

https://dspace.org/

Assessing and comparing the performance of different machine learning regression algorithms in predicting Chlorophyll-a concentration in the Vaal Dam, Gauteng

Mahamuza, Phemelo Hope

2024-03
University of the Witwatersrand, Johannesburg
https://hdl.handle.net/10539/45165

Mahamuza, Phemelo Hope. (2024). Assessing and comparing the performance of different machine learning regression algorithms in predicting Chlorophyll-a concentration in the Vaal Dam, Gauteng. [Master's dissertation, University of the Witwatersrand, Johannesburg]. WIReDSpace. https://hdl.handle.net/10539/45165

Downloaded from a DSpace repository. https://dspace.org/



Assessing and comparing the performance of different machine learning regression algorithms in predicting Chlorophyll-a concentration in the Vaal Dam, Gauteng

This research report is submitted to the University of the Witwatersrand, Faculty of Science Johannesburg, in partial fulfilment of the requirements of my master's degree.

(M.Sc. Geographic information science and Remote sensing)

Phemelo Hope Mahamuza, 1370472

Supervised by: Prof Elhadi Adam

Declaration

I, Phemelo Mahamuza, at this moment, declare that the thesis entitled "**Assessing and comparing the performance of different machine learning regression algorithms in predicting Chlorophyll-a concentration in the Vaal Dam, Gauteng.**" is my original work, conducted without any external assistance. This thesis is being submitted to fulfil the requirements for the M.Sc. in GIS and Remote Sensing degree from the University of the Witwatersrand, Johannesburg. This Thesis has yet to be previously presented for any qualification or examination at any other university.



_____

Candidate Signature

25 March 2024

Abstract

The state of Vaal Dam is influenced by various land uses surrounding the Dam, including agricultural activities, mining operations, industrial enterprises, urban settlements, and nature reserves. Mining activities, farming practices, and sewage outflows from nearby villages led to access contamination within the Dam, increasing algal bloom levels. Sentinel-2 MSI data were utilized to forecast and comprehend the spatial pattern of Chlorophyll-a concentration, indicating algal bloom occurrence in the Vaal Dam. Targeting Sentinel-2 Level-1C, the image was preprocessed on the Google Earth Engine (GEE) with acquisition dates from 25 – 26 October 30, 2016, corresponding to the on-site data collection between October 26 and October 28, 2016. Due to limited resources, up-to-date data on the Vaal Dam could not be collected. However, since this study focuses on applying various machine learning regression models to predict chlorophyll-a levels in waterbodies, the dataset is used to test the models rather than reflect the current state of the Vaal Dam. The dataset, comprising 23 samples, was divided into 70% training and 30% test sets, allowing for comprehensive model evaluation. Band ratio reflectance values were extracted from the satellite image and correlated with in-field Chlorophyll-a values. The highest correlation coefficient values were utilized to train five machine-learning models employed in this study: Random Forest (RF), Support Vector Regression (SVR), Least Absolute Shrinkage and Selection Operator (LASSO), Ridge Regression, and Multilinear Regression (MLR). Each model underwent training with ten iterations each; the best learning iteration was then used to generate the final Chlorophyll-a predictive model. The predictive models were validated using the Sentinel-2 MSI satellite data and in-situ measurements using R^2, RMSE, and MAPE. Among the five machine learning algorithms trained, RF performed the best, with an R^2 of 0.86 and 0.95, an RMSE of 1.38 and 0.8, and MAPE of 15.09% and 10.92% for the training and testing sets, respectively, indicating its ability to handle small, non-linear datasets. SVR also demonstrated a fair performance, particularly in handling multicollinearity in the data points with an R^2 of 0.68 and 0.87, an RMSE of 2.37 and 1.56, and MAPE of 18.13% and 19.28% for the training and testing sets, respectively. The spatial pattern of Chlorophyll-a concentrations, mapped from the RF model, indicated that high concentrations of Chlorophyll-a are along the Dam shorelines, suggesting a significant impact of land use activities on pollution levels. This study emphasizes the importance of selecting suitable machine learning algorithms tailored to the dataset's characteristics. RF and SVR

demonstrated proficiency in handling nonlinearity, with RF displaying enhanced generalization and resistance to overfitting. Limited field data evenly distributed across the Dam and satellite overpass dates may affect result accuracy. Future research should align satellite pass dates with fieldwork dates and ensure an even distribution of in-field samples across the Dam to represent all land uses and concentration levels.

Acknowledgment

I extend my heartfelt gratitude to God for guiding me throughout this journey, as His timing has been perfect: "When the time is right, I, the Lord, will make it happen." I am also deeply thankful to my wonderful family (Matshidiso, and Oratile Mahamuza) and supportive partner Samuel Oladejo for their unwavering encouragement.

Special appreciation goes to Dr. Laven Naidoo and Dr. Samkelisiwe Khanyile from GCRO for their invaluable assistance and meticulous proofreading of my work. Additionally, I thank Mr. Phathutshedzo Ratshiedana for his availability and support during consultations.

Finally, I am immensely grateful to my supervisor, Prof. Elhadi Adam, for consistently challenging me to strive for excellence. I also want to acknowledge the NRF for sponsoring my Master's fees.

Table of Contents

Declaration 2

Abstract 3

Acknowledgment 4

List of Figures 7

List of Tables 8

List of Abbreviations 9

Chapter One 11

Introduction 11

1.1 Background 11

1.2 Problem statement 12

1.3 Aim and Objectives 13

Aim: 14

Objectives: 14

1.4 Research report outline. 14

Chapter Two 15

Literature Review 15

2.1 Algal bloom of freshwater resources 15

2.2 The importance of Algae monitoring in surface water bodies. 17

2.3 The implementation of sentinel-2 in Chlorophyll-a monitoring 18

2.4 Using machine learning algorithms in assessing Chlorophyll-a in water. 20

Chapter 3 23

Methodology 23

..... 24

3.1 Study site 25

3.2 Methodology 26
3.2.1 Remote sensing data acquisition and image pre-processing. 26
3.2.2 In-situ field measurements 27
3.2.3 Band Ratio Correlation 30
3.2.4 Machine Learning Regression Algorithms 31
3.2.5 Model tunning 35
3.2.6. Final model training 38
3.2.7 Accuracy assessment 42
Chapter 4 44
Results 44
4.1.1 Descriptive statistics 44
4.1.2 Chlorophyll-a results 45
4.1.2 Machine Learning Model Outputs 46
4.1.2 Mapping concentration based on best performing model. 49
Chapter 5 51
Discussion and Conclusion 51
5.1 Distribution of in-situ Chlorophyll-a measurements 51
5.2 Relationship between band ratio reflectance and predicted Chlorophyll-a values 51
5.3 Findings obtained from five machine learning models. 52
5.4 Spatial pattern for Chlorophyll-a concentration in the Vaal Dam. 54
5.5 Conclusion 54
5.6 Limitations and recommendation 55
Reference list 57

List of Figures

Figure 3.1: Summary of methodology used in this study. 24
Figure 3.2: Map indicating the Vaal Dam study site. 26
Figure 3.3: 23 sampling locations. 29
Figure 3.4: Random Forest decision trees overview, adapted from Rodriguez-Galiano et al. (2016). 32
Figure 4.1: R-squared comparison for different machine learning models. 48
Figure 4.2: RMSE comparison for different machine learning models. 49
Figure 4.3: MAPE comparison for different machine learning models 49
Figure 4.4: Final Random Forest model Chlorophyll-a concentration maps. 50

List of Tables

Table 2.1: Trophic algal bloom levels adapted from Boyde et al. (1997) and Tanjeck (2019). 15
Table 2.2: Statistical analysis for Landsat 8, adapted from Bande and Adam (2018). 19
Table 2.3: Statistical analysis for Sentinel-2 adapted from Bande and Adam (2018). 20
Table 2.4: A summary of the evaluation metrics obtained from the training and validation sets is presented for all three of the distinct artificial neural network models created for this research adapted from Zhou et al. (2015). 21
Table 3.1: Sentinel-2MSI sensor and spectral properties. 27
Table 3.3: Random Forests parameter tuning 36
Table 3.4: Support vector regression parameter tuning. 36
Table 3.5: LASSO regression tunning. 37
Table 3.6: Ridge regression tunning. 38
Table 4.1: Descriptive statistics for field measurement data. 44
Table 4.2: Chlorophyll-a Sentinel-2 MSI band correlation coefficients (R). 45
Table 4.3: Summary table for Machine learning model outputs. 47

List of Abbreviations

ANN	Artificial Neural Networks
ARC	Agricultural Research Council
CDOM	Colored Dissolved Organic Matter
COD	Chemical Oxygen Demand
DO	Dissolved Oxygen
DT	Decision Tree Regression
EC	Electrical Conductivity
ESA	European Space Agency
ET	Extra Tree Regression
GEE	Google Earth Engine
GEMI	Global Environmental Monitoring Index
GLM	Generalized Linear Models
HAB	Harmful Algal Bloom
LASSO	Least Absolute Shrinkage and Selection Operator
LR	Linear Regression
Lz	Lacustrine Zone
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MLR	Multilinear Regression
MSI	Multi-Spectral Imagery
NIR	Near Infrared
NP	Nutrient-Phosphorus
NTU	Nephelometric Turbidity Unit
PCA	Principal Component Analysis
RBF	Radial Base Function
RF	Random Forest
RMSE	Root Mean Square Error
Rz	Riverine Zone
SVM	Support Vector Machine
SVR	Support Vector Regression

TN	Total Nitrogen
TOA	Top-Of-Atmosphere
TP	Total Phosphorus
TSS	Total Suspended Solids
Tz	Transitional Zone
WT	Water Temperature

Chapter One

Introduction

1.1 Background

Rivers, Lakes, and Dam reservoirs are examples of land surface water sources. They are essential components of the earth's water cycle and are therefore necessary for the worldwide ecosystem and climate change (Du et al., 2016). Papa et al. (2008) emphasize the significance of mapping ground surface water bodies and determining their geographical distribution for studying hydrological processes and managing available water resources. Remote sensing has emerged as a crucial instrument for monitoring inland bodies of water since the data gives Macroscopic, instantaneously, effective, and reasonably priced data that is notably different from conventional in-situ measurement (Singhal et al., 2019).

On June 23rd, 2015, the European Space Agency (ESA) launched the Sentinel-2 MSI, a novel optical satellite with high spatial resolution. Sentinel-2's ability to acquire multispectral images with an acceptable spatial resolution and high return frequency is crucial for developing future operational outputs like landcover change detection maps and geographical variabilities (Asim et al., 2021). Due to its appealing qualities (10-meter (m) spatial resolution and 10-day revisit frequency) and open access, the Sentinel-2 image would undoubtedly be significant in mapping regional water bodies (Caballero et al., 2019). Although specific satellite systems can provide imagery with a higher resolution and shorter revisit time than Sentinel-2 (WorldView spatial resolutions consist of 1.24 centimetres (cm) and 30 cm in multispectral and panchromatic mode, respectively; GeoEye-1 has 41 cm in panchromatic and 1.65 m in multispectral mode; and QuickBird has 2.44 m in multispectral mode and 61 cm in panchromatic), they are usually commercial and costly to acquire. Specifically, when dealing with algal blooms as water quality indicators, Binding et al. (2018) state that algae's nature is temperature and time-sensitive. An algal bloom is the accumulation of algae in freshwater due to the input of nutrients from agricultural fertilizers and sewage systems (Kasprzak et al., 2008). The algal biomass and trophic condition can be determined by measuring and monitoring the Chlorophyll-a concentration (Kasprzak et al.,

2008). As a result, Chlorophyll-a is utilized as an ancillary indication of nutrient levels and algal concentration in water (Basak et al., 2021).

Due to its ability to handle large datasets and complex patterns, machine learning has emerged as a preferred technique for predicting water quality indicators such as Chlorophyll-a in remote sensing when combined with field data. Machine learning models increase the accuracy of predictions by combining data from on-site measurements and remote sensing data. This method provides insightful analyses of temporal and spatial variations and is especially beneficial for accounting for the inherent variability in water quality factors. Furthermore, machine learning can identify intricate patterns and relationships from various datasets; it is a valuable real-time water quality monitoring tool. Machine learning regression models are more adaptable and can handle multicollinearity while capturing non-linear relationships, unlike traditional statistical models that depend on solid theories about the inherent data distribution (Sarker, 2021). This study applies various machine learning regression models to predict Chlorophyll-a concentrations in water bodies, leveraging remote sensing data, particularly from the Sentinel-2 satellite, to provide a more comprehensive analysis of water quality.

1.2Problem statement

In recent years, river water quality prediction modeling has been among the numerous environmental engineering problems for which machine learning models have been broadly used as solutions. Researchers often choose popular regression techniques like Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), Artificial Neural Networks (ANN), Extra tree regression (ET), Decision Tree regression (DT), Random Forest regression (RF), Support vector regression (SVR), and Linear regression (LR)for forecasting and prediction of a variety of environmental problems such as water quality (Rattan et al., 2022). However, based on David Wolpert's "No Free Lunch Theorem" about machine learning and optimization algorithms (Wolpert and Macready, 1997). The theorem states that no algorithm can be optimized to work well on every problem or dataset. Put another way, no algorithm is inherently better than any other and can perform flawlessly in every scenario. As a result, specific algorithms may work well in some contexts but poorly in others (Wolpert and Macready, 1997).

When selecting an algorithm, this theorem emphasizes how crucial it is to comprehend the properties of the data and problem domain. It means that an algorithm's efficiency depends on the context in which it is used, meaning that an algorithm that solves a particular problem well might not solve another one as effectively. Hence, understanding the efficiency of satellite imagery when dealing with machine learning forecasting models is extremely important as remotely sensed data is usually more complex; for example, the data may not be linear for the model to train accurately, Han and Goodenough (2008) stated that most multispectral and hyperspectral data are commonly non-linear, this is attributed by data formulation involving the use of multiple sources, each of which contributes to the final product. The primary source of nonlinearity is frequently thought to be the variations in sun-canopy-sensor geometry combined with multiple scattering between solar radiation and targets. Another source of nonlinearity is a nonlinear attenuating medium, like water. Additionally, nonlinearity is influenced by the heterogeneity of pixel composition (Han and Goodenough 2008; Sardooi et al., 2019).

Since satellite imagery is prone to being non-linear, machine learning models should aim to address this problem. In areas affected by monsoons, Mamun et al. (2019) evaluated the effectiveness of multiple linear regression (MLR) and SVR for predicting Chlorophyll-a concentrations and water clarity in reservoirs. They found that SVR performed better than MLR. The authors proposed that MLR's poor performance was caused by its inability to solve highly complicated nonlinearity problems. This result emphasizes the importance of choosing suitable models to represent non-linear relationships in the data. Other problems facing predictive machine learning models in remote sensing are Multicollinearity (Sardooi et al., 2019), overfitting, and underfitting (Rattan et al., 2022). Furthermore, there is a scarcity of research utilizing machine learning methodologies, like SVR and RF, for evaluating coastal or inland water quality through remote sensing data despite their benefits, such as their ability to predict various metrics without assuming data distribution (Shin et al., 2020). In this study, the application of five predictive algorithms were examined for water quality prediction using algae Chlorophyll-a as an indicator to fully comprehend how various machine learning models handle various data complexities. These methods have been chosen due to their widespread use and consistently high performance in research.

1.3 Aim and Objectives

Aim:

This study aims to assess the efficiency of five machine learning algorithms for Chlorophyll-a concentration prediction within the Vaal Dam, utilizing Sentinel-2 MSI imagery.

Objectives:

1. Investigate the correlation between Sentinel-2 MSI spectral band ratios and measure Chlorophyll-a values to understand their interrelationship.

2. Examine the effectiveness of each machine learning algorithm by employing diverse combinations of band ratios to enhance predictive accuracy in Chlorophyll-a estimation.

1.4 Research report outline.

The research report is structured into five distinct chapters.

Chapter 1 introduces the research background, motivation, and scope. It further provides the aim of the research along with objectives to be investigated. It further outlines the research process. The chapter ends with an outline of the research report structure.

Chapter 2 offers an extensive examination of the existing literature emphasizing the utilization of Chlorophyll-a's role as a water quality indicator for algal blooms, the use of band ratio reflectance in predicting chlorophyll-a, and machine learning algorithms for analyzing Chlorophyll-a using satellite imagery.

Chapter 3 details the study's methodological strategy and process and highlights the study site's location.

Chapter 4 provides a summary of the results obtained in the study; it gives clear, descriptive statistics of the Chlorophyll-a in-situ readings collected on the field, the results of the five machine learning algorithms, and finally, provides a Chlorophyll-a concentration map based on the most effective model.

Chapter 5 is divided into a discussion that compares the research findings to other bodies of work. It presents a thorough conclusion regarding the performance of various machine learning algorithms, identifies the optimal model, and concludes with study limitations and recommended actions.

Chapter Two

Literature Review

2.1 Algal bloom of freshwater resources

An algal bloom is a phenomenon that occurs when excessive amounts of nutrients, such as nitrogen and phosphorus, enter a freshwater system, resulting in an explosive growth of algae (Khan et al., 2019). This excess growth can lead to harmful algal blooms (HABs), negatively impacting freshwater resources and the surrounding ecosystem (Khan et al., 2019). Swanepoel et al. (2017) indicated that South Africa has seen a rise in algal blooms in recent years, with many small Dams experiencing the effects of these blooms. Algal blooms can significantly impact freshwater resources, leading to various adverse effects. One such effect is reduced dissolved oxygen levels, which has resulted in fish and other aquatic animal kills (Khan et al., 2019). Algal blooms can also result in toxins discharge, posing potential risks to human health (Swanepoel et al., 2017). Additionally, algal blooms can reduce water clarity, increase water temperatures, and alter the chemistry of the water, which can impact the entire ecosystem (Khan et al., 2019).

Small Dams are particularly vulnerable to the effects of algal blooms due to their shallow depths and low water volumes (Swanepoel et al., 2017). These factors can lead to higher nutrient concentrations and rapid algal growth (Swanepoel et al., 2017). Furthermore, small Dams are often used for leisure purposes like fishing, swimming, and boating, making them particularly susceptible to the adverse effects of HABs (Swanepoel et al., 2017). According to work done by Boyde et al. (1997) and Walmsley (2000), waterbodies can be categorized on several classifications of trophic levels status based on the level of algal concentration (micrograms per liter (ug/L)), such as Oligotrophic, Mesotrophic, Eutrophic, and Hypertrophic. Summarized in Table 2.1, these trophic levels can be used to determine the level of algal bloom concentrations for better preventative measures.

Table 2.1: Trophic algal bloom levels adapted from Boyde et al. (1997) and Tanjeck (2019).

Trophic Levels	Minimum Chlorophyll-a (ug/L)	Maximum Chlorophyll-a (ug/L)
Oligotrophic refers to an environment characterized by low levels of nutrients, resulting in limited productivity of terrestrial and aquatic vegetation.	≤ 2	≤ 5
Mesotrophic describes an environment with moderate nutrient levels, leading to intermediate terrestrial and aquatic vegetation productivity.	2-5	6-17
Eutrophic denotes an environment abundant in nutrients, fostering vigorous terrestrial and aquatic vegetation growth.	6-17	17-45
Hyperopic describes an environment with an excessive concentration of nutrients, resulting in an overabundance of terrestrial and aquatic vegetation productivity.	>20	>45

Remote sensing has been used to monitor algal blooms in freshwater resources, including small Dams. This technology detects and maps HABs in near-real-time, providing valuable information for water managers and decision-makers (Hu et al., 2021). For example, satellite imagery has been used in the Hartbeespoort Dam, South Africa, to monitor algal blooms. The study by du Preez et al. (2018) found that satellite imagery effectively detects and maps algal blooms, providing valuable information for water managers and the public. In addition to remote sensing, machine learning algorithms have also predicted algal blooms in freshwater resources. These algorithms can use satellite imagery and in-situ water quality data to forecast algal blooms accurately (Deng et al., 2021). du Preez et al. (2018) conducted a study in China where a machine-learning algorithm was employed to predict algal blooms in a large freshwater lake. The research revealed that the algorithm effectively forecasted the occurrence of algal blooms, offering valuable insights for water managers and decision-makers.

Some challenges must be addressed despite the potential benefits of remote sensing and machine learning in monitoring and predicting algal blooms in small Dams. For example, small Dams often have complex bathymetry, making it challenging to interpret remote sensing data accurately (Khan et al., 2019). Moreover, high-quality data inputs are essential for machine learning algorithms to make precise predictions, which can be challenging in small Dams where water quality monitoring is often limited (Deng et al., 2021). In conclusion, algal blooms in freshwater resources, including small Dams, can negatively affect the ecosystem and human health. Remote sensing alongside machine learning algorithms holds promise in aiding the monitoring and prediction of algal blooms, offering near-real-time analysis of these parameters related to water quality (Khan et al., 2019).

2.2 The importance of Algae monitoring in surface water bodies.

The presence of algae in aquatic environments can lead to various negative impacts on both the health of marine ecosystems and human well-being. These effects can include reduced water quality, decreased oxygen levels, and the growth of harmful algal blooms (HABs) (Smith et al., 2018; Tran et al., 2020). HABs can cause fish kills and produce toxins that can harm humans and animals (Feng et al., 2019). Furthermore, algal blooms' ecological impacts can have significant economic and social implications (Jiang et al., 2020). For example, algal blooms can lead to the closure of Dams and the loss of revenue for the tourism and recreation industries (Wang et al., 2019). Zhou et al. (2020) monitored algal blooms in a small lake for over one year using satellite imagery. The study used Sentinel-2 and Landsat 8 images to predict nitrate concentration, one of the main drivers of algal blooms in freshwater systems. The study found that while both satellite images could accurately predict nitrate concentrations, leveraging Sentinel-2 images facilitated a more intricate comprehension of the spatial distribution of nitrate within the lake (Zheng et al., 2020). This data can be instrumental in crafting tailored management approaches to minimize nitrate inputs and alleviate the consequences of algal blooms (Li et al., 2019). Utilizing satellite imagery, such as Sentinel-2 and Landsat 8, can provide valuable information on algae growth in water bodies (Tran et al., 2020). This data can help in enhancing water management practices and safeguard the health and welfare of human populations (Feng et al., 2019). Nonetheless, it is crucial to acknowledge that while satellite imagery can be a valuable resource, it should be complemented by other management strategies, such as reducing nutrient inputs, controlling stormwater runoff,

and implementing best management practices for activities such as agriculture and development (Smith et al., 2018).

In conclusion, Human health can be negatively impacted by algae in water bodies, along with aquatic ecosystems (Jiang et al., 2020). The utilization of satellite imagery, like Sentinel-2 and Landsat 8, aids in valuable understanding of the proliferation of algae within bodies of water (Wang et al., 2019). This data, as highlighted by Zhang et al. (2019), can be instrumental in enhancing water management practices and safeguarding the health and well-being of human communities. Nevertheless, as Zhou et al. (2020) emphasized, it is crucial to carefully weigh the benefits and drawbacks of various satellite imagery options. Additionally, as advocated by Zheng et al. (2020) and Li et al. (2019), human health necessitates integrating additional management strategies to alleviate the adverse impacts of algae on aquatic ecosystems effectively.

2.3 The implementation of sentinel-2 in Chlorophyll-a monitoring

In current years, hyperspectral imaging has been used in conjunction with optical remote sensing algorithms and empirical relationships between reflectance to evaluate water quality, leading to the mapping of Chlorophyll-a and turbidity in inland waters (Tassan, 1994; Doxaran et al., 2002; Lee et al., 2007). Various studies have compared the effectiveness of Landsat 8 and Sentinel-2 in assessing Chlorophyll-a and turbidity in small surface water bodies. One study found that Sentinel-2 had a lower RMSE and higher R^2 values than Landsat 8 for assessing Chlorophyll-a and turbidity in a small reservoir (Feng et al., 2019), while another study found that Sentinel-2 had a higher R^2 and lower RMSE values than Landsat 8 for assessing Chlorophyll-a in a small lake (Chen et al., 2020).

For the assessment of Chlorophyll-a and turbidity, the blue and green bands are critical (Chen et al., 2020). Landsat 8 provides a blue band (B1) and green band (B3), while Sentinel-2's blue band (B2) and green band (B3) at 30 meters (m) and 10 m spatial resolution, respectively, making Sentinel-2 more effective in capturing fine-scale details. Moreover, Sentinel-2's red-edge bands (B5 and B6) effectively assess Chlorophyll-a in small surface water bodies (Feng et al., 2019; Chen et al., 2020). With frequent monitoring facilitated by its higher temporal resolution, there is greater capability to track changes in water quality; sentinel-2 is more effective than Landsat 8, which has a longer revisit time of 16 days (Chen et al., 2020). In conclusion, Sentinel-2 is more

effective than Landsat 8 in assessing Chlorophyll-a and turbidity in small surface water bodies, owing to its higher spatial and temporal resolutions and the additional red-edge bands. However, it is essential to acknowledge that other factors, such as atmospheric correction and the specific characteristics of the water body being assessed, can also affect the accuracy of these assessments.

The study by Bande and Adam (2018) investigated whether Sentinel-2MSI assesses water quality in the Vaal Dam better compared to Landsat 8 using indicators such as Chlorophyll-a and turbidity. The research involved collecting water samples from 23 different points on the Dam, with an effort to incorporate different land-use activities around the Dam. The study, to measure in situ parameters, MS5 Mini Sonde Hydrolab was used, and Chlorophyll-a concentrations were determined through ten depth-specific samples processed with Whatman filters. Turbidity was measured digitally. Sentinel-2 images underwent atmospheric correction using Sen2cor 2.2, a Level-2A processor designed to rectify single-date Sentinel-2 Level-1C to Top-Of-Atmosphere (TOA) products by mitigating atmospheric effects. Regression models for Chlorophyll-a and turbidity utilized band properties in Statistical Package for Social Science (SPSS), with outliers removed for optimization. R^2 and RSME were calculated for model evaluation.

The study by Bande and Adam (2018) found that for Landsat 8, Bands 1, 2, 3, and 4 exhibited a strong correlation with Chlorophyll-a and Global Environmental Monitoring Index (GEMI), while band ratios B4/B2, B4/B1, and B2/B4 were good predictors of turbidity. the study indicated that certain specific bands (B1 and B2) and combinations of bands (such as B1/B2 and B1/B4) from Sentinel-2 imagery were highly indicative of Chlorophyll-a levels. Additionally, other combinations of bands (e.g., B1/B3, B3/B1, etc.) were good turbidity predictors. This suggests that by analyzing these particular bands or combinations of bands in Sentinel-2 imagery, one can estimate Chlorophyll-a and turbidity levels in water bodies. The results for Landsat 8 and Sentinel-2 were summarized in Tables 2.2 and 2.3, respectively.

Table 2.2: Statistical analysis for Landsat 8, adapted from Bande and Adam (2018).

Tested Parameters	**Band index, band ratios, or single bands**	**R^2**	**RMSE**
Chlorophyll-a	GEMI, B3,B4, B1,B2	0.68	1.11 µg/L
Turbidity	B4/B1, B2/B4, B4/B2	0.5	167.9 NTU

Table 2.3: Statistical analysis for Sentinel-2 adapted from Bande and Adam (2018).

Tested Parameter	**Band index, band ratios, or single bands**	**R^2**	**RMSE**
Chlorophyll-a	B1/B2, B1/B4, B1, B2,	0.86	1.01µg/L
Turbidity	B3/B1, B3/B2, B1/B3 B4/B1, B4/B2, B6/B9	0.59	147.69 NTU

The study compared Sentinel-2 over Landsat 8 using Chlorophyll-a and turbidity prediction. Sentinel-2 outperformed Landsat 8 in both cases, emphasizing specific bands. Landsat 8 highlighted bands 1, 2, 3, 4, and 5 for Chlorophyll-a, while Sentinel-2 relied on bands 1, 2, and 4. Landsat 8 identified bands 1, 2, and 4 for turbidity, whereas Sentinel-2 used bands 1, 2, 3, 4, and red edge 2 (B6) and 3 (B7). The dissimilarity is attributed to spatial resolution differences; Sentinel-2's higher resolution contributed to superior outcomes. Both platforms effectively mapped Chlorophyll-a, but Sentinel-2 excelled. Sentinel-2 proved invaluable for turbidity mapping, surpassing Landsat 8 (Bande and Adam, 2018).

The study concludes that hyperspectral imaging, particularly with Sentinel-2, is a promising technique for water quality assessment in shallow waters. Sentinel-2, with higher spatial and temporal resolutions, outperforms Landsat 8 in mapping Chlorophyll-a and turbidity in small surface water bodies. Findings align with Bande and Adam, (2018) study in Vaal Dam, emphasizing Sentinel-2's superiority due to spatial resolution differences. However, accurate assessments should consider atmospheric correction and water body characteristics.

2.4 Using machine learning algorithms in assessing Chlorophyll-a in water.

Zhou et al. (2015) began their study by employing principal component analysis (PCA) in conjunction with artificial neural networks (ANN) to predict the levels of chlorophyll-a in Dian-Shan Lake, located in Shanghai. They emphasized the value of hyperspectral remote sensing for comprehensive spatial and temporal coverage, especially in measuring multiple lakes simultaneously. The objective of the research was to construct an ANN model using hyperspectral measurements, integrate water quality parameters, and enhance model accuracy through a

combination of principal component analysis (PCA) and ANN. Critical nutrient elements such as dissolved oxygen (DO), pH, total nitrogen (TN), total phosphorus (TP), and chemical oxygen demand (COD) were considered influential factors affecting Chlorophyll-a growth. The methodology involved the collection of two datasets in 2010 and 2012, with samples assessed for spectral reflectance and water constituent concentrations. These datasets were utilized to calibrate and validate three ANN models. The initial model (ANN#1) strongly correlated with in situ Chlorophyll-a measurements but displayed deviations within the 20-40 mg/m3 range. Subsequent models (ANN#2 and ANN#3) improved accuracy by incorporating additional input variables and applying PCA to address multicollinearity. The third model (ANN#3) demonstrated superior stability and accuracy, boasting a high coefficient of determination (R^2) and low RMSE, as detailed in Table 2.4. This study contributes valuable insights into refining predictive models for Chlorophyll-a concentration in aquatic environments.

Table 2.4: A summary of the evaluation metrics obtained from the training and validation sets is presented for all three of the distinct artificial neural network models created for this research adapted from Zhou et al. (2015).

Model	**Dataset**	**R^2**	**RMSE**	**ARE**
ANN #1	Training	0.8503	8.41	6.13
	Validation	0.6395	11.85	9.13
ANN #2	Training	0.9091	6.4	4.94
	Validation	0.85	8.36	6.1
ANN #3	Training	0.9184	5.88	4.73
	Validation	0.8648	6.84	4.2

The study concludes by stating that three models based on ANN were devised to approximate the intensity of Chlorophyll-a in lake water. Models ANN#2 and #3 performed better than ANN#1 due to incorporating multiple input variables. Model ANN#3 used Principal Component Analysis (PCA) to address relationships among water parameters, achieving high accuracy with $R^2 = 0.9184$ and RMSE = 5.88. Validation with another dataset confirmed the superiority of ANN#3. The study

suggests that the PCA-ANN hybrid approach is practical for monitoring lake water quality and can be applied in other areas to study algae bloom trends in inland lakes (Zhou et al., 2015).

Another study that utilized a non-parametric machine-learning technique for detecting Chlorophyll-a was conducted by Mamun et al. (2019). The primary objective of their research was to forecast Chlorophyll-a and water clearness in reservoirs characterized by spatial heterogeneity and temporal variability, particularly influenced by the Asian monsoon from 2000 to 2017. Subsequently, the study sought to assess the trophic condition of these reservoirs by employing SVM, MLR, and ANN methodologies. These methodologies were applied to scrutinize spatial distributions across the three main zones: transitional zone (Tz), riverine zone (Rz), and lacustrine zone (Lz) alongside time-based fluctuations observed during pre-monsoon, monsoon, and post-monsoon seasons. The study utilized Monthly physicochemical parameters and precipitation data from 2000 to 2017 to train the machine learning models. They were confirmed through cross-validation. In the pre-and post-validation stages, SVM exhibited superior predictive performance over MLR and ANN. Lower RMSE and MAE values in SVM indicated significantly improved performance compared to MLR and ANN. Additionally, the support vector machine (SVM) displayed a superior coefficient of determination. The elevated levels of total suspended solids (TSS), nutrients (TN and TP), water temperature (WT), conductivity, and algal Chlorophyll-a detected within the riverine zone were attributed to runoff from the surface. Analysis of the trophic state spanning from 2013 to 2017 indicated a transition from mesotrophic to oligotrophic conditions. Factors such as water temperature (WT), total phosphorus (TP), total suspended solids (TSS), total nitrogen (TN), nutrient-phosphorus (NP) ratios, and precipitation directly influenced the levels of Chlorophyll-a and transparency in the reservoir. These predictions offer valuable insights for effective reservoir management strategies.

The study concluded that SVR was the more favorable model, indicating the highest prediction accuracy for Chlorophyll-a detection compared to MLR and ANN. Furthermore, the high performance of SVR compared to MLR is attributed to the fact that MLR was unable to handle complex multicollinearity issues. Compared with ANN, SVR indicated a remarkable ability to deal with nonlinear relationships, further stating that SVR has an excellent ability to minimize error.

Chapter 3

Methodology

This chapter outlines the methodological workflow employed in this study and the specific research methods adapted, as illustrated in Figure 3.1. The study used two primary sources of data: Sentinel-2 imagery for band ratio reflection calculation and in-situ Chlorophyll-a measurements for model training and validation. The methodological workflow provides an overview of how these datasets were utilized to accomplish the study's objectives.

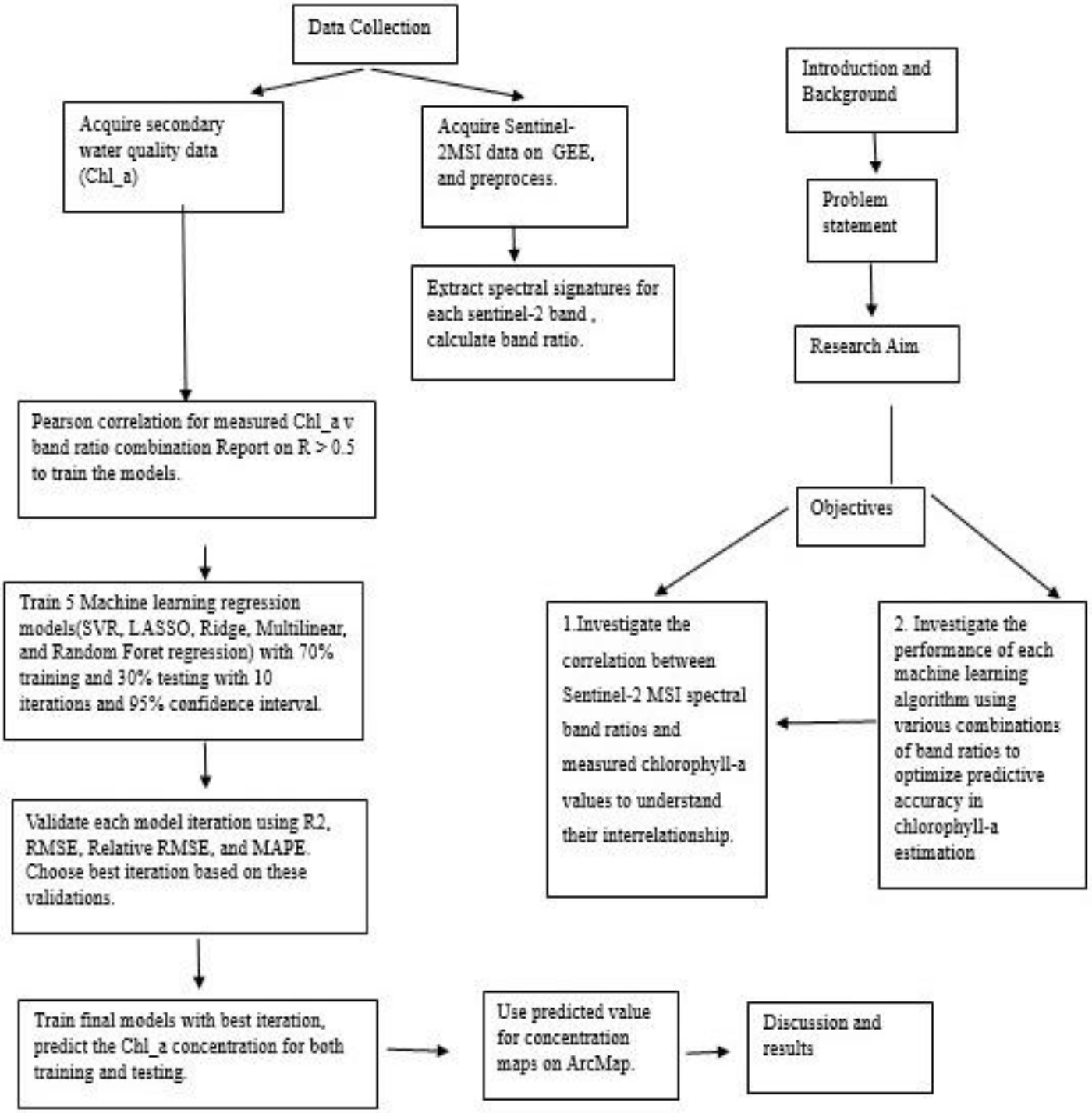


Figure 3.1: Summary of methodology used in this study.

3.1 Study site

The study site depicted in Figure 3.2 is situated at coordinates 26°56'49"S, 28°14'30"E; the Vaal Dam is a prominent interprovincial water reservoir in South Africa. It is a natural border between Gauteng, Free State, and Mpumalanga provinces. The Dam's wall height when it was built in 1938 was 54.2 m. The wall was raised to 63.5 m in 1985 to meet the increasing yearly water capacity of 2188 million cubic meters (m^3). Having an extent of roughly 321 square kilometers (km^2) and an average depth measuring 22.5 m, the Vaal Dam serves as a crucial water source for the people of South Africa (Malahlela et al., 2018). Obaid et al. (2023) indicated in their study that the dominant land cover type around this study site was Grassland, ranging between 52% and 57%; the second dominating landcover type was agricultural lands (including clear and cultivated fields covering 41% of landcover) other landcover included settlement class which largely contributed to population growth round the study site, mining, woody vegetation, and water bodies. The area surrounding the dam comprises agricultural fields, mining operations, industrial facilities, urban developments, and natural reserves. Mining activities, agricultural practices, and outflows of sewage from nearby villages (Braune and Rogers, 1987). These pollutants include fertilizers and pesticides from farming, waste products from acid mine drainage, septic tanks, and toilet effluents. These contaminants enter the Dam and cause higher concentrations of ions (particularly phosphates and nitrates), suspended particles that increase water turbidity, and an increase in algae that raises the amount of chlorophyll a Obaid et al. (2023). At the Vaal Dam, Rand Water has established a voluntary water quality monitoring program that has been in place since March 1998 (DWAF, 2008). This program aims to evaluate the quality of the water for the benefit of final users. The monitoring program involves the assessment of multiple water quality indicators, such as pH, phosphates (PO4), nitrates (NO3), salinity, temperature, turbidity, electrical conductivity (EC), dissolved oxygen (DO), and Chlorophyll-a pigments. Water samples taken from various field sampling locations are analyzed in a lab to determine these properties. In this effort, 23 different sampling sites were used to map and assess the concentrations of turbidity and Chlorophyll-a to find zones with differing degrees of pollution, from non-polluted to very polluted.

.

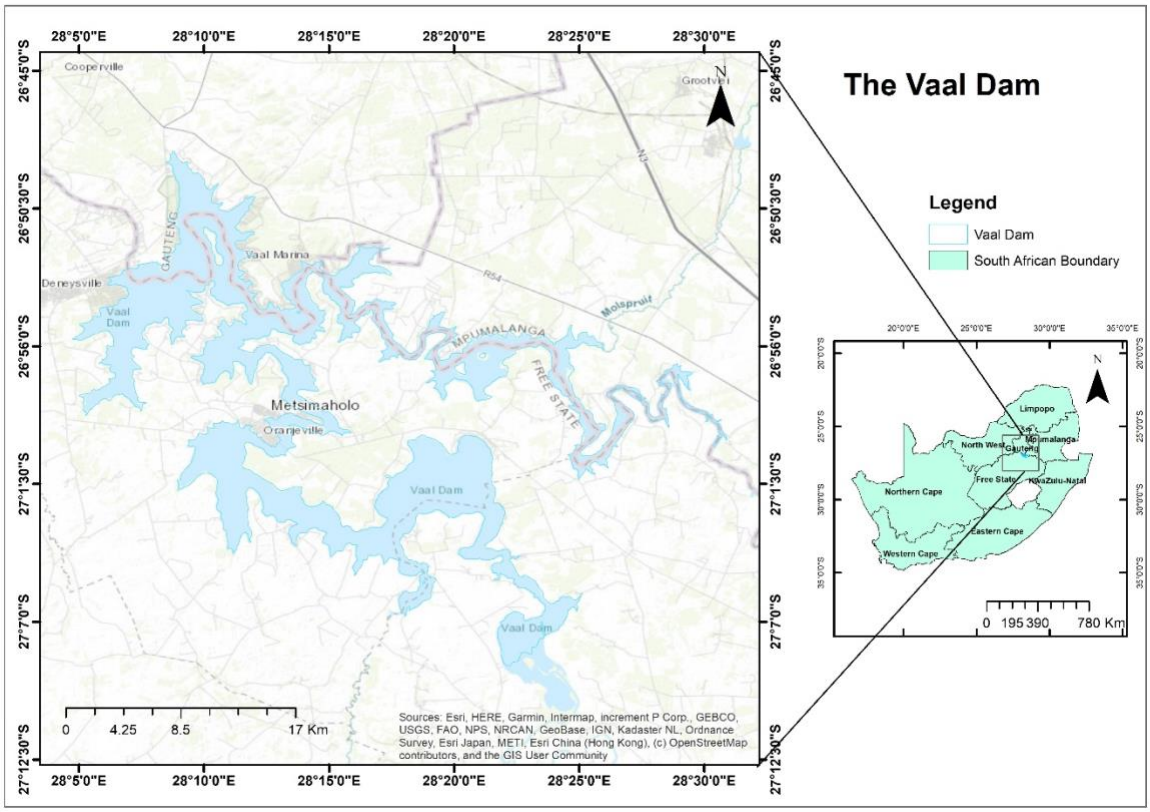


Figure 3.2: Map indicating the Vaal Dam study site.

3.2 Methodology

3.2.1 Remote sensing data acquisition and image pre-processing.

For this study, Sentinel-2 MSI was utilized for predicting and understanding the spatial spreading of chlorophyll-a concentration within the Vaal Dam. The sentinel image was downloaded from the Google Earth engine (Google Earth Engine). JavaScript was used to write the Google Earth Engine script, which included parameters. Using the Quality Assurance band denoted as "*QA60*" in the Sentinel-2 dataset, the maskS2cloud function was implemented to address cloud cover (mask out cloud-affected pixels). There were clear skies since the function set the cloud conditions to zero. It was using the expression ee.ImageCollection('COPERNICUS/S2'_RS), the script loaded the Sentinel-2MSI image. It then specifically targeted the Sentinel-2 Level-1C image between October 25 and October 30, 2016, corresponding to the on-site data collection between October 26 and October 28, 2016. To adjust for atmospheric conditions and scale TOA reflectance, pixel values within an image are first converted to TOA reflectance by dividing the image by 10,000 (i.e.,

separating the image by its QA60 band) and scaling the result by 0.0001, thus obtaining the final product Sentinel-2 L2A image. Sentinel-2 MSI was chosen for this study owing to its advantageous spatial resolution of bands susceptible to Chlorophyll-a, such as Band 2,3,4,5 and Band 8, as displayed in Table 3.1.

Table 3.1: Sentinel-2MSI sensor and spectral properties.

Band	**Description**	**Spatial Resolution (m)**	**Central Wavelength (µm)**
B1	Coastal aerosol Band	60	0.443
B2	Blue Band	10	0.490
B3	Green Band	10	0.560
B4	Red Band	20	0.665
B5	Vegetation Red Edge Band	20	0.705
B6	Vegetation Red Edge Band	20	0.740
B7	Vegetation Red Edge Band	20	0.783
B8	Near Infrared (NIR) Band	10	0.842
B8a	Visible and Near Infrared (VNIR) Band	20	0.865
B9	Water vapor – Cirrus Band	60	0.940
B10	Short Wave Infrared Band	60	1.375
B11	Short Wave Infrared Band	20	1.610
B12	Short Wave Infrared Band	20	2.190

3.2.2 In-situ field measurements

The study employed secondary data that was collected by the Agricultural Research Council (ARC) in partnership with the Japan quality evaluation team in October 2016; the sampling process for the Vaal Dam occurred between 3 days, 26 to 28 October 2016, during optimal summer, cloud-

free conditions, sampling was conducted using a handheld Hydrolab MSS – Multiparameter Mini Sonde sensor which measures up to 10 water quality parameters simultaneously. A total of 23 sampling points were collected (Figure 3.3). Sampling was conducted with a preference towards the diverse land uses surrounding the Vaal Dam, ensuring that the samples accurately represented all land uses in the Dam's vicinity. The sampling area was distributed as follows:

1. The probe sensor was calibrated onsite using the Dam water to ensure the multiprobe manager software could log the data.
2. The probe was submerged 2m deep into the different water sampling stations for approximately 2 minutes for optimal reading.
3. Various sampling sites were assessed for different water quality parameters, including Temperature, Dissolved Oxygen (DO), electric conductivity (EC), Chlorophyll-a, pH, Reduced- oxidation (ORP), salinity, and turbidity.
4. The Geographic Positioning System (GPS) coordinates were collected at each station, and a handheld Trimble GPS receiver was used in conjunction with them to record the coordinates.

However, since this study aims to use satellite imagery to understand the behavior of different machine-learning algorithms, only one water quality parameter (Chlorophyll-a) will be used as a proxy for this study to test against five machine-learning algorithms as indicated in Table 3.2 below. Due to limited resources, such as availability of a Hydrolab, the study relied on this existing dataset. It's important to note that although the dataset is from 2016, it is still applicable to this study as it is used to assess how different machine learning algorithms handle this data. Furthermore, the input satellite imagery is also from 2016, aligning with the dataset's date to ensure consistency in the data analysis.

Out of the 23 sampling points, the data was partitioned into a training set (70%) and a test (validation) set (30%). Specifically for Vaal Dam, this division yielded a training dataset comprising sixteen samples and a testing dataset comprising seven samples.

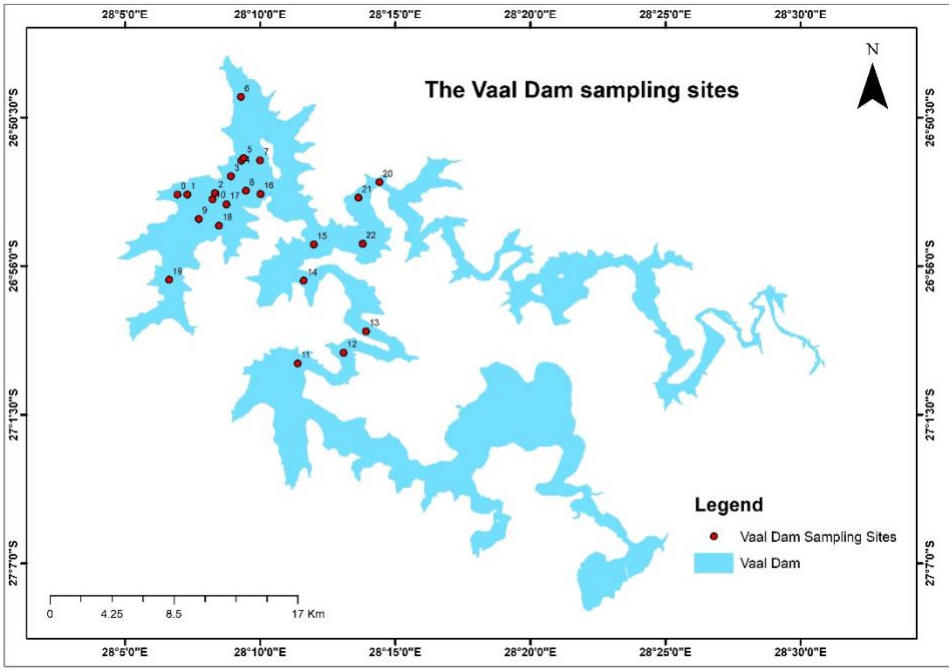


Figure 3.3: 23 sampling locations.

Table 3.2: Summary of all sample locations

Site name	Temp	Chl.a(ug/L)	Latitude	Longitude
1	19.532	3.207	-26.88918	28.115647
2	19.549	2.921	-26.88908	28.121753
3	19.753	3.264	-26.8882	28.138747
4	20.116	2.951	-26.87787	28.148568
5	20.785	2.997	-26.8682	28.155088
6	21.117	3.592	-26.86662	28.156423
7	20.343	3.337	-26.8289	28.154825
8	21.133	4.187	-26.86811	28.166677
9	20.893	3.095	-26.8869	28.157918
10	20.217	4.061	-26.90423	28.128792
11	19.672	3.532	-26.89217	28.137223
12	22.088	5.92	-26.99336	28.189773
13	22.877	6.906	-26.98676	28.218161
14	23.852	6.094	-26.97359	28.231925
15	24.034	6.582	-26.9423	28.193542
16	24.242	12.408	-26.92001	28.199676
17	23.372	3.11	-26.88886	28.166869
18	24.247	4.058	-26.89515	28.14583
19	22.589	4.338	-26.90838	28.14128

20	24.552	13.386	-26.94179	28.110508
21	23.442	13.309	-26.88155	28.240201
22	24.048	12.49	-26.89115	28.227307
23	23.5717	11.871	-26.91957	28.229853

3.2.3 Band Ratio Correlation

The study evaluates the relationship between Sentinel-2 MSI imagery and in-situ measured chlorophyll-a. To determine which band ratio spectral reflectance combination is most responsive to Chlorophyll-a, spectral signatures from Sentinel-2 MSI were examined. The Pearson correlation coefficient (PCC) was formulated by Karl Pearson during the 1880s and is used to evaluate sensitivity. The PCC varies from -1 to 1, which specifies the direction and correlation of the relationship amongst variables. It quantifies the linear correlation between two datasets. There can be three types of correlation between variables: positive (0 to 1), negative (0 to -1), and zero, which indicate no correlation at all, corresponding changes, and opposing changes (Zheng et al., 2016). Research has recommended applying band ratio algorithms to estimate inland and marine environment characteristics (Bukata et al., 1981; Kallio et al., 2001). The blue-green ratio is often utilized to compute chlorophyll-a concentrations. However, Gilerson et al. (2010) stated that in open ocean water, non-algae particles and Colored Dissolved Organic Matter (CDOM) are strongly correlated, leading to the estimation of c concentration based on the reflectance of the blue-green band (Smith et al., 2018). On the other hand, because of the weak correlation between phytoplankton and CDOM, the algorithm's spectral region for retrieving Chlorophyll-a is limited in most inland and marine environments due to elevated turbidity. Robust overlapping absorption features in its spectral region cause the unreliability of blue reflectance (Yacobi et al., 2011). Red edge reflectance and near-infrared (NIR) spectra work well together to estimate Chlorophyll-a concentration in inland and marine environments with greater accuracy (Gitelson et al., 2008). In this combination, the peak height considers both the fluorescence of Chlorophyll-a and the minimum absorption coefficient. Reflectance peaks between 400 and 500 nanometers (nm) result from high Chlorophyll-a and organic matter absorption, while the peak near 560 nm arises from low algae absorption. Reflectance in the range of 620–675 nm, with a minor peak at 650 nm, is attributed to the absorption of Chlorophyll-a. A notable correlation exists between Chlorophyll-a concentration and the peak near 700 nm, as measured in the baseline at 675–730 nm (Gitelson,

1992). Because each band has a different sensitivity level and the information is complex, using band combinations is advised for obtaining reliable results (Dekker and Peters, 1993). The equation of Pearson correlation coefficients is indicated as follows:

$$R = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{(n\sum x^2 - (\sum x)^2)((n\sum y^2 - (\sum y)^2)}}$$

Where,

Pearson correlation coefficient = r

Values in the first set of data = x

Values in the second set of data = y

Total number of values = n.

3.2.4 Machine Learning Regression Algorithms

This study utilized multiple algorithms to evaluate the efficacy of machine-learning models in predicting Chlorophyll-a levels.

3.2.4.1 Random Forest Regression

Regarding modeling, RF has proven more reliable than a single decision tree. Through the combination of several decision trees, they can reduce bias-related errors and overfitting. More solid and trustworthy insights are produced by this cooperative approach (Segawa et al., 2019). RF enhances individual Classification Trees by constructing numerous decision trees using randomly sampled data. Frequently perceived as groups of Decision Tree (DT) algorithms, RF expands on bootstrap aggregation, or "bagging," as it pertains to DT. Using 'bagging,' arbitrary subsets with the same number of data points are generated from the entire dataset. These bootstrap-sampled subsets are then added back to the whole dataset to be used in further tree constructions. RF uses the' bagging' process to select random subsets and guarantee an equal chance of inclusion for every data point. RF is particularly good at handling missing data and can accommodate many predictor variables while excelling at accuracy (Ashphaq et al., 2023). In employing RF, it's essential to specify two primary parameters: the number of trees utilized in the sample (Ntree) and the number of variables selected for each attempt (Mtry). Figure 3.4 indicates the flow chart altered

by Rodriguez-Galiano et al. (2016), which suggests the process of RF trees for each predictor, which is then validated by splitting training data.

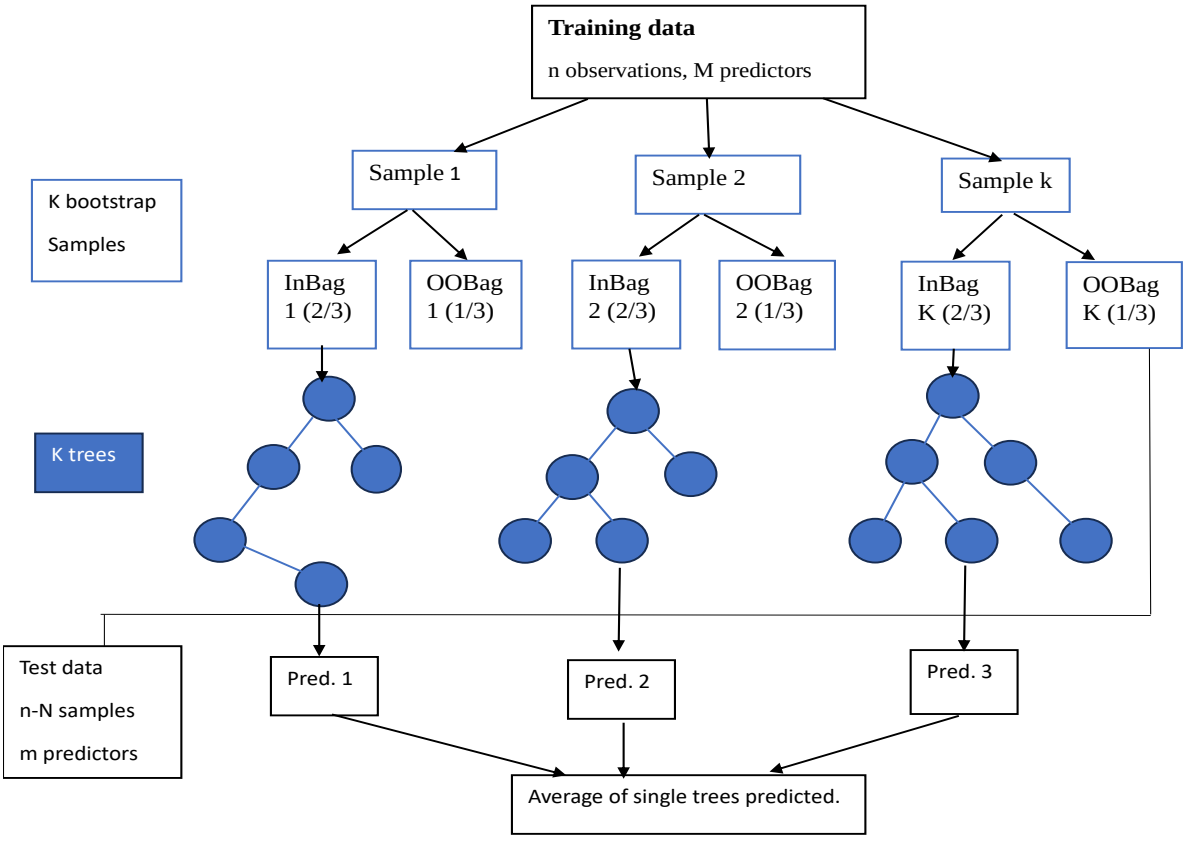


Figure 3.4: Random Forest decision trees overview, adapted from Rodriguez-Galiano et al. (2016).

3.2.4.2 Support vector regression

Support Vector Machine (SVM) algorithms are employed for data processing in Support Vector Regression (SVR), a regression technique. SVR serves as linear binary classifiers at their core, constituting non-parametric supervised classification algorithms. Within this framework, 'support vectors' refer to the points on the boundary margin separating classes or data groups (Mamun., 2019). Utilizing an asymmetric loss function, SVR functions as a supervised learning technique, equally penalizing both high and low misestimates. SVM uses the kernel function to transport the inputs to a high-dimension feature, subsequently establishing a hyperplane with the maximum margin known as the decision boundary. One of its key advantages lies in SVR's computational complexity, which remains independent of the input space size. Furthermore, it demonstrates the

potential for high prediction accuracy and extensive generalization (Ashphaq et al., 2023). The equation for SVR is:

$$f(x) = wTx + b\ (w, x\epsilon Rd)$$

Where;

$x\epsilon Rd$ = Non-linear function that maps input vector into high-dimension feature space.

w = unknown parameter, a weight vector that's normal to the hyperplane.

b = hyperplane bias term.

3.2.4.3 LASSO Regression

Least Absolute Shrinkage and Selection Operator (LASSO) is a method of linear regression in which the objective function includes a regularization term. To discourage overfitting and promote sparse models—those with fewer non-zero coefficients this regularization term acts as a penalty based on the absolute values of the coefficients. LASSO regression is especially helpful with high-dimensional datasets with many features because it can effectively select features by shrinking some coefficients to precisely zero. One benefit of LASSO regression is that it can automatically choose features by setting some coefficients to zero (Hassan and Woo, 2021). However, when there is multicollinearity (high correlation) among the features, it might not perform as well as Ridge regression. Selecting the regularization parameter λ is a critical decision frequently made using methods such as cross-validation. Cross-validation aids in determining the value of λ that offers the best balance between predictive accuracy and model simplicity (fewer features). The LASSO equation can be expressed as:

$$\left[\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{J=1}^{P}|\beta_j|\right]$$

Where;

N = number of observations.

P = number of features.

y_i = Observed value of the I-th observation.

X_{ij} = is the i-th observation of the j-th feature.

B_0 = is the intercept term.

B_j = is the coefficient for the j-th feature.

λ = regularization parameter (controls the strength of the regularized penalty).

3.2.4.4 Ridge Regression (L2 regularization)

Ridge Regression is a linear regression technique intended to reduce multicollinearity and avoid overfitting. In contrast to conventional linear regression, Ridge Regression incorporates a regularization term that stabilizes model performance by penalizing significant coefficients. The regularization term, managed by the hyperparameter (α), pushes the model's coefficients toward zero. As a result, the model becomes more reliable and broadly applicable, especially when working with high-dimensional datasets (Hassan and Woo, 2021). The cost function of Ridge regression is as follows:

$$J(\theta) = MSE + \alpha \sum_{i=1}^{n} \theta_i^2$$

Where,

$J(\theta) =$ is the cost function.

$MSE =$ the Mean Squared Error.

$\theta = represents$ the regression coefficients.

$\alpha =$ is the regularization parameter.

3.2.4.5 Multilinear regression

Multilinear regression (MLR) is one of the most essential linear regression algorithms. It captures the linear relationship between a single dependent variable, continuous independent variables (Known Chlorophyll-a values), and more than one independent variable (Band ratio combination). MLR works by changing the parameter values for cost reduction and the degree of error between the model's predictions and the values within the training dataset Mamun et al. (2019). The

equation below represents that of a hyperplane, which is the line of best fit for the data in 3 or more dimensions.

$$y = b_0 + b_1 x_1 + b_2 x_2 + \cdots b_n x_n$$

Where,

y = dependent variable

b_0 = Dependent variables axis intercept (note model optimizes this parameter

n = number of variables in our dataset

x_1 / x_n = dependent variables within the dataset

b_1/b_n = coefficient parameters (the model also tunes this)

3.2.5 Model tunning

Specific algorithms possess predetermined hyperparameters, whereas other algorithms necessitate extensive fine-tuning. This section discusses the optimal structure of machine learning algorithms, including RF, SVR, MLR, LASSO, and Ridge regression, and some basic definitions of their hyperparameters.

3.2.5.1 Random Forest

This study involved refining several parameters to train the RF model: a) including the number of decision trees (DT); growing more trees improves the model's robustness and generalization ability. In this study. b) the number of randomly selected predictors considered for splitting at each node in a decision tree Mtry, randomly selecting a subset of features for each tree helps improve tree diversity and prevents overfitting. and c) minimum size of terminal nodes (leaves) node size; setting a larger node size helps prevent overfitting by ensuring that each leaf contains a minimum number of training samples. K-fold cross-validation was used; Apologies for that. Here's a rephrased version: this method partitions the dataset into k-folds, iterating through k times for training and validation. Each iteration employs a distinct fold as the validation set while using the rest of the dataset as the training set. " tune.control(sampling = "cross", cross = 5)" cross = 5

indicated the number of folds. Thus, the dataset is divided into five subsets, and the tuning process is repeated five times.

To fine-tune the model below, "tuneRF" was used to specify the ranges of the hyperparameters to try, such as mtry, ntreeTry, and nodesizeTry. The tuneRF function performs a random search over the specified parameter ranges and identifies the best combination to validate the hyperparameter. MSE was used to determine the best hyperparameter. Print (tuned_rf_model) was used to identify the best parameters and access the best RF model using best_rf_model <- tuned_rf_model$finalModel. best_rf_model represents the RF model with the optimal hyperparameters for the training dataset, as indicated in Table 3.3.

Table 3.3: Random Forests parameter tuning.

Hyperparameter	Tuning test values	Final paramter values
Mtry	mtry = c(2, 3, 4)	4
Ntrees	Ntree = c (50, 100, 500)	500
Node size	nodesize = c (5, 10, 15)	10

3.2.5.2 Support Vector Regression

The most crucial hyperparameters of the SVR algorithm are C, which is the penalty classifier and measures the degree of misclassification; kernel, which helps the model predict non-linear data kept at Radial Base Function (RBF); Epsilon (ε) It significantly influences the size of the epsilon-tube, within which no penalty is imposed in the training loss function and lastly, Gamma (γ) is a parameter particularly relevant for the Radial Basis Function (RBF) kernel. It determines the extent of influence exerted by a single training example. Lower values indicate a broader reach, while higher values signify a closer reach. A 5-fold cross-validation was utilized to identify the Optimum configurations for each hyperparameter. To fine-tune the model, “tune control” was used to specify the ranges of the hyperparameters. The final parameter values used for the predictive model are initiated in Table 3.4.

Table 3.4: Support vector regression parameter tuning.

Hyperparameter	Tuning test values/ ranges	Final paramter values
C (regularization constant)	C = c(0.1, 1, 10)	1
Epsilon	Epsilon = c (0.1, 0.5, 1)	0.5
Gamma	gamma = c (0, 0.1, 1)	1
Karnal	Karnal = c("linear," "polynomial," "Radial")	Radial

3.2.5.3 LASSO regression

LASSO Regression extends linear regression by incorporating a regularization penalty into the loss function during training. One commonly used penalty is the L1 penalty, which penalizes the model Determined by the total of absolute coefficient magnitudes. This penalty aims to minimize the magnitude of all coefficients, potentially reducing some coefficients to zero, thereby effectively removing the associated predictors from the model. The main hyperparameter to be tuned is the Alpha (α), which controls the strength of the regularization (L1); it is a tuning parameter that determines the trade-off between fitting the model to the training data and keeping the model coefficients small. A grid was set up to perform a random search over the specified parameter ranges and identify the best combination to validate the hyperparameter. MSE was used to determine the best hyperparameter, as reported in Table 3.5 below.

Table 3.5: LASSO regression tunning.

Hyperparameter	Tuning test values/ranges	Final paramter values
Alpha (α)	α = c(0.01, 0.1, 1)	1

3.2.5.4 Ridge regression

Like LASSO regression, Ridge regression also adds a regularization penalty to the loss function during training; tuning the hyperparameter Lambda (λ) in this model will control the strength of the penalty term. The model tunned λ using the following knowledge:

- If λ = 0 no paramters are eleminated.

- If $\lambda > 0$ (increases), more coefficients will be set to zero and eliminated
- If $\lambda = \infty$, all coefficients are eliminated

Thus, as λ increases, so does the bias, and when λ decreases, so does the variance; therefore, setting λ to a low value is more manageable. The L2 regularization is employed for ridge regression; L2 does not result in sparse models as the same factor shrinks all coefficients, and none are eliminated. The following Table 3.6 summarizes the results.

Table 3.6: Ridge regression tunning.

Hyperparameter	Tuning test values	Final paramter values
lambda (λ)	λ = c(0, 0.1, 1)	0

3.2.5.5 Multilinear regression (MLR)

Unlike other machine learning models, MLR does not need hyperparameter tuning. Instead, it serves as a statistical method for modeling the relationship between the predictor and outcome variables; the main idea of this model is to find a linear equation that best predicts the value of the independent variable on an independent variable.

3.2.6. Final model training

Band ratios exhibiting a strong correlation coefficient ($R > 0.5$) to Chlorophyll-a, derived from Sentinel-2 MSI satellite data collected in the field, were employed in the regression models. The dataset, comprising 23 samples, was divided into a training set (70%) and a test (validation) set (30%). This division resulted in a training set of 16 samples and a test set of 7 samples. All five machine learning models underwent training with ten iterations, yielding ten predictive model outputs for Chlorophyll-a concentrations based on Sentinel-2 MSI satellite data in conjunction with in-situ data. Subsequently, the models were retrained with the best iteration, and the resulting models were used to generate linear regression. The 7 data points earmarked for testing (validation) constituted a subset of the overall 23 points, assessing the model's performance and accuracy specifically for Sentinel-2 MSI satellite data.

3.2.6.1 Random Forest

The' *randomForest*' library was employed to train the RF regression model. This library offers a dedicated function for fitting and analyzing RF regression, facilitating the prediction of continuous numerical outcomes. Similarly, the '*caret*' library was utilized. The dataset to train this model included band ratio reflectance values from Sentinel-2 MSI and in situ Chlorophyll-a values from 23 sampling points across the Dam. The training (70%) and testing (30%) sets of this dataset were separated, employing 500 trees (Ntree) and considering four variables at each split (Mtry). The data were partitioned into 'in-bag' (samples drawn with replacement from the entire dataset) and 'out-of-bag' (excluded data) sets. Bootstrap resampling was performed at a 95% confidence level, and the model underwent training through a loop of 10 iterations. Based on the optimal iteration, the model was re-run. Chlorophyll-a concentrations were then predicted using cross-validation, involving the division of datasets into training and testing sets. Training utilized 'in-bag' samples, while accuracy assessment employed 'out-of-bag' samples. The training dataset was subsequently used to assess the model's accuracy, providing an optimistic estimate of the R^2, RMSE, and MAPE values for predicted Chlorophyll-a.

3.2.6.2 Support Vector Regression

The R Studio Posit Cloud (Posit Cloud)) was essential to training all machine learning models in this study. Several libraries for the modeling process of SVR were used, including the *e1071 library*, containing the kernel function parameters and the SVM() function, the *dplyr library* used for data processing and cleaning, the *caret library* for model training and cross-validation, and *ggplot2* and *lattice libraries*, were used to visualize the results of the SVR model. The dataset to train this model included band ratio reflectance values from Sentinel-2 MSI and in situ Chlorophyll-a values from 23 sampling points across the Dam. This dataset's training (70%) and testing (30%) sets were separated. The SVR hyperplane was created using the training data, and its effectiveness was then confirmed using the testing dataset. The model used the radial basis function (RBF), SVR kernel, to capture non-linear relationships between input features and target variables. SVR excels in determining the optimal regression hyperplane without presuming any specific data distribution. The training dataset establishes the hyperplane in SVR, while an independent testing dataset assesses its efficacy. SVR is a distribution-free algorithm that maximizes the separation between data points and the best hyperplane. This algorithm aims to minimize deviations from actual values by identifying the best decision boundary for regression

tasks. To balance attaining a more significant margin and allowing for misclassification errors when needed, the penalty parameter Cost (C) is added, C = 1, Epsilon = 0.5, and gamma 1. When linear separation is impractical, SVR uses kernel functions like the radial basis function—to map the original input space into a higher-dimensional new feature space, which converts non-linear problems into linear ones. Because the radial basis function outperforms other kernels and has found extensive application in remote sensing research (Xu et al., 2021; Bande, 2020), we specifically chose it as the kernel for our regression study. The dataset to train this model included band ratio reflectance values from Sentinel-2 MSI and in situ Chlorophyll-a values from 23 sampling points around the Dam. This dataset's training (70%) and testing (30%) sets were separated. Ten iterations were used for the training process, and for both the training and testing sets, the reported accuracies (R^2, RMSE, and MAPE) were provided for each iteration. The final model was trained using the best iteration. This extensive cross-validation iterative process was essential to improving the model's performance. The outputs for the final model were used to predict Chlorophyll-a values. The results of the final model were comprehensively reported in the results section.

3.2.6.3 LASSO Regression

To train the LASSO regression model, the following libraries were used: *library(glmnet)* was used for fitting generalized linear models (GLM), which are regularized with various penalty terms proportional to the square of the coefficients to the loss function, which includes LASSO regression, *library(caret)*, and *library(ggplot2*) as described above (SVR). The dataset to train this model included band ratio reflectance values from Sentinel-2 MSI and in situ Chlorophyll-a values from 23 sampling points across the Dam. This dataset's training (70%) and testing (30%) sets were separated. The Alpha (α), the regularizing parameter, was set at 1. The regularization term in LASSO regression is proportional to the absolute values of the coefficients. When alpha is set to 1, it indicates a relatively strong regularization. Ten iterations were used for the training process, and for both the training and testing sets, the reported accuracies (R^2, RMSE, and MAPE) were provided for each iteration. The final model was trained using the best iteration. The outputs for the final model were used to predict Chlorophyll-a values. The results of the final model were comprehensively reported in the results section.

3.2.6.4 Ridge Regression

To train the Ridge regression model, the following libraries were used: *library(glmnet)* was used for fitting generalized linear models (GLM), which are regularized with various penalty terms proportional to the square of the coefficients to the loss function, which includes ridge regression, *library(caret),* and *library(ggplot2)* as described above (SVR). The dataset to train this model included band ratio reflectance values from Sentinel-2 MSI and in situ Chlorophyll-a values from 23 sampling points across the Dam. This dataset's training (70%) and testing (30%) sets were separated. L2 regularization Ridge added a penalty term to the loss function based on the square of coefficients. Lambda (λ), which indicated regularization strength, was kept at 0. The model was also split into training (70%) and testing (30%); ten iterations were used for the training process, and for both the training and testing sets, the reported accuracies (R^2, RMSE, and MAPE) were provided for each iteration. The final model was trained using the best iteration. The outputs for the final model were used to predict Chlorophyll-a values. The results of the final model were comprehensively reported in the results section.

3.2.6.5 Multilinear Regression (MLR)

In this study, an MLR model was trained to forecast Chlorophyll-a concentrations predicted on band ratio reflectance values obtained from Sentinel-2 MSI satellite data across 23 sampling points in the field. The following libraries were used: *library(caret)* as explained in the SVR model *library(olsrr)*. Ten iterations were used in the training process to increase the robustness of the model. The dataset underwent random division into a 70% training set and a 30% testing set (30%) for each iteration. Using band ratio reflectance as the predictor and Chlorophyll-a as the response variable, the model was fitted using R's *lm function*. The model aimed to represent linear correlations between the target variable and the input features. Performance metrics such as R^2, MAPE, and RMSE are computed for each iteration's testing set and training. The best iteration was determined to have the highest R^2 value and lowest MAPE. The final MLR model was then trained using the entire dataset based on the parameters from the best iteration. This iterative, cross-validated process aimed to maximize the model's predictive accuracy and enhance its ability to generalize to new data. The outputs generated by the final model were employed to predict

Chlorophyll-a values. The comprehensive findings of the final model were thoroughly detailed and reported in the results section.

3.2.7 Accuracy assessment

The accuracy assessment process involves evaluating the total number of predictions generated from the band reflectance of satellite data and comparing it with observed Chlorophyll-a concentrations from field measurements. To validate the reliability of the results and the model employed for estimating Chlorophyll-a concentrations, a statistical comparison is conducted between the predicted Chlorophyll-a values and the in-situ measured Chlorophyll-a concentrations. Specifically, the field data collected between October 26th and 28th, 2016, comprising the validation dataset (n=7), was utilized for a more rigorous model validation.

3.2.7.1. Coefficient of Determination

The statistical measure, coefficient of determination, or R-squared, is used in machine learning to evaluate a regression model's efficacy. It assesses the extent to which the model accurately represents or captures the data. Calculate the percentage of the dependent variable's variability that the independent variables account for. A perfect fit is denoted by a value of 1, while the model's failure to account for any unevenness in the dependent variable is indicated by a value of 0. While a smaller number indicates a limited explanatory power, a higher R-squared value indicates a greater ability of the model to explain variability. R squared is calculated as follows where y_i aligns with the observed variable, which describes its mean, and ŷ represents the predicted value

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

3.2.7.2 Root Mean Square Error (RMSE)

Before determining whether a predictive model has been correctly fitted using an R-squared value, consider additional variables like Root Mean Square error (RMSE). The RMSE calculates the deviation among the predictions and the known data. Where N is the number of samples, $y_{predicted}$ is the predicted Chlorophyll-a, and $y_{measured}$ is the measured Chlorophyll-a or turbidity. Though

they must indicate how well or poorly a model fits the data, they can be combined with R-squared to determine its overall quality. It is calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_{predicted} - y_{measured})^2}{N}}$$

3.2.7.3 Mean Absolute Percentage Error (MAPE)

The mean absolute percentage error (MAPE), alternatively referred to as mean absolute percentage deviation (MAPD), assesses the accuracy of a forecasting system by calculating the percentage variation between predicted and actual values. It is calculated by dividing the absolute difference between the expected and actual values by the exact values and then averaging the absolute percent error across all periods. MAPE is frequently employed when assessing and predicting errors, mainly when no extreme data values or zeros exist. Its popularity comes from being a percentage scale, which makes it simple to understand. Regression analysis and model evaluation frequently use it as a loss function. It is calculated as follows:

$$M = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{A_t - F_t}{A_t}\right| * 100\%$$

3.2.7.4 P- values at Confidence interval 95%

The P-value is a crucial measure in hypothesis testing. It represents the significance of the observed value of the Null Hypothesis statistic and serves as evidence for or against a data-related premise. The P-value ranges from 0 to 1, and a lower value indicates more substantial evidence against the null hypothesis. In regression analysis, hypothesis testing focuses on the model's coefficients rather than the target variable, aiming to ascertain their statistical significance to evaluate the strength and reliability of the relationship between predictors and the target variable.

Null hypothesis (H_0) **:** There is no significant correlation exists between the target and predicted variables.

$$H_o : \beta = 0$$

Alternative hypothesis (H_1): There is a significant correlation exists between the target and predicted variables.

$$H_1 : \beta \neq 0$$

Where $\alpha = 0.05$

Chapter 4

Results

4.1.1 Descriptive statistics

Table 4.1 indicates the descriptive statistics for Chlorophyll-a in the Vaal Dam, as shown in Table 4.1 below. The standard deviation (SD) quantifies the degree of variability in a set of 23 values; the SD of this study is 3.81, which describes the spread of individual data points around the mean (5.98), indicating how much data points deviate from the average. The standard error (SE) measures the precision of the sample mean as an estimate of the population; in this study, the SE is at 0.79, which indicates the precision of the mean relative to the number of samples present. The highest recorded Chlorophyll-a value is 13.38 ug/L, and the lowest is 2.92 ug/L.

Table 4.1: Descriptive statistics for field measurement data.

Descriptive statistics	**Chlorophyll-a (ug/L)**
Mean	5.98
Standard Error	0.79
Median	4.06
Standard Deviation	3.81
Sample Variance	14.51
Range	10.46
Minimum	2.92
Maximum	13.38
Count	23

4.1.2 Chlorophyll-a results

Table 4.2 below indicates the top 6 results from the Sentinel -2 MSI band ratio correlation with in-situ Chlorophyll-a data (N = 23). The table indicates the band ratios that performed the best are generally within the same region (0.443- 0.842μm) band 1,2,3,4 and 5 along with band 8. The band ratio that best correlated with Chlorophyll-a is b1/b3 and b1/b4 at 0.75 and 0.74, respectively. Band 1, Coastal aerosol, is significant because it is sensitive to water constituents, including chlorophyll-a, along with its shorter wavelength; it captures variation in Chlorophyll-a absorption as it can distinguish different types of phytoplankton. The presence of Chlorophyll-a in water bodies can impact the blue band (b2 - 0.490µm), which is susceptible to variations in watercolor; elevated Chlorophyll-a levels could result in heightened absorption within the blue spectrum. The green band(b3 - 0.560µm) is also vulnerable to Chlorophyll-a absorption. It is frequently used in vegetation indices and can detect variations in Chlorophyll-a concentrations; elevated values within this band could suggest the existence of robust vegetation encompassing algae. The red band (b4 - 0.665µm) is essential since it shows the maximum absorption of chlorophyll. In general, higher values in this band correspond to higher Chlorophyll-a concentrations. In vegetation indices, the red band is widely used to track the health of the vegetation. Water bodies can be distinguished from other types of surfaces using the Near-Infrared (NIR) band (b5 - 0.705µm), which is sensitive to vegetation's reflectance. Variations in values can indicate algae's presence and water quality changes. Lastly, Band 8 (b8a – 0.842µm) is in the Near-Infrared spectrum, just like Band 5. It is frequently used in conjunction with other bands to improve the discrimination of various land cover types, such as vegetation and water bodies; healthy algae or vegetation may be linked to high values in this band. The different machine learning algorithms used the best-correlated band ratio as predictor values.

Table 4.2: Chlorophyll-a Sentinel-2 MSI band correlation coefficients (R).

Band Ratio	**Sensitivity**	**Reference**	**Correlation Coefficient (R)**
NIR/ Red (b5/b4)	Chlorophyll-a, vegetative greenness	Llodrà-Llabrés et al. (2023)	0.69

NIR/ Red (b8a/b4)	Chlorophyll-a in vegetation, vegetation health	Liu et al. (2021)	0.58
(Green – Red)/(Green + Red -Blue) VARI	Chlorophyll-a, vegetative health	Johnson (1978)	0.53
Coastal / Blue (b1/b2)	Chlorophyll-a, biomass, vegetation health	Malahlela et al. (2018)	0.54
Coastal / Red (b1/b4)	Vegetation greenness, Chlorophyll-a	Bande and Adam. (2018)	0.74
Coastal / Green (b1/b3)	Chlorophyll-a, vegetative greenness	Bande and Adam. (2018)	0.75

4.1.2 Machine Learning Model Outputs

Table 4.3 below indicates the outputs for the five predictive models. We employ bold fonts to emphasize the superior performance outcomes. The RF model performed the best based on an R^2. For the training set, R^2 measures how well the model explains the variance in the target variable. An R^2 of 0.86 implies that the RF model explains 86% of the variance in the training data, and 0.95 in the testing sequence, which is 95% of the variance, suggests a good fit. However, based on the RMSE, the model deviates by approximately 1.38 micrograms per liter (ug/L) from the actual values for the training set and 0.8 ug/L. The low RMSE indicates good predictive accuracy. The MAPE is a percentage-based metric measuring the average absolute percentage difference between predicted and actual values; for the training set on average, the model's predictions have an absolute percentage error of 15.09% and 0.92% on the testing data, which does not indicate a predictive sizeable mistake. These promising results can be attributed to the model's fitting ability. It outperforms other methods, demonstrating smaller prediction errors than alternative approaches. P values = $8.66\,x10^{-11}$ which is less than 0.05 ($8.66\,x10^{-11} < 0.05$); therefore, the Null hypothesis is rejected in support of the alternative hypothesis.

SVR performed second best in terms of R^2 for both training (0.68) and testing (0.87) and has a lower RMSE and MAPE for both training and testing, with only 18.13% on testing and 19.28%

on testing for the predictive error. P values = $5.01\,x10^{-6}$ whichhan 0.05 ($5.01\,x10^{-6} < 0.05$). Therefore, the Null hypothesis is rejected in support of the alternative hypothesis. Further outperforming LASSO regression, LASSO regression had an R^2 for training and testing at 0.6 and 0.70, respectively, explaining 63% and 70% of the variance in the training and testing set, RMSE of 2.21 and 2.13 for the training and testing set. However, average results are seen within the MAPE with 35.5% training and 25.35% for testing, indicating the predictive error between the predicted Chlorophyll-a and measured Chlorophyll-a values. These results for LASSO regression perform very poorly compared to RF and SVR because of issues like Variable selection. While LASSO performs automatic feature selection by driving some coefficients to exactly zero, it may arbitrarily select one variable over another in the presence of highly correlated predictors. Lastly, MLR and Ridge's regression performed the worst because of MLR's inability to deal with complex issues such as multicollinearity. It is also better than Ridge regression as it has fewer hyperparameters and does not require any tuning of regularization parameters (lambda). Additionally, due to the collection dates of the field data, which are dated to the flight bypass dates, the selected closest date might have potential effects on the model's predictive accuracy, leading to observed outliers. RF training and testing set data followed the line of best fit better with fewer outliers than the other predictive models, further indicating its predictive power. The correlation coefficients between the predicted Chlorophyll-a and measured Chlorophyll-a values for the five predictive machine-learning models exceed 0.5. RF exhibits the highest correlation coefficient, with values of 0.855 during training and 0.945 during testing. Followed by SVR, LASSO, MLR, and lastly, Ridge regression.

Table 4.3: Summary table for Machine learning model outputs.

	Training (70%)			**Testing set (30%)**			
Model	**R^2**	**RMSE**	**MAPE**	**R^2**	**RMSE**	**MAPE**	**P-value**
Randon forest regression	**0.86**	**1.38**	**15.09%**	**0.95**	**0.8**	**10.92%**	**$8.66\,x10^{-11}$**
SVR	0.68	2.37	18.13%	0.87	1.56	19.28%	$5.01\,x10^{-6}$

LASSO	0.63	2.21	35.53%	0.70	2.13	25.35%	$2.77\ x10^{-6}$
MLR	0.63	2.21	35.86%	0.69	2.14	25.38%	$3.48\ x10^{-6}$
Ridge regression (L2 regularization)	0.57	2.38	27.95%	0.77	1.84	29.36%	$7.69\ x10^{-6}$

In this study, five machine learning algorithms were used to predict Chlorophyll-a as an independent variable to validate these models R^2, RMSE, and MAPE were employed as evaluation metrics for model performance. Visual comparisons of the performance of these models can be made using Figure 4.1 through Figure 4.3, respectively. Figure 4.1 indicates that based on R^2, the best model is RF, followed by SVR, LASSO, and MLR, and the worst-performing model is Ridge regression. Figure 4.2 validates the accuracy of RF with the lowest RMSE and MAPE percentage (Figure 4.2).

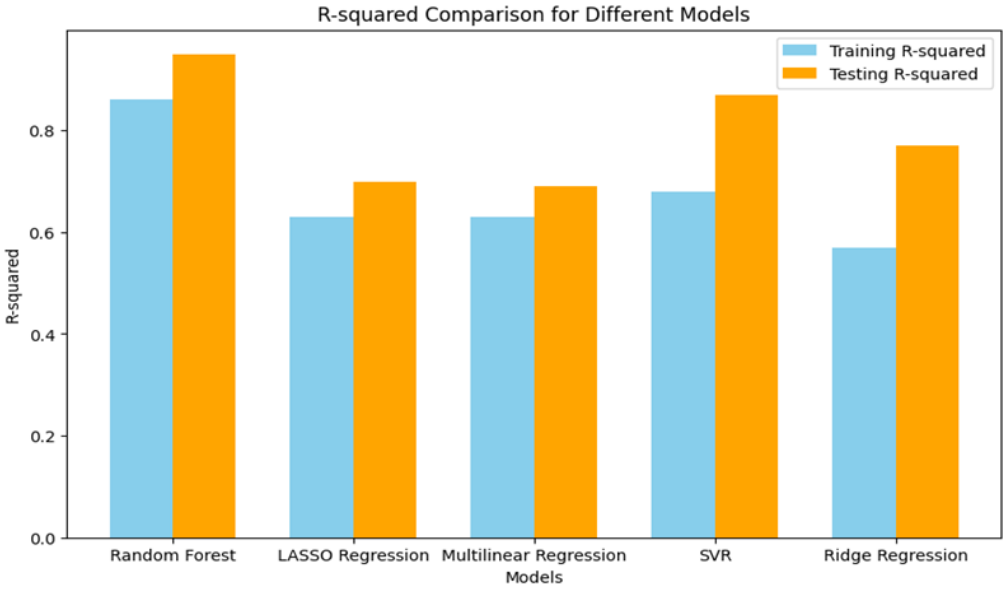


Figure 4.1: R-squared comparison for different machine learning models.

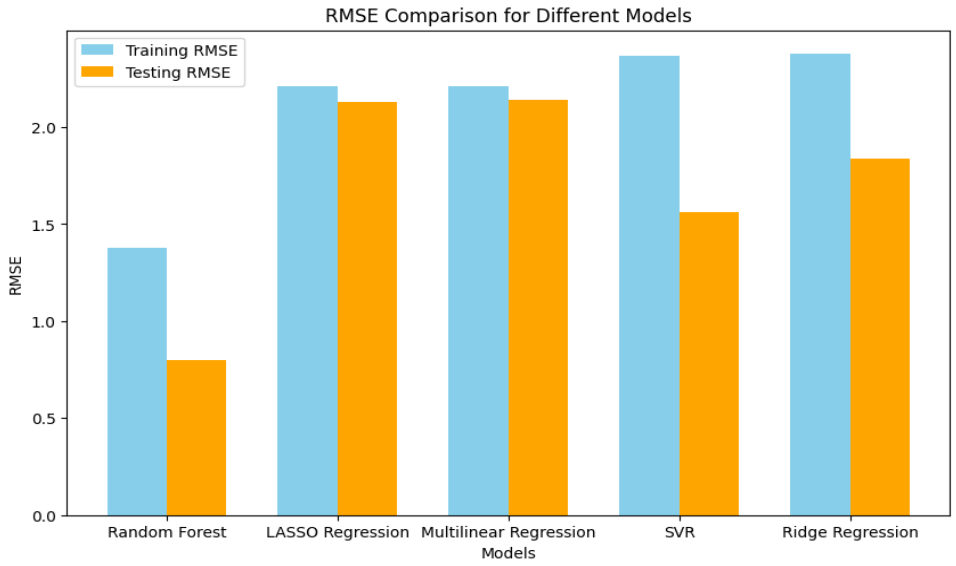


Figure 4.2: RMSE comparison for different machine learning models.

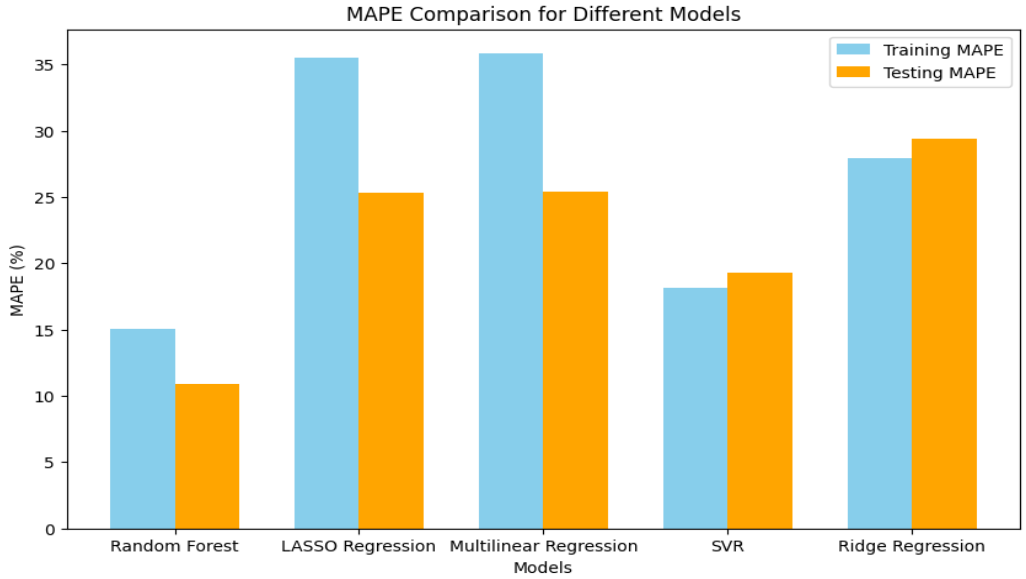


Figure 4.3: MAPE comparison for different machine learning models

4.1.2 Mapping concentration based on best performing model.

The Chlorophyll-a concentration map in Figure 4.4 was based on the RF model, which performed significantly better than SVR, LASSO, MLR, and Ridge regression. The model's Variance of Importance (IncNodePurity) percentage was used to plot the final concentration map. The variance of importance typically refers to the variability in feature importance scores across different runs

or subsets of data. Feature importance in RF rates: The degree to which each feature (band ratio) influences the model's decision-making process. Figure 4.4 indicates the final Chlorophyll-a concentration maps; the distribution patterns of Chlorophyll-a concentrations derived from Sentinel-2 data and compared with on-site measurements indicate elevated values primarily along the Dam's shorelines and regions farther from the reservoir, which had limited sampling locations. Conversely, lower Chlorophyll-a concentrations are predominantly observed in the inner sections of the Dam. This suggests an outward increase in Chlorophyll-a concentrations within the Dam. The likely explanation for this pattern is eutrophication occurring along the Dam's shorelines, where substantial amounts of organic and inorganic deposits from agricultural and urban areas accumulate. The deposition of sediments into the Dam is primarily influenced by wind and wave currents, resulting in a higher concentration of Chlorophyll-a towards the Dam's periphery.

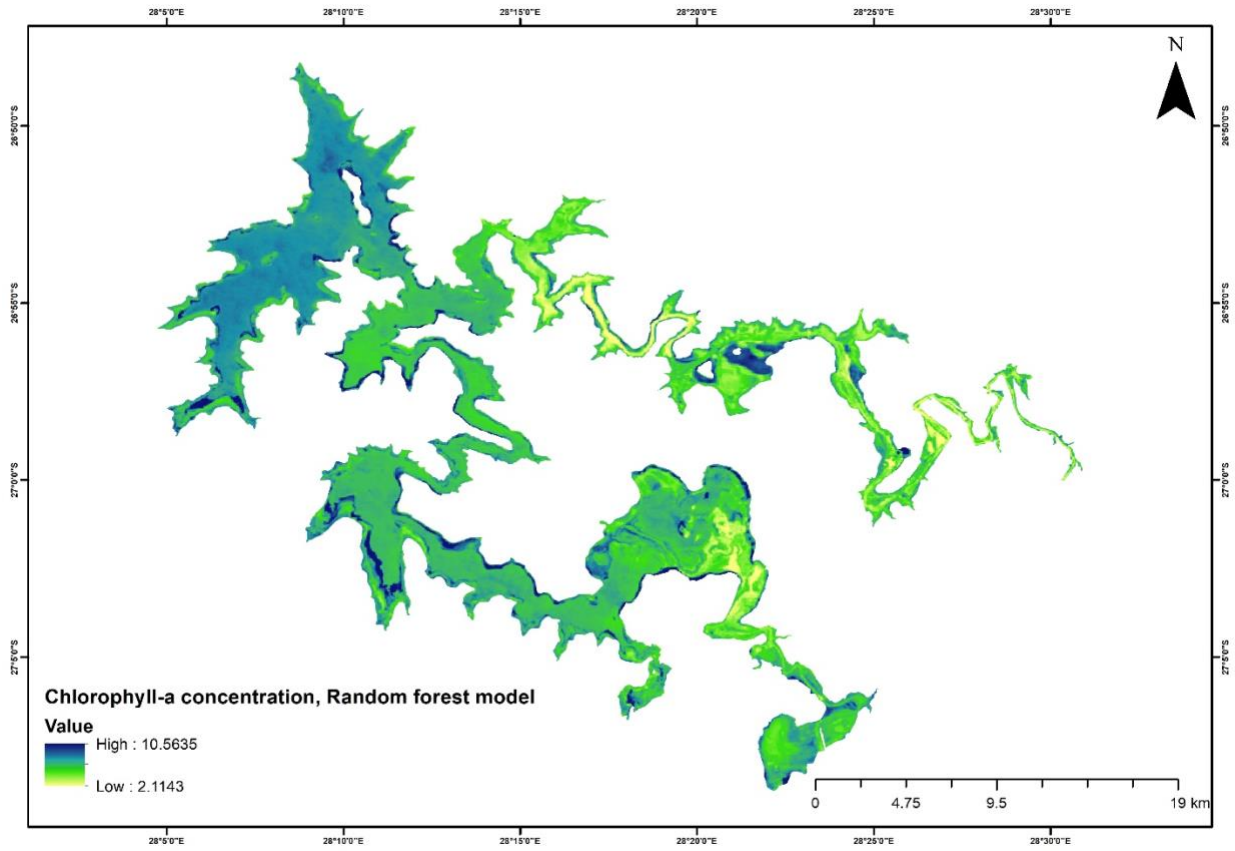


Figure 4.4: Final Random Forest model Chlorophyll-a concentration maps.

Chapter 5

Discussion and Conclusion

5.1 Distribution of in-situ Chlorophyll-a measurements

As indicated in Table 4.1, the results of Chlorophyll-a show the minimum and maximum concentration distribution of Chlorophyll-a, which ranges from 2.92 ug/L to 13.38 ug/L, respectively. This indicates that the water quality is slightly contaminated in one part of the Dam. Additionally, the smallest and largest values of Chlorophyll-a indicated that contamination is not even through the Dam as it shows high difference in Chlorophyll-a concentration across the Dam. Based on the trophic status index by Boyd et al. (1997) and Tanjeck (2019) indicated in Table 2.1, the results of the in-field Chlorophyll-a concentration, the Vaal Dam concentration ranges from low to average, which can be categorized as Mesotrophic, This suggests the presence of algae, leading to a decline in the water's aesthetic appearance and quality, along with some degree of oxygen depletion. These high concentrations can be seen along the shoreline of the Dam, which is accounted for by the land use practices in the area, ranging from Agricultural practices to mining activity. This is further supported by Malahlela et al. (2018, who found that the average concentration of Chlorophyll-a in the Vaal Dam in 2015 – 2016 ranged from 2.5 ug/L to an excessive 35 ug/L (Mesotrophic) along the Dam shoreline. These measured concentrations indicate the continued effect of land use and how they affect the Dam's water quality status. This knowledge is essential in assisting management authorities in identifying the direct sources of contamination within the Dam and finding better strategies to treat and conserve the Vaal Dam.

5.2 Relationship between band ratio reflectance and predicted Chlorophyll-a values.

Sentinel-2 MSI bands were used to predict Chlorophyll-a (as seen in Table 4.2). The best band ratio correlation was seen for Coastal aerosol/ Red (Band 1/Band 4), Coastal / Green (Band 1/Band 3), and NIR/Red (Band 5/Band 4), with correlation coefficients of 0.75, 0.74, and 0.69, respectively. These results indicated the effectiveness of the Visible region of the electromagnetic spectrum in Chlorophyll-a detection, with Band 1(Coastal) aerosol effective for atmospheric correction, Band 2 (Blue) sensitivity to vegetation, Band 3 (Green) where the green peak indicating sensitivity to total Chlorophyll-a in vegetation, Band 4 (Red) indicating maximum Chlorophyll-a

absorption, and Band 5 (VNIR) which is sensitive to Chlorophyll-a as high content reflects significantly in this region. This Observation is Further supported by Bande (2020), who used Coastal aerosol, Blue, Green, Red, and NIR band ratios to train her SVR model to predict Chlorophyll-a. Bande and Adam. (2018) also indicated band ratio; Band 1, Band 2, Band 1/Band 4, and Band 1/Band 4 had the highest correlation and were used in the Partial Least Square (PLS) model in detecting Chlorophyll-a. Moreover, in many studies, Band 1 has been indicated to be a good indicator of Chlorophyll-a. This is because Band 1 is critical in atmospheric correction. After all, it accurately recovers surface reflectance values. Reducing atmospheric effects allows for more precise measurements of target signals such as Chlorophyll-a concentration in water bodies. Furthermore, Band 1 in the blue segment of the electromagnetic spectrum is sensitive to water constituents such as Chlorophyll-a, which has blue absorption peaks. This band effectively captures the reflectance characteristics associated with chlorophyll, a concentration in water. The observed high correlation coefficient of Band 1 with Chlorophyll-a concentration across multiple studies (Bande and Adam. (2018); Bande, 2020; Tanjeck, 2019; Malahlela et al., 2018) emphasizes its importance in this domain. This strong correlation highlights the close relationship between Band 1 spectral features and Chlorophyll-a concentration, making it a valuable indicator for this specific application.

5.3 Findings obtained from five machine learning models.

This study used five machine learning models to predict Chlorophyll-a concentration in the Vaal Dam. These models consisted of RF, SVR, and three linear models: LASSO, MLR, and Ridge Regression. The performance of these models was then assessed using metrics like R^2, RMSE, MAPE, and P-values. Figures 4.1–4.3 show visual comparisons of individual model performance. A thorough comparison revealed that RF outperformed all other models significantly, followed by SVR, LASSO, and MLR, with Ridge Regression performing the worst. In addition, this study demonstrated the ability of RF and SVR to handle nonlinearity. Furthermore, the ensemble model RF showed improved generalization due to its stochastic nature, which reduced the likelihood of overfitting. Notably, RF had an R^2 of 0.86 for the training set and 0.96 for the testing set, indicating a well-tuned model that accurately captured the relationship between band reflectance and in-field Chlorophyll-a measurements. This is supported by Xu et al. (2021), who showed the efficiency of ensemble models like RF over linear models (Ridge) and Deep learning models like Multilayer

Perceptron (MLP) in predicting variables for water quality predictions. They stated that RF is an ensemble learning method that uses decision trees, which are simpler models than MLP's complex architecture. Decision trees are intuitive and straightforward, making them more robust and less prone to overfitting, particularly with limited data. MLP, on the other hand, maybe more sensitive to outliers if not adequately regularized.

Despite SVR's second-best performance among the models, the testing set had a higher R^2 of 0.87 than the training set's 0.68. This strong performance on the testing set suggests that the SVR model generalizes well to new data, demonstrating its ability to predict unseen data points accurately. This indicates that the model effectively captures underlying patterns and relationships in the data without overfitting. SVR did outperform LASSO, MLR, and Ridge regression (Linear models). SVR outperformed LASSO, MLR, and Ridge regression (linear models) due to its ability to capture nonlinearity and increased robustness to outliers. Furthermore, SVR employs a unique regularization technique based on the margin of support vectors, which is more effective in preventing overfitting than LASSO and Ridge regression. This margin-based regularization is especially useful in high-dimensional spaces with many features relative to the number of samples, where linear models frequently encounter issues such as multicollinearity and overfitting. These findings are supported by Mamun et al. (2019), who compared the effectiveness of SVR and MLR in predicting algae and water clarity; the study indicated that SVR performed significantly better than MLR, MLR's poor performance was attributed to its inability to handle more complex non-linear datasets. Furthermore, Lui et al. (2022) supported the concept that linear machine learning models typically perform poorly and struggle to outperform RF and SVR. However, the study concluded that while RF and SVR may outperform linear models individually, SVR's main limitation is its need for a more extensive training dataset. This is due to SVR's reliance on identifying the optimal hyperplane that maximizes the data point margin, aided by kernel function regularization.

When analyzing the linear models (LASSO, MLR, and Ridge), LASSO performed significantly better, followed by MLR and Ridge, which was the worst. This is because LASSO uses L1 regularization, which penalizes the absolute size of the coefficients, allowing it to effectively shrink coefficients of less important features to zero, resulting in feature selection for irrelevant or redundant features. MLR, on the other hand, seeks to minimize the residual sum of squares

between observed and predicted values; the lack of a regularization parameter makes it more prone to overfitting when the number of features in the dataset is large compared to the number of samples. Ridge regression, unlike LASSO, does not perform feature selection; instead, it reduces the coefficients of correlated features, failing to address issues with redundant or irrelevant features in the training dataset.

5.4 Spatial pattern for Chlorophyll-a concentration in the Vaal Dam.

The spatial pattern of Chlorophyll-a concentration in the Vaal Dam was mapped using the best machine-learning model, RF, to represent the concentration distribution visually (Figure 4.4). The analysis was performed on an atmospherically corrected Sentinel-2 MSI image. The results indicated that the distribution of algae is from low (in the middle region of the Dam) to high (along the Dam shoreline), with the east-southern side of the Dam being more polluted (depicted as light green in Figure 4.4). These findings can be attributed to land use activities along the Dam. These findings align with Obaid et al. (2023), who indicated that the dominant land use activities around the Dam include agricultural practices, outflow of sewage from nearby villages, and mining activity. Malahlela et al. (2018) further elaborated that these pollutants flowing into the Dam show high concentrations of fertilizers and pesticides resulting from farming activities, mining waste products, and the influx of acid mine drainage (AMD), as well as wastewater from septic tanks and toilet effluents. These contaminants cause high concentrations of ions (phosphate and nitrate), heavy metals, and suspended particles, which directly result in increased pollution and algal bloom, thus affecting water clarity, reducing oxygen levels, causing the death of aquatic life, and diminishing the aesthetics of the Dam.

5.5 Conclusion

This study has indicated that Sentinel-2 MSI sensors' spatial and temporal characteristics were sufficient in providing reflectance data suitable for predicting Chlorophyll-a concentration in the Vaal Dam. This is advantageous as Sentinel-2 MSI is free and easy to preprocess. Among the five machine learning algorithms used to predict Chlorophyll-a concentration, the RF ensemble model performed exceptionally well compared to the rest of the models, showing good R^2 values for both the training and testing datasets. Furthermore, RF produced low RMSE and MAPE, indicating its

ability to handle small data points without overfitting, as it is insensitive to outliers and handles non-linearity well. SVR performed slightly better than the linear models (LASSO, MLR, and Ridge Regression) due to its ability to handle non-linear data points. The basic principle of SVR is based on the formation of a hyperplane; thus, small data points may have been a limitation of the SVR model. However, SVR produced fair results in both the training and testing datasets, with testing showing a slightly higher R^2, indicating that SVR is good at learning unseen data. Among the linear models, LASSO outperformed MLR and Ridge Regression mainly due to its L1 regularization parameters, which allow for feature selection by shrinking less significant coefficients to zero. Consequently, Ridge Regression does not conduct feature selection, resulting in the inability to handle redundant or extreme outliers in the dataset and deal with non-linear data points.

The spatial distribution of the Chlorophyll-a concentration map from the RF model indicates that the highest concentration of algal blooms is along the Dam shorelines rather than towards the center. This observation directly supports the notion that land use activities along the Dam shoreline may contribute to the Dam's pollution levels.

5.6 Limitations and recommendation

The study's primary limitation was the need for more adequate field data samples, particularly in the southeastern region of the Dam, which exhibited the highest concentration, as depicted in Figure 4.4. This research recommends that future studies incorporate more field data samples that are equally distributed along the Dam. Another limitation of this study was the difference between the field sampling date and satellite overpasses, which affected the accuracy of the results. Therefore, this study suggests that fieldwork should be coordinated with the date of the next satellite pass over the study area. Furthermore, due to the nature of algal blooms, more than a temporal resolution of a 5-day revisit interval might be needed for this study. Hence, drone imagery is recommended since drones can provide real-time data with higher spatial resolution than Sentinel-2 MSI imagery. Lastly, this study suggests that Band 1 (coastal aerosol) of Sentinel-2MSI be explored expansively as an indicator of Chlorophyll-a's presence within surface water as the best band ratio correlation had band one and other researchers have also used Band 1 individually to map Chlorophyll-a in surface water bodies.

Reference list

Ashphaq, M., Srivastava, P. K., & Mitra, D. (2023). Preliminary examination of the influence of Chlorophyll, Total Suspended Material, and Turbidity on Satellite Derived-Bathymetry estimation in coastal turbid water. Regional Studies in Marine Science, 62, 102920. https://doi.org/10.1016/j.rsma.2023.102920.

Asim, M., Hasan, M. A., & Islam, M. R. (2021). Improving Chlorophyll-a Estimation From Sentinel-2 (MSI) in the Barents Sea Using Machine Learning. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 5529-5549. doi: 10.1109/JSTARS.2021.3074975.

Bande, P., Adam, E., Abd Elbasit, M. A. M., & Adelabu, S. (2018). Comparing Landsat 8 and Sentinel-2 in Mapping Water Quality at Vaal Dam. In 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS) (pp. 9280-9283). IEEE. doi: 10.1109/IGARSS.2018.8518195.

Bande, P., (2020) A comparison of Landsat 8 and Sentinel 2 in mapping Chlorophyll-a and turbidity in the Vaal Dam, and Lakes Shinji and Nakaumi of Japan (Doctoral dissertation, Wits University).

Basak, R., Wahid, K.A. and Dinh, A., 2021. Estimation of the Chlorophyll-a concentration of algae species using electrical impedance spectroscopy. Water, 13(9), p.1223.

Binding, C. E., Bunting, L., Grönfors, O., & Selle, B. (2018). An analysis of satellite-derived chlorophyll and algal bloom indices on Lake Winnipeg. Journal of Great Lakes Research, 44(3), 436-446. doi: 10.1016/j.jglr.2018.04.001.

Boyd, C.E., Diana, J.S., Szyper, J.P., Batterson, T.R. and Piedrahita, R.H. (1997). Water quality in ponds. In: Dynamics of Pond Aquaculture, eds. H.S. Egna and C.E. Boyd, pp. 53–71.

Bukata, R.P., Bruton, J.E., Jerome, J.H., Jain, S.C. and Zwick, H.H., (1981). Optical water quality model of Lake Ontario. 2: Determination of chlorophyll-a and suspended mineral concentrations of natural waters from submersible and low-altitude optical sensors. Applied optics, 20(9), pp.1704-1714.

Braune, E. and Rogers, K.H., (1987). Vaal River catchment: problems and research needs. National Scientific Programmes Unit: CSIR.

Caballero, I., Ruiz, J., & Navarro, G. (2019). Sentinel-2 Satellites Provide Near-Real Time Evaluation of Catastrophic Floods in the West Mediterranean. Water (Basel), 11(12), 2499. doi: 10.3390/w11122499.

Chen, J., Chen, S., Yang, C., He, L., Hou, M. and Shi, T., (2020). A comparative study of impervious surface extraction using Sentinel-2 imagery. European Journal of Remote Sensing, 53(1), pp.274-292.

Dekker, A.G. and Peters, S.W.M., (1993). The use of the Thematic Mapper for the analysis of eutrophic lakes: a case study in the Netherlands. International Journal of Remote Sensing, 14(5), pp.799-821.

Deng, T., Chau, K.W. and Duan, H.F., (2021). Machine learning based marine water quality prediction for coastal hydro-environment management. Journal of Environmental Management, 284, p.112051.

Doxaran, D., Froidefond, J.-M., Lavender, S. and Castaing, P. (2002) Spectral Signature of Highly Turbid Waters. Application with SPOT Data to Quantify Suspended Particulate Matter. Remote Sensing of Environment, 81, 149-161. https://doi.org/10.1016/S0034-4257(01)00341-8.

Du, Y., Xia, J., Chen, X., & Qin, Q. (2016). Water Bodies' Mapping from Sentinel-2 Imagery with Modified Normalized Difference Water Index at 10-m Spatial Resolution Produced by Sharpening the SWIR Band. Remote Sensing (Basel, Switzerland), 8(4), 354. doi: 10.3390/rs8040354.

du Preez, G.C., Wepener, V., Fourie, H. and Daneel, M.S., (2018). Irrigation water quality and the threat it poses to crop production: evaluating the status of the Crocodile (West) and Marico catchments, South Africa. Environmental monitoring and assessment, 190, pp.1-14.

DWAF Department of Water Affairs and Forestry (DWAF).(2008) Strategic framework for water for sustainable growth and development. Department of Water Affairs and Forestry, Pretoria, South Africa.

Feng, X., Li, Y., and Chen, J. (2019). Predicting total suspended solids concentration in a small river using Landsat 8 images. International Journal of Remote Sensing, 40(9), pp. 3113-3126.

Gilerson, A.A., Gitelson, A.A., Zhou, J., Gurlin, D., Moses, W., Ioannou, I. and Ahmed, S.A., (2010). Algorithms for remote estimation of Chlorophyll-a in coastal and inland waters using red and near infrared bands. Optics Express, 18(23), pp.24109-24125

Feng, X., Li, Y., and Chen, J. (2019). Predicting total suspended solids concentration in a small river using Landsat 8 images. International Journal of Remote Sensing, 40(9), pp. 3113-3126.

Gitelson, A., (1992). The peak near 700 nm on radiance spectra of algae and water: relationships of its magnitude and position with chlorophyll concentration. International Journal of Remote Sensing, 13(17), pp.3367-3373.

Gitelson, A.A., Dall'Olmo, G., Moses, W., Rundquist, D.C., Barrow, T., Fisher, T.R., Gurlin, D. and Holz, J., (2008). A simple semi-analytical model for remote estimation of Chlorophyll-a in turbid waters: Validation. Remote Sensing of Environment, 112(9), pp.3582-3593.

Han, T. and Goodenough, D.G., (2008). Investigation of nonlinearity in hyperspectral imagery using surrogate data methods. IEEE Transactions on Geoscience and Remote Sensing, 46(10), pp.2840-2847.

Hassan, N. and Woo, C.S., (2021), August. Machine learning application in water quality using satellite data. In IOP Conference Series: Earth and Environmental Science (Vol. 842, No. 1, p. 012018). IOP Publishing.

Hu, M., Zhang, Y., Ma, R., Xue, K., Cao, Z., Chu, Q. and Jing, Y., (2021). Optimized remote sensing estimation of the lake algal biomass by considering the vertically heterogeneous chlorophyll distribution: Study case in Lake Chaohu of China. Science of The Total Environment, 771, p.144811.

Jiang, X., Li, X., and Wang, J. (2020). Using drone and Sentinel-2 images to predict turbidity in a small lake. Remote Sensing, 12(8), pp. 1-10.

Johnson, R.W., (1978). Mapping of chlorophyll a distributions in coastal zones. Photogrammetric engineering and remote sensing, 44(5), pp.0617-0624.

Kallio, K., Kutser, T., Hannonen, T., Koponen, S., Pulliainen, J., Vepsäläinen, J. and Pyhälahti, T., (2001). Retrieval of water quality from airborne imaging spectrometry of various lake types in different seasons. Science of the Total Environment, 268(1-3), pp.59-77.

Kasprzak, P., Szarek-Gwiazda, E., & Koschel, R. (2008). Chlorophyll a concentration across a trophic gradient of lakes: An estimator of phytoplankton biomass?. Limnologica, 38(3), 327-338. doi: 10.1016/j.limno.2008.07.002.

Khan, S., Jahan, R., Rahman, M.A. and Haque, M.M., (2019). Eutrophication enhances phytoplankton abundance in the Maheshkhali channel, Bay of Bengal, Bangladesh. Aust. J. Sci. Techno, 3, pp.141-147.

Lee, Z., Casey, B., Arnone, R., Weidemann, A., Parsons, R., Montes, M.J., Gao, B.C., Goode, W., Davis, C.O. and Dye, J. (2007) Water and Bottom Properties of a Coastal Environment Derived from Hyperion Data Measured from the EO-1 Spacecraft Platform. Journal of Applied Remote Sensing, 1, 011502.

Li, X., Chen, Z., and Wang, Y. (2019). Mitigating the effects of algal blooms in small lakes: A review of management strategies. Journal of Water Management, 57(3), pp. 123-132.

Liu, N., Townsend, P.A., Naber, M.R., Bethke, P.C., Hills, W.B. and Wang, Y., (2021). Hyperspectral imagery to monitor crop nutrient status within and across growing seasons. Remote Sensing of Environment, 255, p.112303.

Llodrà-Llabrés, J., Martínez-López, J., Postma, T., Pérez-Martínez, C. and Alcaraz-Segura, D., (2023). Retrieving water Chlorophyll-a concentration in inland waters from Sentinel-2 imagery: Review of operability, performance and ways forward. International Journal of Applied Earth Observation and Geoinformation, 125, p.103605.

Malahlela, O.E., Oliphant, T., Tsoeleng, L.T. and Mhangara, P., (2018). Mapping Chlorophyll-a concentrations in a cyanobacteria-and algae-impacted Vaal Dam using Landsat 8 OLI data. South African Journal of Science, 114(9-10), pp.1-9.

Mamun, M., Kim, J.J., Alam, M.A. and An, K.G., (2019). Prediction of algal Chlorophyll-a and water clarity in monsoon-region reservoir using machine learning approaches. Water, 12(1), p.30.

Obaid, A., Adam, E. and Ali, K.A., (2023). Land Use and Land Cover Change in the Vaal Dam Catchment, South Africa: A Study Based on Remote Sensing and Time Series Analysis. Geomatics, 3(1), pp.205-220.

Papa, F., Prigent, C. and Rossow, W.B. (2008) 'Monitoring Flood and Discharge Variations in the Large Siberian Rivers From a Multi-Satellite Technique', Surveys in Geophysics, 29(4-5), pp. 297-317. doi: 10.1007/s10712-008-9036-0.

Rattan, P., Penrice, D.D. and Simonetto, D.A., (2022). Artificial intelligence and machine learning: what you always wanted to know but were afraid to ask. Gastro Hep Advances, 1(1), pp.70-78.

Rodriguez-Galiano, V. F., Sánchez Castillo, M., Dash, J., & Atkinson, P. M. (2016). Modelling interannual variation in the spring and autumn land surface phenology of the European forest. Biogeosciences, 13(11), 3305-3317. DOI: 10.5194/bg-13-3305-2016. [CC BY 3.0 License].

Segawa, T., Yamashita, Y., Okumura, T. and Yamanokuchi, T., (2019). Satellite derived bathymetry using machine learning and multi-temporal satellite images, Remote Sens., 11, 1155.

Sardooi, E.R., Azareh, A., Choubin, B., Barkhori, S., Singh, V.P., and Shamshirband, S. (2019). Applying the remotely sensed data to identify homogeneous regions of watersheds using a pixel-based classification approach. Applied Geography, 111, p.102071.

Sarker, I.H., (2021). Machine learning: Algorithms, real-world applications and research directions. SN computer science, 2(3), p.160.

Shin, Y., Kim, T., Hong, S., Lee, S., Lee, E., Hong, S., Lee, C., Kim, T., Park, M.S., Park, J. and Heo, T.Y., (2020). Prediction of Chlorophyll-a concentrations in the Nakdong River using machine learning methods. Water, 12(6), p.1822.

Singhal, G. et al. (2019) 'Chlorophyll estimation using multi-spectral unmanned aerial system based on machine learning techniques', Remote Sensing Applications, 15(Journal Article), p. 100235. doi: 10.1016/j.rsase.2019.100235.

Smith, J., Brown, L., and Jones, P. (2018). Mitigating the effects of algae and turbidity in small water bodies. Journal of Water Management, 56(3), pp. 123-132.

Smith, M.E., Lain, L.R. and Bernard, S., (2018). An optimized chlorophyll a switching algorithm for MERIS and OLCI in phytoplankton-dominated waters. Remote Sensing of Environment, 215, pp.217-227.

Swanepoel, A., Du Preez, H.H. and Cloete, N., (2017). The occurrence and removal of algae (including cyanobacteria) and their related organic compounds from source water in Vaalkop Dam with conventional and advanced drinking water treatment processes. Water Sa, 43(1), pp.67-80.

Tanjeck, A.E., (2019). Using Multispectral Remote Sensing for Mapping and Monitoring Water Quality at the Vaal Dam (Doctoral dissertation, Faculty of Science, University of the Witwatersrand).

Tassan, S. (1994) Local Algorithms Using SeaWiFS Data for the Retrieval of Phytoplankton, Pigments, Suspended Sediment, and Yellow Substance in Coastal Waters. Applied Optics, 33, 2369-2378. https://doi.org/10.1364/AO.33.002369.

Tran, T., Nguyen, T., and Tran, D. (2020). Using a drone and Sentinel-2 images to predict cyanobacteria density in a small lake. Remote Sensing, 12(4), pp. 1-10.

Walmsley, R., (2000). Perspectives on Eutrophication of surface waters. policy/research document. Water research commission project.

Wang, Y., Chen, Z., and Li, X. (2019). Predicting Chlorophyll-a concentration in a small lake using Landsat 8 images. Journal of Applied Remote Sensing, 13(1), pp. 1-10.

Wolpert, D.H. and Macready, W.G., (1997). No free lunch theorems for optimization. IEEE transactions on evolutionary computation, 1(1), pp.67-82.

Xu, J., Xu, Z., Kuang, J., Lin, C., Xiao, L., Huang, X. and Zhang, Y., (2021). An alternative to laboratory testing: Random forest-based water quality prediction framework for inland and nearshore water bodies. Water, 13(22), p.3262.

Yacobi, Y.Z., Moses, W.J., Kaganovsky, S., Sulimani, B., Leavitt, B.C. and Gitelson, A.A., (2011). NIR-red reflectance-based algorithms for Chlorophyll-a estimation in mesotrophic inland and coastal waters: Lake Kinneret case study. Water research, 45(7), pp.2428-2436.

Zheng, W., Zhang, Y., Ruan, J., & Hu, G. (2016). Improved Species-Specific Lysine Acetylation Site Prediction Based on a Large Variety of Features Set. PLoS One, 11(5), e0155370.

Zhang, Y., Liu, X., and Li, Y. (2019). Predicting dissolved oxygen concentration in a small river using Sentinel-2 images. Remote Sensing, 11(11), pp. 1-10.

Zheng, Y., Liu, X., and Li, Y. (2020). The use of satellite imagery for predicting and managing harmful algal blooms in freshwater systems: A review. Journal of Water Management, 58(3), pp. 123-132.

Zhou, X., Huang, X., and Wang, J. (2020). Comparing the use of drone and Landsat 8 images for predicting nitrate concentration in a small lake. Journal of Applied Remote Sensing, 14(1), pp. 1-10.

Zhou, L., Ma, W., Zhang, H., Li, L. and Tang, L., (2015). Developing a PCA–ANN model for predicting chlorophyll a concentration from field hyperspectral measurements in Dianshan Lake, China. Water Quality, Exposure and Health, 7, pp.591-602.