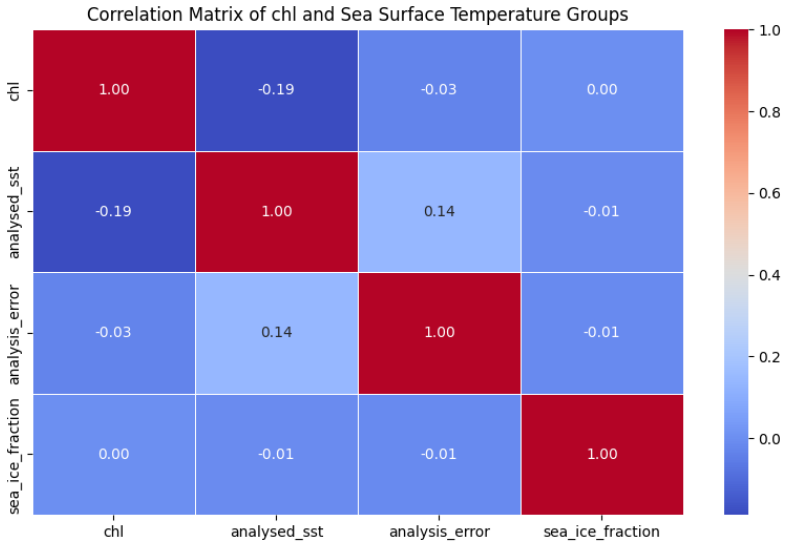

Figure 4.9: Correlation Matrix of chl and Sea Surface Temperature Groups

Lastly, Figure 4.10, plotting wave height parameters, shows more decidedly stronger negative correlations between chlorophyll-a and the different wave height parameters, most notably VHM0 (r = -0.06) and VTPK (r = -0.10):

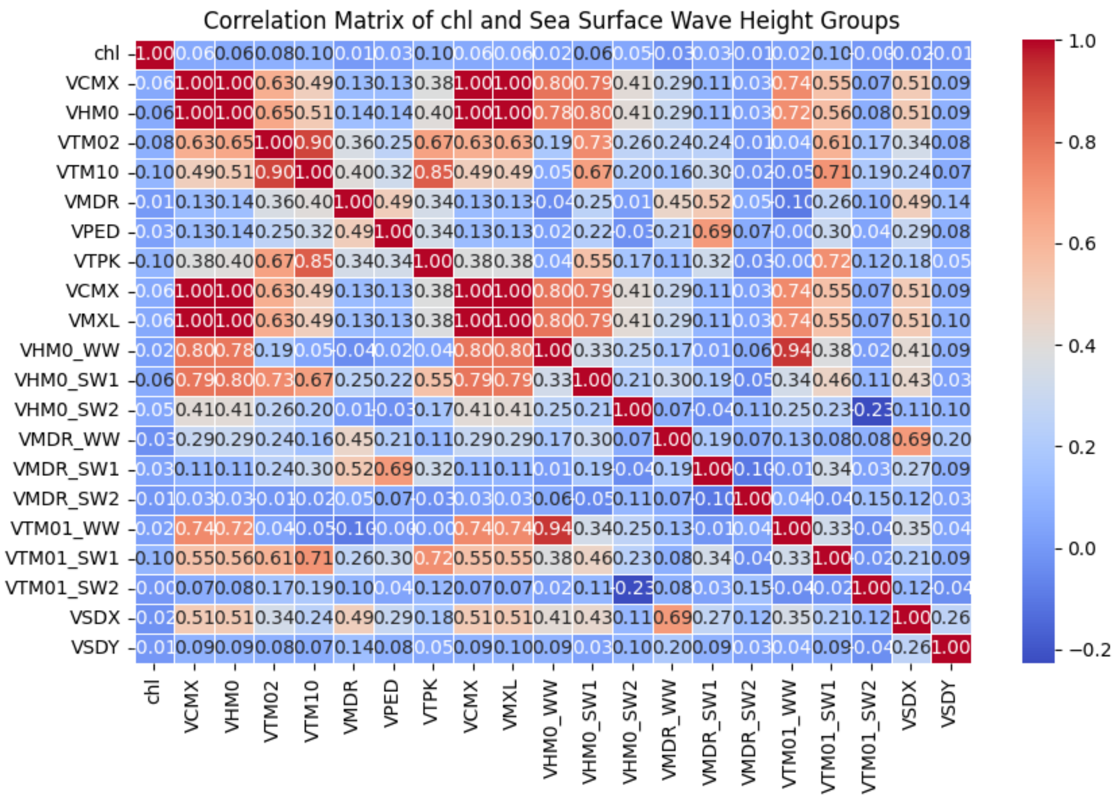

Figure 4.10: Correlation Matrix of chl and Sea Surface Wave Height Groups

The findings show that increased wave activity has the effect of inhibiting chlorophyll-a buildup through increased mixing of the water column capable of fragmenting stratification and dispersing nutrient levels.

Collectively, the matrices state that while nutrients, particularly phosphate, silicate, and primary productivity, are dominant controls on chlorophyll-a variability, physical oceanographic factors like SST and VHM0 play strong but less overt or more oblique roles as regulators of phytoplankton dynamics.

As an extra quality control on the correlation matrix test, Figure 4.11 provides a better sense of chlorophyll-a concentration correlation with two of the most important environmental variables, SST and VHM0, in the form of Pearson correlation scatterplots. The plots allow us to see data distributions and linear trends visually.

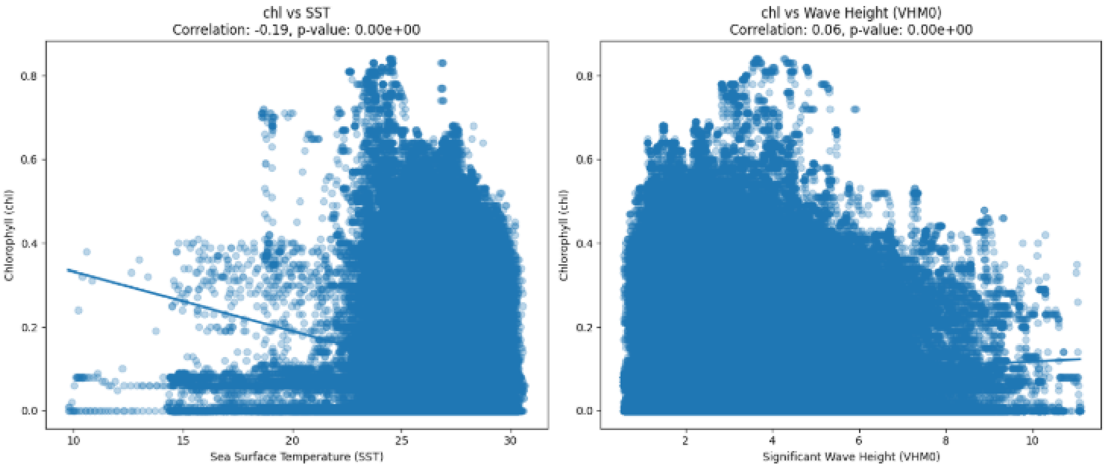

Figure 4.11: Pearson Correlation Between chl and Key Variables

The left plot shows the SST-chlorophyll-a relationship, which is negative ($r = -0.19$) and significant ($p < 0.001$). While there is some clustering of high chlorophyll-a values at mid-range SST values (26–28°C), overall there is a tendency for rising SST to be weakly correlated with falling levels of chlorophyll-a. This agrees with earlier findings that the SST-phytoplankton productivity relationship is nonlinear and a function that depends on the concurrent availability of resources such as nutrients or processes of stratification.

The plot on the right illustrates the correlation of chlorophyll-a with VHM0 revealing a positive correlation ($r = 0.06$) of high statistical significance ($p < 0.001$). The outcome is that high frequency of chlorophyll-a tends to occur in calmer wave conditions, which is in agreement with the assertion that calmer seas can enhance bloom development by offering the source of water column stability and nutrient retention in the euphotic zone.

Together, these scatterplots lend graphical support to the above statistical analysis: while both SST and VHM0 have effects on chlorophyll-a dynamics, these effects are weak and are likely interactive with biological and chemical processes, i.e., nutrient inputs, in controlling phytoplankton bloom dynamics.

4.3 Data Recovery and Modeling

Figure 4.12 displays the data recovery and modeling approach applied during the course of the dissertation, presenting a stepwise process from raw data acquisition to predictive modeling and result interpretation:

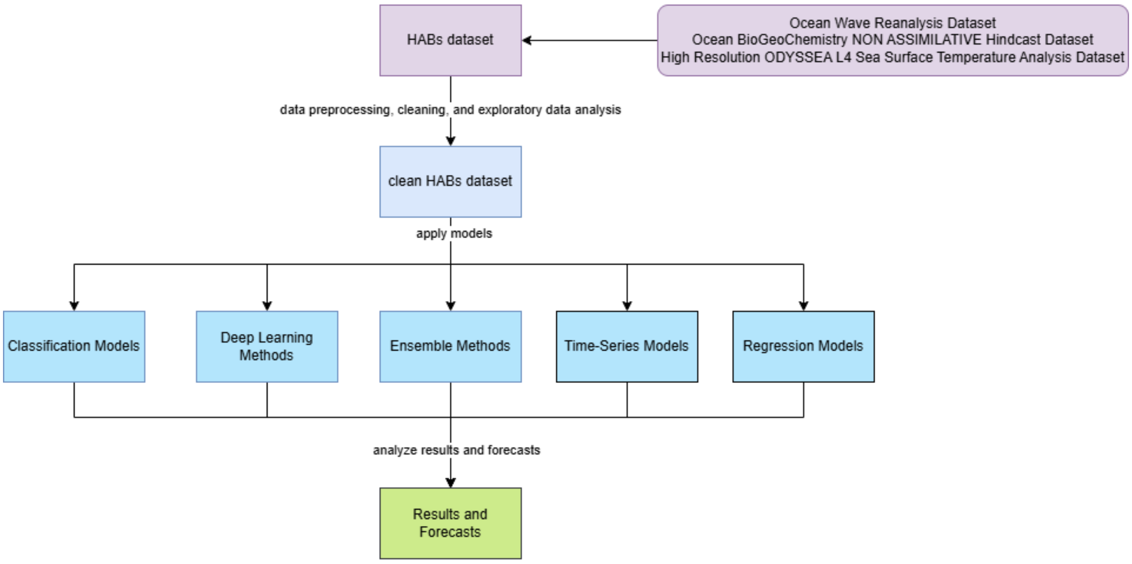

Figure 4.12: Data Recovery and Modeling Strategy

The figure presents the methodological pipeline followed to describe and predict chlorophyll-a concentration and other oceanographic parameters.

It begins with the HABs dataset which was gathered from the datasets mentioned in Chapter 3.2. They constitute an adequate basis of biogeochemical and physical variables to apply when performing simulations of HABs.

Raw data gets cleaned, preprocessed, and EDA is done in order to organize a clean, formatted data set to model. Clean data gets input into various types of analytical and predictive models in five general categories:

- Classification Models: To make bloom or non-bloom status prediction.
- Deep Learning Techniques: Most likely to be used to extract nonlinear complex patterns and high-dimensional relations.
- Ensemble Methods: For improved prediction by leveraging the strengths of more than one model.
- Time-Series Models: For predicting future trends based on trends over time.
- Regression Models: For predicting chlorophyll-a explanatory variable relationships.

Finally, model outputs are combined and examined at the Results and Forecasts step and provide informative data about chlorophyll-a change and hopefully informing bloom management practice. It is an admirable integrated effort demonstrating a good integrated attempt filling the gap of disjointed modeling practices and hard data processing to provide useful environmental data.

Table 4.1: Mapping of Method Categories to Implemented Models

Method Category	**Implemented Models**
Classification Models	Logistic Regression, Naïve Bayes, K-Nearest Neighbors (KNN), Decision Tree
Deep Learning Methods	Neural Network (Multi-Layer Perceptron), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Convolutional Neural Network (CNN)
Ensemble Methods	Random Forest, Bagging, AdaBoost, Gradient Boosting, XGBoost, LightGBM
Time-Series Models	ARIMA, SARIMA, SARIMAX, LSTM, GRU, CNN
Regression Models	Linear Regression, Random Forest Regressor, XGBoost Regressor

A variety of statistical and machine learning models were selected in this dissertation to enable diverse challenges of ecological data, such as the challenge of predicting chlorophyll-a concentration based on seasonal, spatial, and non-linear environment variables. Basic models such as Logistic Regression and Naïve Bayes were selected because they are simple, transparent, and easy to understand and therefore can serve as baselines for benchmarking. Logistic Regression is a linear assumption between features and target, and while it is simple in decision boundaries, it is often not sufficient for modeling complex patterns. Naïve Bayes, depending on probabilistic assumptions, is computationally efficient but is based on the relatively unrealistic feature independence assumption.

Instance-based and rule-based machine learning algorithms like K-Nearest Neighbors (KNN) and Decision Trees were also considered due to their intuitive reasoning and ability to handle non-linear feature spaces effortlessly. KNN is extremely sensitive to distance metrics and feature scaling, and local data distributions, and therefore less reliable in high-dimensional or noisy domains. Decision Trees are interpretable and can handle both numerical and categorical features without preprocessing but risk overfitting unless regularized or used as a component within ensemble methods.

To address the limitations of more primitive models and to increase robustness and generalization, some ensemble methods were used including RF, Bagging, AdaBoost, Gradient Boosting, XGBoost, and LightGBM. These models combine multiple base learners, generally decision trees, to overcome variance and bias and improve predictive performance. They are best applied to heterogeneous datasets with feature interaction. Their main weaknesses are increased training time, reduced interpretability compared to single models, and hypersensitivity towards the hyperparameter choice.

Deep models such as the Multi-Layer Perceptron (MLP), Long Short-Term Memory (LSTM), and GRU were applied in DL to learn non-linear relationships and time constraints. MLPs are capable of learning any function but must be finely tuned and trained on large data sets to avoid overfitting. LSTM and GRU models are particularly designed to learn long-range dependencies in sequence data and hence are ideal for time series modeling where remembering past values is important. However, these models are computationally intensive, require enormous training sets, and are architecturally, learning rate wise, and sequence length sensitive. CNNs were also tried out in sequence modeling to find out how they can leverage local patterns and reduce the model when temporal windows are involved, even though their efficiency can be destroyed if not paired with proper sequence design or regularization.

In time series forecasting and regression, conventional models such as Linear Regression, ARIMA, SARIMA, and SARIMAX were utilized due to their explanatory nature and prior application in environmental modeling. These models are based on linearity and stationarity, and perform optimally when the processes underlying them are well-behaved and seasonally patterned. Their advantage is ease of interpretation and transparency, but with the need for strong assumptions regarding data properties and a tendency to fail in cases involving sudden or non-linear changes. For these, tree-based regression algorithms like RF Regressor and XGBoost Regressor were selected due to their ability to learn complex non-linear patterns, handle missing values, and operate with minimal feature engineering. They are robust and adaptive but computationally expensive and less interpretable in the absence of feature attribution tools.

The selection of these models was driven by the need to balance interpretability, model flexibility, temporal awareness, and computational tractability in various phases of data analysis, ranging from static classification to dynamic forecasting. Each class of modeling has specific theoretical strengths and weaknesses that make it better or worse suited depending on the nature of data and the topology of prediction task. The detailed evaluation of these models and their comparative outcomes is thoroughly discussed in Chapter 5.

4.4 Modeling Overview

To allow for proper prediction of the levels of chlorophyll-a concentration, the data was pre-filtered to remove the lowest 25% of chlorophyll-a values that were likely to be low-variability or non-bloom cases and would lower the classification problem. The rest of the samples were partitioned into three ecologically significant classes, Low, Medium, and High, based on the 50th and 75th percentiles of the chlorophyll-a values. This process allowed the models to focus on the variable and bloom-relevant portion of the dataset. After target variable encoding and scaling of numerical features, a stratified 70-30 train-test split, was performed to preserve class distribution and ensure equitable evaluation. Various classification models, i.e., simple learners, ensemble approaches, and neural networks, were then trained to observe how they could classify patterns among environmental variables and chlorophyll-a level categories. Most importantly, no hyperparameter tuning or cross-validation was employed here. This was by design, as comparative general performance of model classes using default or base parameters was of more interest than incurring the computational cost or difficulty of tuning. It was an early benchmarking step to determine which algorithms were worth optimizing in the subsequent stages of the inquiry.

There were some challenges faced in preparing and training the model. First, the data set contained missing values, which were eliminated in the preprocessing in order to preserve model training integrity. This is because most machine learning algorithms lack support for missing values natively. Second, feature scaling could not be circumvented as a result of having predictors with varying units and magnitudes (e.g., wave height, and temperature). Standardization allowed every feature to contribute equally to distance-based or gradient-based models, i.e., KNN, SVM, and neural networks. The ecological character of the data also meant that some of the forms of chlorophyll-a naturally appeared more frequently, most significantly in the middle range, so class imbalance was a problem. This was somewhat alleviated by percentile-based bucketing and stratified sampling when carrying out the train-test split. Lastly, caution was taken in a way that only numeric and pertinent features were kept in the feature selection process, discarding non-numeric columns or metadata that could influence the modeling process.

For the prediction of future chlorophyll-a concentration, a large ensemble of regression models was calibrated to temporal series data from environmental and biological predictors. Data were built by inserting measurements of chlorophyll-a in chronological order and pairing them with temporally co-occurring environmental predictors. Alternative modeling techniques were attempted, from basic statistical models such as Linear Regression, ARIMA, SARIMA, and SARIMAX to complex machine learning models such as RF Regressor, XGBoost Regressor, and various DL models such as LSTM, GRU, and CNN of varying sequence sizes. The model performances were compared with traditional regression metrics, MAE, MSE, RMSE, and R^2, to ascertain their predictive ability and generalizability. Different from the classification notebook, this notebook had hyperparameter tuning using 'GridSearchCV' specifically for the SARIMA and RF Regressor. Cross-validation was not feasible due to the nature of the data being time-based (where shuffling is irrelevant), but the tuning was performed with validation splits that retained chronological order. Instead of 'train_test_split', the data was manually split across time, holding out the latest observations to test with in order to best mimic a real-world forecasting environment, being the first 80% of data dedicated for training and the remaining for testing. Each model's best hyperparameters are listed in Table 4.2.

Table 4.2: Best Hyperparameters for Time Series Regression Models

Model	**Best Hyperparameters**
SARIMA	`order=(2, 0, 1), seasonal_order=(1, 1, 1, 12)`
Random Forest Regressor	`n_estimators=200, max_depth=15, min_samples_split=4`
XGBoost Regressor	`n_estimators=100, learning_rate=0.1, max_depth=6` (default used, no tuning)
LSTM	`units=64, activation='relu', optimizer='adam', learning_rate=0.001, epochs=50, batch_size=32`
GRU	`units=64, activation='relu', optimizer='adam', learning_rate=0.001, epochs=50, batch_size=32`
CNN (Sequence=10)	`conv_filters=64, kernel_size=3, activation='relu', optimizer='adam', batch_size=32`
Linear Regression	No hyperparameters (closed-form solution)
ARIMA	`order=(2, 0, 1)` (manually set based on AIC analysis)
SARIMAX	`order=(2, 0, 1), seasonal_order=(1, 1, 1, 12), exog=external variables`

Various technical and methodological issues were faced in time series modeling. Among the most obvious issues was the presence of missing values, characteristic with environmental data as a result of breaks in the sensors or malfunctioning of retrieval by the satellite. These were addressed by cleaning and interpolation to maintain continuity to the time series. Numerical predictors were feature-scaled with MinMaxScaler and StandardScaler depending on the sensitivity of the model to the input magnitude. One of the biggest challenges was tuning the hyperparameters of the DL models, particularly LSTM, GRU, and CNNs, because these models are very sensitive to input sequence length, architecture, and training conditions, generally due to it being very time-consuming. Although the 10 length of the CNN was selected, single-layer models were applied for LSTM and GRU with the same optimizer and activation parameters, the models remained vulnerable to overfitting and unable to capture the more extreme seasonal peaks present in earlier chlorophyll-a history. ARIMA and SARIMAX statistical models also required cautious parameterization for stability, especially under non-stationarity. Besides, the scarcity of high-quality, recent data and natural bloom intensity variability made learning and generalization more complicated. Despite these, the notebook provides an effective context for experimenting with both neural and conventional prediction techniques under practical environmental modeling conditions.

Chapter 5

Results and Discussion

This chapter explains and presents the results of experimentation involving the application of the ML and DL models mentioned in Chapter 4 to classify and predict chlorophyll-a concentration in aquatic environments. The purpose is to contrast the predictive capability of each model and determine the underlying factors contributing to their success or failure, with particular attention to ecological patterns and environmental drivers.

The chapter has two main sections: 5.1 and 5.2. The initial section discusses a broad array of classification models including traditional approaches (e.g., Logistic Regression, Naïve Bayes), instance-based learning methods (e.g., KNN), Neural Networks, and ensemble methods (e.g., RF, Bagging, XGBoost, LightGBM) using metrics such as precision, recall, F1-score, accuracy, Matthews Correlation Coefficient (MCC), and Receiver Operating Characteristic Area Under the Curve (ROC AUC). A comparative analysis selects the models which work best to discriminate different chlorophyll-a level classes (Low, Medium, High), and highlights the most significant features which are contributing most towards the decision.

In the second half, the focus is placed on time series forecasting models. These incorporate some statistical models (Linear Regression, ARIMA, Seasonal Autoregressive Integrated Moving Average (SARIMA), SARIMA with Exogenous Variables (SARIMAX)), DL models (LSTM, GRU), and CNNs of varying sequence lengths. Performance against future chlorophyll-a values in prediction of each model is compared using regression metrics like MSE, Root Mean Squared Error (RMSE), MAE, and R-squared (R^2). Visualization of the prediction outcome provides qualitative assessment, especially capture of seasonality and bloom occurrence.

These solutions give a general impression of model performance across paradigms and set the strengths and limitations of each approach applied to difficult, seasonally fluctuating marine data. The argument combines model accuracy with ecological interpretability, informing the selection of methodologies for upcoming environmental surveillance and decision-support systems.

5.1 Classification Models

This section describes a comparison of some of the ML models used to classify the level of chlorophyll-a into three ecological classes: High, Medium, and Low. Classifiers used in this situation range from the basic ones such as Logistic Regression and Naïve Bayes to the advanced ones such as Neural Networks and ensemble classifiers such as RF, XGBoost, and LightGBM.

The performance of each model is validated employing primary evaluation metrics: precision, recall, F1-score, accuracy, MCC, and ROC AUC. They are the metrics that give an equal-weighted score of the prediction ability and robustness of the models, especially for multiclass classification tasks. Feature importance rank is also presented for tree-based models to indicate the most crucial variables engaged in classifying choices. As a whole, this section describes the top-performing models of ecological classification and displays the most influential features of model performance. The use of ensemble models such as RF and Bagging has also been confirmed as beneficial in other HAB prediction studies. For example, Molares-Ulloa et al., (2023) demonstrated that ensemble strategies delivered higher robustness and predictive accuracy across multiple water quality environments. Their study emphasized that these models better handled variability and uncertainty across spatially distributed data.

Tables 5.1 and 5.2 shows that ensemble models, i.e., Random Forest, Bagging, and XGBoost, provide the best classification with precision and recall above 0.95 on all chlorophyll-a classes. Logistic Regression and KNN are comparatively good but degrade on the Medium and Low classes. Naïve Bayes, however, has the worst performance, particularly on the Low class, most likely due to its simplifying assumptions. In general, the tree-based ensemble models are highly predictive and stable for this multi-class ecological prediction. In the work of Sujatha et al., (2023) Random Forest showed excellent classification performance on algal bloom toxicity levels (F1 score of 0.97 on the train-test dataset and 0.96 on the validation dataset; AUROC = 0.998).

Table 5.1: Class-Level Precision, Recall, and Support for Top-Performing Models

Model	**Class**	**Precision**	**Recall**	**Support**
Random Forest	**High**	**0.99**	**0.99**	**146752**
	Low	**0.99**	**0.98**	**132884**
	Medium	**0.98**	**0.98**	**152936**
Bagging	High	0.99	0.99	146752
	Low	0.98	0.98	132884
	Medium	0.98	0.98	152936
XGBoost	High	0.99	0.98	146752
	Low	0.97	0.96	132884
	Medium	0.95	0.96	152936
Decision Tree	High	0.99	0.99	146752
	Low	0.98	0.98	132884
	Medium	0.97	0.97	152936
Neural Network	High	0.98	0.98	146752
	Low	0.97	0.95	132884
	Medium	0.93	0.95	152936

Table 5.2: Class-Level Precision, Recall, and Support for Remaining Models

Model	**Class**	**Precision**	**Recall**	**Support**
LightGBM	High	0.98	0.98	146752
	Low	0.96	0.95	132884
	Medium	0.93	0.95	152936
Gradient Boosting	High	0.96	0.96	146752
	Low	0.95	0.91	132884
	Medium	0.88	0.92	152936
KNN	High	0.96	0.95	146752
	Low	0.92	0.93	132884
	Medium	0.91	0.91	152936
Adaboost	High	0.94	0.93	146752
	Low	0.92	0.85	132884
	Medium	0.82	0.88	152936
Logistic Regression	High	0.95	0.90	146752
	Low	0.87	0.86	132884
	Medium	0.81	0.86	152936
Naïve Bayes	High	0.87	0.82	146752
	Low	0.79	0.53	132884
	Medium	0.63	0.85	152936

The ROC AUC comparison from Figure 5.1 gives us a useful and significant comparison of the classification worth of the set of ML algorithms.

The most surprising outcome is the outstanding performance of the ensemble and boosting techniques of RF, Bagging, XGBoost, LightGBM, and the Neural Network with an AUC value of the ideal value of 1.00. Ensemble methods like RF and Gradient Boost showed high AUROC values (up to 0.998) when classifying microcystins (Salah et al., (2023)). This implies that the models have an extremely high ability to distinguish between the target classes with nearly zero false positive and false negative rates. Their ROC curves follow the top-left of the plot closely, implying that they can achieve the highest sensitivity as well as highest specificity. For this top-level course, some other algorithms like KNN, Gradient Boosting, and Decision Tree also worked very well with AUC scores of 0.99 and 0.98, respectively. These algorithms, though not perfect, are still quite stable in classification performance and therefore still good options if interpretability or computational resources are an issue.

More traditional models like Logistic Regression and Naïve Bayes, however, had comparatively middle-of-the-pack performance. Logistic Regression was 0.97 on the AUC, and also excellent, but whose ROC curve isn't quite so strongly increasing as the very best ones, so maybe harder to mislead with boundary samples or class-aliased distributions. Naïve Bayes, at 0.91 on the AUC, was one of the worst of them.

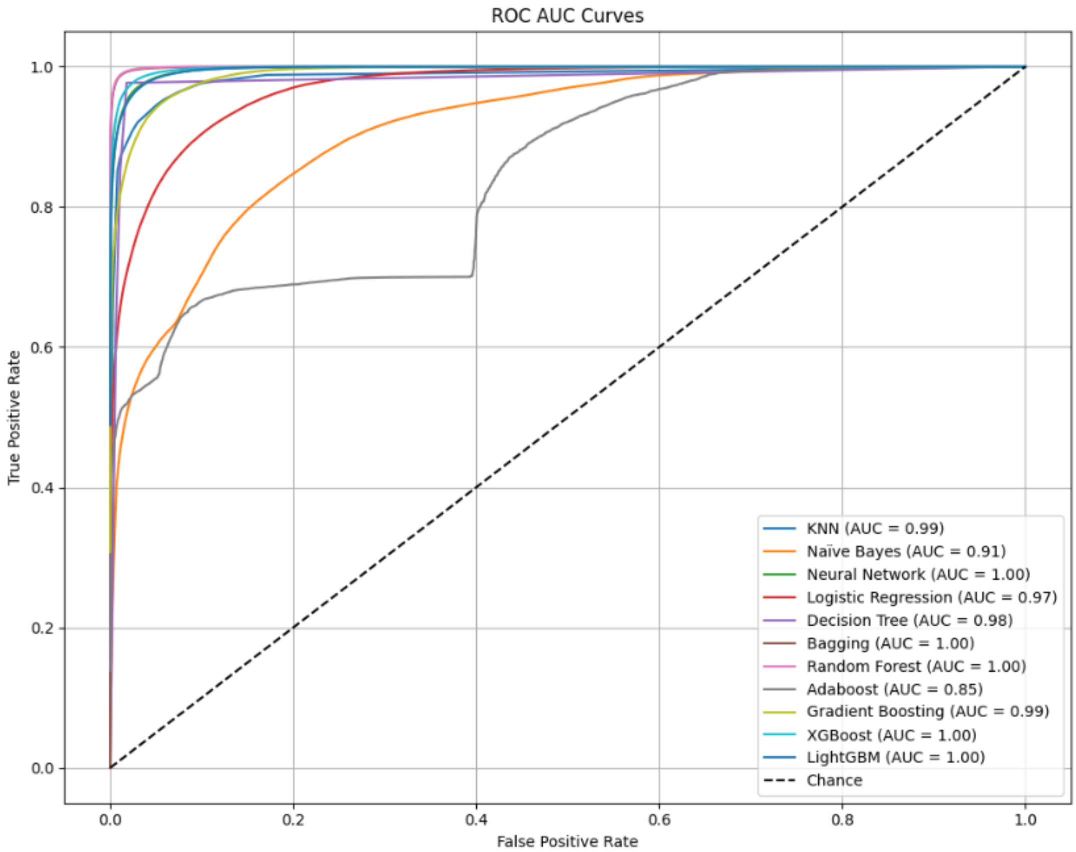

Figure 5.1: Area Under Curve Score for Each Model

Adaboost, which was a method designed to boost the performance with weak learners, only had an AUC of 0.85 and therefore might suggest possibly that it did not generalize as the other ensemble algorithms. In total, these results demonstrate the higher predictiveness and stability of current state-of-the-art ensemble algorithms being employed, especially when applied to difficult and high-dimensional data. For highly accurate and regular classification tasks, RF, XGBoost, and LightGBM are the top picks, but for simpler models, they could be the prefered ones if employed as baselines of comparison or when interpretability of the model is of major interest. This aligns with broader findings in recent literature, where hybrid and deep learning models such as N-BEATS and Conv-Attention networks outperformed traditional approaches in capturing spatiotemporal dependencies and predicting harmful algal blooms with greater precision (Martín-Suazo et al., (2024), Chen et al., (2022)).

Table 5.3 also shows the overall ranking of the classification models based on accuracy, MCC, ROC AUC, and weighted F1-Score. Random Forest is similarly illustrated in Figure 5.2 to be the best among all and then Bagging, XGBoost, and Decision Tree. Neural Network and LightGBM are also doing very well with high accuracy and consistent ROC AUC values. CNN-based models like B1D-CNN have demonstrated superior chlorophyll-a estimation from satellite data using deep learning (Salah et al., (2023)). On the contrary, Adaboost, Logistic Regression, and Naïve Bayes perform poorest. While Adaboost has good performance, its lower accuracy suggests model instability in extracting the complicated feature interactions. While Logistic Regression is popular for its simplicity and interpretability, it is not a good method to identify the intrinsic non-linear relation in this data.

Table 5.3: Model Performance Ranking Including Accuracy, MCC, ROC AUC, and F1-Score

Model	Accuracy	MCC	ROC AUC	F1-Score (Weighted)
Random Forest	**0.9844**	**0.9766**	**0.9995**	**0.98**
Bagging	0.9841	0.9760	0.9990	0.98
XGBoost	0.9686	0.9528	0.9977	0.97
Decision Tree	0.9766	0.9649	0.9825	0.98
Neural Network	0.9597	0.9395	0.9966	0.96
LightGBM	0.9583	0.9374	0.9964	0.96
Gradient Boosting	0.9285	0.8928	0.9900	0.93
KNN	0.9278	0.8916	0.9853	0.93
Adaboost	0.8894	0.8342	0.8450	0.89
Logistic Regression	0.8720	0.8080	0.9702	0.87
Naïve Bayes	0.7393	0.6176	0.9109	0.74

Finally, Naïve Bayes performs the poorest because its feature independence assumption in underlying data holds a violation in actual environmental data. These results determine that although the simpler models are still helpful for baseline or interpretability purposes, more sophisticated ensemble and boosting algorithms provide better predictive accuracy and robustness.

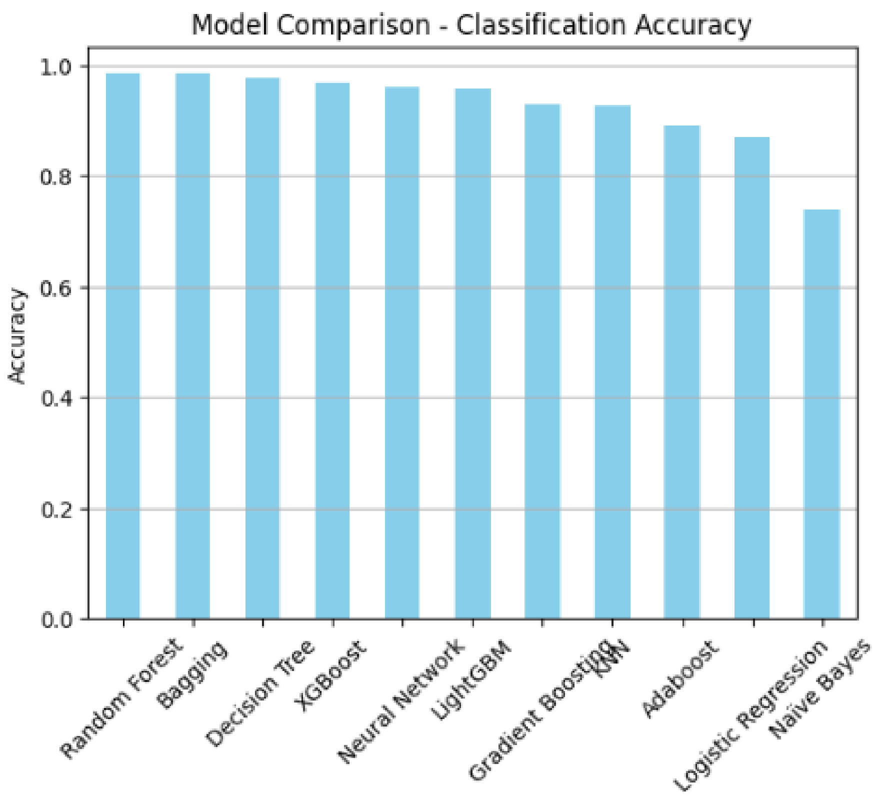

Figure 5.2: Accuracy Results Score

These models are still viable choices based on performance vs. interpretability vs. computation cost trade-offs. Adaboost, Logistic Regression, and Naïve Bayes are on the lower end. Adaboost is average in performance but with decreased accuracy that can be symptomatic of overfitting or instability when dealing with sophisticated feature interactions. Logistic Regression, so popular as it is due to its interpretability and simplicity, seems to lag behind since it's just not very good at modeling non-linear, high-dimensional relationships. Finally, Naïve Bayes is last since one might predict based on its noted feature independence assumptions that never actually take place in real-world data. These results validate the fact that even though lower-order models are useful for interpretability or as a baseline, advanced ensemble and boosting techniques result in much better accuracy and general prediction capability.

The Gradient Boosting feature importance plot from Figure 5.3 illustrates concentrated dependence on a small subset of features. Most prominently, nppv (Net Primary Productivity, per unit Volume) emerges as the dominant variable, with an importance value of more than half of the cumulative importance value. This suggests that Gradient Boosting puts considerable emphasis on productivity data in predictions, most probably due to the fact that it is strongly correlated with biotic activity such as algal blooms. Other moderately important attributes include 'zeu' (euphotic depth), 'latitude', and 'chl', which are ecologically relevant measures of water condition and biotic potential.

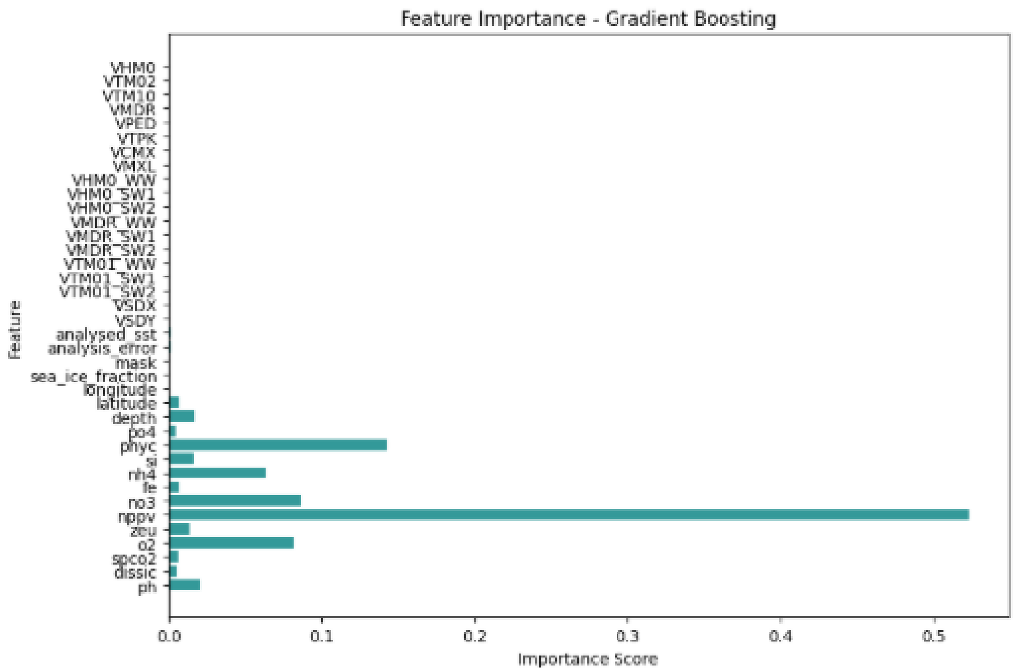

Figure 5.3: Feature Importance - Gradient Boosting

Notably, most of the physical attributes like wave parameters ('VHM0', 'VTPK', etc.) and wind parameters contribute minimally to the model, substantiating that this model's forecasts are primarily driven by biological and spatial inputs.

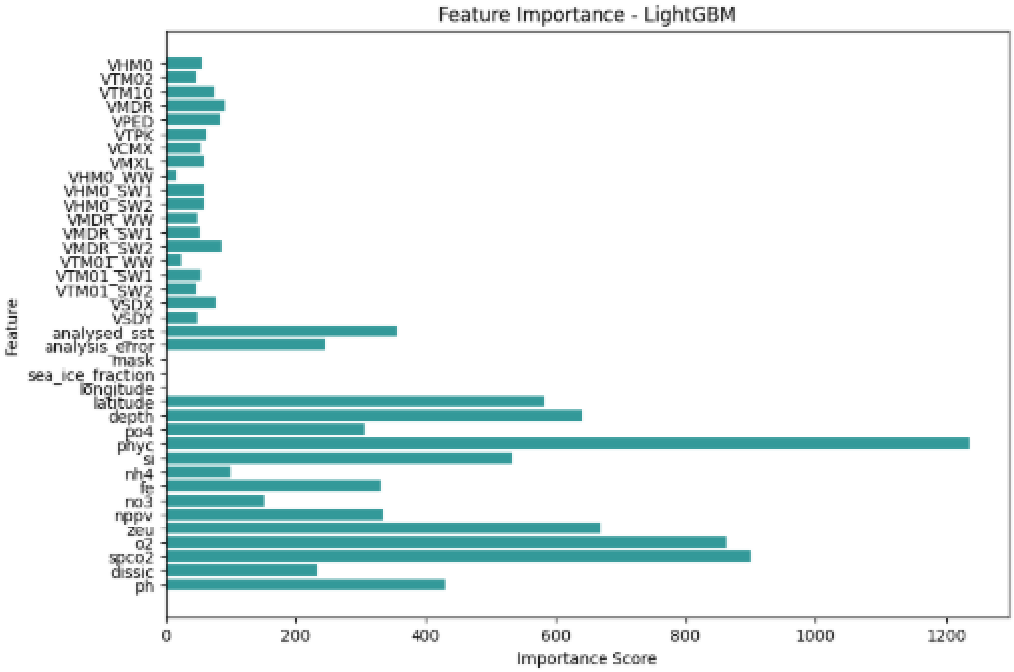

Figure 5.4: Feature Importance - LightGBM

In Figure 5.4, the LightGBM model is more balanced and distributed in feature importance distribution, so it takes a higher number of variables into consideration while deciding. The most dominant feature, 'phyc', is above 1200, and that tells us it is most important in the model's classification decision, probably because it most directly relates to water quality or bloom presence. 'spco2', 'nppv', 'zeu', and 'latitude' are also among top contributors, increasing the sensitivity of the model to geospatial and biological signals. LightGBM also highly ranks 'depth', 'no3', and even 'sea_ice_fraction', demonstrating its capability to learn from the widest range of biological, chemical, and climatic features. It is this wide consideration of inputs that enables LightGBM to perform well and consistently so at prediction tasks, as also shown in hybrid ensemble-based studies on harmful algal bloom forecasting (Molares-Ulloa et al., (2023)).

In the RF model, feature importances from Figure 5.5 are less biased than in Gradient Boosting but less daunting than in LightGBM. nppv is again the most significant feature, and zeu, spco2, and 'chl' are high contributors to model predictive ability. Geographical features like latitude and depth also have high importance, suggesting that spatial variation is a high contributor to classification. The RF's ability to read noise and interaction between features highly likely allows it to capitalize on more predictors without fitting on any individual one. As much as there are features that are largely meaningless such as wind variables and wave, their infinitesimally but real contributions point to the strength of the model in reading involved relationships without censoring weaker ones entirely.

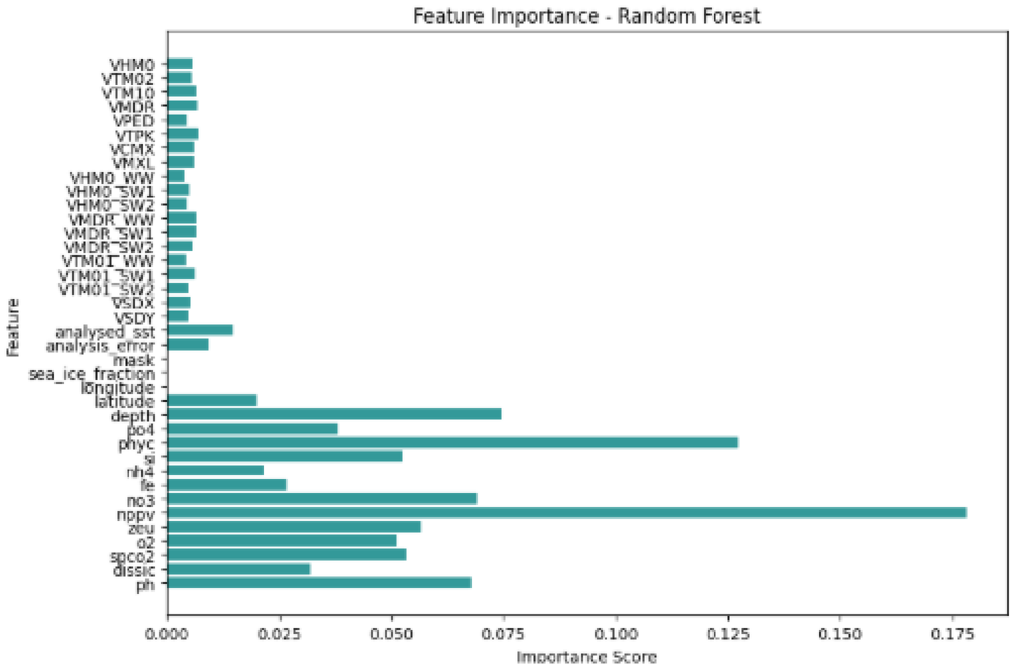

Figure 5.5: Feature Importance - Random Forest

The XGBoost model displays a more concentrated feature importance plot visible in Figure 5.6, with nppv dominating by far as the most significant feature, reaching an importance value well over 5000:

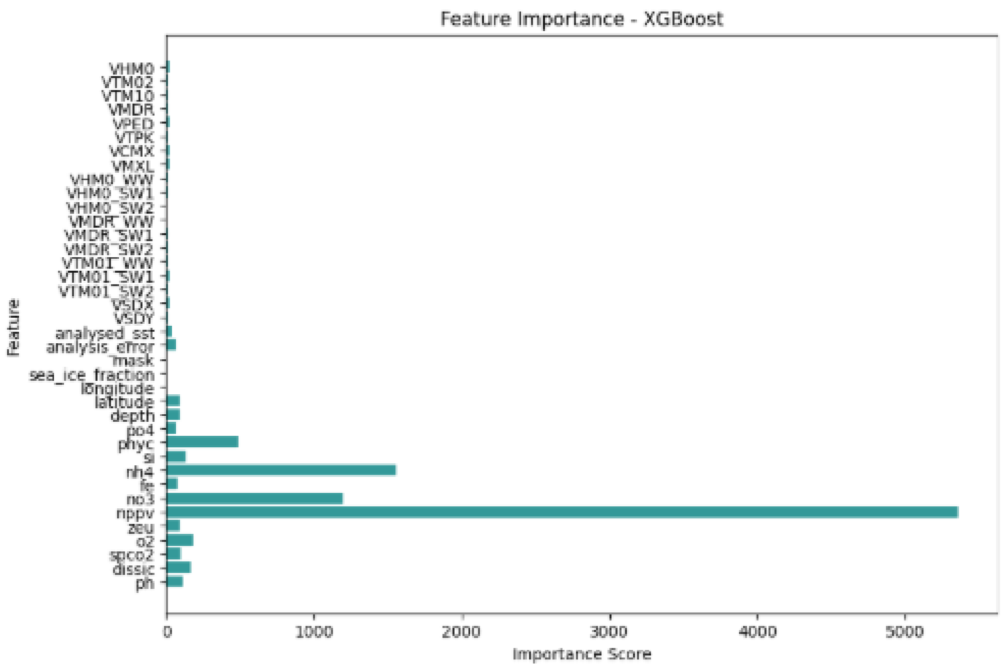

Figure 5.6: Feature Importance - XGBoost

This is reflective of the model's high reliance on net primary productivity information in making decisions. Other features such as 'no3', 'chl', and 'zeu' have less contributions but are meaningful in the model as well. The rapid fall-off of importance following the top few features shows that XGBoost concentrates on a small set of biologically important variables and pays little heed to others, particularly physical variables like wave and wind measurements. This prioritization allows XGBoost to perform well and avoid overfitting, though sometimes at the expense of missing subtle interactions captured by more evenly weighted models like LightGBM.

The RF confusion matrix in Figure 5.7 shows the degree to which it succeeded in correctly classifying each of the classes (High, Low, and Medium). By the high values along the diagonals, the majority of instances of all classes were accurately predicted. Specifically, according to the actual High instances of 143,558, barely any (just 7 of them as Low, 3,187 of them as Medium) were misclassified. This is found to be highly low false negative rate for High class and concludes the high reliability of the model in distinguishing high chlorophyll-a level or such consequential ecological states. This result is consistent with findings by Aslan et al., (2022), who reported that RF was particularly effective in modeling eutrophication risk in coastal lagoons, especially in identifying high-chlorophyll events with strong ecological relevance.

For Low class, it correctly predicted 125,762. It predicted 7,122 incorrectly as Medium, but none as High, indicating a potential overlap between the Low and Medium classes. This could be due to similar features or boundary feature values shared between these classes, indicating that the boundary may be more nuanced than the model can cleanly split. A similar pattern was observed by Ramgirkar et al., (2024), where RF models used to predict chlorophyll accumulation in oceans showed clear boundaries between extreme and mild states but struggled somewhat in separating adjacent or overlapping classes like Low and Medium.

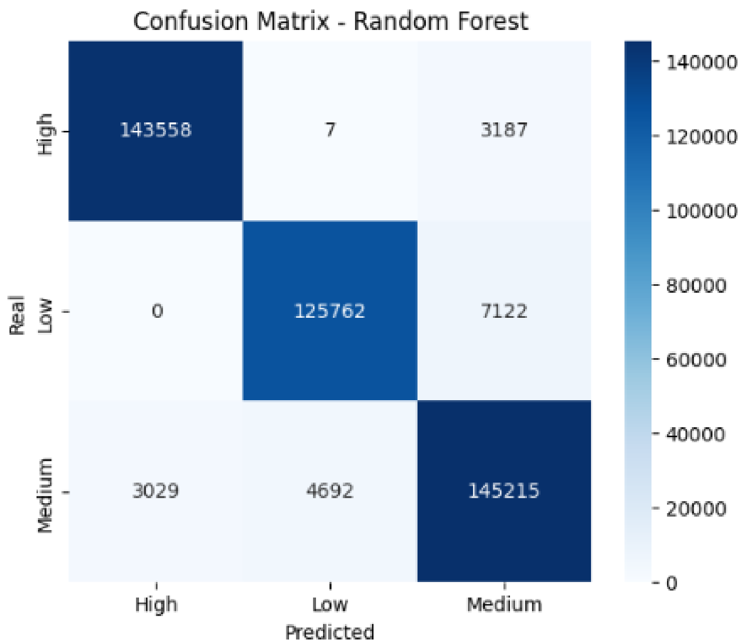

Figure 5.7: Confusion Matrix for Random Forest Model

For Medium class, RF model classified 145,215 correctly but incorrectly classified 3,029 as High and 4,692 as Low. Despite all misclassifications, overall accuracy remains very high and error distribution fairly balanced. The confusion matrix confirms that the model is optimum in discriminating between the High class with residual confusions still within Low and Medium classifications. In general, the matrix increases the stability of the model and confirms the results of previous metrics like accuracy, MCC, and ROC AUC, which all indicated RF as one of the top-performing classifiers.

5.2 Time Series Models Analysis

The following section is a comparative analysis of time series forecasting models employed in the prediction of chlorophyll-a concentration. This is done with the view to assess each model's capacity to describe seasonality, bloom dynamics, and impacts of environmental drivers such as SST and SSWH.

The models tested include the basic statistical models (Linear Regression, ARIMA, SARIMA, SARIMAX), Recurrent Neural Networks (RNN) (GRU and LSTM), and CNNs with varying sequence lengths to evaluate depth sensitivity in time. The models are evaluated using regression performance metrics like MSE, RMSE, MAE, and the coefficient of determination (R^2).

By qualitative and quantitative analysis, in this section, the advantages and disadvantages of both classification and time series methods are highlighted, and which models best fit to adequately forecast trends in chlorophyll-a in intricate marine systems are selected.

Figure 5.8 illustrates the temporal dynamics of chlorophyll-a concentration, SST, and SSWH, capturing how physical and chemical conditions relate to phytoplankton biomass:

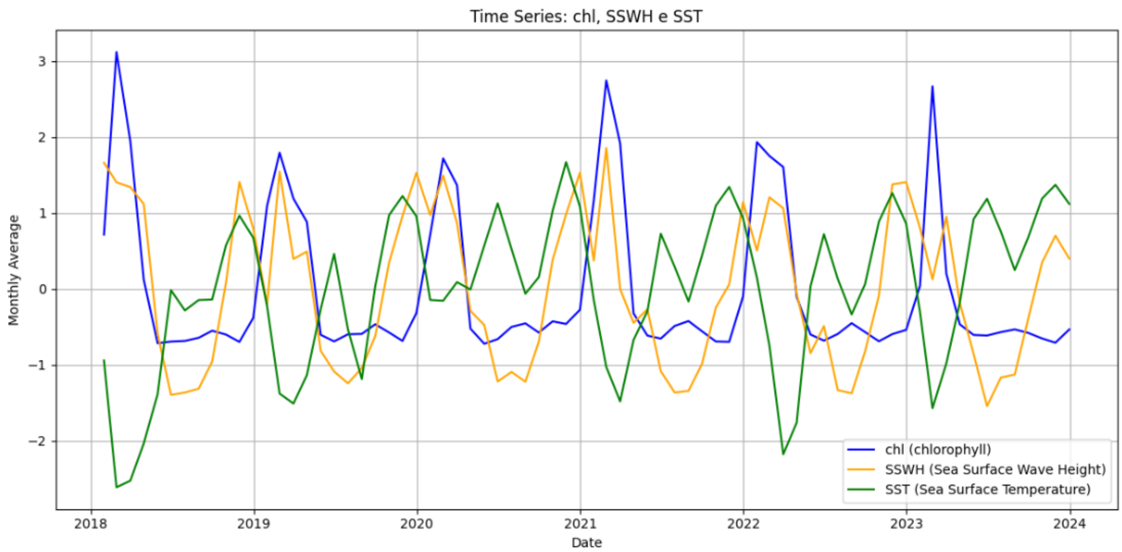

Figure 5.8: Time Series Representation of chl, SSWH and SST

There exists a clear and strong seasonality in all three variables, and chlorophyll-a (blue line) experiences most extreme values at peaks and lower ones, especially during early 2018, 2020, 2021, and 2023. Such peaks are probably coinciding with phytoplankton bloom events that typically occur during spring or early summer when the optimal environmental conditions prevail.

The SSWH (orange line) shows moderate seasonality, with fluctuations loosely aligning with some chlorophyll peaks, though not consistently. Periods of reduced wave height often precede or coincide with increased chlorophyll-a, supporting the hypothesis that calmer waters promote bloom formation by reducing water mixing. SST (green line) demonstrates smoother, cyclical changes, reaching annual peak values in summer and lowest values in winter. Interestingly, chlorophyll-a peaks often gets slightly behind SST rises, suggesting temperature-induced biological activity.

The seasonal changes in chlorophyll-a levels seen in the Han River indicate significant ecological processes that correlate with the graph shown in Figure 5.8. According to Ly et al., (2021), chlorophyll-a concentrations reached their highest in spring and were at their minimum in summer, even though summer is the season with the most rainfall. This inverse connection indicates that intense rainfall and resultant runoff in summer might lower nutrient levels or heighten hydrodynamic disturbances, like increased flow rate and wave action, complicating phytoplankton buildup. Although the Han River is a freshwater system, analogous dynamics occur during times of heavy rain and flow rates, which replicate the disturbance impacts of heightened wave action, ultimately leading to the noted inverse correlation between chlorophyll-a and hydrodynamic intensity (e.g., elevated flow or SSWH). Conversely, the somewhat dry spring probably offers more consistent conditions, enhancing nutrient retention and light accessibility, which collectively encourage algal proliferation. These patterns support the notion that algal blooms, indicated by elevated chlorophyll-a, are more likely to occur under low-disturbance conditions, where both temperature and organic matter availability align to support phytoplankton proliferation.

Figure 5.9 represents the time series of chlorophyll-a concentration with predictions by a linear regression model:

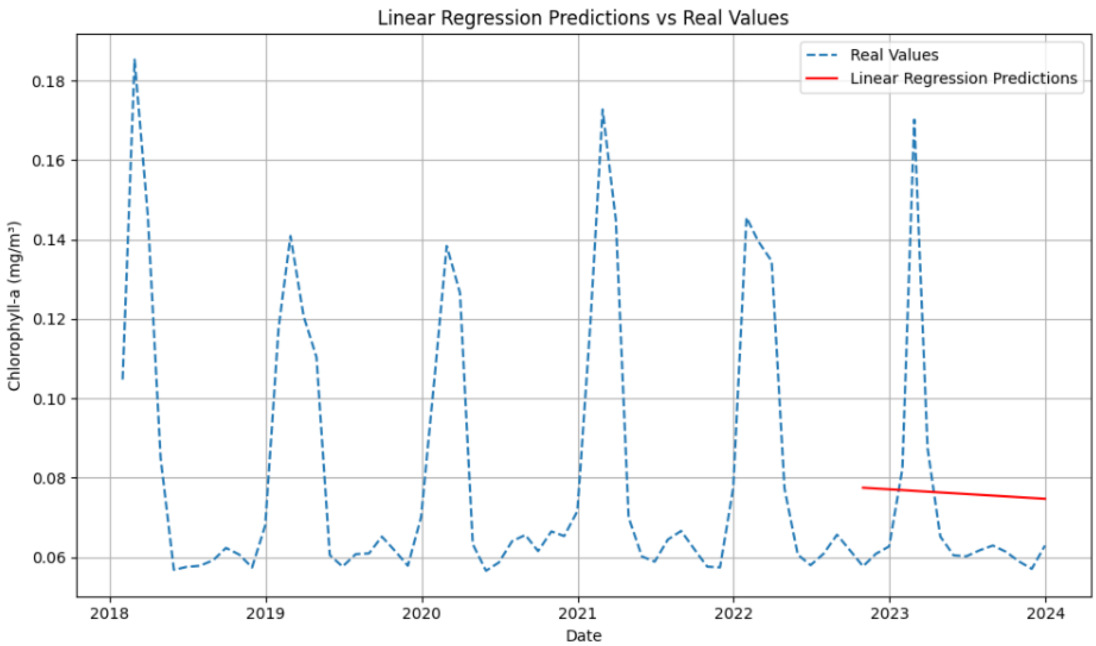

Figure 5.9: Time Series of Chlorophyll-a Using Linear Regression Model

The true values (blue dashed line) are highly seasonal with repeated peaks each year, which corresponds to the timing of algal blooms. The linear regression model (solid red line) does not have this periodic behavior but instead displays an almost flat trend that is very far from the data. This incompatibility refers to the limitations of linear regression in the estimation of complex, circular environmental processes like chlorophyll-a dynamics where non-linearity and seasonality are two prominent data features.

Figure 5.10 presents forecasts of chlorophyll-a concentration according to an ARIMA(2,0,1) model, plotted over the time series data observed:

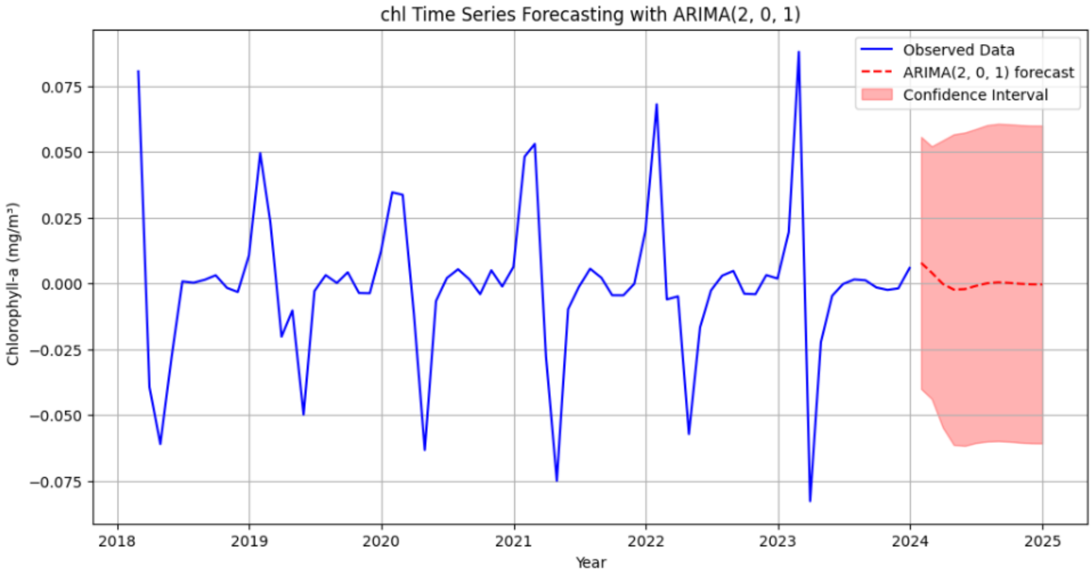

Figure 5.10: Time Series of Chlorophyll-a Using ARIMA Model

The historical record (blue line) once more exhibits a strong seasonal pattern with steep yearly peaks. The ARIMA model prediction (dotted red line) appears to remove this seasonality, failing to capture either the timing or size of the seen peaks. Further, the confidence interval (shaded red area) spreads out quite a lot out into the forecast period, reflecting increasing uncertainty. This would mean that while ARIMA can model some underlying trends, it fails to model the strong seasonality involved in chlorophyll-a dynamics and thus leads to reduced predictive consistency.

Figure 5.11 shows the chlorophyll-a concentration forecast through a SARIMA(2, 0, 1) model with added seasonal elements not present in the ARIMA(2, 0, 1) model in Figure 5.10.

Although both graphs depict the historical data that was observed (in blue), the SARIMA model indicates a more refined and seasonally intelligent prediction. In contrast to ARIMA, which produced a pretty flat line of prediction with hardly any variation, SARIMA follows the periodic peak and downfall cycle characteristic of annual chlorophyll-a behavior. This is evident in SARIMA's attempt to follow the steep peaks and downfalls that occur annually in the series in question. Additionally, the red shaded interval of confidence in Figure 5.11, although also expanding with time, is still closer to the anticipated range of seasonal fluctuation. This indicates that SARIMA not only captures the center of tendency better but also preserves sensible limits of prediction uncertainty. The larger, fuzzier interval of the ARIMA model in

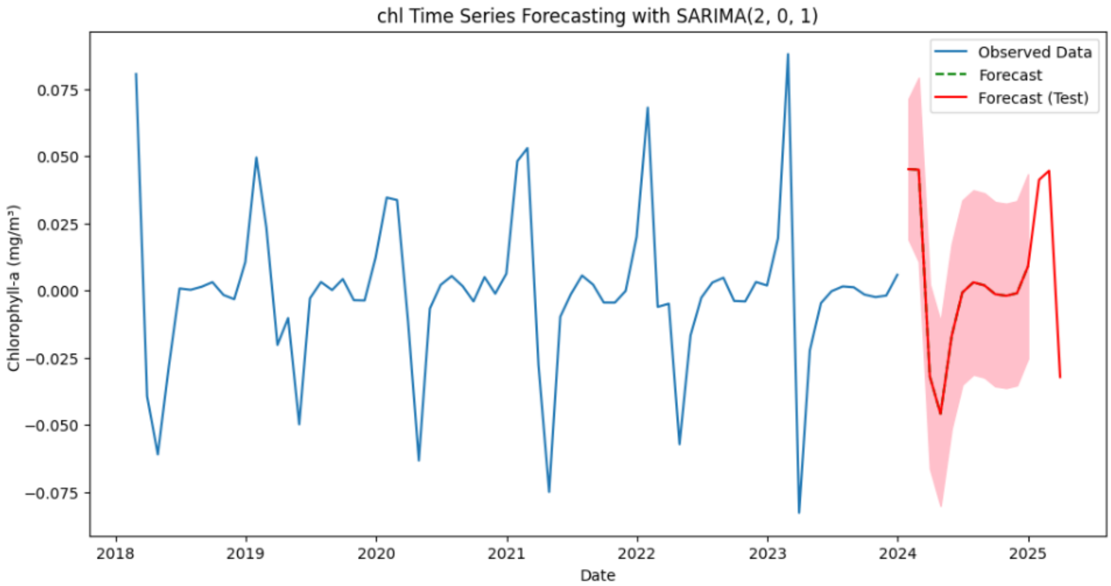

Figure 5.11: Time Series of Chlorophyll-a Using SARIMA Model

The ARIMA plot indicates larger uncertainty and lower confidence in its forecast. In consequence, by completely capturing seasonal behavior, the plot in Figure 5.11 produces a closer fit to chlorophyll-a measurements, with higher-quality short-term forecasting ability and more understandable confidence intervals, both of overarching concern in environmental forecasting and planning.

The plot shown in Figure 5.12 shows the chlorophyll-a concentration forecast through a SARIMAX(2, 0, 1) model with added seasonal elements not present in the SARIMA(2, 0, 1) model in Figure 5.11.

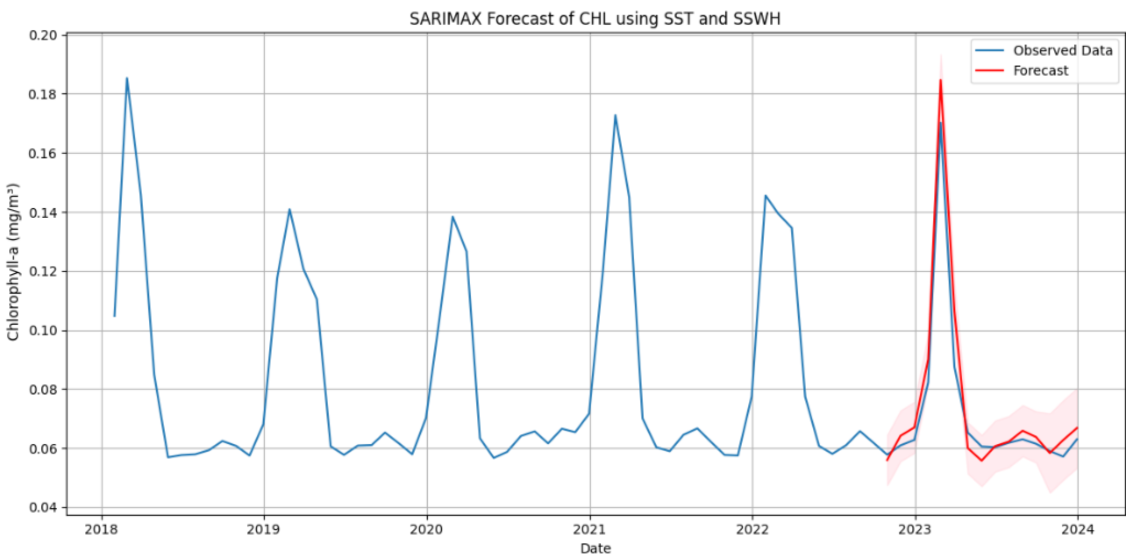

Figure 5.12: Time Series of Chlorophyll-a Using SARIMAX Model

Although both graphs depict the historical data that was observed (in blue), the SARIMA model indicates a more refined and seasonally intelligent prediction. In contrast to ARIMA, which produced a pretty flat line of prediction with hardly any variation, SARIMA follows the periodic peak and downfall cycle characteristic of annual chlorophyll-a behavior. This is evident in SARIMA's attempt to follow the steep peaks and downfalls that occur annually in the series in question, a behavior that aligns with the findings of Xia et al., (2024), who emphasized that forecasting models which account for periodic extremes provide more insight into ecological fluctuations during bloom-prone periods.

Additionally, the red shaded interval of confidence in the plot although also expanding with time, is still closer to the anticipated range of seasonal fluctuation. This indicates that SARIMA not only captures the center of tendency better but also preserves sensible limits of prediction uncertainty. The larger, fuzzier interval of the ARIMA model in Figure 5.10 indicates larger uncertainty and lower confidence in its forecast. Ramgirkar et al., (2024) similarly observed that while ARIMA was effective at modeling general trends, it lacked the seasonal sensitivity required for chlorophyll-specific forecasts, which SARIMA addressed through its better handling of temporal variation. In consequence, by completely capturing seasonal behavior, SARIMA produces a closer fit to chlorophyll-a measurements, with higher-quality short-term forecasting ability and more understandable confidence intervals, both of overarching concern in environmental forecasting and planning.

The plot in Figure 5.13 shows chlorophyll-a concentration predicted by the time series forecasting based on an LSTM model where all the data observed is in grey, the test set is in blue, and model predictions are in green:

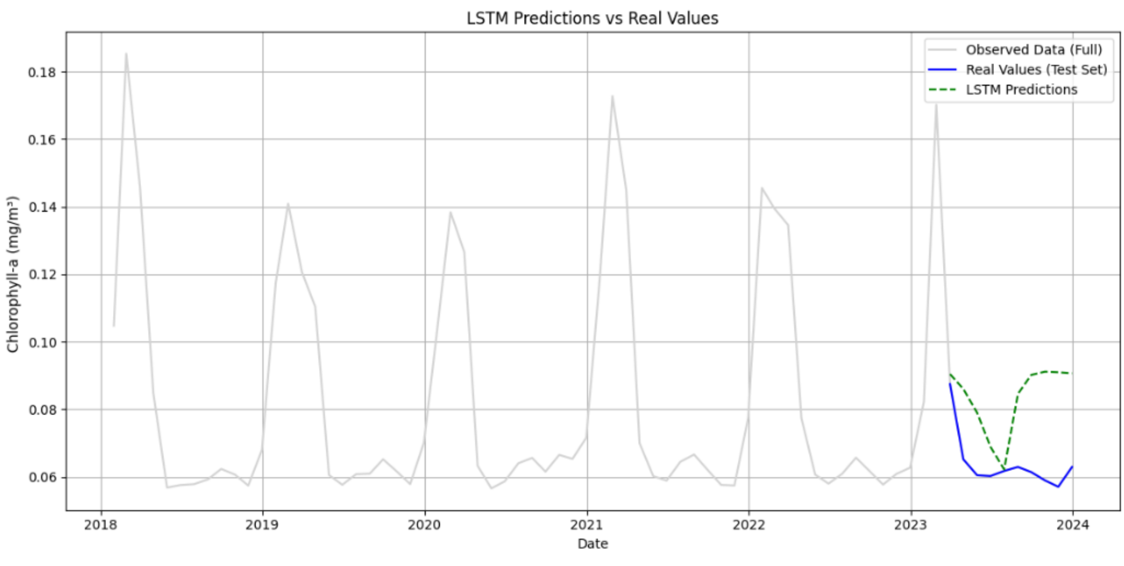

Figure 5.13: Time Series of Chlorophyll-a Using LSTM Model

The time frame of the graph is between 2018 and early 2024, putting the window for prediction into context against the complete history series. The raw data shows the extreme seasonal cycles with steep peaks in nearly all the years, indicating persistent algal bloom episodes. While the LSTM model captures the overall downtrend for the test period (2023–2024), its prediction does not capture the characteristic peaks and downfalls that characterize the past years.

The prediction is rather very smoothed with a curved behavior missing the steep seasonal swings. This means that the model, trained on a multi-year times series, failed to learn periodicity in chlorophyll-a concentration. The mismatch between predicted and observed values reflects the challenge in recreating complex seasonally-determined ecological processes with RNN using univariate or brief-sequence input. Echo state networks and similar DL methods for algal bloom prediction highlight difficulties in handling seasonality unless architecture is carefully adapted (Zhao et al., (2022)).

Figure 5.14 shows chlorophyll-a concentration predictions on a GRU model put into context within all history data available from 2018 up to the beginning of 2024. Grey observed data are visible with apparent seasonality having yearly maxima as one would anticipate for the dynamics of algal bloom. Test set value (blue line) shows most recent minimum in season, and GRU model prediction (green dashed line) attempts to replicate this.

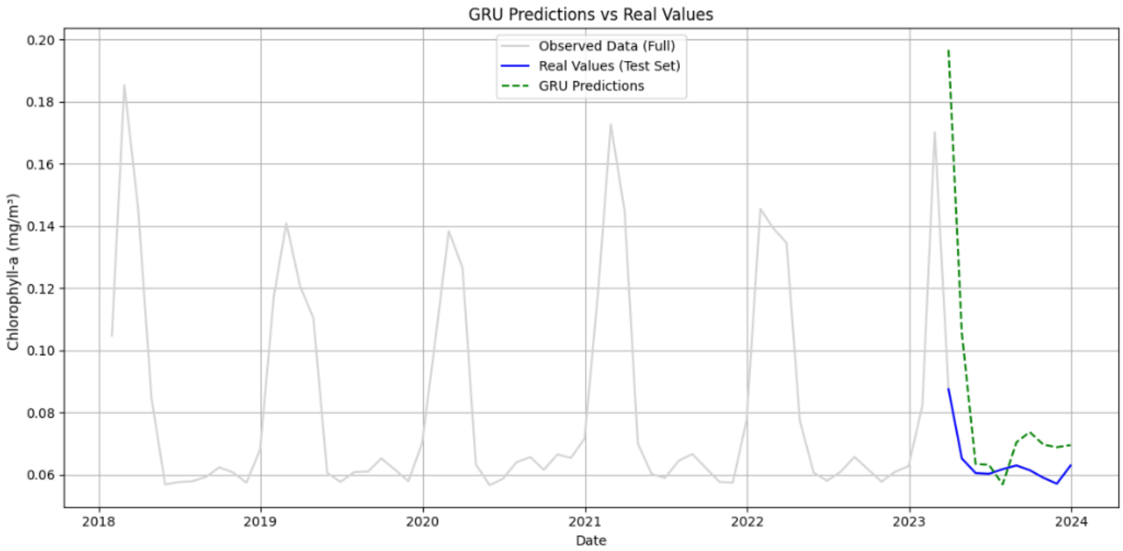

Figure 5.14: Time Series of Chlorophyll-a Using GRU Model

There is, however, a divergence from history right away, with an inordinately large amount of chlorophyll-a at the start of the forecast period as the model drastically overpredicts levels. This early peak is not consistent with recent seasonal historical trend and suggests GRU had been exaggerating the extent of the last bloom or resisting the recent shift in seasonal timing. While it does a little better in the second half of the prediction, the model fails to reproduce fine-scale chlorophyll-a concentration variability.

The fact that the GRU lagged, smoothed prediction is unable to pick up the seasonally abrupt trend that was witnessed means that, like the LSTM, this model struggled to keep up with the complexities of chlorophyll-a over time. This means that recurrent models can be helped by better training protocols, maybe through additional temporal or environmental data, in trying to forecast ecological variables.

Figure 5.15 depicts the prediction of chlorophyll-a concentrations using the RF Regressor model, with every historical data value in grey, test set values in blue, and predicted model values in red.

Evidently, the model is excellent to capture the steep peak in 2023 both in timing and magnitude as well as to indicate the applicability of the model to respond to non-linear, sudden changes, a trend that occurs fairly often within chlorophyll-a dynamics due to recurrent seasonally regular blooms.

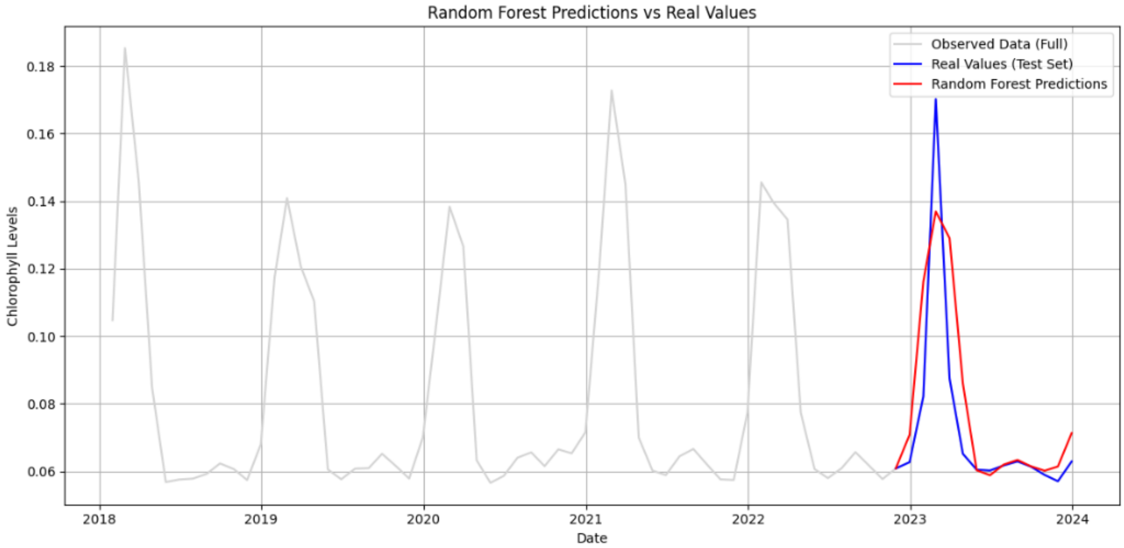

Figure 5.15: Time Series of Chlorophyll-a Using Random Forest Regressor Model

In contrast to the LSTM and GRU models (Figures 5.13 and 5.14), which struggled with under- or overestimation and smearing of important events, the RF Regressor gives a highly sensitive prediction. Its estimation is very close to the actual values throughout the test period, including the sudden increase and drop of chlorophyll-a levels. This proximity implies that this problem is highly appropriate for RF's ensemble approach since it can utilize composite patterns most effectively without assuming linearity and temporal contiguity.

Additionally, the model's ability to generalize from historical seasonal trends while adjusting for recent shifts places it among the most precise techniques of those tested. Its accuracy supports the importance of tree-based ensemble models in environmental time series prediction, especially when recording irregular but periodic biological phenomena such as algal blooms.

For the XGBoost model, Figure 5.16 illustrates a comparison between predicted and actual values produced for chlorophyll-a concentration, plotted against a sample index.

The outcome shows near-perfect congruence of the predicted and actual values, both of which are static at or near zero across all sample points. While this initial suggests a perfect prediction, it is likely to be a breakdown of the model to capture meaningful variation or dynamics in the data.

This static result suggests that the XGBoost model has probably underfit the data, falling back on predicting the mean or mode of the training data, instead of picking up on the patterns or features of chlorophyll-a change. This may be due to a number of factors, including poor feature engineering, too simplistic target representation, or limitation in incorporating temporal dependencies in the input representation. As a result, despite XGBoost's high performance reputation on the majority of regression tasks, its application here leads to uninformative and biologically unrealistic predictions, supporting the argument for better feature design or model tuning in time series use. While XGBoost worked well for classification, it is sensitive to input representation and may underperform in regression settings without adequate tuning (Sujatha et al., (2023)).

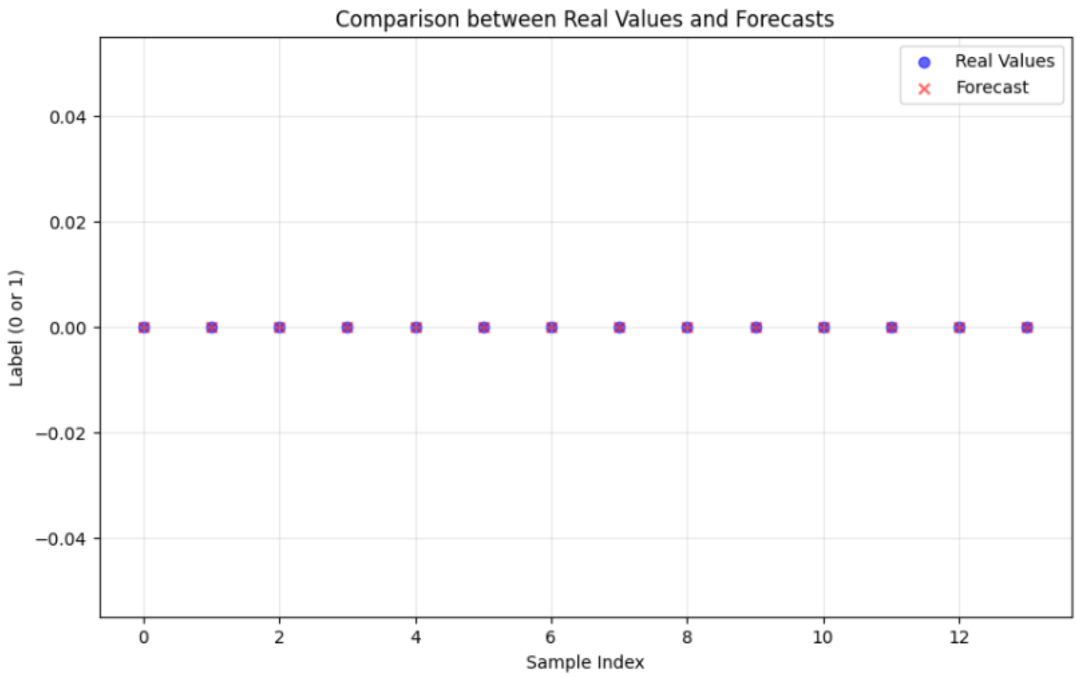

Figure 5.16: Time Series of Chlorophyll-a Using XGBoost Model

Figures 5.17 to 5.20 present chlorophyll-a concentration prediction results generated by CNN models learned with varied sequence lengths of 10, 20, 30, and 50 time steps, respectively. The plots illustrate how the temporal window size input to the model influences its ability to make predictions, particularly when handling intricate ecological time series data such as chlorophyll-a dynamics:

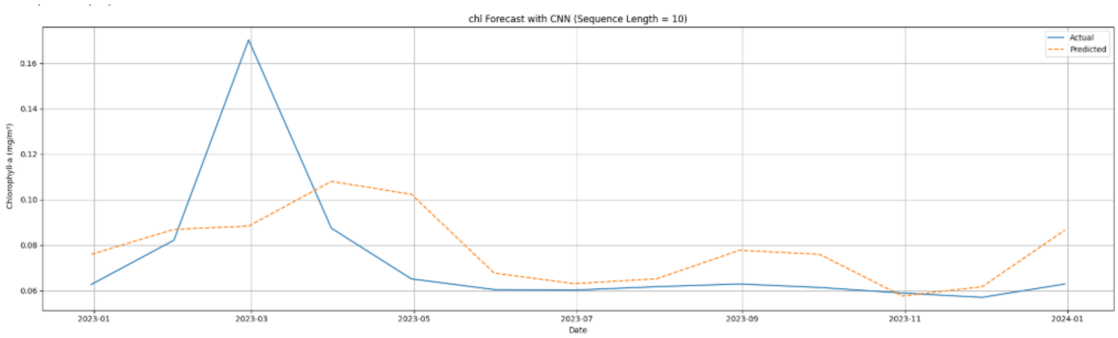

Figure 5.17: Time Series of Chlorophyll-a Using CNN Model - 10 Sequences

In Figure 5.17, the CNN model that is trained on a 10-step sequence makes predictions that reasonably match the general trend of the true chlorophyll-a values. It is, nonetheless, incapable of capturing the actual variability (i.e., missing peaks and overestimating lower values). This is characteristic of an over-smoothed response, which has a tendency to occur when models learn trends rather than nuanced fluctuations.

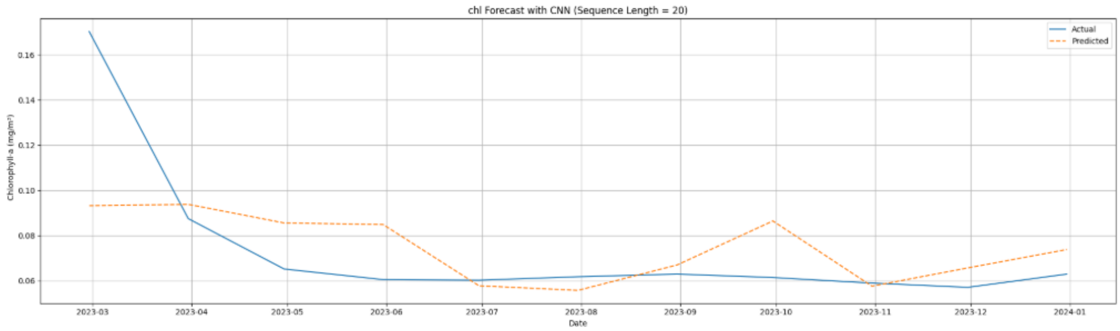

Figure 5.18: Time Series of Chlorophyll-a Using CNN Model - 20 Sequences

When the sequence length is further raised to 20 time steps (Figure 5.18), the model also suffers from the same problem, only with greater loss of accuracy. The model reacts less to short-term variations in chlorophyll-a levels, suggesting that it cannot properly leverage the additional temporal information.

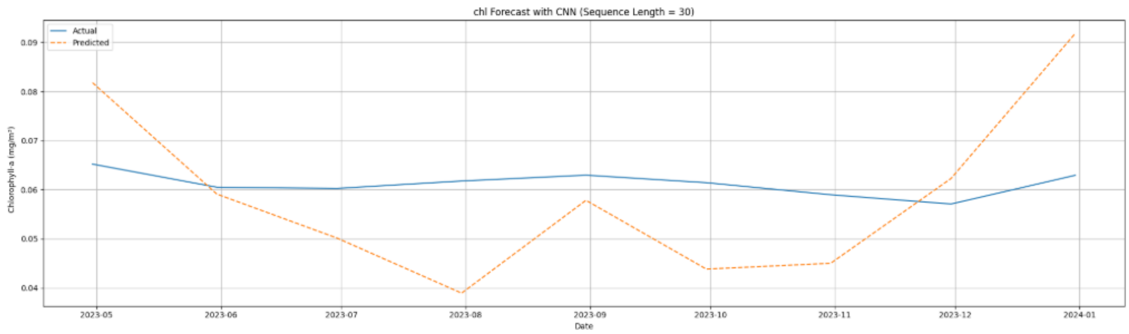

Figure 5.19: Time Series of Chlorophyll-a Using CNN Model - 30 Sequences

The same trend continues in Figure 5.19 as the model trained with 30 sequences begins to make highly volatile predictions. Rather than following the smoother trend of actual chlorophyll-a concentrations across the seasons, the CNN output oscillates wildly and is likely due to the model overfitting noise or collapsing in coherence on the longer input sequences.

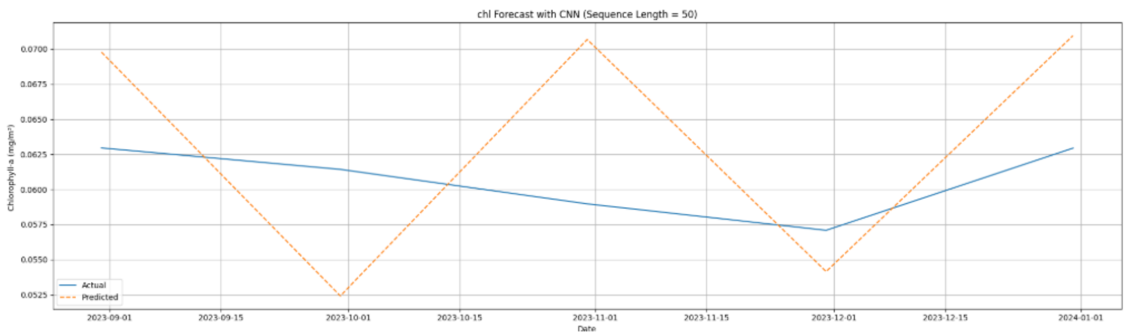

Figure 5.20: Time Series of Chlorophyll-a Using CNN Model - 50 Sequences

The instability peaks in Figure 5.20, where the 50-sequence model shows gross oscillations in a zigzag pattern that is far removed from the actual signal. These unrealistic predictions show a failure in the generalization ability of the model, a common issue when too much past data is given without enough regularization or optimization of the architecture.

In short, while Figures 5.17 to 5.20 indicate how differing sequence length within CNN models influences performance for the prediction of chlorophyll-a, it is possible to observe that longer sequences do not necessarily perform best for improved prediction. In fact, an input window too long destabilizes the model and yields nonsensical results. This reinforces the requirement for accuracy when tuning sequence length in an attempt to create models that are accurate, and interpretable at a biological level. CNN-based architectures, while effective in static chlorophyll-a retrieval, require careful tuning when applied to dynamic time-series modeling (Salah et al., (2023)).

Table 5.4 shows the performance of the group of forecasting models. RF Regressor was found to have the overall best performance among all, with the least MSE (0.0003), RMSE (0.0181), and MAE (0.0110), and the highest R^2 value (0.5940), showing very good predictive accuracy. Recurrent models GRU and LSTM, although possessing relatively low error values, had very negative R^2 values, suggesting poor generalization. The CNN models exhibited very fluctuating performance depending on sequence length, with much lower R^2 values (e.g., -168.94 for sequence length 50), suggesting sensitivity to input layout.

Model	**R^2**	**MSE**	**RMSE**	**MAE**
Random Forest	**0.5940**	**0.0003**	**0.0181**	**0.0110**
CNN (Seq 10)	0.07	0.00	0.03	0.02
XGBoost	0.0000	0.0004	0.0189	0.0189
Linear Regression	-0.0131	0.000775	0.0278	0.0193
CNN (Seq 20)	-0.13	0.00	0.03	0.02
ARIMA	-6.0827	0.0065	0.0804	0.0740
LSTM	-6.8788	0.0005	0.0229	0.0199
SARIMA	-7.7243	0.0067	0.0817	0.0677
SARIMAX	-7.7243	0.0067	0.0817	0.0677
GRU	-15.4540	0.0011	0.0331	0.0165
CNN (Seq 30)	-40.84	0.00	0.01	0.01
CNN (Seq 50)	-168.94	0.00	0.03	0.03

Table 5.4: Performance Metrics for Different Models

5.3 Summary

This chapter conducted an extensive review of a wide range of machine learning and DL prediction and classification algorithms for chlorophyll-a concentrations as a proxy for algal biomass in water bodies. Classification and time series prediction approaches were attempted based on their performance, comprehension, and applicability to real-world scenarios while working with ocean data under transformation and seasonality.

In the classification problem, ensemble models such as RF, Bagging, and XGBoost were both stable and very accurate with ROC AUC values of around or almost 1.00. They were more accurate than standard algorithms such as Logistic Regression and Naïve Bayes, and also more stable than the boosting algorithm such as Adaboost. Feature importance analysis indicated that biogeochemical and spatial attributes (i.e., nppv, phyc, chl, zeu) contributed the most towards predicting chlorophyll-a concentration, while physical attributes such as wave height and wind contributed little.

DL, tree-based, and statistical frameworks were compared on the task of time series forecasting. Conventional statistical models such as ARIMA and SARIMA could not capture seasonality behavior, and SARIMA offered incremental gain with seasonals. Recurrent networks (LSTM, GRU) and CNNs performed poorly, not registering bloom timing and amplitude by smoothing bias or sequence length sensitivity. RF Regressor made the most predictable predictions and tracked season maxima best, and performed better on all error metrics and R^2 score. XGBoost performed best on classification and underfitting for regression tasks and could not meaningfully capture chlorophyll-a dynamics on this timescale. Briefly, the results show that ensemble methods are the best for chlorophyll-a classification tasks and offer great predictive accuracy and identification of major drivers in the environment.

Statistical forecast models fail due to the inflexibility of seasonality, and recurrent models must be trained and still can fail to impress. RF Regressor was a robust contender in the prediction of chlorophyll-a in marine ecosystems with meaningful, seasonally fluctuating output that is suitable for ecological monitoring and decision-support systems.

Chapter 6

Conclusions

This dissertation showed the application of different environmental data sets to predict cyanoHABs using ML approaches. As cyanoHABs are aggressive ecological destabilizers, public health hazards, and economic destabilizers for sectors such as aquaculture and tourism, its predictive performance improvement is not only a scientific necessity but also a socio-environmental necessity. This research shows that the utilization of heterogeneous data sources from satellites, ocean reanalysis, biogeochemical parameters, and so on, are far better than the traditional single-source approaches.

The research possessed a robust pipeline, structured following the CRISP-DM methodology. It began with a systematic literature review, following PRISMA guidelines, to find the most relevant trends and gaps in current ML-based prediction of cyanoHABs. This informed the data acquisition strategy, utilizing three independent Copernicus Marine Service datasets: SST, SSWH, and a range of diverse biogeochemical indicators such as chlorophyll-a, dissolved oxygen, nutrients, and iron. Once the data was cleaned, joined, and explored, predictive models were applied on classification and time-series forecasting tasks.

In terms of classification, ensemble models turned out to be the optimal ones. RF, XGBoost, and LightGBM all consistently delivered high accuracy, ROC AUC, and F1-scores. Not only did these models work well with traditional metrics, but they were also robust in terms of class imbalance as well as complicated feature dependencies. Feature importance comparison among ensemble models picked net primary production, euphotic depth, chlorophyll-a concentration and geospatial variables such as latitude and depth as leading predictors, showing the central role played by biological productivity and light levels in bloom processes.

For time-series prediction, RF Regressor outperformed DL models (e.g., LSTM, GRU, CNN) and conventional statistical models (e.g., ARIMA, SARIMA, SARIMAX). It alone was able to estimate annually the seasonality of peak algae blooming with an acceptable degree of accuracy, reflected in its relatively high R^2 value and low MSE. Recurrent neural networks, however, were not able to accurately simulate seasonal fluctuation and gave smoothed or biologically unrealistic estimates instead. CNN architectures illustrated instability as the sequence length increases, even more reason to possess correct hyperparameter tuning in environmental time-series application.

However, despite the demonstrated benefit of ensemble models, some constraints restrict their global applicability. Environmental data were of good quality and varied but not specifically tailored to Portuguese coastal waters. This geographical mismatch might introduce subtle biases or regional bloom-specific under-representations. Certain variables, especially dissolved micronutrients such as iron, also contained temporal gaps or sparse coverage and likely compromised the completeness of input feature space.

The second is ensemble model interpretation. While more predictive, their internal explanations are effectively clear in comparison to more elementary mechanistic models of domain knowledge. The "black-box" nature could restrict their use among environmental managers, marine biologists, and policy makers requiring clear and explainable tools facilitating informed decision-making.

To reduce these limitations and extend on the basis of this dissertation, some avenues for future research are proposed. First, the collection of high-quality, locally relevant data from Portuguese monitoring effort, e.g., via the Portuguese Institute for Sea and Atmosphere (IPMA) database, would significantly enhance model specificity. Local alliances can provide access to shellfish toxicity reports, bloom frequency histories, and land-based nutrient runoff records, providing further richness to the predictor domain.

Second, expanding the environmental variable scope to incorporate wind direction and speed, solar radiation, precipitation, and turbidity may better describe bloom drivers. All these variables have a tendency to synergistically interact with one another in the development of blooms, particularly in the scenario of stratified or nutrient-rich water.

Third, inclusion of Explainable AI (XAI) methods or attention mechanisms would render the models interpretable and allow hypothesis testing more suitable for the domain. XAI Methods are techniques that describe the extent to which every input feature played a role in a model's prediction. They help the users understand why the model made a specific decision. Attention Mechanisms are neural network components that assign different weights to inputs (for example, some variables or time steps), allowing the model to focus on the most significant information during prediction. It is as essential to understand why a model forecasts a bloom event under specific circumstances as to know if it forecasts one, particularly where the forecast guides aquaculture operations or public health intervention.

Fourth, attempts must be made to develop hybrid model paradigms that merge domain models and ML. Such integration may potentially lead to models that are not only predictive but mechanistically insightful and transferable.

Finally, operational deployment of the models in an operational decision support system would constitute a significant step towards usability. A web or API-based prediction portal in turn fed with real-time Copernicus or in-situ sensor observations could issue actionable alerts to shellfish farms, coastal administrations, and environmental agencies.

Bibliography

Abdullah, Z. et al. (2022). “Multiclass-Classification of Algae using Dc-GAN and Transfer Learning”. In: *2022 2nd International Conference on Image Processing and Robotics (ICIPRob)*. Colombo, Sri Lanka, pp. 1–6. doi: 10.1109/ICIPRob54042.2022.9798730.

Ananias, Pedro Henrique M., Rogério G. Negri, Adriano Bressane, Marilaine Colnago, et al. (2023). “ABD: A Machine Intelligent-Based Algal Bloom Detector for Remote Sensing Images”. In: *Software Impacts* 15, p. 100482. issn: 2665-9638. doi: 10.1016/j.simpa.2023.100482.

Ananias, Pedro Henrique M., Rogério G. Negri, Adriano Bressane, Maurício A. Dias, et al. (2023). “ABF: A Data-Driven Approach for Algal Bloom Forecasting Using Machine Intelligence and Remotely Sensed Data Series”. In: *Software Impacts* 17, p. 100518. issn: 2665-9638. doi: 10.1016/j.simpa.2023.100518.

Aranay, O. M. and P. K. Atrey (Oct. 2022). “Deep Active Genetic Learning-Based Assessment of Lakes’ Water Quality Using Climate Data”. In: *IEEE Transactions on Sustainable Computing* 7.4, pp. 851–863. doi: 10.1109/TSUSC.2022.3163229.

Aslan, Sinem et al. (2022). “Recurrent Neural Networks for Water Quality Assessment in Complex Coastal Lagoon Environments: A Case Study on the Venice Lagoon”. In: *Environmental Modelling & Software* 154, p. 105403. issn: 1364-8152. doi: 10.1016/j.envsoft.2022.105403. url: https://doi.org/10.1016/j.envsoft.2022.105403.

Balcacer, A. et al. (2023). “Mechanics of a Drone-based System for Algal Bloom Detection Utilizing Deep Learning and LLMs”. In: *2023 IEEE MIT Undergraduate Research Technology Conference (URTC)*. Cambridge, MA, USA, pp. 1–5. doi: 10.1109/URTC60662.2023.10534955.

Barrientos-Espillco, Fredy et al. (2023). “Semantic Segmentation Based on Deep Learning for the Detection of Cyanobacterial Harmful Algal Blooms (CyanoHABs) Using Synthetic Images”. In: *Applied Soft Computing* 141, p. 110315. issn: 1568-4946. doi: 10.1016/j.asoc.2023.110315.

Bonnet, Sophie et al. (2018). “Influence of Iron, Cobalt, and Vitamin B_12 Supply on Phytoplankton Growth in the Tropical East Pacific During the 2015 El Niño”. In: *Global Biogeochemical Cycles* 32.9, pp. 1393–1411. issn: 0886-6236. doi: 10.1029/2018GB005968.

Chegoonian, A. M. et al. (2022). “Regional Upscaling of Chlorophyll-A Retrieval from Small Eutrophic Lakes Via Sentinel-2”. In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia, pp. 5622–5625. doi: 10.1109/IGARSS46834.2022.9883793.

Chen, Xiaoqian, Yonggang Fu, and Honghua Zhou (2022). “Conv-Attention Model Based on Multivariate Time Series Prediction: The Cyanobacteria Bloom Case”. In: *Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering (EITCE ’21)*. New York, NY, USA: Association for Computing Machinery, pp. 1018–1023. doi: 10.1145/3501409.3501591. url: https://doi.org/10.1145/3501409.3501591.

Cho, Sohyun et al. (2014). "Factors Affecting Algal Blooms in a Man-made Lake and Prediction Using an Artificial Neural Network". In: *Measurement* 53, pp. 224–233. issn: 0263-2241. doi: 10.1016/j.measurement.2014.03.044.

Choi, Yongjin et al. (2021). "Using Artificial Intelligence Techniques for Evidence-Based Decision Making in Government: Random Forest and Deep Neural Network Classification for Predicting Harmful Algal Blooms in New York State". In: *Proceedings of the 22nd Annual International Conference on Digital Government Research (dg.o '21)*. New York, NY, USA: Association for Computing Machinery, pp. 27–37. doi: 10.1145/3463677.3463713. url: https://doi.org/10.1145/3463677.3463713.

Du, Yichen et al. (2024). "Cyanobacterial Blooms Prediction in China's Large Hypereutrophic Lakes Based on MODIS Observations and Bayesian Theory". In: *Journal of Hazardous Materials* 480, p. 136057. issn: 0304-3894. doi: 10.1016/j.jhazmat.2024.136057.

Fang, Z. et al. (2022). "Deep Learning Based Algae Detection Method". In: *2022 6th International Conference on Universal Village (UV)*. Boston, MA, USA, pp. 1–5. doi: 10.1109/UV56588.2022.10185530.

Fonseca, Joana et al. (2021). "Algal Bloom Front Tracking Using an Unmanned Surface Vehicle: Numerical Experiments Based on Baltic Sea Data". In: *OCEANS 2021: San Diego–Porto*, pp. 1–6. isbn: 978-1-6654-3868-9. doi: 10.1109/OCEANS44145.2021.9705825.

Grendaitė, D. and L. Petkevičius (2024). "Identification of Algal Blooms in Lakes in the Baltic States Using Sentinel-2 Data and Artificial Neural Networks". In: *IEEE Access* 12, pp. 27973–27988. doi: 10.1109/ACCESS.2024.3366491.

Hong, Seok Min et al. (2023). "Autonomous Calibration of EFDC for Predicting Chlorophyll-a Using Reinforcement Learning and a Real-Time Monitoring System". In: *Environmental Modelling & Software* 168, p. 105805. issn: 1364-8152. doi: 10.1016/j.envsoft.2023.105805. url: https://doi.org/10.1016/j.envsoft.2023.105805.

Inokuchi, Y. et al. (2024). "Development of a Red Tide Early Detection System Using Satellite Images". In: *2024 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*. Miri Sarawak, Malaysia, pp. 159–163. doi: 10.1109/GECOST60902.2024.10474956.

Instituto Politécnico do Porto (n.d.). *Regulamento do Código de Boas Práticas e de Conduta do Instituto Politécnico do Porto*. Accessed April 2025. url: https://www.iscap.ipp.pt/regulamentos/CodigoboaspraticasedecondutaIPP.pdf.

Jia, Y. et al. (2024). "Machine Learning Analysis of CYGNSS Data for Chlorophyll Concentration Monitoring in Algal Blooms". In: *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*. Athens, Greece, pp. 1025–1028. doi: 10.1109/IGARSS53475.2024.10642292.

Karmakar, Jay et al. (2024). "Analyzing Spatio-Temporal Variability of Aquatic Productive Components in Northern Bay of Bengal Using Advanced Machine Learning Models". In: *Ocean & Coastal Management* 251, p. 107074. issn: 0964-5691. doi: 10.1016/j.ocecoaman.2024.107074. url: https://doi.org/10.1016/j.ocecoaman.2024.107074.

Khan, R. M., B. Salehi, and M. Mahdianpari (2022). "Machine Learning Methods for Water Quality Monitoring Over Finger Lakes Using Sentinel-2". In: *IGARSS 2022 - 2022 IEEE International Geoscience and Remote Sensing Symposium*. Kuala Lumpur, Malaysia, pp. 6316–6319. doi: 10.1109/IGARSS46834.2022.9883230.

Lanuza, R. V., R. D. F. Magday, and A. N. Yumang (2024). "Development of a Database for the Classification of the Severity of Algal Bloom with the Implementation of Support Vector Machine". In: *2024 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET)*. Kota Kinabalu, Malaysia, pp. 418–423. doi: 10.1109/IICAIET62352.2024.10730345.

Li, B. et al. (2023). "Identification of Cyanobacteria for Harmful Algal Blooms Research Using the YOLO Framework". In: *2023 IEEE 14th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*. New York, NY, USA, pp. 0407–0415. doi: 10.1109/UEMCON59035.2023.10316078.

Luo, X. et al. (2024). "HabLSTM: A Nonstationary Feature Focusing LSTM for Spatiotemporal Prediction of Harmful Algal Bloom". In: *IEEE Transactions on Geoscience and Remote Sensing* 62.Art no. 5638817, pp. 1–17. doi: 10.1109/TGRS.2024.3416293.

Ly, Quang Viet et al. (2021). "Application of Machine Learning for Eutrophication Analysis and Algal Bloom Prediction in an Urban River: A 10-Year Study of the Han River, South Korea". In: *Science of the Total Environment* 797, p. 149040. issn: 0048-9697. doi: 10.1016/j.scitotenv.2021.149040.

Lyu, X. et al. (2023). "Classification With Unbalanced Samples by Self-Sampling and Semi-correlated Co-Training—An Application to Algal Bloom Detection". In: *IEEE Transactions on Geoscience and Remote Sensing* 61.Art no. 4206612, pp. 1–12. doi: 10.1109/TGRS.2023.3299312.

Marrone, B. L. et al. (2024). "Toward a Predictive Understanding of Cyanobacterial Harmful Algal Blooms through AI Integration of Physical, Chemical, and Biological Data". In: *ACS ES&T Water* 4.3, pp. 844–858. doi: 10.1021/acsestwater.3c00369.

Martín-Suazo, Silvia et al. (2024). "Deep Learning Methods for Multi-Horizon Long-Term Forecasting of Harmful Algal Blooms". In: *Knowledge-Based Systems* 301, p. 112279. issn: 0950-7051. doi: 10.1016/j.knosys.2024.112279. url: https://doi.org/10.1016/j.knosys.2024.112279.

Martino, Simone, Fatima Gianella, and Keith Davidson (2020). "An Approach for Evaluating the Economic Impacts of Harmful Algal Blooms: The Effects of Blooms of Toxic *Dinophysis* spp. on the Productivity of Scottish Shellfish Farms". In: *Harmful Algae* 99, p. 101912. issn: 1568-9883. doi: 10.1016/j.hal.2020.101912.

McKinley, Emma et al. (2024). "The Human Dimensions of Harmful Algal Blooms: An Evolving Research Agenda". In: *Ocean & Coastal Management* 259, p. 107432. issn: 0964-5691. doi: 10.1016/j.ocecoaman.2024.107432.

Molares-Ulloa, Andres, Enrique Fernandez-Blanco, et al. (2022). "Machine Learning in Management of Precautionary Closures Caused by Lipophilic Biotoxins". In: *Computers and Electronics in Agriculture* 197, p. 106956. issn: 0168-1699. doi: 10.1016/j.compag.2022.106956. url: https://doi.org/10.1016/j.compag.2022.106956.

Molares-Ulloa, Andres, Daniel Rivero, et al. (2023). "Hybrid Machine Learning Techniques in the Management of Harmful Algal Blooms Impact". In: *Computers and Electronics in Agriculture* 211, p. 107988. issn: 0168-1699. doi: 10.1016/j.compag.2023.107988. url: https://doi.org/10.1016/j.compag.2023.107988.

Moore, C. M., S. Dutkiewicz, et al. (2023). "Global analysis of ocean phytoplankton nutrient limitation reveals high prevalence of co-limitation". In: *Nature Geoscience* 16, pp. 697–704. issn: 1752-0894. doi: 10.1038/s41561-023-01286-4.

Moore, C. M., M. M. Mills, et al. (2013). "Processes and Patterns of Oceanic Nutrient Limitation". In: *Nature Geoscience* 6.9, pp. 701–710. doi: 10.1038/ngeo1765. url: https://doi.org/10.1038/ngeo1765.

Nikoo, Mohammad Reza et al. (2025). "A review of machine learning, remote sensing, and statistical methods for reservoir water quality assessment". In: *Journal of Hydrology* 659, p. 133323. issn: 0022-1694. doi: 10.1016/j.jhydrol.2025.133323.

O'Donncha, Fearghal et al. (2024). "Forecasting closures on shellfish farms using machine learning". In: *Aquaculture International* 32, pp. 5603–5623. doi: 10.1007/s10499-024-01438-y.

Ordem dos Engenheiros (n.d.). *Código de Ética e Deontologia*. Accessed April 2025. url: https://www.ordemengenheiros.pt/fotos/editor2/regulamentos/codigo_ed.pdf.

Pacilio, E., A. Silvarrey, and A. Pardo (2024). "Sentinel 3 OLCI and Machine Learning for Cyanobacteria Bloom Detection Over Small Inland Water Target". In: *2024 IEEE 15th Latin America Symposium on Circuits and Systems (LASCAS)*. Punta del Este, Uruguay, pp. 1–5. doi: 10.1109/LASCAS60203.2024.10506157.

Painter, Stuart C. (2024). "Observations of phytoplankton productivity and growth rates in the Malin shelf break environment". In: *Continental Shelf Research* 279, p. 105281. issn: 0278-4343. doi: 10.1016/j.csr.2024.105281.

Paul, Jane T., N. Ramaiah, and S. Sardessai (2008). "Nutrient regimes and their effect on distribution of phytoplankton in the Bay of Bengal". In: *Marine Environmental Research* 66.3, pp. 337–344. issn: 0141-1136. doi: 10.1016/j.marenvres.2008.05.007.

Penne, C. et al. (2023). "Independent Component Analysis: A Tool for Algal Bloom Detection". In: *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. Athens, Greece, pp. 1–5. doi: 10.1109/WHISPERS61460.2023.10430844.

Rahaman, Md Hibjur et al. (2024). "Surface Water Quality Prediction in the Lower Thoubal River Watershed, India: A Hyper-Tuned Machine Learning Approach and DNN-Based Sensitivity Analysis". In: *Journal of Environmental Chemical Engineering* 12.3, p. 112915. issn: 2213-3437. doi: 10.1016/j.jece.2024.112915. url: https://doi.org/10.1016/j.jece.2024.112915.

Ramaraj, M. and R. Sivakumar (2023). "Integration of Band Regression Empirical Water Quality (BREWQ) Model with Deep Learning Algorithm in Spatiotemporal Modeling and Prediction of Surface Water Quality Parameters". In: *Modeling Earth Systems and Environment* 9.1, pp. 3279–3304. issn: 2363-6203. doi: 10.1007/s40808-023-01695-5.

Ramgirkar, Atharva et al. (2024). "Mining of Heterogeneous Time Series Information for Predicting Chlorophyll Accumulation in Oceans". In: *Sustainable Computing: Informatics and Systems* 42, p. 100980. issn: 2210-5379. doi: 10.1016/j.suscom.2024.100980. url: https://doi.org/10.1016/j.suscom.2024.100980.

Rostam, N. A. P. et al. (2021). "A Complete Proposed Framework for Coastal Water Quality Monitoring System With Algae Predictive Model". In: *IEEE Access* 9, pp. 108249–108265. doi: 10.1109/ACCESS.2021.3102044.

Salah, M. et al. (2023). "B1D–CNN: A Novel Convolution Neural Network-Based Chlorophyll-A Retrieval Algorithm for Sentinel–2 Data". In: *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*. Pasadena, CA, USA, pp. 3950–3953. doi: 10.1109/IGARSS52108.2023.10281795.

Schröer, Christoph, Felix Kruse, and Jorge Marx Gómez (2021). "A Systematic Literature Review on Applying CRISP-DM Process Model". In: *Procedia Computer Science* 181, pp. 526–534. issn: 1877-0509. doi: 10.1016/j.procs.2021.01.199.

Sohrabi, C. et al. (2021). "PRISMA 2020 Statement: What's New and the Importance of Reporting Guidelines". In: *International Journal of Surgery* 88, p. 105918. issn: 1743-9191.

doi: 10.1016/j.ijsu.2021.105918. url: https://doi.org/10.1016/j.ijsu.2021.105918.

Srilatha, K. et al. (2023). “Prediction of Algal Bloom and Its Effects on Aquaculture in Coastal Area Using MODIS Dataset and Machine Learning Techniques”. In: *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*. Chennai, India, pp. 1–9. doi: 10.1109/ACCAI58221.2023.10200562.

Stow, Craig A., Yuntao Guan, and Douglas R. Dolan (2003). “Short-term Lake Erie algal bloom prediction by classification and regression tree models”. In: *Journal of Great Lakes Research* 29.4, pp. 627–638. issn: 0380-1330. doi: 10.1016/S0380-1330(03)70466-3.

Sujatha, M. G., P. Ranganathan, and S. Korom (2023). “Classification of Microcystins Based on Nutrient Concentration Using Machine Learning Models”. In: *2023 IEEE World AI IoT Congress (AIIoT)*. Seattle, WA, USA, pp. 0535–0541. doi: 10.1109/AIIoT58121.2023.10174561.

Tiwary, Sanjeeb et al. (2023). “Prediction of Algae Growth: A Machine Learning Perspective”. In: *Proceedings of the 2023 Fifteenth International Conference on Contemporary Computing (IC3-2023)*. New York, NY, USA: Association for Computing Machinery, pp. 109–114. doi: 10.1145/3607947.3607967. url: https://doi.org/10.1145/3607947.3607967.

Wang, Binbin et al. (2024). “Algae Species Classification Based on Few-Shot Learning: A Case Study of Macau Freshwater”. In: *Proceedings of the 2023 9th International Conference on Communication and Information Processing (ICCIP ’23)*. New York, NY, USA: Association for Computing Machinery, pp. 133–137. doi: 10.1145/3638884.3638904. url: https://doi.org/10.1145/3638884.3638904.

Wang, H., E. Galbraith, and M. Convertino (2023). “Risky Blooms: Space-Time Chlorophyll-a Analysis and Forecasting”. In: *2023 22nd International Symposium on Communications and Information Technologies (ISCIT)*. Sydney, Australia, pp. 252–256. doi: 10.1109/ISCIT57293.2023.10376115.

Wang, Li et al. (2023). “Remote Sensing Image Analysis and Cyanobacterial Bloom Prediction Method Based on ACL3D-Pix2Pix”. In: *Desalination and Water Treatment* 297, pp. 138–159. issn: 1944-3986. doi: 10.5004/dwt.2023.29646.

Weir, Michael J., Melina Kourantidou, and Di Jin (2022). “Economic Impacts of Harmful Algal Blooms on Fishery-Dependent Communities”. In: *Harmful Algae* 118, p. 102321. issn: 1568-9883. doi: 10.1016/j.hal.2022.102321.

Wen, Jiabao et al. (2022). “Harmful Algal Bloom Warning Based on Machine Learning in Maritime Site Monitoring”. In: *Knowledge-Based Systems* 245, p. 108569. issn: 0950-7051. doi: 10.1016/j.knosys.2022.108569. url: https://doi.org/10.1016/j.knosys.2022.108569.

Wen, Bi-Yao et al. (2024). “Multi-Scale Feature Extraction and TrasMLP Encoder Module for Ocean HABs Segmentation”. In: *Ocean Engineering* 311.Part 1, p. 118947. issn: 0029-8018. doi: 10.1016/j.oceaneng.2024.118947. url: https://doi.org/10.1016/j.oceaneng.2024.118947.

Wu, H. et al. (2022). “Deep Learning to Model the Complexity of Algal Bloom”. In: *2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. Suzhou, China, pp. 114–122. doi: 10.1109/CyberC55534.2022.00027.

Xia, Wei, Ilija Ilievski, and Christine Ann Shoemaker (2024). “Enhancing Algal Bloom Forecasting: A Novel Framework for Machine Learning Performance Evaluation During Periods of Special Temporal Patterns”. In: *Environmental Modelling & Software* 180, p. 106164. issn: 1364-8152. doi: 10.1016/j.envsoft.2024.106164. url: https://doi.org/10.1016/j.envsoft.2024.106164.

Yang, C. et al. (2023). “A Comparative Analysis of Machine Learning Methods for Algal Bloom Detection Using Remote Sensing Images”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 16, pp. 7953–7967. doi: 10.1109/JSTARS.2023.3310162.

Yang, M. et al. (2021). “Automatic Recognition of Harmful Algae Images Using Multiple CNNs”. In: *2021 International Conference on Computer Engineering and Artificial Intelligence (ICCEAI)*. Shanghai, China, pp. 276–280. doi: 10.1109/ICCEAI52939.2021.00055.

Yoon, Joo-Eun et al. (2018). “Reviews and syntheses: Ocean iron fertilization experiments – past, present, and future looking to a future Korean Iron Fertilization Experiment in the Southern Ocean (KIFES) project”. In: *Biogeosciences* 15, pp. 5847–5889. doi: 10.5194/bg-15-5847-2018. url: https://doi.org/10.5194/bg-15-5847-2018.

Zhang, Huiyan, Bo Hu, Xiaoyi Wang, Li Wang, et al. (2022). “An Echo State Network Based Adaptive Dynamic Programming Approach for Time-Varying Parameters Optimization with Application in Algal Bloom Prediction”. In: *Applied Soft Computing* 122, p. 108796. issn: 1568-4946. doi: 10.1016/j.asoc.2022.108796. url: https://doi.org/10.1016/j.asoc.2022.108796.

Zhang, Huiyan, Bo Hu, Xiaoyi Wang, Jiping Xu, et al. (2021). “Self-Organizing Deep Belief Modular Echo State Network for Time Series Prediction”. In: *Knowledge-Based Systems* 222, p. 107007. issn: 0950-7051. doi: 10.1016/j.knosys.2021.107007. url: https://doi.org/10.1016/j.knosys.2021.107007.

Zhang, Sheng et al. (2024). “Spatial and Seasonal Variations in Iron and the Response of Chlorophyll-a in Zhanjiang Bay, China”. In: *Water* 16.16, p. 2338. issn: 2073-4441. doi: 10.3390/w16162338.

Zhu, Liudi et al. (2024). “Robust Remote Sensing Retrieval of Key Eutrophication Indicators in Coastal Waters Based on Explainable Machine Learning”. In: *ISPRS Journal of Photogrammetry and Remote Sensing* 211, pp. 262–280. issn: 0924-2716. doi: 10.1016/j.isprsjprs.2024.04.007.