HIGH SCHOOL EDITION
Journal of Student Research

An AI-Driven Multi-Layer Perceptron Model for Early Detection of Lake Eutrophication

Seyoung Lim[1] and Jimin Choi[#]

[1]Korean Minjok Leadership Academy, Republic of Korea
[#]Advisor

ABSTRACT

Early detection of eutrophication is crucial for water body protection because it requires substantial time and resources to restore the water quality once polluted. Conventional methodology of determining water quality mandated repetition and protracted manual labour. Many on-field sensors and monitoring systems have been proposed to ease the burden. However, though in-situ systems could reduce manual chemical analyses, AI models classifying the quality of lakes have drawn relatively little attention as a solution to simplifying the process. Therefore, this study suggests a multi-layer perceptron (MLP) model for early detecting eutrophic lakes. MLP models are easily adaptable to different environments and features, assisting the researchers by greatly shortening the detection time. Water assessment data received from WaterAtlas were sparse and therefore were preprocessed using linear regression. The data was randomly split to tune the number of epochs and the depth of the model. Afterwards, the model was tested by lake-based datasets to observe the model performance for unknown lakes only by existing data. The model had classified the dataset with 0.8911 accuracy and precision, recall, and f1-score of eutrophic lakes all reaching above 0.90, proving that the model could accurately evaluate the water quality. The macro-ROC and micro-ROC curves had the AUC scores of 0.96 and 0.95, respectively. This result with a narrow gap indicated that the model can precisely estimate the quality of lakes though the input data type is not evenly distributed. Therefore, this study shows that MLP is a suitable approach for early detection of lake eutrophication.

Introduction

Eutrophication of water bodies poses a serious threat to the aquatic ecosystem. When the aquatic environment becomes enriched with nutrients, plant and algae growth increases due to the excessive availability of one or more factors needed for photosynthesis. Not only do the increased microorganisms deplete the dissolved oxygen in water bodies, but they also produce large amounts of CO_2 while decomposing. This is followed by the decrease of pH, blocking of sunlight, and giving off more nutrients into the water. The resulting environment is hostile for fish and bivalve molluscs. Death of blooming algae produces toxic chemicals such as cytotoxins, biotoxins, and anatoxins and threatens human health while limiting access to safe drinking water (Yang et al., 2008).

Though eutrophication occurs naturally over decades, human activities such as agriculture and factories have greatly accelerated the process speed (Chrislock et al., 2013). Fertilisers, industrial water, and animal waste enter the water, providing more nutrients and destroying the ecosystem. According to the UNEP (United Nations Environment Protection), 30 to 40% of lakes and water reservoirs have been affected by eutrophication (Yang et al., 2008). Since the treatment of water bodies becomes substantially more difficult as the pollution deteriorates, early detection of eutrophication is crucial to provide sufficient time for water resource protection.

Conventionally, eutrophication is detected by measuring various chemical factors such as biological oxygen demand (BOD), total nitrogen (TN), or total phosphorus (TP). However, these conventional methods require a long examination period and manual chemical preprocessing procedures. The preprocessing methods are different for every factor as well. For instance, BOD is traditionally measured using the 5-day cultivation method (BOD_5). This method requires cultivating the microorganisms in a water sample at 20℃ and measuring the difference in the amount of dissolved oxygen (DO). Preprocessing such as seeding or adding water to dilute the sample is included for a more accurate outcome. Seeding refers to injecting additional aerobic microorganisms when the sample doesn't contain enough to measure DO. Water is added when the opposite situation occurs (Zhou et al., 2024). The alkaline persulfate digestion method is widely used to determine TN. Nitrogenous compounds are completely oxidised to nitrate by alkaline persulfate at 120°C. Then they pass through the cadmium-copper reduction column to reduce nitrate to nitrite. After preprocessing, TN is measured by UV spectrophotometry. It is necessary to make solutions for preprocessing including an alkaline persulfate solution and ammonium chloride buffer solutions. The overall method requires many steps and chemicals (Republic of Korea's National Institute of Fisheries Science, 2007). Repeating these same procedures is time-consuming and inefficient, but regular monitoring is mandatory to ensure the water quality.

To improve the traditional procedure, several in-situ monitoring systems and data collecting and processing methods have been suggested. Shivaanivarsha et al. (2022) proposed a low-cost multiparameter eutrophication monitoring system. By using different types of sensors and a microcontroller unit, the study measured parameters such as pH, turbidity, and DO and succeeded in sending the data to the database using LoRaWAN technology. However, though in-situ monitoring greatly reduces the time needed for chemical procedures, data analyses are still inevitable. Vázquez-Burgos et al. (2019) used AHP (Analytical Hierarchy Process) to weight relevantly important factors. They also used fuzzy logic to deal with great fuzziness and limited data. However, the data was monitored inside real culture tanks, which was a limited environment and resulted in subjective outcomes. Lin et al. (2020) approached the problem based on the hybrid method of TOPSIS (the technique for order preference by similarity to an ideal solution) and MCS (Monte Carlo simulation) to overcome potential data collection errors and limitations to data. This hybrid approach improved the reliability of the outcome but included factors such as chlorophyll-α and COD_{Mn}, usually measured by traditional methods. It also uses the data from a single lake, Lake Erhai, and therefore cannot be certain if the model will be applicable to other lakes.

This study presents an approach for the early detection of eutrophication by using an AI model based on a multi-layer perceptron (MLP). This approach goes beyond selecting important features for detection and shows moderate accuracy even only with factors that can be measured on the field. Even with three factors, a small number compared to the previous studies, the MLP model can predict eutrophication with reasonable accuracy. Additionally, unlike past studies, this research was based on data from eight different lakes. The model performance was tested when introduced to a completely different lake data source by dividing the testing and training dataset based on lakes. This ensured the accuracy of the model to be general in various adaptations.

The model introduced in this study will enable quicker eutrophication detection though the number of factors might be limited and be adaptable to versatile environments. The model will provide an efficient and simple way for early eutrophication detection.

Materials

This section introduces and describes the dataset used for the model training. The dataset source, features, and preprocessing methods are mentioned below. The dataset source is given to ensure that the data are not biased and are collected from varying environments to provide generality to the result of the model prediction. Features are provided to show that the model functions well only with data measured on the field. Lastly, the preprocessing method is stated to describe how the sparsity of the dataset was handled.

Dataset

The data for model training were retrieved from Water Atlas (USF Water Institute, School of Geosciences, University of South Florida, n.d.). This study created the water quality assessments of eight lakes (Lake Apopka, Dora, Eustis, George, Harris, Johns, Louisa, and Minneola) from different sampling spots into separate datasets. The above lakes were either currently eutrophic or experienced eutrophication so that the model would be sufficiently provided with eutrophic values. On the other hand, lakes that were regularly monitored and treated were not examined since they maintained a clean water quality, which was inadequate for model training. The features were selected assuming the model would be combined with a real-time monitoring system: pH, total organic carbon (TOC), and nitrogen by NO_3-NO_2 (NO_x). pH can be easily determined by a pH meter and is commonly chosen for on-field measurements. TOC is an important factor for detecting lacustrine eutrophication because it is strongly relevant to human activities (Sun et al., 2022) and can be measured by an optical sensor. Nitrogen acts as a direct nutrient for microorganisms and is one of the key factors of algae bloom. It also increased rapidly as rapid urbanisation and excessive usage of fertilisers occurred (Dong et al., 2023).

TSI (Trophic State Index) was used to distinguish eutrophication levels (oligotrophic, mesotrophic, and hypereutrophic) of samples and enable supervised learning of the model. Water Atlas provided the index formulae, and the result is given by a number ranging from 0 to 100. 0-59 indicates good lake condition, 60-69 fair, and 70-100 poor (USF Water Institute, School of Geosciences, University of South Florida, n.d.). The specific formulae are described below. First, we calculate indexes based on chlorophyll-α, total nitrogen (TN), and total phosphorus TP.

$$TSI_{(chl\ a)} = 16.8 + [14.4 \times ln(chl\ a)]$$
$$TSI_{(TP)} = 18.6 \times [ln(TP \times 1000)] - 18.4$$
$$TSI_{(TN)} = 56 + 19.8 \times ln(TN)$$
$$TSI_{2\ (TP)} = 10 \times [2.36 \times ln(TP \times 1000) - 2.38]$$
$$TSI_{2\ (TN)} = 10 \times [5.96 + 2.15 \times ln(TN + 0.001)]$$

Then, based on the nutrient ratio of the lake, we select one of the three formulae for the final calculation.

1. Nutrient Balanced Lakes ($10 \leq TN/P \leq 30$):
$$TSI = \{TSI_{(chl\ a)} + [TSI_{(TN)} + TSI_{(TP)}] / 2\} / 2$$

2. Phosphorus-Limited Lakes ($TN/TP > 30$):
$$TSI = [TSI_{(chl\ a)} + TSI_{2\ (TP)}] / 2$$

3. Nitrogen-Limited Lakes ($TN/TP < 10$):
$$TSI = [TSI_{(chl\ a)} + TSI_{2\ (TN)})] / 2$$

Data Preprocessing

The datasets of eight lakes were sparse since the sampling spots and the target for measurement varied depending on the time the sample was collected. The specific dataset size and the number of non-null values of each feature are presented in Table 1. It shows that the total dataset size is 13146, but those of each feature are considerably smaller. Therefore, data preprocessing to assign a heuristic value to the null values can improve model performance.

Table 1. The total dataset size and the number of non-null values of pH, TOC, and NO_x.

Total size (=TSI)	pH	TOC	NO_x
13146	5078	3138	4227

While linear regression is not an adequate method to interpolate the null values of the dataset when the data is categorical, it is appropriate for lake eutrophication since the data is continuous and shows a certain definite trend as the lake changes from an oligotrophic state to a eutrophic state as time passes. The correlations between TSI and three factors of pH, TOC, and NO_x were observed by drawing separate scatterplots. The outcome is shown in Figure 1, 2, and 3.

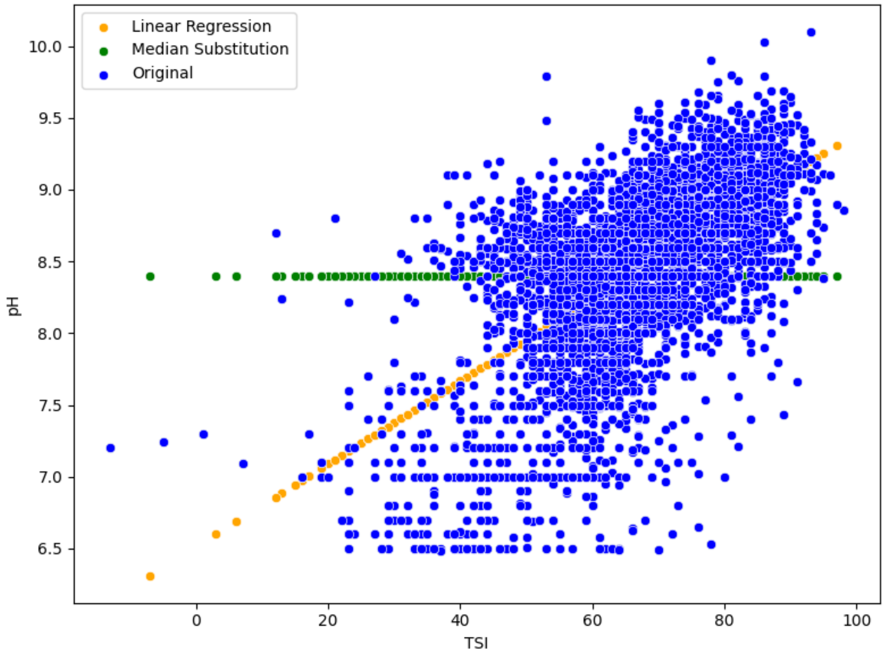

Figure 1. The scatterplot of pH and TSI

pH and TOC had distinct positive correlations. NO_x had an ambiguous correlation compared to the two factors, nearly close to none. TOC was regarded to have the clearest correlation because it is directly related to the amount of organic compounds dispersed in water.

Next, linear regression was used for the interpolation method. The interpolation was also done with the median values to compare the accuracy of the two models trained separately with each dataset. The scatterplots between the original and the median substituted value are shown in Figure 1, 2, and 3. Filling in all the null values with the median resulted in high discrepancies and contradicted the correlation.

The TSI column was labelled as the control variable (X) and pH, TOC, and NO_x as the dependent variable (y) respectively to perform linear regression and replace the null values with the outcome for each of the factors. The resulting equations are as follows. The numbers were rounded off with three significant figures.

$$pH: y = 0.0288X + 6.513$$
$$NOx: y = -0.000151X + 0.0208$$
$$TOC: y = 0.436X + -8.350$$

Corresponding to the scatterplots, TOC showed the highest slope and NO_x had the lowest, almost close to zero. This slope value seems to have occurred since the TSI values of many of the selected lakes relied on the TP levels; Lake Apopka, Lake Harris, and Lake Eustis are included in the Harris Chain of Lakes, and phosphorus levels were the major contribution to their eutrophication (St. Johns River Water Management District, n.d.). Lake Dora was connected to Lake Beauclair, a lake included in the Harris Chain. Dominant nutrient sources flowed from Lake Beauclair though there also were runoffs from nearby residential areas (Lake County Water Authority, n.d.). Due to this phosphorus-focused nutrient composition, it is likely that the TN values did not have a significant effect on the TSI values and resulted in a near-zero slope value.

As observable from the equations and the scatterplots, the correlations between TSI and the three factors prove linear regression is an appropriate approach for heuristic data interpolation. After the interpolation, the dataset was divided into three subsets depending on the TSI value ranges stated above.

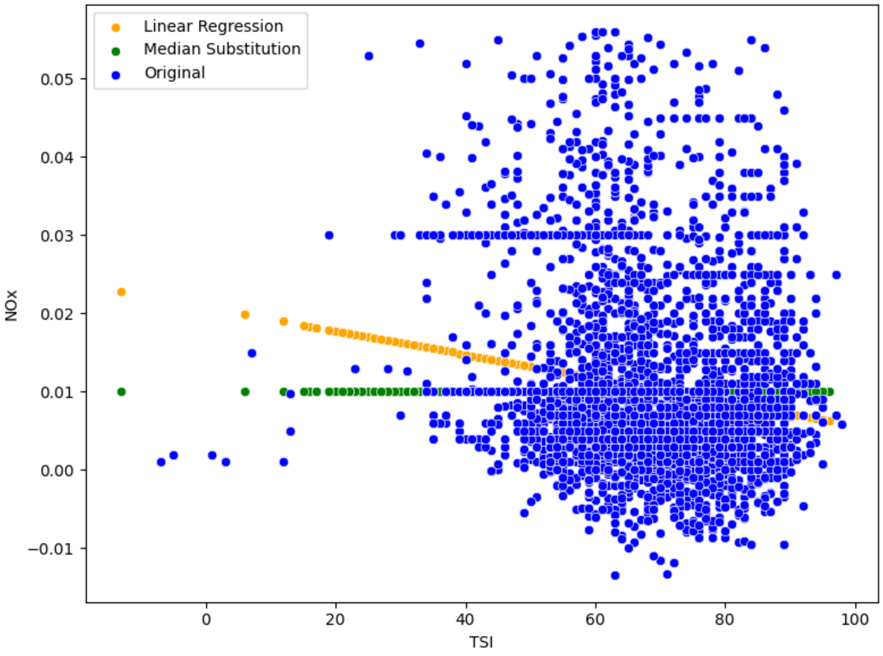

Figure 2. The scatterplot of NO_x and TSI

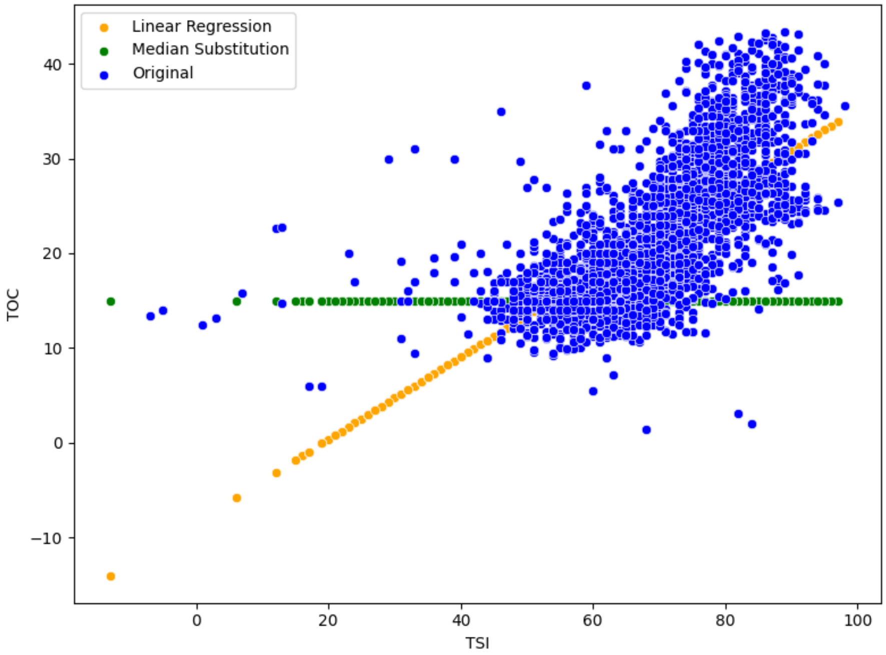

Figure 3. The scatterplot of TOC and TSI

Data splitting took two approaches: (a) random splitting and (b) splitting by lakes. For random splitting, the train set and the test set were divided in an 8:2 ratio. For splitting data by lakes, numbers from 1 to 8 were given to each lake dataset before merging them. After undergoing the same preprocessing process, 7 lakes were classified as the train set and the remaining 1 as the test set using the given numbers. The lakes were grouped considering the dataset size and the proportion of hypereutrophic, mesotrophic, and oligotrophic data in each dataset to make the ratio even. The training and testing dataset sizes of the two respective methods are presented in Table 2. It can be observed that the two methods have similar size divisions. Therefore, the possible difference in the model training outcome by the dataset size can be minimised.

Table 2. The dataset size of two methods: random splitting and splitting by lakes

Random Splitting		**Splitting by Lakes**	
Train	**Test**	**Train**	**Test**
10517	2629	11492	1654

Methods

Eutrophication Detection Model

For early eutrophication detection, we devised a deep-learning model using a multi-layer perceptron (MLP). MLP is an adequate deep-learning model for early detection for several reasons: MLP is specialised in solving non-linear problems, which is useful since there are various relationships between factors determining lake eutrophication, and MLP is capable of modelling them. The model also allows flexibility in choosing features, which is practical since the types of collected data (pH, Chlorophyll-α, DO, salinity, etc.) normally differ depending on the region and the available infrastructure. Between continuous evaluation and categorical evaluation, the latter was selected due to the facile interpretation of the outcome, the prediction of the model resulting in one of the three labels assigned to the dataset earlier depending on the TSI value. This model will be used as the first-phase sensing tool to help the researchers identify whether or not the lake is eutrophic. After input data passes through two hidden layers, each node in the output layer presents the class probability for each class. Then the model predicts the status of water bodies by choosing the class with the highest probability. The overall model structure is shown in Figure 4.

Table 3. MLP structure for eutrophication detection

Layer (type)	Output shape	Param #	Activation function
dense (Dense)	(None, 128)	512	ReLU
dense_1 (Dense)	(None, 64)	8256	ReLU
dense_2 (Dense)	(None, 3)	195	Softmax

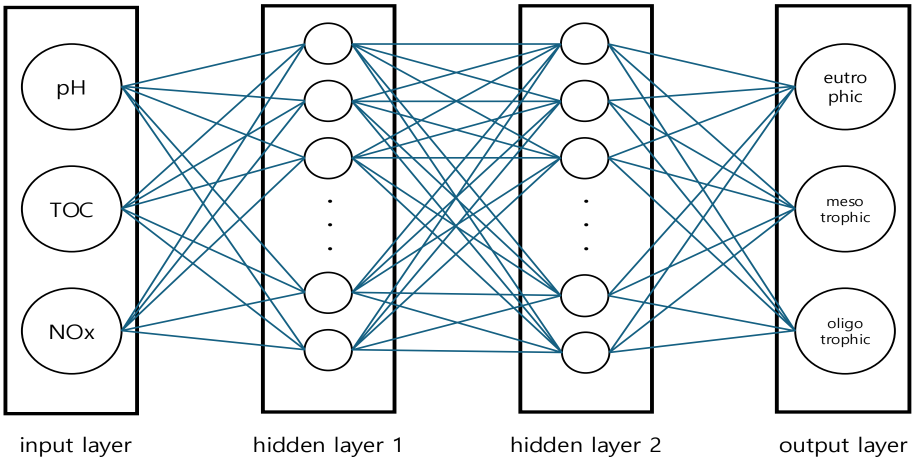

Figure 4. The structure outline of the MLP model

The training and testing were conducted in two separate ways as mentioned above in the data preprocessing. The approach of splitting data by lakes was taken to achieve model generalisation; observing the detection ability of a completely random lake only by existing data would correspond to the actual situation when they proceeded to the real-life application of the model.

Hyperparameter Selection

The three labels were encoded into 0 (hypereutrophic), 1 (mesotrophic), and 2 (oligotrophic) for model prediction since it was a supervised learning, and the model needed to know which data is classified as which water quality. The model was trained for up to 250 epochs, selected as the number that showed the best accuracy without overfitting. After testing multiple models, the study chose the optimal number of hidden layers and the number of epochs to specify the combination with the highest accuracy. The process and specific model evaluations are stated in the Results section. The specific results are shown in Table 4.

Table 4. Hyperparameters and hidden layer numbers for training

Hyperparameter	**Value**
Optimizer	Adam
Activation function	ReLU
Batch size	32
Epochs	250
Loss function	Cross Entropy

Results

Evaluation Metrics

The model performance was evaluated by calculating precision, recall, f1-score values, and overall accuracy. Precision (P) counts the proportion of the true positive predictions among all positive predictions. Recall (R) is attained by calculating the proportion of true positives in all actual positive data within the set. F1-score (F) is the harmonic mean of precision and recall values. The exact formulae are given below. TP is an abbreviation of true positive, TN of true negative, FP of false positive, and FN of false negative. These scores provide more insight into the classification outcome than their accuracy only, since the results are divided into four types and show whether the model is bold or cautious in making predictions.

$$P = \frac{TP}{TP + FP}$$
$$R = \frac{TP}{TP + FN}$$
$$F = \frac{2 * P * R}{P + R}$$

Moreover, the model performance was visualised by generating a ROC curve. The ROC curve is plotted by featuring a true positive rate (TPR) on the y-axis and a false positive rate (FPR) on the x-axis. AUC, the area beneath the ROC curve, indicates the accuracy of the model classification. Thus, a perfect classification model has an AUC value of 1, since TPR is 1.0 and FPR is 0.0. A random classifier denotes a value of 0.5. Furthermore, by plotting the micro-average ROC and macro-average ROC curves, we can observe how balanced the model performance is. Micro-average ROC takes the imbalance in the dataset into account, and the result is based on the frequency of each class. A class with a bigger size has more impact on the outcome. Macro-average ROC, however, does not consider the imbalance and gives every class the same weight. By comparing the AUC values of the two curves, we can evaluate the model performance for infrequent classes and the overall balance.

Parameter Study

The epochs and the depth of the model were tuned to generalise the model. The two parameters were chosen because they are both relevant to the time needed for model training and overfitting. Since the model aims to be applied in real-time lake water quality management, short training time is better for implementation. Also, since the risk of overfitting should not be ignored, the number of epochs and the depth would have to be tuned beforehand to prevent the problem. The parameters were determined using the random split data model. Then, the model using those parameters was trained with the lake-based split dataset to see if it would also show moderate accuracy when a completely new lake was tested.

The number of epochs was tuned by testing out models with varying numbers of epochs and selecting the highest number before the accuracy started to decrease. For each number of epochs, the model was trained 5 times. Precision (P), recall (R), F1-score (F), and the accuracy of the models concerning epochs are shown in Table 5. The values are the average of the 5 tests. For each score, the first row indicates the value of that score for the hypereutrophic label, the second for the mesotrophic label, and the third for the oligotrophic label.

Table 5. P, R, F, and accuracy data for different epochs

Epochs	100	150	200	250	300
P	0.95	0.95	0.96	0.97	0.94
	0.79	0.88	0.87	0.87	0.83
	0.92	0.90	0.91	0.91	0.95
R	0.90	0.92	0.91	0.91	0.93
	0.86	0.81	0.85	0.85	0.86
	0.92	0.97	0.96	0.96	0.93
F	0.92	0.93	0.93	0.94	0.93
	0.82	0.84	0.86	0.86	0.84
	0.92	0.93	0.93	0.93	0.93
Accuracy	0.8839	0.9031	0.9053	0.9135	0.9059

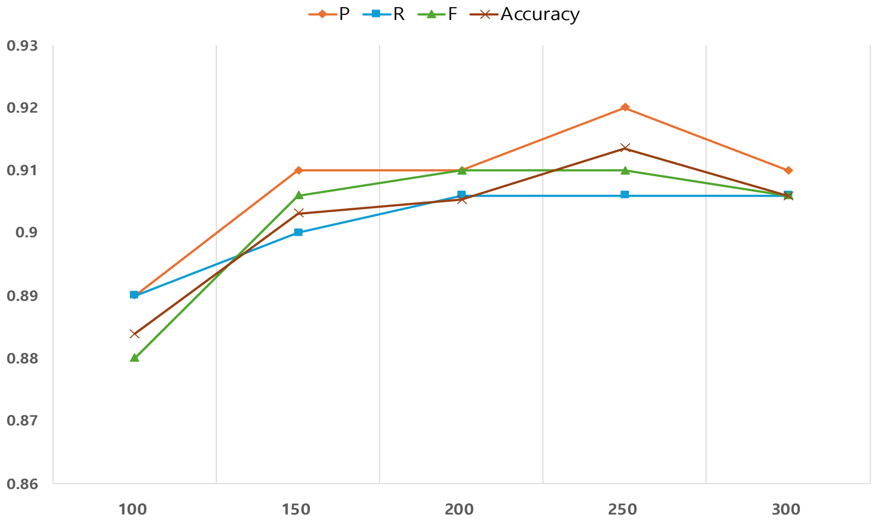

Figure 5. The average P, R, and F values and accuracy per epoch

Among all the trials, as Figure 5 shows, the trial with the 250 epochs displayed the highest P and accuracy values, and from 300 epochs, the values started to decline. For all scores, mesotrophic lakes had the lowest values. The reason is expected to be that the size of the mesotrophic lake data was the smallest of the three. For hypereutrophic values, the precision score was higher than the recall score, and for oligotrophic values the opposite. This means the model is conservative in predicting eutrophication and bold in predicting oligotrophic lakes. This could have occurred since the size of the oligotrophic data was the largest, and therefore the model had more oligotrophic data to learn compared to the other two classes. Providing different class weights and shifting the threshold value are likely to give more positive predictions for eutrophic lakes.

The model trained with 250 epochs showed the most stable and constant data trend for loss and accuracy. The specific graphs of the loss and accuracy are shown in Figure 6. The fluctuation between the successive data did not surpass 0.05 at the most and the range was narrowed as the number of epochs reached 250. Moreover, the loss and accuracy displayed constant decreases and increases. This indicates that the model wasn't overfitted and generalised the dataset well. Overall, the model with the epoch 250 had the highest accuracy, P, R, and F scores for hypereutrophic lakes, which were the priority the model should focus on. Therefore, it was selected as the parameter.

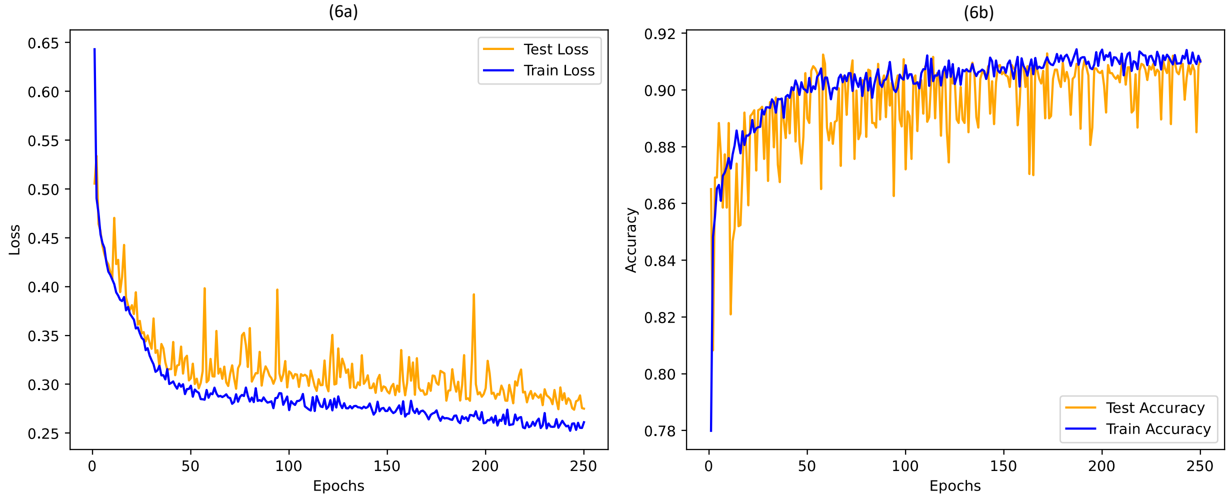

Figure 6. Train and test loss & accuracy history of epoch 250. (a) loss per epoch (b) accuracy per epoch

After tuning the epochs, the depth of the model was tuned using the same method as the epochs. The table with the same format is presented below.

Table 6. P, R, F, and accuracy data for different depths for epochs 250

Layers	**3**	**4**	**5**
P	0.97	0.95	0.93
	0.87	0.88	0.79
	0.91	0.90	0.94
R	0.91	0.91	0.91
	0.85	0.83	0.88
	0.96	0.97	0.88
F	0.94	0.93	0.92
	0.86	0.85	0.83
	0.93	0.93	0.91
Accuracy	0.9135	0.9041	0.8676

There were distinctive constantly decreasing value trends throughout P, R, F, and accuracy from depth 3 to depth 5. The underestimation of eutrophic lakes and the overestimation of oligotrophic lakes were also seen in these trials, corresponding to the reason stated above.

Among the three trials, the model with 3 hidden layers showed the least fluctuation in the test loss and accuracy as shown in Figure 7. Similar to Figure 6, the gap of fluctuation gradually decreased as the number of epochs reached 250. As the depth increased, the fluctuation of the test loss and accuracy also increased, indicating that the model was not generalising the data properly as the model deepened. Furthermore, the accuracy was the highest when there were three hidden layers and noticeably decreased as the layers were added. Thus, was chosen to be the final depth for the model.

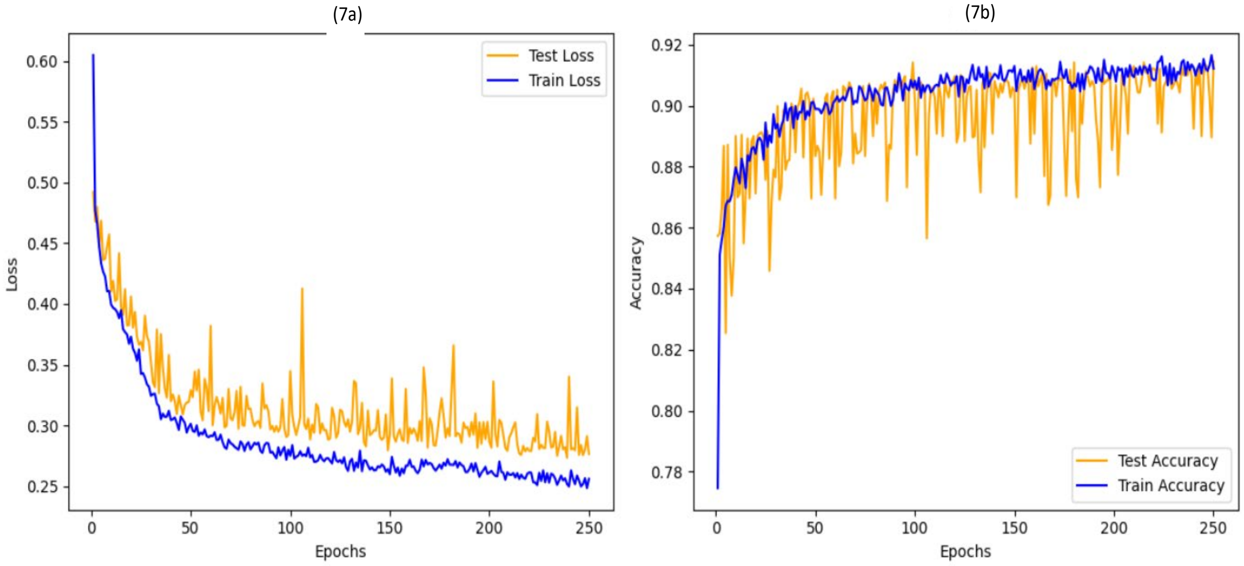

Figure 7. Train and test loss & accuracy history of depth 3. (a) loss per epoch (b) accuracy per epoch

These variations in P, R, and F values and the accuracy prove that the MLP model can potentially result in different performances when the parameters are not tuned to demonstrate its best outcomes. Therefore, through these model training trials, we can conclude that tuning these parameters is an essential step to enhance the model performance.

Observation from the Model

Table 7. The number of outliers for each feature

pH	TOC	NO_x
277	71	597

Before data preprocessing and model training, the outliers of the three factors had to be removed from the original dataset since we could not overlook the possible risk of mismeasurement. The number of outliers in the dataset for each feature is shown in Table 5. The null values were excluded from the outliers since they were to be interpolated. pH had an outlier percentage of 5.4%, TOC of 2.2%, and NO_x of 14.1%. Although the outlier percentage of NO_x was higher than the other two features, the dataset's scatterplot below in Figure 8 showed that NO_x outliers did not correlate with the TSI values. Generally, 0.15mg/L of TN can become an indicator of algae blooms in lakes, and it is highly unlikely for the value to exceed 10mg/L at most. (Nutrient Criteria Development Document: Lakes and Reservoirs | US EPA, 2023) Thus, the outliers were removed to improve the performance of the model.

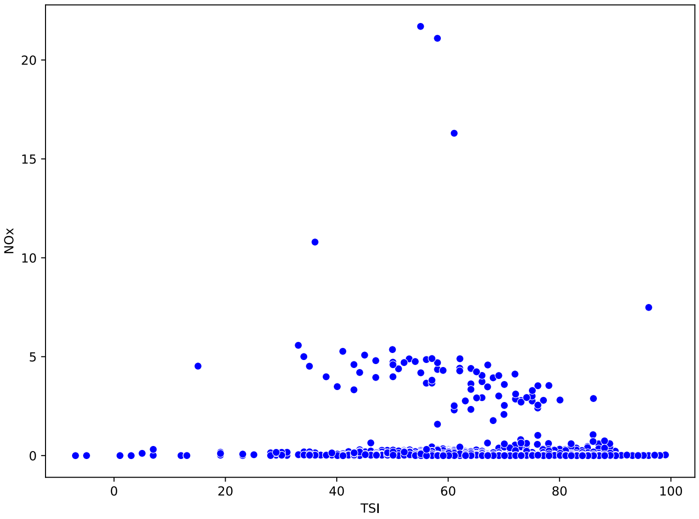

Figure 8. Scatterplot of original NO_x and TSI value

The sizes of hypereutrophic, mesotrophic, and oligotrophic datasets are presented in Table 6.

Table 8. The dataset size for each label

Hypereutrophic	**Mesotrophic**	**Oligotrophic**
4028	3265	4936

The oligotrophic dataset had the largest size, followed by the hypereutrophic and mesotrophic datasets. This corresponds to Table 5 and 6, where the P, R, and F scores of the three labels follow the dataset size order. However, though there were differences in scores, the model did not fail to make a big discrepancy between the predictions of the three classes and showed moderate performance for all of them. This indicates that the MLP model can accurately predict the status of lakes even though the input data is imbalanced. The specific outcomes are explained by the ROC curves below.

Model Performance

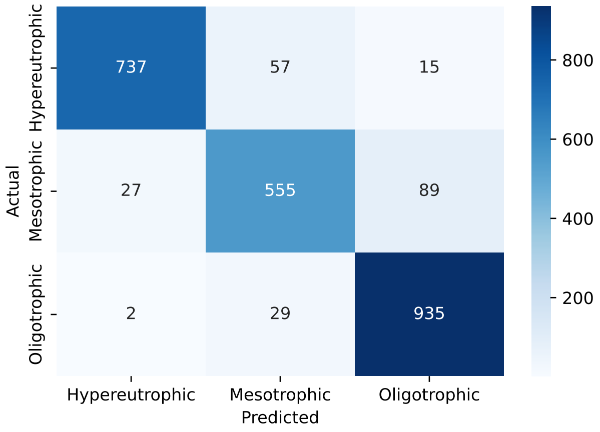

Figure 9. The confusion matrix of the random data split model

The classification results for the test data are shown by the confusion matrices in Figure 9 and 10. Since there were 3 labels to classify, the matrices had a size of 3x3. Both matrices show higher values for the main diagonal than the rest of the elements, which indicates that the model was thoroughly trained and could perform the classification well. Also, the lake-based data split model showed a decent accuracy of 0.8911 and P, R, and F values as listed in Table 9. Thus, it proves that the model can achieve high performance when classifying data from new environments with only existing data.

Compared to the random data split model, the lake-based data split model showed higher accuracies for eutrophic and oligotrophic lakes, but it declined for mesotrophic lakes. Instead, the number of mesotrophic lakes being classified into oligotrophic lakes increased. This seems to have happened due to the input order of data. The order of the random data split model is arbitrary; therefore, the three data types are mixed. On the other hand, for lake-based split data, the data were aligned by lakes and types before input. This may have caused a different decision boundary and classified mesotrophic data similar to oligotrophic data into the latter.

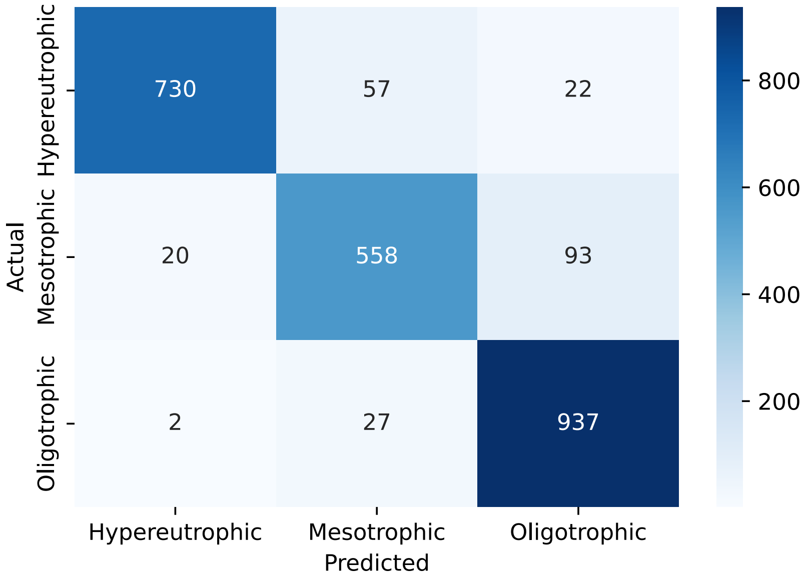

Figure 10. The confusion matrix of the lake-based data split model

Table 9. P, R, F scores and accuracy of lake-based data split model

Lake-based data split model	
P	0.90
	0.81
	0.96
R	0.91
	0.88
	0.89
F	0.91

	0.84
	0.92
Accuracy	0.8911

However, like the model trained by randomly split data, the cases of underestimating the status of lakes, the elements above the main diagonal, are more frequent than overestimation, the values below the main diagonal in Figure 10. If the prediction fails to be accurate, among underestimation and overestimation, the latter would be better since it can alert the researchers on the field beforehand and make them prevent potential pollution. This problem could be improved by lowering the classification threshold or customising the loss function to weigh the two choices differently, resulting in a higher state of eutrophication than that of the actual situation.

The ROC curve shown below also clearly presents the performance of the lake-based data split model. Since the model performed a multiclass classification, a one-versus-rest ROC curve was drawn. Hypereutrophic values were selected to be compared with the other two classes because of the reason stated in parameter studies.

It can be inferred from the AUC score of 0.98 shown in Figure 11. that hypereutrophic lakes were detected well from mesotrophic or oligotrophic lakes. For Figure 12., the AUC scores for all three classes appeared to be high. Specifically, the oligotrophic lakes showed the best AUC score of 0.98. This could have occurred since the oligotrophic data is the largest among the three as mentioned above. The mesotrophic AUC score was the lowest, as observable by other evaluation methods. The main reason behind this seems to be the range of TSI values that divides the lake status: mesotrophic lakes have the smallest range (60-69) and oligotrophic the biggest (0-59). It is easier for lakes to be included in the oligotrophic values than the other two types.

Nonetheless, the micro-average ROC curves had the AUC score of 0.95 and the macro-average ROC curves had that of 0.96, which proves the model has balanced its performance across the three classes despite the difference in dataset size. It further indicates that the model's performance in detecting eutrophic lakes is successful. The confusion matrices show high P values, and the ROC curves demonstrate even AUC values. These prove that the suggested MLP model can precisely detect eutrophic lakes from the new environment while maintaining a balanced performance even when provided with biased datasets.

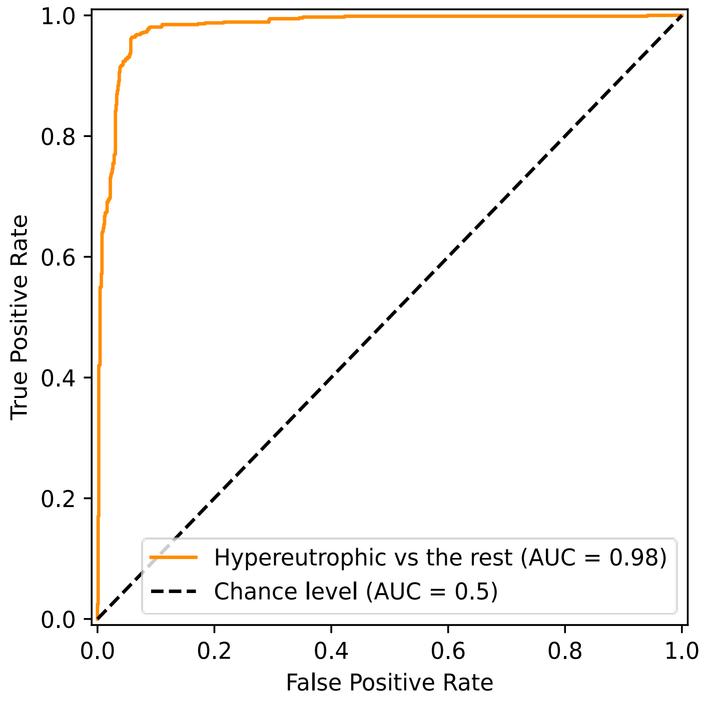

Figure 11. The One-vs-Rest ROC curve of hypereutrophic versus mesotrophic and oligotrophic lakes

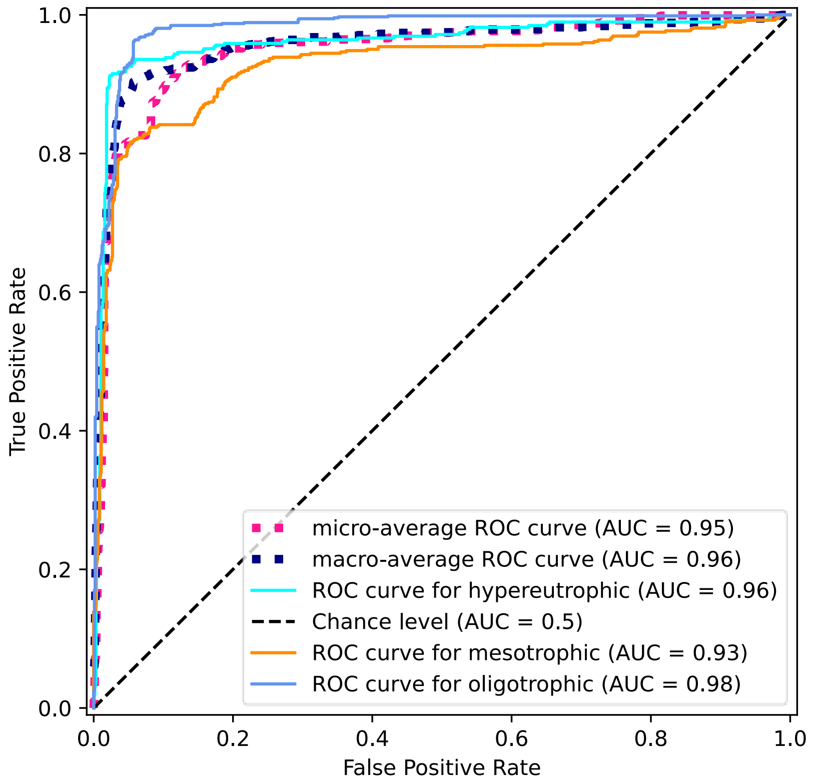

Figure 12. The 1 versus rest ROC curves of hypereutrophic, mesotrophic, and oligotrophic lakes, comparing each class against all the others

Conclusion

This study presents an approach for the early detection of eutrophication of lakes based on an MLP deep-learning model. Unlike traditional methods of determining eutrophication which require a long examination period and manual chemical preprocessing procedures, this model can distinguish eutrophied lakes with features measurable at the field, shortening the determination time. It also has flexibility in the inclusion and exclusion of factors, making it easier to adapt to lakes with different chemical or biological environments. The sparse data was filled using linear regression. Between mean substitution and linear regression, the latter showed better correlations with the original data. Afterwards, the number of epochs and the depth of the model were tuned to generalise the dataset most effectively. The dataset was first randomly split, and then divided by lakes to observe its performance when tested by completely new lake data. The result showed that the model classified eutrophic lakes with high accuracy and precision, both for random data split cases and the lake-based data split case. The final AUC score of eutrophic lakes was 0.96, indicating that the model detected eutrophication with high proficiency despite the differences in the size of each data type.

There are several aspects to explore to enhance the model's performance. Firstly, collecting data from more lakes could improve the accuracy of the model, since the smaller number of the null values is better for linear regression and model training. Moreover, as mentioned above, adjusting the threshold values or providing weight to the eutrophic lake data could increase the frequency of the eutrophication prediction. Though this is a false positive prediction, it could help researchers prevent the lake from pollution and start early treatment on nearly eutrophied water bodies. Finally, parameters other than the number of epochs and the depth could also be tuned using methods such as grid search. This would allow us to use more adequate parameters for the model.

Acknowledgments

We would like to thank Professor Caesar for his invaluable guidance and support throughout this research project.

References

Yang, X., Wu, X., Hao, H., & He, Z. (2008). Mechanisms and assessment of water eutrophication. *Journal of Zhejiang University SCIENCE B*, *9*(3), 197–209. https://doi.org/10.1631/jzus.b0710626

Chislock, M. F., Doster, E., Zitomer, R. A. & Wilson, A. E. (2013) Eutrophication: Causes, Consequences, and Controls in Aquatic Ecosystems. *Nature Education Knowledge* 4(4):10

Zhou, Y., Zheng, S., & Qin, W. (2024). Electrochemical biochemical oxygen demand biosensors and their applications in aquatic environmental monitoring. *Sensing and Bio-Sensing Research*, *44*, 100642. https://doi.org/10.1016/j.sbsr.2024.100642

Republic of Korea's National Institute of Fisheries Science. (2024). Marine environmental standard test methods [Appendix 1] Standard test methods for seawater, Chapter 4. Evaluation by Categories, Section 13 Total Nitrogen. https://www.nifs.go.kr/board/actionBoard0052List.do?BBS_CL_CD=A

Shivaanivarsha, N., Selvaraj, D. V., Vigita, S., Santhini, V., & Vijayendiran, A. G. (2022). Low-Cost Multi-Parameter lake monitoring system for early detection of eutrophication. *2022 IEEE International Power and Renewable Energy Conference (IPRECON)*. https://doi.org/10.1109/iprecon55716.2022.10059562

Vázquez-Burgos, J. L., Carbajal-Hernández, J. J., Sánchez-Fernández, L. P., Moreno-Armendáriz, M. A., Tello-Ballinas, J. A., & Hernández-Bautista, I. (2019). An Analytical Hierarchy Process to manage water quality in white fish (Chirostoma estor estor) intensive culture. *Computers and Electronics in Agriculture*, *167*, 105071. https://doi.org/10.1016/j.compag.2019.105071

Lin, S., Shen, S., Zhou, A., & Xu, Y. (2020). Approach based on TOPSIS and Monte Carlo simulation methods to evaluate lake eutrophication levels. *Water Research*, *187*, 116437. https://doi.org/10.1016/j.watres.2020.116437

Sun, W., Niu, X., Teng, H., Ma, Y., Ma, L., & Liu, Y. (2022). A 133-year record of eutrophication in the Chaihe Reservoir, Southwest China. Ecological Indicators, 134, 108469. https://doi.org/10.1016/j.ecolind.2021.108469

Dong, Y., Cheng, X., Li, C., Xu, L., & Lin, W. (2023). Characterization of nitrogen emissions for freshwater eutrophication modelling in life cycle impact assessment at the damage level and urban scale. Ecological Indicators, 154, 110598. https://doi.org/10.1016/j.ecolind.2023.110598

USF Water Institute, School of Geosciences, University of South Florida. (n.d.). Learn more: Trophic State Index (TSI) - Sarasota County Water Atlas - Sarasota.WaterAtlas.org. https://sarasota.wateratlas.usf.edu/library/learn-more/learnmore.aspx?toolsection=lm_tsi

St. Johns River Water Management District. (n.d.). Managing the Harris Chain of Lakes. In *Lake County Water Atlas*. https://lake.wateratlas.usf.edu/upload/documents/fs_ocklawahachain.pdf

Lake County Water Authority. (n.d.). Lake Dora EcoSummary, 2016-2017. In *Orange County Water Atlas*. https://orange.wateratlas.usf.edu/upload/documents/Lake-Dora-Ecosummary-LCWA-2016-2017.pdf

Nutrient Criteria Development Document: Lakes and Reservoirs | US EPA. (2023, November 30). US EPA. https://www.epa.gov/nutrientpollution/nutrient-criteria-development-document-lakes-and-reservoirs