Using Multispectral Remote Sensing Image Data and Neural Networks to Automatically Predict Optically Active Parameters for Inland Water Quality Analysis

Rohan Sikand

Entry into the Stockholm Junior Water Prize 2020

New Jersey

I. Abstract Inland water systems are essential to our environment because they are vital ecosystems that are biodiverse. However scientists have found that monitoring inland water quality through *in situ* testing can be costly, time consuming and spatially limited. In this paper, I present in detail, a machine learning algorithm that takes in multispectral remote sensing data from the AquaSat data set as input to predict optically active water quality parameters as desired output. Several data preprocessing steps were performed. A neural network model consisting of five layers was implemented taking in the preprocessed multispectral data as input. Several water quality parameters that were deemed optically active based on their spectral signatures were predicted as the output of the neural network model. The parameters are Chlorophyll-A, Dissolved Organic Carbon, Total Inorganic Sediment, and Total Suspended Sediment. Multiple metrics were used to calculate the models performance for each parameter predicted including mean absolute error, mean square error, and root mean squared error. The results proved to be statistically significant–especially for Chlorophyll-A, and Total Inorganic Sediment. However, several steps can be taken to improve the performance of the model introduced in this experiment such as utilizing more data or introducing other machine learning algorithms. Monitoring water quality is indeed compulsory to determine trends in the fluctuation of aquatic environments and how their nutrient and pollutant contents are affected by surrounding human activity. With the advancements in technology, remote inland water quality assessment and conservation strategies will be available at a fraction of the cost of their parallel *in situ* evaluation. This will lead to effective remote sensing capabilities in vast inland aquatic systems, especially in communities with limited resources. My experiment shows that the use of machine learning is applicable to predict water quality parameters using remote sensing data without the need for *in situ* testing.

II. Table of Contents

III. Key Words Water Quality, Remote Sensing, Neural Networks, Multispectral Imaging, Regression, Machine Learning, Chlorophyll a, Dissolved Organic Carbon, Total Inorganic Sediment, Total Suspended Sediment

IV. Abbreviations and Acronyms Chlorophyll a - chl_a, Dissolved Organic Carbon - doc, Total Inorganic Sediment - tis, Total Suspended Sediment - tss, Near-infrared Spectroscopy - nir, Artificial

neural network - ANN, Mean squared error - MSE, Mean absolute error - MAE, Root mean squared error - RMSE

V. Acknowledgements The work in this paper was done solely by the author with little assistance from other people. To learn the necessary information required to complete this research project, online resources, textbooks, and primary literature were utilized. Special thanks to Dr. Alison Appling of the United States Geological Survey for assistance with manipulating and using the AquaSat data set. Furthermore, appreciation to Mr. James Gill, for teaching me the fundamentals of computer science and for the endless inspiration. Also, appreciation to Ranney School's administrative staff, including guidance counselor Mr. Adam Materasso, for flexibility to allow me to independently pursue this project during the school year.

VI. Biography Rohan Sikand is a senior at Ranney School in Tinton Falls, New Jersey, and a rising freshman at Stanford University, where he hopes to pursue a B.S. in Computer Science with a concentration in artificial intelligence. He published a 77-page history research paper in The Concord Review on the Chesapeake-Leopard Affair of 1807 and a 30-page philosophy of the mind paper titled "Defining Humanity in the 21st Century" in the ISSCY. In addition to his humanities achievements, Rohan completed the Rutgers WISE program where he isolated, sequenced, and discovered duckweed DNA sequences using bioinformatics software. Rohan's passion lies machine learning in which he has pursued several other research projects including "EffuseNet - Transfer Learning with Deep Convolutional Neural Networks for Differentiating Exudative and Transudative Pleural Effusion Through Ultrasounds" and "Convolutional Neural Networks for Computer-Aided Detection of Musculoskeletal Abnormalities". Rohan believes that with the increase in data sets since the turn of the century, applying machine learning to various problems will be pivotal to reaching his goal of making a large impact on the world.

1 Introduction

Undoubtedly, humanity and our global environment are dependent on freshwater resources. A large amount of freshwater is stored in inland aquatic systems. The water quality of these inland systems are affected by population growth, economic changes and the utilization of surrounding land-forms. However,

freshwater management has become a global challenge and the evaluation of water quality is vital to the protection of freshwater resources. Monitoring water quality properties remotely has long been utilized by scientists throughout the world for over 50 years. However, previously airborne and satellite sensors were used to examine and interpret water quality components for oceanic productivity. Remote water quality analysis of inland aquatic systems have been scarce partially due to limited technology available for inland water quality remote sensing.

Recently with the emergence of machine learning, using this multispectral imaging data can be implemented into machine learning algorithms to predict inland water quality parameters for water quality analysis. Research trends have highlighted improved dataset availability and processing knowledge leading to a deeper understanding of the water quality of our vast water systems.

Parameters that were deemed optically active based on their spectral signatures and wave lengths emitted that contribute to water clarity are Chlorophyll-A (chl-a), Dissolved Organic Carbon (doc), Total Inorganic Sediment (tis), Total Suspended Sediment (tss). These water quality parameters are to be predicted and explored in this experiment. These components, when combined, can affect water clarity and ultimately impact water quality. By monitoring these parameters remotely, conclusions about water quality can be drawn more effectively and efficiently than *in situ* data alone [1].

One of the most vital components of plants are chlorophylls which are active compounds responsible for photosynthesis and the conversion of light into energy. Most of a plant's chlorophyll is Chlorophyll a (chl_a) which can be found in all plants, algae and cyanobacteria. In water systems, chl_a is used as an indicator for total algal biomass. This is critical because the algal biomass of an aquatic system impacts its overall productivity known as a trophic state. Chlorophyll measurements reflect the amount of algae growing in a specific waterbody. Although algae naturally lives in inland water bodies, an abundance can cause decreased levels of dissolved oxygen [2]. Water quality degradation occurs with an increase in algal biomass as measured by the concentration of chl_a. Water systems with artificially abundant levels of certain nutrients from fertilizers, septic systems, sewage treatment plants and generalized urban spillover could potentially have excessive amounts of chl_a. More importantly a water system's algal biomass reflects its integrity. Although some algal blooms are non-toxic, algae containing phycocyanin producing cyanobacteria are toxic to livestock, wildlife and humans. Unfortunately, climate change and artificially driven nutrient alterations have increased the prevalence of these toxic algal blooms. The optical spectral signature of chl_a is altered by the composition of the phytoplankton present in the aquatic system ana-

lyzed. Thus, a change in water quality due to a change in chl_a concentration, will be emitted remotely through multispectral spectral signatures.

Dissolved organic carbon (doc) is a source of energy and carbon for heterotrophic organisms and can potentially influence the metabolism and balance of an ecosystem. Generally dissolved organic carbon can be produced by the decomposition of plant and animal material and by the soluble particles released by algae and bacteria. Visually dissolved organic matter causes the dark colors of many surface water systems, including wetlands. However, doc is significant in its impact on nutrient sequestration and a valuable source of carbon for microorganisms. Visually dissolved organic matter causes the dark colors of many surface water systems. These changes corresponds to changes in its respective spectral signatures in remote sensing multispectral imaging.

Total suspended solids refer to both Total Inorganic Sediment (tis) and Total Suspended Sediment (tss). The composition of these particles vary geographically influenced by phytoplankton and inorganic sediment. Surprisingly, the carbon deposition into inland lakes and reservoirs is twice the deposition into oceans, despite lakes only comprising 4% of total land area [3]. Previous remote sensing studies have examined tis and tss in coastal systems with delivery into oceans along with impact of reservoirs on sediment concentration, and the impact of changes in land use. Because of this, these two parameters are deemed optically active and are used in this experiment.

The combination of Chlorophyll-A, Dissolved Organic Carbon, Total Inorganic Sediment, and Total Suspended Sediment collectively influence water clarity. It has been established that water clarity, and hence water quality, ultimately control the viability of freshwater ecosystems. Therefore, there is a global need for improving the accuracy and efficiency of water quality monitoring methods. Conventional *in situ* monitoring of inland water quality are spatially limiting, costly and time consuming.

Although satellite based measurements have been commonly used in monitoring oceans, only recently have they been used to study inland waters. Because inland water systems are inhomogeneous, multispectral data images are often implemented to collect data for bodies of inland water from satellite systems, resulting in a more thorough and efficient evaluation of our aquatic systems. In this paper, I present in detail how I utilize a machine learning algorithm that takes in multispectral remote sensing data from the AquaSat data set as input to predict the optically active water quality parameters as described as desired output.

2 Materials and Methods

To test the performance of the use of remote sensing for water quality parameter prediction, a machine learning model was implemented. First, the selected dataset was modified to fit the needs of this experiment. Then, several data preprocessing steps were applied to enhance the performance of the machine learning models. Since the goal of this experiment is to predict water quality parameters through multispectral remote sensing, several water quality parameters were chosen to be predicted based on the spectral signatures. Subsequently, a machine learning neural network model was implemented and tested for this experiment. Figure 1 shows an overall schematic of the procedure of the experiment visualizing the steps taken.

2.1 Data Set

The first step in conducting this experiment was to identify a data set that was applicable for this task. With growing trends in remote sensing and multispectral data, more and more data sets are available open-source. One of the main drawbacks to machine learning is not having enough data which can subsequently cause overfitting. Thus, a large amount of data points was needed for this experiment. Furthermore, both *in situ* field test data, consisting of water quality parameters, and remote sensing multispectral image data points were needed. Introduced in [4], AquaSat is a dataset that was constructed to enable the use of remote sensing of water quality for inland waters. AquaSat provides matchup data between remote sensing multispectral image data and *in situ* field test data. Because of this, AquaSat was utlized for this experiment.

2.2 Data Preprocessing

Since the AquaSat data set is made for general purpose use, several preprocessing tasks were performed to modify the data set for optimal use in this experiment. First, several columns and parameters were not needed for this experiment and were thus deleted from the data set. These deleted columns included parameters such as date, time, landsat id, SiteID, and path. Secondly, all rows of data that contained a NaN value were deleted from the data set and not included in the experiment. Third, some columns included categorical information rather than numerical information (i.e. landsat satellite number). To fix

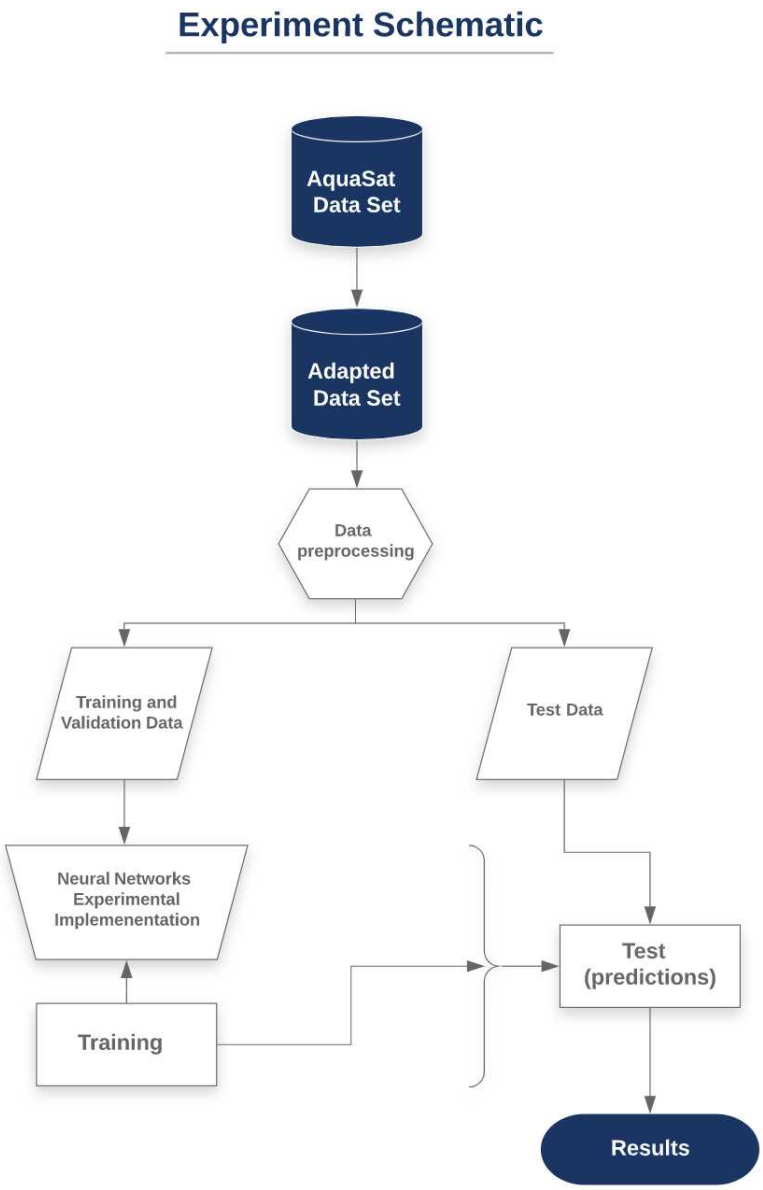

Figure 1: Complete schematic of the experiments procedure, showcasing the steps and methodology taken. The original data set was adapted to fit the needs for this experiment. The data set was split into training, validation, and testing. The data was then preprocessed using several different techniques. A neural network model was experimentally implemented and trained. The model then made test predictions which formulate the results listed in section 3.

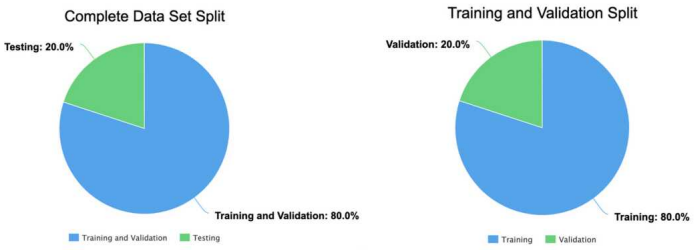

Figure 2: A visual representation of the data split shown in pie charts. The data set was first split into 80% training and validation and 20% testing. Then, the training and validation data was split into 80% training and 20% validation.

this, all categorical parameters were one hot encoded. Fourth, certain *in situ* test parameters did not help improve the overall precision of the model because of their optically inactive multispectral signatures.

A key point to recognize is that not all parameters used in *in situ* field tests to measure water quality are useful when using multispectral remote sensing. Certain parameters can be deemed optically inactive. That is, these parameters vary minimally in multispectral wave length for different water quality levels. Because of this, it is difficult to accurately predict these parameters using multispectral remote sensing without *in situ* field testing and are thus not useful in this experiment. However, other parameters are optically active. For these parameters, different multispectral wave lengths are radiated for different water quality levels. Because of this, machine learning can be used to accurately predict these optically active parameters and are therefore explored in this experiment. Table I shows the adjusted AquaSat data set including all of the parameters used in this experiment and content descriptions of each parameter.

Since this experiment was a machine learning task, the data set needed to be split into training, validation, and testing. Once the complete data set was preprocessed, it was split into 80% training and 20% testing, the standard amounts for a machine learning task. Furthermore, while undergoing training, the machine learning model needs to be tested against new data so that the models performance can be measured during training. Thus, the training data, which consists of 80% of the original data set, was divided into 80% for training purposes and 20% for validation purposes. A visual representation of the data set splits is shown in figure 2.

Table 1: The adjusted AquaSat data set used in this experiment. The content descriptions are obtained from the AquaSat paper [4]. * indicates categorical data that was one hot encoded during preprocessing

Name	Contents
blue	Median blue reflectance
blue_sd	Standard deviation of blue
green	Median green reflectance
green_sd	Standard deviation of green
nir	Median nir reflectance
nir_sd	Standard deviation of nir
pixelCount	Number of water pixels that are averaged into each median and sd value
qa	The quality assesment band indicating clouds,land, and other classifications
qa_sd	Standard deviation of the quality band
red	Median red reflectance
red_sd	Standard deviation of red
sat*	Landsat satellite (5,7,or 8)
swir1	Median shortwave infrared reflectance at 1550-1750 nm
swir1_sd	Standard deviation of shortwave infrared
swir2	Median of shortwave infrared reflectance at 2000-2350 nm
swir2_sd	Standard deviation of shortwave infrared
chl_a	Chlorophyll a concentration in ug/L
doc	Chlorophyll a concnentration in ug/L
tis	Total Inorganic Sediment in mg/L
tss	Total Suspended Sediment in mg/L
type*	Type of waterbody, either lake, river, or estuary

2.2.1 Z-score Normalization

An intrinsic trait in machine learning algorithms is to attempt to find trends by comparing features of data points [5]. Subsequently, if the features the model learns from are on widely different scales and contain different ranges, the less accurate the model becomes. In this case, some features are naturally weighted heavier than others creating an unbalanced prediction. However, normalization, a common data preprocessing step, can be used to adjust the scales and ranges of the features to similar levels for all features effectively balancing the data [6]. To avoid any outliers, z-score normalization was performed on all of the data points and is defined below:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where x is the selected feature, μ is the mean value of the feature, and σ is the standard deviation of the feature.

2.3 Water Quality Parameters to be Predicted

The goal of this experiment is to predict optically active water quality parameters directly through remote sensing instead of the currently used *in situ* field tests. Thus, within the adjusted data set, the parameters that are to be predicted were chosen based on their optically active spectral signatures. These parameters are Chlorophyll a (chl_a), Dissolved Organic Carbon (doc), Total Inorganic Sediment (tis), and Total Inorganic Sediment (tss).

2.4 Water Quality Parameter Predictions through Regression

Since the intended output of the machine learning model is a single numerical value, the task in this experiment is considered a regression task. Several water quality parameters listed in section 2.3 are each predicted separately using the same machine learning model and data. Thus, for each parameter predicted, the other *in situ* water quality parameters were deleted leaving only the water quality parameter that is intended to be predicted and the respective multispectral data. This ensures that only the multispectral data is used to predict the selected water quality parameter. This process was repeated for each water quality parameter respectively.

2.4.1 Neural Networks

To predict these water quality parameters, a neural network, the most prominent deep learning algorithm, was created and fitted to the data. A neural network is inspired from the synapses of neurons in the human brain. In biological neurons, each neuron contains dendrites and a single axon. Both the input and the output is in the form of electrical impulses. The input is received from the dendrites which is then transmitted as output through the axon. The neuron formulates the sum of all of its inputs to determine if the sum is greater than the neuron's firing threshold. If it is, the neuron fires off a new electrical impulse along its axon. The axon then distributes the signal along its branches of synapses and which dispurses the signal upon thousands of other neurons [7].

In an artificial neural network (ANN), a node or artificial neuron aims to replicate this same function that a biological neuron performs. Instead of the dendrites producing the input, the input is simply variables $(x1, x2, x3, \ldots, xN)$, which are weighted differently using weights $(w1, w2, w3, \ldots, wN)$. These variables are then summed by the artificial neuron in a similar fashion to how a biological neuron sums its inputs. This weighted sum is the output of an artificial neuron which, in turn, is analogous to a biological neuron's action potential [8].

In mathematical notation, the weighted sum, s, is calculated as follows:

$$s = \sum_{i=1}^{n} w_i x_i \tag{2}$$

where s the weighted sum of all the inputs, n is the total number of inputs, x_i is a vector of inputs, and w_i is a vector of weights.

In classification problems, a nonlinear activation function is attached to the end of each artificial neuron which is used to classify given data to certain classes. However, since this experiment is a regression task, the neural network that is used in this experiment is designed to output a single numerical value–that is, one of the water quality parameters to be predicted.

An ANN is formed when multiple artificial neurons are combined forming complex connections between the neurons where the output of one neuron serves as input for another neuron. Layers are formed by stacking multiple connected groups of neurons, giving an ANN the visual representation seen in figure 3.

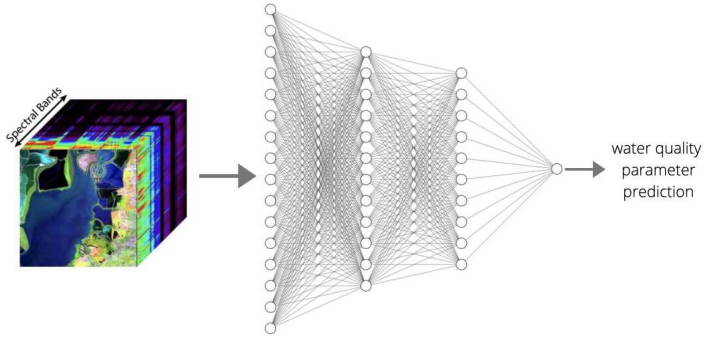

Figure 3: A general visual representation of an ANN using multispectral data as input to predict a water quality parameter prediction as output.

Neural Network Layers:

Layer (type)	Output Shape	Param #
dense (Dense)	(None, 64)	640
dense_1 (Dense)	(None, 64)	4160
dense_2 (Dense)	(None, 64)	4160
dense_3 (Dense)	(None, 64)	4160
dense_4 (Dense)	(None, 1)	65

Total params: 13,185
Trainable params: 13,185
Non-trainable params: 0

Figure 4: A detailed summary of all of the layers for the neural network model used in this experiment.

2.4.2 Neural Network Experimental Implementation

The neural network model used this experiment was built from scratch and consisted of five layers. The first layer is the input layer which takes in the multispectral image data. The second, third, and fourth layers of the neural network are all dense (nonlinear) layers consisting of 64 neurons and 4160 trainable neural network parameters. The last and final layer of the neural network has a shape of 1 so that it outputs a single numerical value: the water quality parameter that is to be predicted. Figure 4 shows in detail a visual representation of the neural network's layers.

2.4.3 Compiling and Training

Before training can begin, several metrics need to be defined which constitutes the compiling step. In specific, an optimizer, a loss function, and several metrics for validation purposes need to be specified. In this experiment, the RMSProp optimizer, introduced in [9], was implemented and used. For the loss function, mean squared error (MSE) was used as defined below:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2 \tag{3}$$

where N is the total number of data points, f_i is the actual water quality parameter value and y_i is the predicted value. The metrics used for validation purposes during training were MSE, as defined above, and mean absolute error (MAE), as defined below:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|f_i - y_i| \tag{4}$$

where N is the total number of data points, f_i is the actual water quality parameter value and y_i is the predicted value.

The standard mechanism to train a neural network is to use the backpropogation algorithm [10]. The neural network model used in this experiment was trained using backpropogation for a total of 100 epochs (iterations of the data).

3 Results

The purpose of this experiment is to use multispectral remote sensing data to predict water quality parameters using machine learning. To generate these results, the neural network model, described in 2.4.2, predicted a numerical value for each water quality parameter in the test set–which represents 20% of the total data and data that the model was not trained or validated on. This section details the results for these predictions in a multitude of ways.

3.1 Training and Validation Graphs: MSE (Loss) and MAE

When the neural network is trained using backpropogation, the RMSProp optimizer undergoes an iterative method for optimizing a loss function. As specified, the loss function used in this experiment

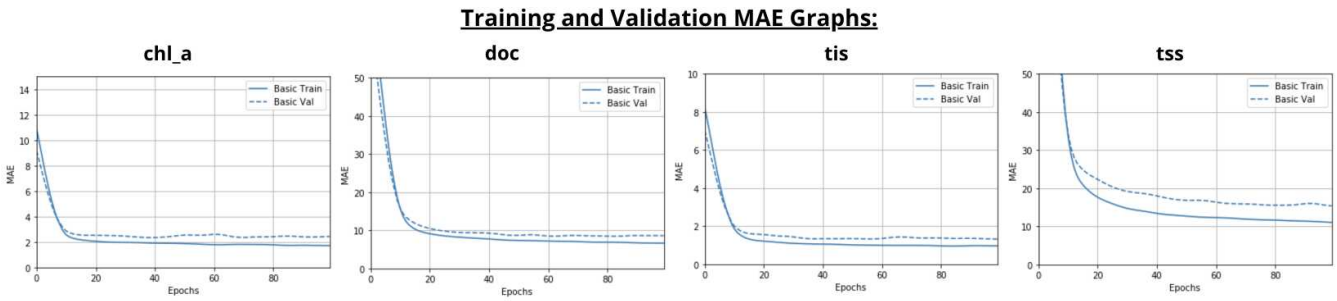

Figure 5: Training and validation MAE graphs of the neural network for all water quality parameters.

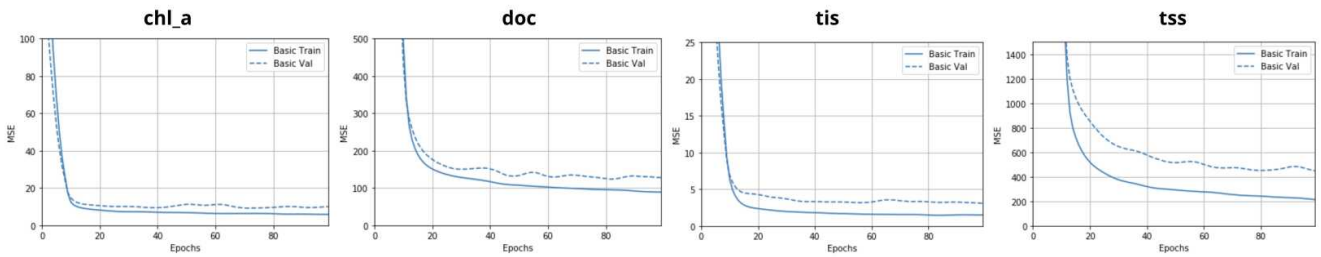

Figure 6: Training and validation MSE (loss) graphs of the neural network for all water quality parameters.

is mean squared error and is defined in equation 3. Furthermore, the other metric that was specified in the compiling step was mean absolute error (defined in equation 4) and was thus also measured during training and validation. The graphs of training and validation MAE and training and validation MSE (loss) are shown for the neural network model for all water quality parameters in figure 5 and figure 6 respectively. These graphs are important since it shows how the model is improving over each iteration of the data. Additionally, these graphs can be used to detect model overfitting. Overfitting is a problem that is detrimental to a machine learning models success and is analyzed in greater detail in section 4.

3.2 Predicted vs. Actual Values Scatter Plots

To visually show the difference between predicted and actual parameter values, a scatter plot is made. On the x-axis of each graph is the true actual water quality parameter value and on the y-axis of each graph is the predicted water quality parameter value. Furthermore, in each graph, a line with a slope of 1 is drawn which represents perfect prediction. Thus, the more linear the scatter plot points are, the better the model is predicting. Figure 7 shows a scatter plot showing the relationship between the actual and

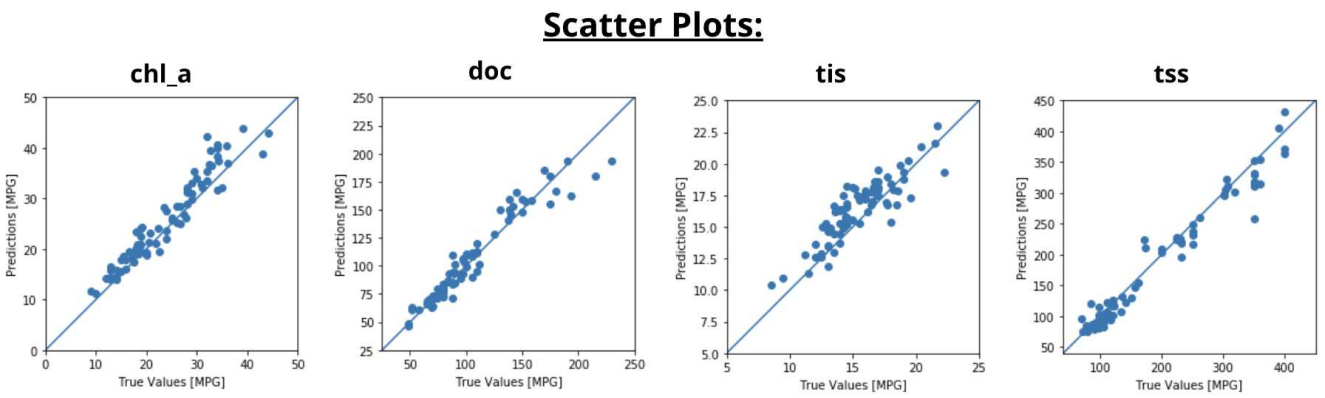

Figure 7: Scatter plots, showing the difference between predicted and actual parameter values, for all water quality parameters predicted using the neural network model.

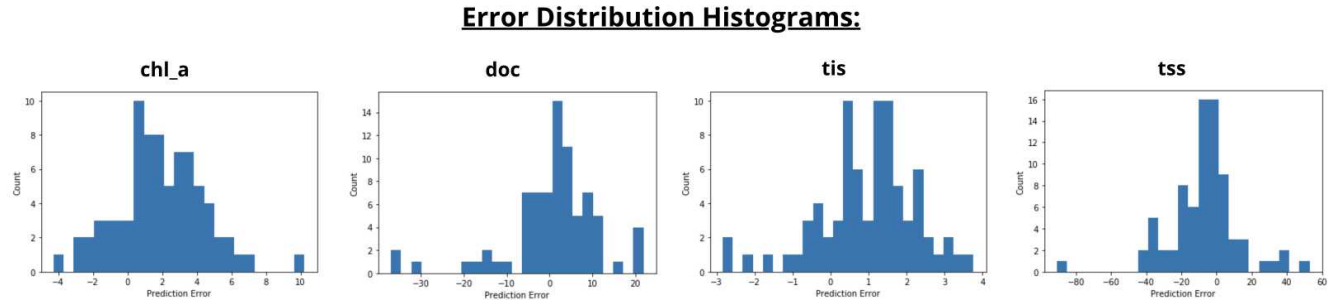

Figure 8: Error distribution histograms for all water quality parameters that were predicted using the neural network model.

predicted values for all water quality parameters predicted using the neural network model.

3.3 Error Distribution Histograms

Error distribution histograms count the measures of differences between predicted and actual values (error). All of the generated error points are placed in 25 class interval bins. The total count for each bin represents the amount of data within each bin. Histograms are used to graph these amounts. On the x-axis, the error value is listed. On the y-axis, the total count is shown. These graphs are important for analyzing the error distribution. If the histogram forms a bell-shape, then the error distribution is normal (also known as Gaussian) and the model is performing well. Figure 8 shows error distribution histograms for all water quality parameters that were predicted using the neural network.

3.4 Regression Measures

In addition to the charts and graphs listed above, several different numerical statistics can be calculated to measure the performance of a regression task. These include, mean squared analysis, mean absolute analysis, and root mean squared error (RMSE). Both MSE and MAE were defined in section 2.4.3 as equation 3 and 4 respectively. RMSE is the square root version of MSE and is useful for scaling the ranges of predicted values to that of actual values. It is defined below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(f_i - y_i)^2}{N}} \tag{5}$$

where N is the total number of data points, f_i is the actual water quality parameter value and y_i is the predicted value.

Table 2 shows a tabular representation of the calculated regression measures for all water quality parameters predicted using the neural network model.

Table 2: Tabular representation of regression measures, detailed in section 3.4, calculated for each water quality parameter predicted using the neural network.

Parameter	**MAE**	**MSE**	**RMSE**
chl_a	2.55	9.91	3.15
doc	10.48	155.71	12.48
tis	1.15	2.05	1.43
tss	14.41	442.48	21.04

4 Discussion

Since the goal of this experiment is to test the ability of machine learning by using a neural network model that uses multipsectral remote sensing image data as input data to predict water quality parameters, the results listed in section 3 shed light on the performance of machine learning for this task. The results are portrayed in a variety of different ways. Thus, specific comparisons can be made between each water quality parameter predictions.

4.1 Training and Validation Graphs: MAE and MSE

Overfitting, when the model does not generalize well on new data, is a common problem in machine learning tasks. It can be caused by a variety of reasons including lack of data points or a high number of trainable neural network parameters. The training and validation MSE and MAE graphs are important to analyze for each water quality parameter at each epoch. As mentioned, overfitting can be diagnosed through these graphs initially. Since the training data was split into training and validation data, the neural network model was able to be tested on new validation data at each epoch for performance evaluation. Since the MSE and MAE graphs both measure a type of error, the lower the values are, the better the model is performing. This can be seen in all graphs as most epochs consisted of a drop in error–therefore shaping the curve to be decreasing as expected. Furthermore, since overfitting is when machine learning models don't generalize well on new data, the further apart the training and validation curves are, the more likely overfitting is occurring. In the MAE graphs, the neural network models that were trained to predict chl_a, doc, and tis fit well to both the training and validation data. However, this is not the case for the neural network model used to predict tss. The training and validation MAE curves have a significant gap, therefore identifying clear overfitting. Since another metric was specified during training, the training and validation MSE graphs can also be analyzed for overfitting. As with the MAE graphs, both chl_a and tis fit well. However, the MSE graphs for doc and tss show significant overfitting due to large gaps between the training and validation curves.

4.2 Scatter Plots

As noted in section 3.2, scatter plots are useful for showing the relationship between two variables. In this case, those variables were the true actual water quality parameters and the predicted water quality parameter value. Furthermore, the graphs show a line with slope 1 representing a perfect fit. Thus, the more linear the data points were on each scatter plot, the better the model performed. The scatter plot for chl_a showed data points that made a linear figure as most data points were marginally close to the perfect fit line. For doc, some values were far from the perfect fit line and the overall shape the data points made did not resemble a line as much as chl_a did. The scatter plots for tss showed similar results to that of doc. The tss scatter plot contains values that closely resemble the perfect line, as well as as several others that are far off from the line. The scatter plot for tis showed that the data points make a general linear

representation and that most values predicted close to the perfect fit line.

4.3 Error Distribution Histograms

The error distribution shows the error range where error is defined as the distance between predicted and actual values. A total of 25 bins were created on the x-axis for each graph. The histogram represents the count. The most common type of distribution is Gaussian distribution in which the distribution graphs are bell-shaped. Thus, the more bell-shaped the graph is, the better the model performed. Both the chl_a and tis graphs showed a general bell-curve shaped with most error points ranging from [0-4] for chl_a and [0-3] for tis. The distribution graphs for doc and tss showed much higher error levels with the majority of data points residing between [-5-10] for doc and [-40-20] for tss.

4.4 Regression Measures Analysis

Lastly, several regression measures were calculated for each water quality parameter predictions. These values were MAE, MSE, and RMSE and are defined in section 3.4. All three values measure error. Thus, the smaller the error, the closer the fit to the data. In specific, MSE measures how close a perfect fit line is to the predicted data points [11]. The distance (residual) between each data point and the corresponding true value on the perfect fit line is summed up for all data points [12]. This value is then squared which constitutes MSE and represents the average of the residual errors. RMSE is the square root value of MSE and represents the standard deviation of the residual errors [13]. MAE refers to the mean of the absolute values of all error points which differs from MSE in that MAE does not take in the direction of the error [14].

Tis and chl_a were predicted the best in terms of MAE with an MAE score of 1.15 for tis and 2.55 for chl_a. Doc and tss were predicted the worst in which the MAE values of each predictions were not as low as chl_a and tis. This was the same case for MSE in which tis again was predicted the best and chl_a as second best predicted. This was a reoccurring trend as the tis, chl_a, doc, tss order stayed the same throughout all measures including RMSE.

Based on a thorough analysis of the results as a whole, it is clear that, using only multispectral remote sensing data, the neural network model performed the best when predicting chl_a and tis.

5 Conclusion

Inland water systems are essential resources vital to satisfy human needs globally in all socioeconomic communities. It is therefore imperative that scientists evaluate and manage these bio-diverse ecosystems to sustain productivity and longevity. Water quality of inland aquatic systems is dependent on surrounding populations and land use. However, with population growth occurring locally and internationally, industrial wastewater is contributing to the pollution of these precious water systems. It is therefore critical that we develop efficient, cost-effective and accurate methods to analyze these water systems available to all communities regardless of their resources. With the advent of remote sensing technology, large data sets evaluating important water quality parameters are being accumulated and made globally available to scientists.

Over the past 10 years, there has been rapid growth in the resources available to remotely evaluate inland water quality. With a dramatic shift to open access to data sets, the number of research studies have significantly increased, piquing an interest in earth observation data. The availability of open-access satellite imagery to scientists is vital to the advancement and thorough comprehension of water quality. For example, in Europe, water quality data for coastal and inland waterways is being collected into a synchronized dataset called Waterbase, made available to researchers and civilian initiatives [15]. Furthermore, the novel AquaSat data set, which was used in this experiment, uses various remote sensing satellites to produce matchup coincident reflectance values for *in situ* water quality values scraped from the Water Quality Portal [16]. It is clear that data sets such as these will provide accurate data in a more cost-effective and efficient manner, while maintaining valid sample outcomes for model development.

With the advancement of data sets and the airborne multispectral sensors, scientists are able to devise solutions to measure optically active components with more precision, therefore, facilitating geochemical analysis of inland water systems. Fundamentally, machine learning relies on data and because of the advancement of these various data sets, machine learning can be accurately applied and explored.

The intention of my research is to analyze the use of machine learning for using a large amount of multispectral remote sensing data as input to predict water quality parameters as output. The large amounts of match-up data provided by AquaSat allow a machine learning algorithm to identify and match the reflective values associated with a respective water quality parameter concentration prediction. As shown in the results and analyzed in the discussion, machine learning models can be used to accurately

predict water quality parameter concentrations to either assist or replace current protocol *in situ* tests. Another main point is that this approach is entirely data-driven meaning that a lack of expert domain knowledge wouldn't have a large affect in the utilization of this procedure. Because of this, the impact of this research experiment is largely scalable.

Monitoring water quality is indeed compulsory to determine trends in the fluctuation of aquatic environments and how their nutrient and pollutant contents are affected by surrounding human activity. Therefore, the availability of spatial images through globally connected databases allow researchers to determine how water quality differs in various geographic regions in relation to utilization and quality regulation. This progression will indeed help local ecologists implement policies that will enhance sustainability and conservation of biodiverse ecosystems.

With these advancements in technology, remote inland water quality assessment and conservation strategies will be available at a fraction of the cost of their parallel *in situ* evaluation. This will undoubtedly lead to effective remote sensing capabilities in vast inland aquatic systems, especially in communities with limited resources.

Water quality is a key factor in determining the ability of water systems to remain sustainable and biochemically viable. Surface inland water systems are dependent on natural factors and vary according to climate change and human intervention. Water quality is not only altered by dams, draining of wetlands and diversion of natural flow, but also by pollution, discharge of wastewater into natural water systems and the runoff of chemicals into natural drainage basins. The unforgiving treatment of our natural ecosystems may disturb potential water quality elements and ultimately disrupt the sustainability and longevity of Earth's aquatic counterparts.

Although this research experiment produced statistically significant results, there are several ways in which the results can be improved. Even though a large amount of multispectral data was used as input data in this experiment, since this is a machine learning task, the more data that becomes available, the more accurate a model will likely perform. This is true for doc and tis which is generally available only in small bodies of water that are not able to be accurately captured by remote sensing data. With the improvement in remote sensing technology, it is inevitable that more data will become available in which experiments like this one can build off of. Furthermore, only one machine learning model was created and implemented for this experiment. In the future, other machine learning algorithms such as support vector machines, decision trees, and gradient boosting algorithms, can be explored and implemented for possible

improvement.

With the increased availability of data sets, it is ultimately up to the pioneers in the field of machine learning to partner with ecologists to develop remote, cost effective and accurate methods to evaluate water quality trends in our inland water systems, therefore creating a path to cultivating robust essential ecosystems.

References

[1] S. N. Topp, T. M. Pavelsky, D. Jensen, M. Simard, and M. R. Ross, "Research trends in the use of remote sensing for inland water quality science: Moving towards multidisciplinary applications," *Water*, vol. 12, no. 1, p. 169, 2020.

[2] J. Bartram, R. Ballance, W. H. Organization, and U. N. E. Programme, "Water quality monitoring : a practical guide to the design and implementation of freshwater quality studies and monitoring programs / edited by jamie bartram and richard ballance," 1996.

[3] F. Pu, C. Ding, Z. Chao, Y. Yu, and X. Xu, "Water-quality classification of inland lakes using landsat8 images by convolutional neural networks," *Remote Sensing*, vol. 11, no. 14, p. 1674, 2019.

[4] M. R. V. Ross, S. N. Topp, A. P. Appling, X. Yang, C. Kuhn, D. Butman, M. Simard, and T. M. Pavelsky, "Aquasat: A data set to enable remote sensing of water quality for inland waters," *Water Resources Research*, vol. 55, no. 11, pp. 10012–10025, 2019.

[5] S. Patro and K. K. Sahu, "Normalization: A preprocessing stage," *arXiv preprint arXiv:1503.06462*, 2015.

[6] W. Mendenhall and T. Sincich, *Statistics for engineering and the sciences*. Prentice Hall, 1995.

[7] J. Roell, "From fiction to reality: A beginner's guide to artificial neural networks," Jun 2017.

[8] K. Suzuki, *Artificial neural networks: methodological advances and biomedical applications*. BoD–Books on Demand, 2011.

[9] A. Graves, "Generating sequences with recurrent neural networks," *CoRR*, vol. abs/1308.0850, 2013.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[11] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and nse performance criteria: Implications for improving hydrological modelling," *Journal of hydrology*, vol. 377, no. 1-2, pp. 80–91, 2009.

[12] D. Wallach and B. Goffinet, "Mean squared error of prediction as a criterion for evaluating and comparing system models," *Ecological modelling*, vol. 44, no. 3-4, pp. 299–306, 1989.

[13] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.

[14] E. J. Coyle and J.-H. Lin, "Stack filters and the mean absolute error criterion," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 8, pp. 1244–1254, 1988.

[15] T. Srebotnjak, G. Carr, A. [de Sherbinin], and C. Rickwood, "A global water quality index and hot-deck imputation of missing data," *Ecological Indicators*, vol. 17, pp. 108 – 119, 2012. Indicators of environmental sustainability: From concept to applications.

[16] E. K. Read, L. Carr, L. De Cicco, H. A. Dugan, P. C. Hanson, J. A. Hart, J. Kreft, J. S. Read, and L. A. Winslow, "Water quality data for national-scale aquatic research: The water quality portal," *Water Resources Research*, vol. 53, no. 2, pp. 1735–1745, 2017.