Towards a New Protocol for Enhanced Hyperspectral Regression in the Wild: A Lough Neagh, UK Case Study

Andrew Roberts[1], Christopher J. Holder[1], Ankush Prashar[2], Maria-Valasia Peppa[3], Abhinav Galodha[3], Sam Wilson[2], Deepayan Bhowmik[1]
[1]School of Computing, Newcastle University, Newcastle upon Tyne, UK
[2]School of Natural & Environmental Sciences, Newcastle University, Newcastle upon Tyne, UK
[3]School of Engineering, Newcastle University, Newcastle upon Tyne, UK
a.roberts18@newcastle.ac.uk, chris.holder@newcastle.ac.uk, ankush.prashar@newcastle.ac.uk, maria-valasia.peppa@newcastle.ac.uk, srz208737@sire.iitd.ac.in, sam.wilson@newcastle.ac.uk, deepayan.bhowmik@newcastle.ac.uk

***Abstract*—Lough Neagh is experiencing severe ecological pressure from nutrient enrichment, resulting in recurrent harmful algal blooms (HABs). To address this we investigated the feasibility of using UAV-based hyperspectral imaging, combined with in situ water sampling and laboratory analysis, as a quantitative monitoring solution to track algal biomass via its spectral signature. Hyperspectral images and ground truth chlorophyll concentrations were collected and physical algae samples were collected and propagated in a lab. These were later analysed using a different hyperspectral camera to determine the isolated effect of concentration on the spectral curves. By comparing field spectra to lab spectra we were able to identify that some field curves were mislabelled and design a systematic outlier rejection method to filter incorrect curves whilst ensuring balanced and fair predictions. Using our protocol we trained models including: Stochastic Gradient Descent Regression (SGDR), Histogram Gradient Boosting Regression (HGBR), 1D Convolutional Neural Network (1DCNN) as well as running AutoGluon (AG) for best model evaluation. Our results show reasonable results across all models, with lab data always performing better, and near-infrared (NIR) data performing better than visible (VIS) in most cases. Best results were with the AG WeightedEnsemble_L2 (WEL2) with an R^2 of 0.9717 and MAE of 6.83 $\mu g/L$ for the lab NIR data and an R^2 of 0.8153 and MAE of 12.45 $\mu g/L$ for the field VIS+NIR data. These results support the feasibility of UAV-based hyperspectral monitoring as a scalable quantification tool. Strong correspondence between remotely sensed and laboratory spectral features shows the utility of combining these datasets to develop calibrated spectral indices and regression models for estimating algal biomass and bloom extent.**

***Index Terms*—Algae, Lough Neagh, Hyperspectral, Spectral Processing, Machine Learning**

I. Introduction

Lough Neagh, the largest freshwater lake in the British Isles (shown in Fig. 1), is experiencing increasing ecological pressure from nutrient enrichment associated with agricultural run-off, higher temperatures, and other anthropogenic discharges [1]. Sustained inputs of nitrogen and phosphorus have accelerated eutrophication processes, resulting in recurrent harmful algal blooms (HABs) that threaten water quality, aquatic biodiversity, and regional ecosystem services. In order to monitor and accurately measure the effects of preventative actions, methods for large-scale algae concentration measurement are required. Due to the large area involved, remote sensing is a highly desirable approach.

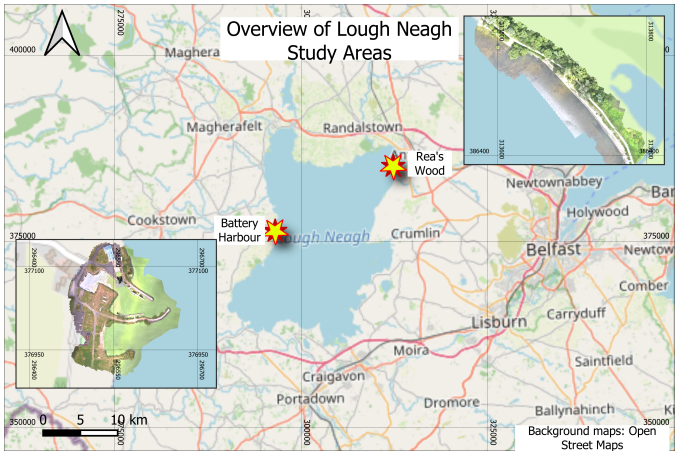


Fig. 1: Map showing locations of study on Lough Neagh. RGB images collected using DJI Mavic UAV, orthomosaics generated for reference purposes.

Hyperspectral cameras, as well as capturing spatial data, enable the spectrum of light per pixel to be captured as well. In contrast to the three channels of RGB images, hyperspectral cubes can consist of hundreds of channels, known as spectral bands, spread over a wider, contiguous wavelength range. Analysis of plant matter via hyperspectral sensors is highly linked to the chlorophyll concentrations in the subject [2]. Chlorophyll is highly absorbent in the visible region (VIS, 400-700 nm) but highly reflective in the near-infrared region (NIR, 700-1400 nm) in which structural features of the target dominate. For chlorophyll-containing subjects this results in a characteristic ”red-edge”, beginning at around 700 nm, where the spectral curve reflectance increases significantly as shown in our data in Fig. 4. This same phenomenon is seen in algal

species as well as land-based plants.

A key advantage of hyperspectral imaging is the ability to recover the inherent reflectance of the material being imaged. This reflectance is recovered through complex radiometric processing [3] [4] and represents an intrinsic property of the material which can therefore be used to compare the same types of material across different illumination environments, such as between field and lab.

When comparing hyperspectral data from sensors with different numbers of bands it is crucial to apply spectral harmonisation. This process involves reconstructing the bands of one of the sensors to match those of the other and requires use of the original Spectral Response Functions (SRF), which accounts for band crosstalk in the original image. The SRF's quantify which and how much non-target wavelength light is picked up by each band due to the imperfect nature of the filters being used. Depending on how much cross-talk occurs this can significantly change the resulting spectral curve when downsampling compared to simple linear interpolation or matching band selection.

Due to the high dimensional nature of the hyperspectral curves, machine learning (ML) models are well suited to leverage the complexity of the data. Previous studies [5] have used a range of ML algorithms in the regression prediction of algae concentration in lakes and rivers such as Random Forest (RF) [6], Multilayer Perceptron (MLP) [7], Light Gradient Boosting Machine (LGBM) [8], and 1D Convolutional Neural Networks (1D-CNN) [9]. These papers report R^2 scores in the range of 0.28 - 0.96 showing that a high regression score is not a given for field data collection due in part to the complexity of the natural environment. A lot of these studies have invested significant time and resources in field visits to collect data, but have not coupled that field data with lab analysis.

In order to determine the best protocol for monitoring algae levels this study integrates UAV-based hyperspectral remote sensing, in situ water sampling, and laboratory analysis of retrieved algal specimens. Specifically, we present:

1) A protocol for harmonised lab-field hyperspectral regression.
2) A systematic mislabel/outlier handling and imbalance-aware evaluation.
3) An empirical comparison across VIS/NIR/combined and model families.

II. Methodology

Field data was collected from two sites over four days in August 2025, as shown in Fig. 1. The camera used for field acquisition was a UAV-mounted Headwall Nano nNP Hyperspectral Imager from the National Environment Research Council's (NERC) Field Spectroscopy Facility (FSF) [10]. This is a push-broom sensor with 343 bands over 400 - 1000 nm. These were pre-processed by NERC FSF involving the following stages to produce ortho-rectified, geo-referenced images:

1) Raw digital signal to radiance.
2) Radiance to reflectance.

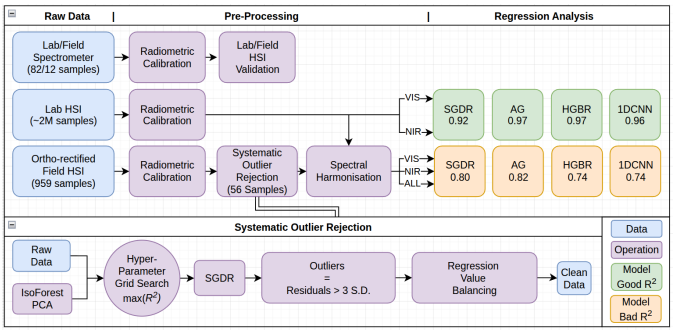


Fig. 2: Hyperspectral Regression Protocol. Overview showing the pre-processing, systematic outlier rejection method and the suite of individual regression experiments. In the regression section the colour (green-orange) indicates the success of the model, via the highest scoring R^2 value.

3) Digital Elevation Model (DEM) creation.
4) Single Best Estimate of Trajectory (SBET) file creation.
5) Ortho-rectification of the data.

Further processing was conducted at Newcastle University, which involved projection from the provided EPSG:4326 co-ordinate reference system to NTM:29903. Ground truth water content measurements were taken with an Aquaprobe AP-6000 [11] and GPS locations taken with a Leica GS 18 Real Time Kinematic sensor. During sampling, multiple readings were taken to provide a measurement of the variability in the sampling method. Where sampling was conducted in especially thick algal regions the sensor was cleaned between subsequent measurements. Data was post-processed, removing erroneous readings such as negative readings, and 500-pixel average spectral curves were extracted corresponding to the GPS locations of ground truth samples. Large pixel-averages were taken due to the high levels of noise present in the hyperspectral images that averaging helped eliminate. The resulting spectral curves are shown in Fig. 4.a.

For the lab data two Ximea XiSpec [12] cameras were used covering the VIS and NIR ranges, with 40 bands in total. Each camera was of snapshot design with 16 and 25 sub-pixels coated in specific bandpass fabry-perot filters. These are arranged in mosaic fashion across the CMOS sensor in order to capture a snapshot-style hyperspectral image. We collected white and dark references, captured spectral images across multiple concentrations. Halogen bulbs were used and pre-heated for 15 minutes before imaging. When capturing images the HSI Mosaic software provided by IMEC was used for initial capturing and post processing via a full coverage pipeline into reflectance images.

In more detail, the lab process involved an initial period of camera streaming to allow their sensors to reach a steady temperature state. Once the plateau was reached the various reference images were taken. Dark reference was taken with the lens cover in place. White reference was taken by placing a Zenith Lite 95% reflectance panel covering the entire field of view of the camera. Lighting in the lab was constant with very little effect from natural light so only a single white reference

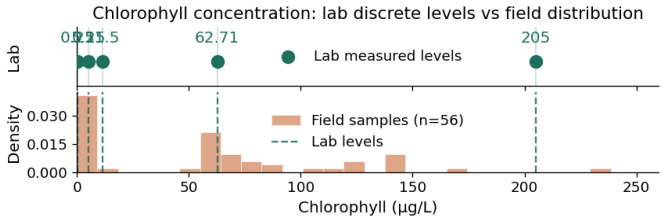


Fig. 3: The field and lab chl-a concentrations after field data has been cleaned.

was taken at the start of the session and applied to all images. Integration time on the camera was set to the highest value whilst viewing all band images to ensure no over-saturation occurred. In general the integration time was around 14ms, with F# set to 10 to allow a wide depth of focus. Images were taken in sequence, first with the VIS camera then with the NIR camera due to the inability to image both at the same time on our hardware. This meant both datasets should be considered separately for individual pixels but can be incorporated into a single dataset when taking the averages over each image. All images were taken from directly above the beaker. All samples were distilled from a main sample that was collected from Lough Neagh during the August 2025 visit and propagated in the lab for a period of 4 months. The initial sample was taken as the maximum concentration then diluted roughly 1:1 with lab water to get the lower concentrations. At each distillation the chlorophyll content was measured with algae presence only becoming visible to the naked eye around the 62.71 $\mu g/L$ sample and above.

The post-processed reflectance images were cropped to the central square region which contained no glint from the beaker edges. The VIS and NIR spectral curves were appended to provide a spectral curve of the full range of the two cameras. Figure 4.b shows the resulting curves. Note that the cameras do not cover the region between 590.02 nm - 672.78 nm, this is a hardware limitation that we account for by truncating the field VIS and NIR to matching regions for our ML comparisons.

A. Pre-processing

For our method, dimensionality reduction has been limited to resampling down from the 343 Headwall Nano bands to the 40 Ximea Xispec camera bands via convolution using the original camera sensor SRF. Further dimension reduction was not applied in this case as we wanted first to investigate best performance. Qualitative analysis of the spectral curves using the lab samples as a reliable reference revealed that some field sample labels did not match the collected spectral curve from the images, a common issue when collecting unconstrained aquatic field data. We postulate that there were two main sources of this error depending on the location: in high-flow areas, the time difference between collecting the samples and taking the HSI means that the imaged area is no longer the same as it was during sampling; in low-flow areas, the algae was more likely to clump up, and this resulted in high-variance readings with the Aquaprobe device. To this end, we required a systematic method to remove the mislabelled curves. The method involved the following, and a flowchart is shown in Fig. 2: label-agnostic outlier removal via PCA [13] and Isolation Forest [13] (IF) methods with a grid search of hyperparameters to maximise R^2. The remaining data was passed through a label-aware method; a model was trained using a Stochastic Gradient Descent Regressor [13] (SGDR), and any residuals larger than three standard deviations were removed as outliers. Finally, concentration value binning and balancing was applied to bring the lower regression values to the same order as the range of higher regression values. All data was passed through the SciKit-Learn StandardScaler [13] before model training, which adjusts for curve variance within the dataset. The field data was split into the two wavelength ranges corresponding to the visible (VIS, 458–594 nm) and near-infrared (NIR, 673–950 nm), to match the division of data from the two hyperspectral snapshot cameras used in the lab.

B. Regression Analysis

Experiments were performed running three ML methods on both field and lab data. SGDR [13], a linear regression method using Stochastic Gradient Descent (SGD) in order to deal with larger datasets like our lab data, provided a starting baseline with which to compare other models. We then used AutoGluon [14] (AG), an automated machine learning package, which runs various models with hyperparameter searches in order to find the best model based on the provided data and target metric. We ran with R^2 as the target metric with 600 seconds per model. Based on the AG results we then selected a single model from Scikit-Learn that performed well on all datasets and was easier and more flexible to build into other pipelines than the AG models. We selected HistGradientBoostingRegressor [13] (HGBR), a tree-based regression algorithm optimised for handling large datasets. All tests were run with 5-fold cross-validation on the lab and cleaned field datasets. R^2 and MAE were recorded as key metrics.

C. 1D-CNN

The abundance of the lab data pixels (2M) allows us to apply more complex deep-learning methods. Although classic tree-based machine learning is frequently quoted in literature as the most effective methods, these are sometimes chosen based on lack of a large enough dataset. 1D-CNNs have been used previously as suitable regression models for Chl-a; we took inspiration from Pyo et al [9] in constructing a simple 1D-CNN. Two convolutional layers (32 and 64 filters, kernel size 3) with ReLU activations, followed by adaptive global average pooling and a fully connected regression head with a 32-unit hidden layer, 20% dropout, and a single output neuron for regression prediction. We trained the 1D-CNN on both field and lab data.

III. Experimental Results and Discussions

Initial field spectral curves before outlier rejection are shown in Fig. 4.a. The mean lab data spectral curves are shown in

Fig. 4.b. Analysis of the lab data shows strong correlation between chl-a concentration and spectral magnitude; however, in the field samples the concentration does not seem to follow that trend for many samples. For this reason the systematic outlier rejection step was created. We also compare spectrometer readings (in red) to their camera-collected counterparts, the spectrometer readings are more reliable than the hyperspectral image data due to the absence of spectral cross-talk, high signal-to-noise ratio (SNR), and radiometric simplicity. We can see that in the field the spectrometer readings display similar spectral features and magnitudes. For the lab data we see a difference in magnitude between the HSI and the spectrometer data, which could indicate a problem with the lab calibration procedure. A key observation is the significant difference in reflectance magnitude for similar chl-a values between the lab and field data. The small difference between the lab HSI and spectrometer values does not explain this large difference. The cause remains unclear, though we expect sample depth, illumination angle, and bidirectional reflectance distribution function (BRDF) effects to play a role.

The data in Fig. 4.a was then passed through our outlier removal method. The PCA/IF grid search resulted in keeping 30% of PCA variance and a 6% contamination removal value for IF. Though 30% PCA variance sounds drastic, this left 93.95% of the original data. Subsequently, the label-aware SGDR rejection step left 92.07% of the original data. Finally, due to a large imbalance in the concentration values (85% $<$ 10 $\mu g/L$), binning and balancing was applied, leaving a concentration distribution shown in Fig. 3. Note that the previous outlier removals acted to increase the R^2 score, whereas the act of balancing the classes actually resulted in a significant reduction in performance when run with the ML models. This was due to the models' bias to predict from the most numerous class and thus inflated results, hence the importance of the balancing regression values step. The remaining data was then spectrally harmonised to the 40 bands of the lab data, resulting in the spectral curves shown in Fig. 4.c, which we can see qualitatively align with the magnitude versus concentration relationship that we would expect from the lab data.

The cleaned field data and lab data were then used to train the SGDR, AG, HGBR, and 1D-CNN models; the summary of the results is shown in Table. I. SGDR is the simplest approach, using linear regression to model the data, and as such provides an initial baseline to judge other results. SGDR shows a strong performance on lab data but substantially weaker performance on the field data. After the initial linear results, AG was run to determine the best-performing models for this data. The top model was *WeightedEnsemble_L2* (WEL2) for every dataset, an ensemble model that leverages the strength of individual models. Although WEL2 achieved the highest validation performance, the Histogram Gradient Boosting Regressor (HGBR) was selected for subsequent experiments. Gradient-boosted decision trees scored similarly well to WEL2, so HGBR provided competitive predictive performance while offering a single, well-defined model architecture that enabled consistent evaluation. As Table. I shows, HGBR showed similar performance to WEL2 for the lab data but a drop in performance for the field data, possibly due to the small dataset. Finally, the 1D-CNN proved almost as effective as the AG method and about the same as HGBR for the lab data but was the worst overall models for the field data. This was expected due to the small dataset size. We can also see another expected result in that, for our target of predicting algae concentration, in most cases the NIR range scores beat that of the VIS or even VIS+NIR range. This is likely because although the VIS range contains green that will be highly correlated with algae growth, it is more informative what is going on with the red-edge, as explained in Section I. As a

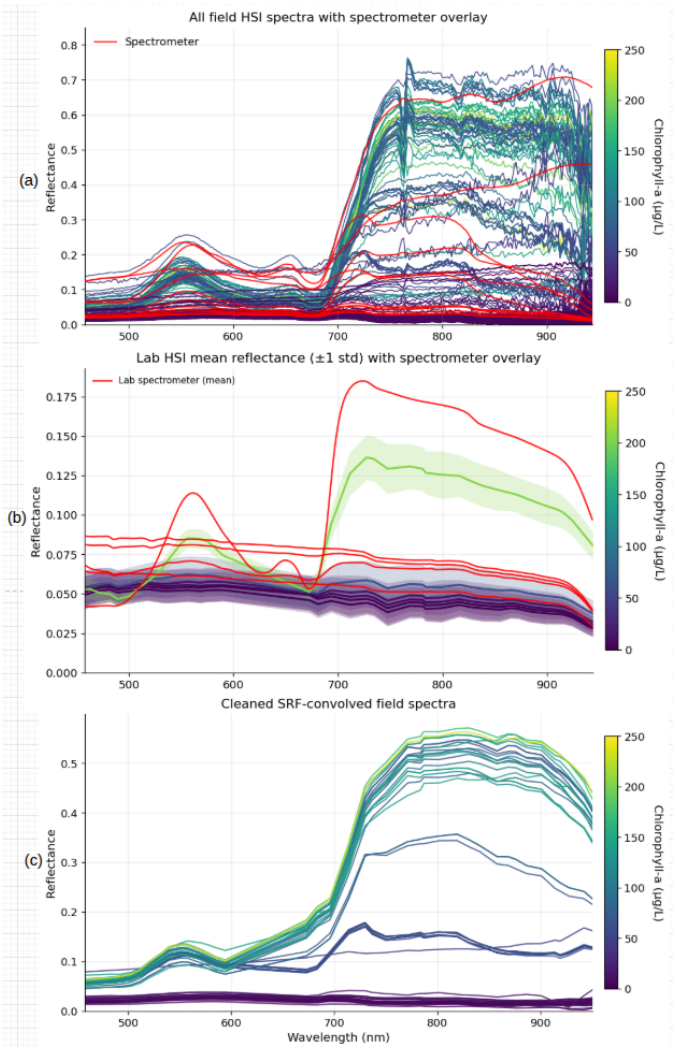


Fig. 4: Plots of field and lab spectral curves, corroborated with spectrometer readings, colour indicates Chl-a concentration: (a) All 959 field spectra overlaid with spectrometer readings in red. (b) Averages of the VIS+NIR combined lab data overlaid with spectrometer spectra. (c) Field data after cleaning and spectral harmonisation with lab data. Note that all data is on the same x-axis but the y-axis scale varies, which shows one aspect of the domain gap between lab-field.

TABLE I: Best performance comparison of Stochastic Gradient Descent Regression (SGDR), AutoGluon (AG), Histogram Gradient Boosting Regression (HGBR), and 1D CNN regression models across the laboratory and field datasets. Each method was run with 5-fold cross-validation, 20 repeats (except AG and 1D-CNN), random state incremented with each run, with the results shown from the best R^2 score.

		SGDR		**AG**		**HGBR**		**1D-CNN**	
Dataset	**Samples**	**MAE (μg/L)**	R^2	**MAE (μg/L)**	R^2	**MAE (μg/L)**	R^2	**MAE (μg/L)**	R^2
Lab VIS	1.2M	16.56	0.9190	**10.57**	**0.9469**	11.21	0.9437	14.72	0.9247
Lab NIR	812k	15.30	0.9241	**6.83**	**0.9717**	7.53	0.9683	9.57	0.9623
Field VIS+NIR	56	15.33	0.8014	**12.45**	**0.8153**	17.38	0.7384	15.96	0.7367
Field VIS	56	22.13	0.6562	**14.70**	**0.7742**	25.32	0.5667	21.78	0.5823
Field NIR	56	16.29	0.7161	**14.87**	**0.7503**	16.31	0.7394	23.57	0.5829

final step beyond the summarised results, the better performing lab models were run on the field counterpart data to judge the usefulness of these models as replacements for intensive field data. The results were consistently worse than the field-specific models, showing the need for more complex adaption methods to properly leverage the lab data.

IV. Conclusion

We have described in detail a protocol for processing hyperspectral images with corresponding Chl-a concentrations in both lab and field environments in order to build the most effective models for regression. Our approach avoids common pitfalls of field data collection like unbalanced classes inflating performance scores and mislabelled data decreasing performance. It provides a systematic method for cleaning data, using lab data wavelengths to inform band selection for the field data thereby not requiring dimension reduction beyond spectral harmonisation. The results show that a reasonable performance can be obtained for the field data and that in fact both VIS and NIR contain sufficient features to still retain most of the performance compared to the full VIS + NIR range, though NIR most often slightly outperforms the VIS range. The lab data with its significantly higher sample number performs impressively though is limited to only 6 concentration levels, acting more like discrete data than the truly continuous field data. As such our next steps would involve collecting more lab data to fill in the gaps between concentration levels. Additionally, based on the gap between lab-field data performance our next step is to investigate scaling/domain adaption/fine-tuning methods to leverage the more numerous and reliable lab data and adapt the lab models to the field data for more reliable predictions without the costly effort of collecting more field data.

V. Acknowledgements

We gratefully acknowledge the use of instruments, field support and post-processing from NERC [10]. Funding came from project:"ALGAE-Watch: Blue Algae Surveillance; A remote sensing-based digital twin framework at Lough Neagh, NI" and the Department of Agriculture, Environment and Rural Affairs (DAERA) and the UK Space Agency's Small Business Research Initiative (SBRI).

References

[1] Northern Ireland Water, "Lough neagh blue-green algae," https://www.niwater.com/about-your-water/bga, 2026, Accessed: 27 July 2026.

[2] Edward B. Knipling, "Physical and physiological basis for the reflectance of visible and near-infrared radiation from vegetation," *Remote Sensing of Environment*, vol. 1, no. 3, pp. 155–159, 1970.

[3] Muhammad Saad Shaikh, Keyvan Jaferzadeh, Benny Thörnberg, and Johan Casselgren, "Calibration of a hyper-spectral imaging system using a low-cost reference," *Sensors*, vol. 21, no. 11, 2021.

[4] Paul Geladi, Jim Burger, and Torbjörn Lestander, "Hyperspectral imaging: calibration problems and solutions," *Chemometrics and Intelligent Laboratory Systems*, vol. 72, no. 2, pp. 209–217, 2004, Advances in Chromatography and Electrophoresis - Conferentia Chemometrica 2003, Budapest.

[5] Abhinav Galodha, Maria Peppa, Sam Wilson, Sanya Anees, Brejesh Lall, and Shaikh Ziauddin Ahammad, "Spatio-temporal and dynamic prediction of water quality changes (turbidity, chlorophyll-a and colour dissolved organic matter) using multi-sensor data fusion: satellite, UAV, and cloud-based modeling," in *AGU Fall Meeting Abstracts*, Dec. 2025, vol. 2025 of *AGU Fall Meeting Abstracts*, pp. H23D–06.

[6] Neha Joshi, Jongmin Park, Kaiguang Zhao, Alexis Londo, and Sami Khanal, "Monitoring harmful algal blooms and water quality using sentinel-3 olci satellite imagery with machine learning," *Remote Sensing*, vol. 16, no. 13, 2024.

[7] Wonjin Jang, Yongeun Park, JongCheol Pyo, Sanghyun Park, Jinuk Kim, Jin Hwi Kim, Kyung Hwa Cho, Jae-Ki Shin, and Seongjoon Kim, "Optimal band selection for airborne hyperspectral imagery to retrieve a wide range of cyanobacterial pigment concentration using a data-driven approach," *Remote Sensing*, vol. 14, no. 7, 2022.

[8] Doyun Kim, KyoungJin Lee, SeungMyeong Jeong, MinSeok Song, ByeoungJun Kim, Jungsu Park, and Tae-Young Heo, "Real-time chlorophyll-a forecasting using machine learning framework with dimension reduction and hyperspectral data," *Environmental Research*, vol. 262, pp. 119823, 2024.

[9] JongCheol Pyo, Seok Min Hong, Jiyi Jang, Sanghun Park, Jongkwan Park, Jae Hoon Noh, and Kyung Hwa Cho, "Drone-borne sensing of major and accessory pigments in algae using deep learning modeling," *GIScience & Remote Sensing*, vol. 59, no. 1, pp. 310–332, 2022.

[10] Alex Merrington and Zoe Halem, "Natural environment research council, headwall nano-hyperspec® uav imaging system," https://fsf.nerc.ac.uk/instruments/uav/headwallnano/, 2025, Accessed: 28 July 2026.

[11] Aquaread Ltd., "AP-6000 Advanced Multiparameter Water Quality Probe," 2026, Accessed: 28 July 2026.

[12] XIMEA, "xispec hyperspectral miniature cameras," https://www.ximea.com/products/hyperspectral-imaging/xispec-hyperspectral-miniature-cameras, 2026, Accessed: Jul. 27, 2026.

[13] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *J. Mach. Learn. Res.*, vol. 12, no. null, pp. 2825–2830, Nov. 2011.

[14] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola, "Autogluon-tabular: Robust and accurate automl for structured data," *arXiv preprint arXiv:2003.06505*, 2020.