Water Research 197 (2021) 117073

Contents lists available at ScienceDirect

Water Research

journal homepage: www.elsevier.com/locate/watres

Machine learning for anomaly detection in cyanobacterial fluorescence signals

Husein Almuhtaram[a,*], Arash Zamyadi[b,c], Ron Hofmann[a]

[a] Department of Civil and Mineral Engineering, University of Toronto, Toronto ON M5S 1A4 Canada
[b] Water RA Melbourne based position hosted by Melbourne Water, 990 La Trobe St, Docklands VIC 3008, Australia
[c] BGA Innovation Hub and Water Research Centre, School of Civil and Environment Engineering, University of New South Wales (UNSW), Sydney, NSW 2052, Australia

ARTICLE INFO

Article history:
Received 27 November 2020
Revised 6 February 2021
Accepted 17 March 2021
Available online 19 March 2021

Keywords:
Phycocyanin
Drinking water treatment
Cyanobacteria
Monitoring
Artificial intelligence
Chlorophyll a
CCchHlo C

ABSTRACT

Many drinking water utilities drawing from waters susceptible to harmful algal blooms (HABs) are implementing monitoring tools that can alert them to the onset of blooms. Some have invested in fluorescence-based online monitoring probes to measure phycocyanin, a pigment found in cyanobacteria, but it is not clear how to best use the data generated. Previous studies have focused on correlating phycocyanin fluorescence and cyanobacteria cell counts. However, not all utilities collect cell count data, making this method impossible to apply in some cases. Instead, this paper proposes a novel approach to determine when a utility needs to respond to a HAB based on machine learning by identifying anomalies in phycocyanin fluorescence data without the need for corresponding cell counts or biovolume. Four widespread and open source algorithms are evaluated on data collected at four buoys in Lake Erie from 2014 to 2019: local outlier factor (LOF), One-Class Support Vector Machine (SVM), elliptic envelope, and Isolation Forest (iForest). When trained on standardized historical data from 2014 to 2018 and tested on labelled 2019 data collected at each buoy, the One-Class SVM and elliptic envelope models both achieve a maximum average F1 score of 0.86 among the four datasets. Therefore, One-Class SVM and elliptic envelope are promising algorithms for detecting potential HABs using fluorescence data only.

1. Introduction

Cyanobacteria are increasingly threatening drinking water supplies worldwide (Fernández et al., 2015). There is a need for improved monitoring to trigger responses by stakeholders. Traditional monitoring relies on visual observation of the source water and cell counting by microscopy (Chorus and Bartram, 1999; EPA Office of Water, 2015; Health Canada, 2016). However, visual monitoring of the water surface does not necessarily capture the conditions at the intake of a drinking water treatment plant, and microscopy is a labor-intensive and slow technique. Consequently, approaches including gene quantification by quantitative polymerase chain reaction (qPCR), remote sensing, cell imaging, and real-time fluorescence monitoring have been developed (Pacheco et al., 2016; Srivastava et al., 2013). However, qPCR kits must be used by treatment plant staff at appropriate measurement frequencies; remote sensing is limited to capturing the conditions at or near the water surface; and automated cell imaging and identification techniques are promising but are often highly dependent on the quality of the model calibration (Jin et al., 2018). Of these, only cell imaging and fluorescence monitoring can be implemented online at the drinking water intake.

Fluorescence monitoring probes measure the fluorescence of the cyanobacteria-specific photosynthetic pigment phycocyanin and chlorophyll a, present in all photosynthetic organisms. There is a need to find a better way for utilities to use fluorescence data to trigger a response to mitigate the effects of a developing algal bloom. In their response, a utility can also determine whether the bloom is a HAB that poses a potential toxin or taste and odor risk. The primary approach to interpreting phycocyanin fluorescence data in the literature is to correlate it to cell counts or biovolume determined by microscopy (Bertone et al., 2018). The resulting coefficients of determination in field samples have ranged from $R^2 = 0.41$ to 0.87, indicating that a linear correlation generally exists between phycocyanin fluorescence and cell counts or biovolume (Almuhtaram et al., 2018). Threshold values for early warnings for cyanobacteria blooms can be set based on guideline values for cell counts or biovolumes given by various juris-

* Corresponding author:
E-mail addresses: husein.almuhtaram@mail.utoronto.ca, husein1226@hotmail.com (H. Almuhtaram).

https://doi.org/10.1016/j.watres.2021.117073
0043-1354/

dictions including the World Health Organization (WHO). However, the correlations can be site-specific and, if the composition of the cyanobacteria community changes, season-specific, requiring periodic validation of their accuracy by additional cell counting in field samples or raw water samples spiked with cyanobacteria cell cultures (Chang et al., 2012; Loisa et al., 2015; Symes and van Ogtrop, 2016). Consequently, using monitoring probes in this way may require considerable continued effort to ensure the threshold values remain relevant given changing local conditions (Symes and van Ogtrop, 2016).

Moreover, in practice, utilities do not regularly enumerate cyanobacteria by microscopy. Consequently, monitoring data are often used without quantitative correlations to cell counts by interpreting the fluorescence pattern in a qualitative way, to trigger a response. For example, (Zamyadi et al., 2016b) set an arbitrary fluorescence threshold of 10% above the baseline phycocyanin readings to trigger permanganate dosing in a full-scale trial to oxidize cyanobacteria cells and microcystins. However, this approach is subjective and may be prone to bias and inefficiency. Therefore, there is a need to interpret real-time monitoring data such that a utility would be able to determine when to initiate their HAB response strategy without relying on slow and laborious manual methods.

The potential to use machine learning for anomaly detection to interpret phycocyanin fluorescence data has not yet been investigated despite successful applications for other types of water quality data (Hou et al., 2013; Jin et al., 2019; Liu et al., 2020; Shi et al., 2018. Anomalies in fluorescence data inform utilities of when to investigate possible cyanobacteria blooms. An anomalous data point could be due to either a change in the actual cyanobacteria concentration where the monitoring probe is installed, or due to interference from chlorophyll a, turbidity, or temperature, which can be significant (Chang et al., 2012; Choo et al., 2018, 2019; Zamyadi et al., 2016a). Therefore, utilities need to determine the cause of the anomaly by analyzing samples, such as for cell counts or toxin or taste and odor compound concentrations.

In machine learning applications, it is recommended to evaluate multiple algorithms and compare their performance since each has its own assumptions about the underlying structure of the data (Wolpert, 1996). Anomaly detection algorithms can be categorized based on their mechanisms of operation, which include classification, clustering, density, distance, isolation, and prediction, among others (Celebi and Aydin, 2016; Hodge and Austin, 2004; Liu et al., 2020; Mehrotra et al., 2017). Previous studies have successfully implemented various types of algorithms to identify outliers in environmental data including dissolved oxygen (Samuelsson et al., 2019), groundwater levels (Azimi et al., 2018; Jeong et al., 2017), windspeed (Hill and Minsker, 2010), and surface water quality (Jin et al., 2019). However, none have applied machine learning to detect anomalous fluorescence signals that might correspond to an increase in cyanobacteria activity. There are several promising types of algorithms that might be applicable for such data based on previous studies. Deep learning algorithms are promising for detecting anomalies in water quality data (Dogo et al., 2019), but they require hundreds of thousands or millions of data points (Namuduri et al., 2020). Moreover, they are not necessarily better than traditional machine learning algorithms for detecting anomalies in univariate time-series data (Braei and Wagner, 2020). Therefore, this study focuses on traditional machine learning algorithms that have been used successfully in related applications.

Specifically, local outlier factor (LOF), a density-based algorithm, has been used to identify outliers in water quality sensor data collected online in an aquaculture application (Gao et al., 2019). Isolation Forest (iForest), an isolation-based algorithm, was used by Liu et al., (2020) to detect anomalies in surface water quality parameters collected using handheld sensors. A classifier-based algorithm, One-Class Support Vector Machine (SVM), has been used to identify anomalies in wastewater treatment plant influent quality (Cheng et al., 2019) and operating conditions (Harrou et al., 2018). In studies evaluating multiple algorithms, One-Class SVM has been used alongside iForest and elliptic envelope to detect outliers in groundwater and water tank levels with either iForest or One-Class SVM resulting in the best performance, depending on the test conditions (Azimi et al., 2018; Tan et al., 2020). Nonetheless, elliptic envelope, a distance-based algorithm, has been successfully used to identify outliers in multivariate lake water quality data collected over several years (Alameddine et al., 2010). Thus, the success of these algorithms in previous studies warrants an investigation into whether they might be useful for detecting anomalies in cyanobacteria fluorescence data.

Therefore, the objective of this study is to illustrate a proof-of-concept for LOF, One-Class SVM, elliptic envelope, and iForest to identify potential HABs from monitoring data without the need for corresponding cell count data. The algorithms are evaluated on data collected in Lake Erie from 2014 to 2019, and the models with the best average performance are identified as potential tools for future use. To the best of these authors' knowledge, this study is the first to implement unsupervised machine learning on phycocyanin fluorescence data to identify cyanobacteria activity. This approach may benefit many drinking water utilities that employ online probes to monitor phycocyanin but lack corresponding cyanobacteria cell count data with which a correlation could be established.

2. Materials and methods

2.1. Site and data description

Western Lake Erie is particularly susceptible to harmful algal blooms due to nutrient delivery from its tributaries, especially the Maumee River (Harke et al., 2016). The National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) therefore monitors water quality in western Lake Erie using several buoys. Four of these buoys (WE2, WE4, WE8, and WE13) are equipped for continuous monitoring of phycocyanin and chlorophyll a fluorescence using YSI EXO2 (YSI, Yellow Springs, OH, USA) multiparameter water quality sondes equipped with Total Algae sensors (Fig. 1). Hourly data collected in 2014 at the WE2 and WE4 buoys and data collected every 15 min in 2015–2019 at all four buoys was obtained from the GLERL open source archives (NOAA/GLERL, 2020).

The parameter of interest in this application is phycocyanin fluorescence, and it can be pre-processed to improve classification via standardization. Standardization involves setting the mean of the data to zero and the variance to one. An algorithm's performance may be more consistent when tested on standardized datasets because the data among them will be made more similar than if they were unmodified. Therefore, it is expected that the optimized algorithm models will be applicable to multiple datasets. Standardization was applied using the StandardScaler preprocessing tool available in Python .

2.2. Machine learning algorithms

Four machine-learning algorithms for unsupervised anomaly detection were applied to the Lake Erie monitoring data: local outlier factor (LOF), One-Class Support Vector Machine (SVM), elliptic envelope, and Isolation Forest (iForest) although the approach described in this study is not limited to only these algorithms (Chalapathy and Chawla, 2019; Goldstein and Uchida, 2016; Hodge and Austin, 2004; Mladenov et al., 2013). These algorithms

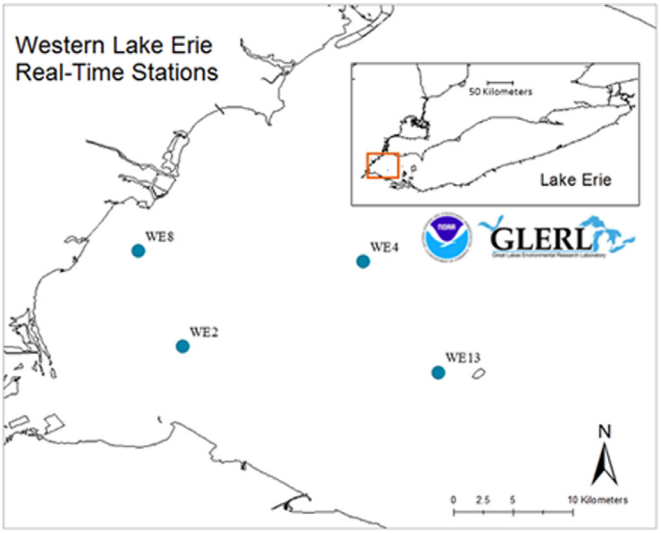

Fig. 1. Locations of the four NOAA buoys in Lake Erie that continuously monitor general air and water quality parameters including temperature, turbidity, chlorophyll a, phycocyanin, nitrogen, phosphorus, dissolved oxygen, and pH. Weekly sample collection for microcystins, extracted chlorophyll and phycocyanin, turbidity, and temperature also occurs at these and four other monitoring sites. The buoys are located near the mouth of the Maumee River (WE2), near the center of the Lake Erie western basin (WE4), near the edge of the western basin (WE8), and near the water intake for the City of Toledo, Ohio (WE13) (Meyer et al., 2017).

are unsupervised because they use unlabelled data. That is, the algorithms do not know whether any of the data points inputted to them are normal or anomalous. All of the data analysis was conducted in the Python programming language (V. 3.7.3) using the scikit-learn machine-learning package (V. 0.20.3) (Pedregosa et al., 2011).

The LOF algorithm, first described by Breunig et al., (2000), computes a score for every point in a dataset based on the distance to its k nearest neighbors. The LOF score represents the degree to which a point is outlying. A point is identified as an outlier if its LOF score exceeds a threshold, which is determined based on a user-defined contamination rate.

In contrast, the One-Class SVM algorithm estimates a function that returns positive for normal (non-outlier) data points and negative for outliers using a user-defined probability that any data point is not an outlier (i.e., the contamination rate). This is accomplished by mapping the data points into a feature space corresponding to the radial basis function kernel (commonly used in SVM algorithms) and separating them from the origin with a maximum margin hyperplane in a higher dimension feature space using a minimization formulation (Schölkopf et al., 2001).

Unlike the maximum margin method of the One-Class SVM, the elliptic envelope models the data to a Gaussian distribution and identifies an ellipse that contains most of the data. A data point outside the ellipse is anomalous. The size and shape of the ellipse are determined by the FAST-Minimum Covariance Determinant algorithm (Rousseeuw and Driessen, 1999), which iteratively computes the Mahalanobis distance (a measure of how many standard deviations a data point is from the mean) of subsamples from the data until the determinant of the covariance matrix converges (Hoyle et al., 2015). A contamination rate is used to define the approximate proportion of data points that lie outside the ellipse. It has been reported that the contamination rate does not necessarily need a high degree of accuracy, so it can be estimated initially and adjusted in subsequent runs of the algorithm (Hoyle et al., 2015).

Lastly, a fourth and fundamentally different anomaly detection approach, iForest, is used. The preceding three methods rely on building a profile of the data and identifying outliers by various metrics. In iForest, anomalous points are explicitly isolated based on the fact that they are few and different, and no profile of the normal data is constructed (Liu et al., 2008). Instead, the algorithm calculates an anomaly score based on the path length required to isolate a data point in binary trees containing all the data points. The points with the shortest path lengths that are below a threshold determined by a user-defined contamination rate are identified as anomalies. Further details for all four algorithms are provided in SI Text S1 and Figure S1.

Each algorithm's ability to classify data can be optimized by adjusting its contamination rate. The LOF algorithm additionally requires the number of nearest neighbors to be inputted and optimized. Contamination rate ranges from 0 to 0.5 for all the algorithms except One-Class SVM, for which it ranges from 0 to 1, and provides the algorithms with approximate starting points for the proportion of outliers in the data. Thus, contamination rates ranging from 0.01 to 0.45 are evaluated for all the algorithms in this study in increments of 0.05, which is appropriate because the effect of small changes (e.g., 0.01) in contamination rate does not significantly impact an algorithm's performance, as reported in previous studies (Hoyle et al., 2015; Tan et al., 2020).

2.3. Validation

The four algorithms are used to detect outliers in fluorescence data, but it is not known if the outliers identified correspond to real harmful algal bloom (HAB) events. The detection of an outlier by the algorithm alone is not sufficient to initiate a management response if the outlier cannot be reliably interpreted as representing a real HAB (Rudin, 2019). Real HAB events in Lake Erie in 2019 were identified using the satellite-based NOAA GLERL Experimental Lake Erie HAB Tracker and NOAA HAB Forecasts by overlaying the buoy positions onto the processed satellite images of Lake Erie. The HAB Tracker provides hourly satellite images of western Lake Erie processed to show the concentration of cyanobacterial chlorophyll a. The absence and presence of cyanobacterial activity at the buoy locations were used to label the 2019 fluorescence monitoring data as normal or anomalous, respectively. The HAB Tracker tracks the movement of cyanobacterial blooms hourly whereas the monitoring probes take measurements every 15 min. Therefore, the normal or anomalous condition of each hour in the satellite data was used to label the four corresponding data points measured by the monitoring probe for that hour (i.e., on the hour, 15 min past, 30 min past, and 45 min past). Although these are two different types of data, both the monitoring probes and the satellite images capture surface water conditions. Thus, if cyanobacterial activity occurs in the location of the buoy, it would be detected by both the satellite images and the probe sensors. Furthermore, the satellite data is used only to qualitatively label the monitoring data as normal or anomalous and not to compare actual pigment concentrations. The algorithm outputs are compared to the labels when trained and tested only on the phycocyanin measurements of each of the four buoy probes (NOAA/GLERL, 2020). Chlorophyll a is omitted from the training and testing data because it represents not only cyanobacteria but also green algae, whereas the HAB Tracker is specifically designed to identify cyanobacteria blooms. Nonetheless, chlorophyll a can interfere with the measurement of phycocyanin fluorescence (Zamyadi et al., 2016a), so clorophyll a values are considered to identify periods where interference may have affected either the optical sensor or the HAB Tracker and HAB Forecast.

Numbers of true positives, false positives, and false negatives were determined by comparing the algorithm outputs to the labelled data. A true positive occurs when both the model output and the remote sensing data indicate a bloom is present; a false positive occurs when the model indicates a bloom is present but

Table 1
Unmodified and standardized data used to train the unsupervised learning algorithms. Training was conducted using this data with contamination rates ranging from 0.01 to 0.45. The mean and standard deviation of the standardized datasets are 0 and 1, respectively.

	Number of data points	Unmodified rangerr (RFU)	Standardized range rRaange (RFU)	Unmodified mean (RFU)	Unmodified standard deviation (RFU)
WE2	61,456	-0.76 to 68.1	-1.1 to 58.0	0.61	1.2
WE4	61,262	-0.75 to 49.9	-1.5 to 72.3	0.31	0.7
WE8	51,873	-0.31 to 48.3	-0.72 to 34.0	0.70	1.4
WE13	46,065	-0.04 to 18.1	-0.83 to 29.0	0.46	0.6

the remote sensing data does not; and a false negative occurs when the model output indicates there is no cyanobacteria activity while the remote sensing data does. Two important metrics for comparing algorithm performance can be calculated from these values: precision, the proportion of positive results that are actually correct, and recall, the proportion of positive results that are correctly identified, as follows (Alla and Adari, 2019):

$$Precision = \frac{TP}{TP + FP} \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. Finally, the harmonic mean of precision and recall, the F1 score, is used to compare the classification accuracy of the models, as follows (Abou-Moustafa and Schuurmans, 2015):

$$F1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

The F1 score ranges from 0 to 1. An F1 score of 1 indicates perfect precision and recall (best performance) while a score of 0 indicates that either precision or recall are 0 (worst performance). Therefore, the optimal conditions are those that result in the highest F1 score. For example, Tan et al., (2020) found that a remarkable F1 score of 0.99 could be achieved for anomaly detection in water level data using the One-Class SVM algorithm with a contamination rate of 0.3. However, the F1 score decreases to 0.69 if the contamination rate is set to 0.1 due to a drop in recall from 0.99 to 0.52 despite precision remaining at 0.99 (Tan et al., 2020). Thus, the F1 score is the most representative metric of a model's performance because it penalizes losses in either precision or recall, or both (Muharemi et al., 2019).

2.4. Training data

Unsupervised algorithms have no knowledge of the correct classes for each data point, but they derive internally generated error measures to classify data based solely on the statistics of the training data (Kyan et al., 2014). Therefore, a training dataset should be abundant and diverse (Gong et al., 2019). As such, it is common practice to use the bulk of an available dataset for training, leaving a smaller portion behind for testing. The algorithms were trained using different hyperparameter values (contamination rates and k nearest neighbors) on four datasets: 2014 – 2018 WE2 and WE4 data and 2015 – 2018 WE8 and WE13 data, described in Table 1, and tested on each buoy's 2019 data. The selection of the hyperparameter values is as important as the training data itself, and although their use is specific to each algorithm, the general concept is that they provide an approximate starting point for the classifier.

The same process was repeated for the standardized datasets, described in Table 1. Reducing the means from 0.31 – 0.70 to 0 effectively makes the algorithms more sensitive. Thus, the sensitivity of the algorithms trained on standardized data will begin to increase at lower contamination rates. More importantly, by making the four datasets more similar in terms of mean and standard deviation, the performance of each algorithm is expected to be more consistent at each contamination rate for all the datasets compared to without standardization. If so, then the models with the best performance identified in this study should be applicable to standardized data collected at other sites, that is, the models are expected to be generalizable.

The optimal hyperparameter selection varies for every dataset. In this study, the optimal contamination rates and k nearest neighbors (k-NN) were determined by identifying the conditions that result in the highest F1 score when tested on the labelled 2019 datasets and trained on the unmodified or standardized training data, as described by Xu et al., (2019).

3. Results

3.1. Local outlier factor

The optimal k-NN for LOF varied depending on the dataset being trained and on whether the data were standardized. K-NN values ranging from 1 to 30 were evaluated at the previously selected contamination rates for each of the unmodified and standardized datasets. Figure S2 shows the results of the k-NN optimization for the standardized WE2 dataset, and Table S1 summarizes the results of all the k-NN optimizations. For the standardized WE2 dataset, the optimum k-NN values are 1, 5, 10, and 19 for contamination rates of 0.01, 0.05, 0.1, and 0.15, respectively. A low k-NN value emphasizes the detection of outliers located within smaller clusters of data compared to high k-NN values. Therefore, more outliers are detected when the k-NN is small so the k-NN value tends to increase (i.e., detect fewer outliers) as the contamination rate increases to offset the increase in sensitivity by the higher contamination rate to ensure the F1 score is as high as possible. However, increasing k-NN with contamination rate did not always improve the F1 score, and in all cases the optimal k-NN remained unchanged beyond a contamination rate of 0.2. This indicates that k-NN has a greater influence on the algorithm output than contamination rate, so it is more important for k-NN to be optimized.

Despite optimizing both hyperparameters, the performance of the LOF algorithm was adequate at best, reaching a maximum F1 score of 0.81 when trained on the unmodified 2014 – 2018 WE4 data and tested on the 2019 WE4 umodified data (Fig. 2). For all the datasets, the F1 score plateaued beyond a contamination rate of 0.05. The algorithm performs significantly better using the unmodified training data, with an F1 score improvement of up to 0.36 (WE4 raw vs. WE4 std for contamination rates > 0.1). The average F1 score among the unmodified results is only 0.69. Therefore, even when optimized in terms of the training data, k-NN, and contamination rate, this algorithm is not suitable for anomaly detection in phycocyanin fluorescence signals.

3.2. One-class SVM

The One-Class SVM algorithm performed significantly better than the LOF algorithm when the training data was standardized. The maximum average F1 score for the four datasets increases

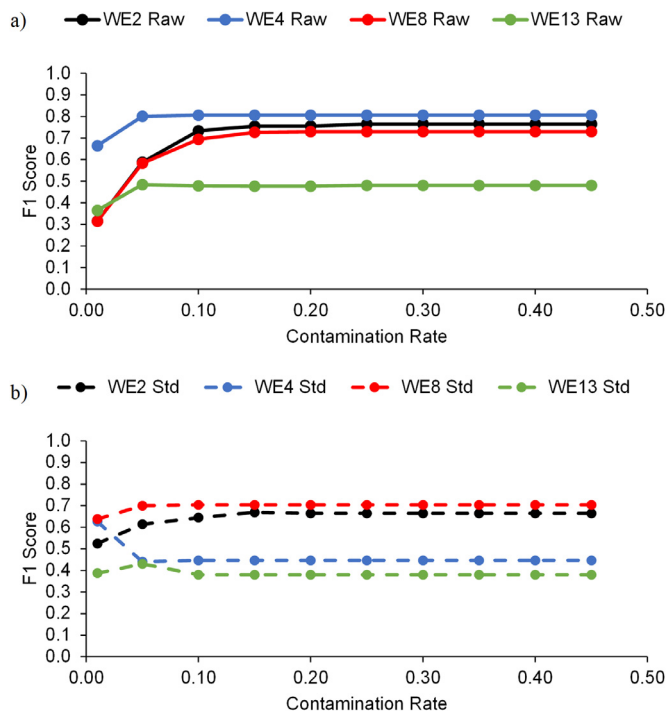

Fig. 2. F1 scores obtained using the LOF algorithm using the optimized k-NN value for each contamination rate in the (a) unmodified data and (b) standardized data. Raw means the training data is unmodified and std means it is standardized (mean = 0, standard deviation = 1). Little to no variation in the F1 score is observed for contamination rates of 0.05 and above, indicating that the optimized k-NN value has a stronger effect on the algorithm's performance.

from 0.72 at a contamination rate of 0.45 in unmodified data to 0.86 at contamination rates of 0.3 and 0.35 in standardized data (Fig. 3). Fig. 3a shows that the F1 score increases gradually with contamination rate for the WE2, WE8, and WE13 datasets but that the performance for the WE13 data is significantly worse. Without standardization, the peak performance for the WE2, WE4 and WE8 datasets is similar, although they occur at much lower contamination rates for the standardized data. However, the algorithm achieves a maximum F1 score in the range of 0.62 to 0.67 for the standardized WE13 data compared to a maximum of only 0.35 for the unmodified data. Therefore, standardization is an important pre-processing step for One-Class SVM.

3.3. Elliptic envelope

Like the One-Class SVM algorithm, the elliptic envelope algorithm's peak average performance occurred at a lower contamination rate for the standardized data that the unmodified data (Fig. 4). However, for both types of data the performance decreases sharply as contamination rate increases. This suggests that this algorithm is more sensitive to changes in contamination rate than the One-Class SVM algorithm, making it less forgiving from a user's perspective when a labelled dataset is not available for optimization. When the data are unmodified, the peak performance is acceptable at 0.84 and occurs at a contamination rate of 0.25 (Fig. 4a). For the standardized data, the maximum average F1 score is equal to that of the One-Class SVM algorithm, 0.86, and it occurs at a contamination rate of 0.15 (Fig. 4b).

3.4. Isolation forest

The change in performance with contamination rate of the iForest algorithm closely resembles that of the elliptic envelope algorithm. The peak average F1 score in the unmodified data occurs

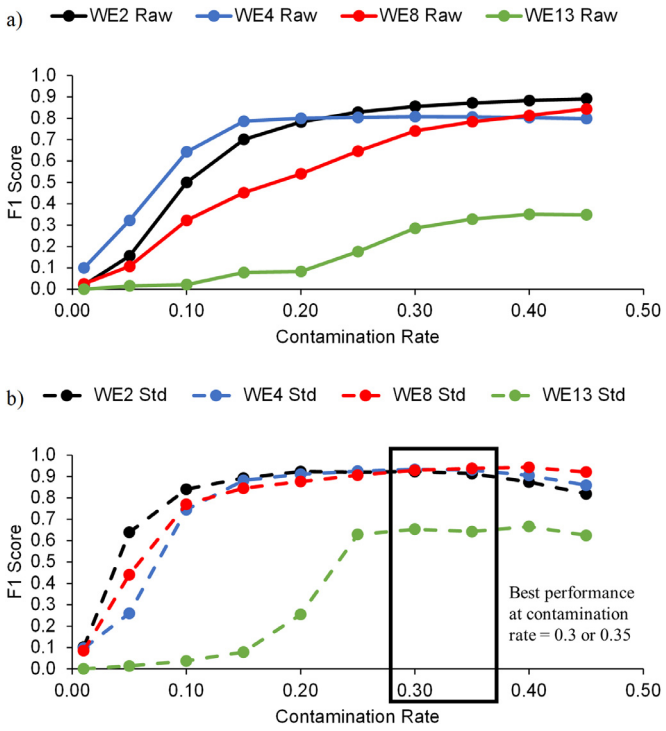

Fig. 3. Performance of the One-Class SVM algorithm on the (a) unmodified and (b) standardized phycocyanin fluorescence data collected at the four buoys. The peak average F1 score for the unmodified data is 0.72 and occurs at a contamination rate of 0.45 while the peak average score for the standardized data is 0.86 and occurs at contamination rates of 0.3 and 0.35.

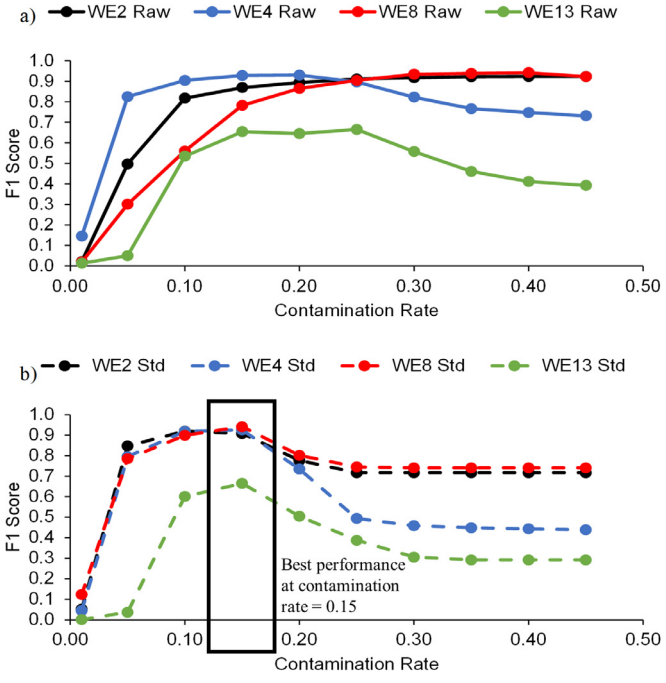

Fig. 4. Performance of the elliptic envelope algorithm on the (a) unmodified and (b) standardized phycocyanin fluorescence data collected at the four buoys. The peak average F1 score for the unmodified data is 0.84 and occurs at a contamination rate of 0.25 while the peak average score for the standardized data is 0.86 and occurs at a contamination rate of 0.15.

Table 2
Maximum and average peak F1 scores achieved for each algorithm using their optimum contamination rates. The maximum scores for the LOF algorithm were achieved in the unmodified datasets while the other three algorithms performed best in the standardized data. Maximum F1 scores for the WE13 data are significantly lower than for the other datasets.

	WE2	WE4	WE8	WE13	Average	Standard deviation
LOF	0.76	0.81	0.73	0.48	0.69	0.13
One-Class SVM	0.92	0.93	0.94	0.67	**0.86**	0.11
Elliptic Envelope	0.92	0.93	0.94	0.66	**0.86**	0.12
iForest	0.92	0.91	0.92	0.64	0.84	0.12

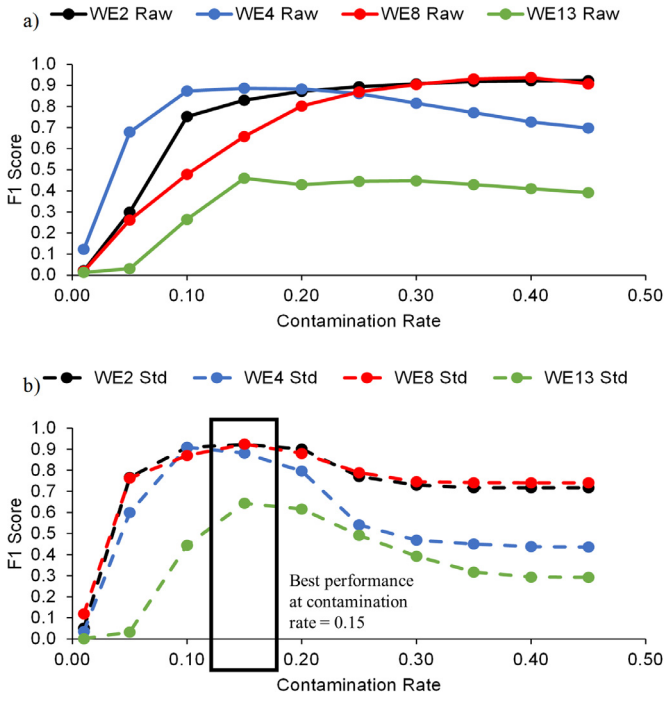

Fig. 5. Performance of the iForest algorithm on the (a) unmodified and (b) standardized phycocyanin fluorescence data collected at the four buoys. The peak average F1 score for the unmodified data is 0.77 and occurs at contamination rates of 0.25 and 0.3 while the peak average score for the standardized data is 0.84 and occurs at a contamination rate of 0.15.

at contamination rates of 0.25 and 0.3, and it occurs at 0.15 in the standardized data (Fig. 5). However, the actual peak scores are lower, at 0.77 and 0.84, respectively, than for the elliptic envelope method. Therefore, to achieve an F1 score above 0.8, standardization of the training data is required.

3.5. Algorithm optimization and validation

Table 2 summarizes the peak F1 scores for each algorithm and dataset when the optimal hyperparameters and preprocessing conditions are used. Standardization of the training data resulted in the highest maximum F1 scores for all the algorithms except LOF. For LOF, the F1 score was affected primarily by the k-NN value, which was optimized for every test condition (Table S1). Maximum F1 scores greater than 0.9 were achieved by the One-Class SVM, elliptic envelope, and iForest algorithms for the standardized WE2, WE4, and WE8 datasets. Fig. 6 shows the anomaly detection results for the optimized One-Class SVM, elliptic envelope, and iForest algorithms on the standardized WE4 dataset, as an example (Figures S3 and S4 show the results for the WE2 and WE8 datasets). These three algorithms were able to correctly detect the two major cyanobacterial blooms that occurred in the WE4 buoy location in 2019, shaded in yellow, with F1 scores above 0.9.

Chlorophyll a interference was significant for the WE13 dataset, causing all the F1 scores to be below 0.7. Fig. 7 shows the prediction results for the One-Class SVM, elliptic envelope, and iForest algorithms on the WE13 dataset. For all three, it is clear that false positives occurred primarily from September 3 to 11, 2019. These dates coincide with a rise in chlorophyll a fluorescence from about 2 to 6 RFU. It is possible that the elevated chlorophyll a in the water column interfered with the probe's measurement of phycocyanin despite no cyanobacteria being detected by the HAB Tracker (Zamyadi et al., 2016a). Conversely, it is also possible that cyanobacterial activity did occur during those dates but was not detected by the HAB Tracker due to being masked by green algae or other reasons such as cloud cover (Erickson et al., 2012).

4. Discussion

The findings of this study demonstrate for the first time that anomaly detection using unsupervised machine learning can be used to detect cyanobacteria activity from phycocyanin fluorescence data. Cyanobacteria activity was detected accurately in four Lake Erie datasets using three of the four algorithms evaluated. Across the datasets, the optimized One-Class SVM (contamination rate = 0.3 or 0.35) and elliptic envelope (contamination rate = 0.15) algorithms exhibited the best performance, each with an average F1 score of 0.86. The optimized iForest algorithm followed closely with an average F1 score of 0.84, although this difference is not statistically significant. LOF achieved an average score of only 0.69. This was likely due to the gradual separation of outliers from normal data whereas the LOF algorithm's strength is in identifying local clusters of outliers (Goldstein and Uchida, 2016).

An advantage to using these models is that they may allow utilities to interpret phycocyanin fluorescence without the need for microscopy data. Additionally, online monitoring data can be added to the training datasets to ensure the models make predictions based on the most recent information. For example, Gao et al., (2019) developed and implemented an aquaculture water quality monitoring model that identifies anomalies and predicts water quality in real time. Another advantage is that these algorithms can be applied to any source water provided that historical data is available and that it is distributed like the datasets shown in Table 1. This is supported by the finding that the three models are generalizable for all four datasets, that is, they were trained and tested on four different datasets and still achieved high average F1 scores. Similar results were reported by Tan et al., (2020) who found that their dual-stage One-Class SVM algorithm is applicable to different scenarios having achieved an F1 score of 0.99 for water level data using a contamination rate of 0.3. A potential way to implement this approach in practice is to include the algorithms in probe monitoring software to detect anomalies in real time data (e.g., every 15 min as a new data point is logged), triggering a response. Liu et al., (2020) similarly recommend that decision mak-

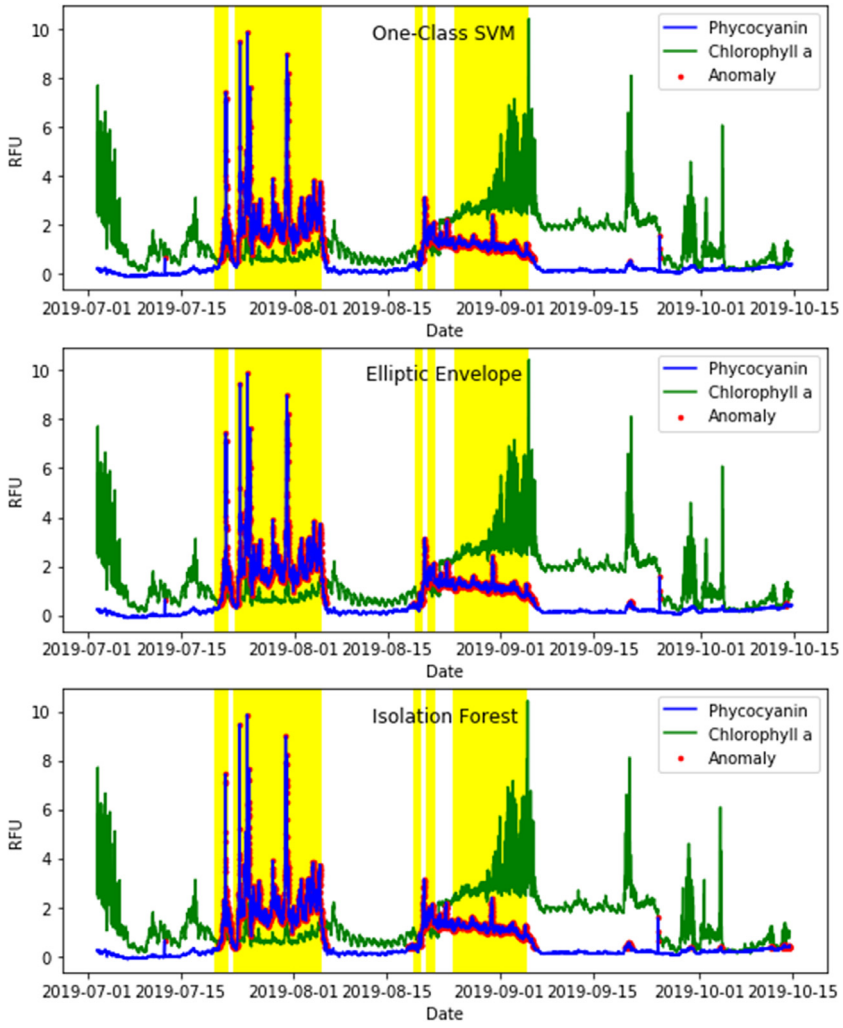

Fig. 6. Anomaly detection results for the WE4 dataset using the One-Class SVM, elliptic envelope, and iForest algorithms with contamination rates of 0.3, 0.15, and 0.1, respectively. Their corresponding F1 scores are 0.93, 0.93, and 0.91, indicating that these optimized models correctly predicted cyanobacterial activity with a high degree of accuracy. The yellow shaded regions indicate periods where cyanobacterial activity was detected by the HAB Tracker. A few apparent false positives were identified, but it is possible that they were due to chlorophyll a interference. The performance of the algorithms on the WE2 and WE8 datasets is similar. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

ers carry out emergency responses when anomalies in river water quality are detected by their machine learning framework.

However, there are important limitations to this approach that must be considered before it is adopted. First, it does not forecast how a HAB will develop or what the future implications are for the most recent predictions. In particular, anomalies in phycocyanin fluorescence data cannot identify successional changes to the species composition of the cyanobacteria community (Chorus, 2012). So, without additional data the potential for toxin or taste and odor compound production cannot be determined by this approach alone (Bastien et al., 2011; Bertone et al., 2018), although some studies have reported correlations between phycocyanin and microcystins production (Aragão et al., 2020; Francy et al., 2016; Izydorczyk et al., 2009).

Hence, this has the potential to more accurately identify when a utility should respond to a cyanobacteria bloom (e.g., by collecting samples or reactively dosing oxidants) compared to the conventional practice of correlating phycocyanin fluorescence with cell counts. For example, a recent study determined a phycocyanin fluorescence threshold value of 3.6 RFU using the same model of probe as this study, corresponding to the World Health Organization Alert Level 1 (0.2 mm^3/L), for Great Lakes region plants including a plant whose source water is Lake Erie (Almuhtaram et al., 2018; Chorus and Welker, 2021). Although Alert Level 1 is based on the potential for cyanobacteria to produce 1 µg/L microcystins, it has been used as an early warning threshold by some researchers (Brient et al., 2008; Izydorczyk et al., 2009; Macário et al., 2015; Zamyadi et al., 2012). If that threshold value were to be applied to the four datasets used in this study, it would achieve an average F1 score of only 0.05, although in practice a threshold determined this way can continue to be adjusted based on knowledge of the system. This simple example demonstrates that the machine learning approach may be more sensitive than the conventional approach for setting an early warning threshold and can be an important part of a drinking water utility's HAB monitoring system.

From a utility perspective, it may be relevant to use the algorithms to detect both phycocyanin and chlorophyll a anomalies as green algae can also be problematic from a treatment perspective (Kommineni et al., 2009; Zamyadi et al., 2013). This was not attempted in this study due to the lack of a labelled dataset that can be used to validate the anomalies detected with real green algae blooms. Nonetheless, the algorithms used in this study are capable of handling multiple input variables, so this could be theoretically implemented in the future. Alternatively, because chlorophyll a in-

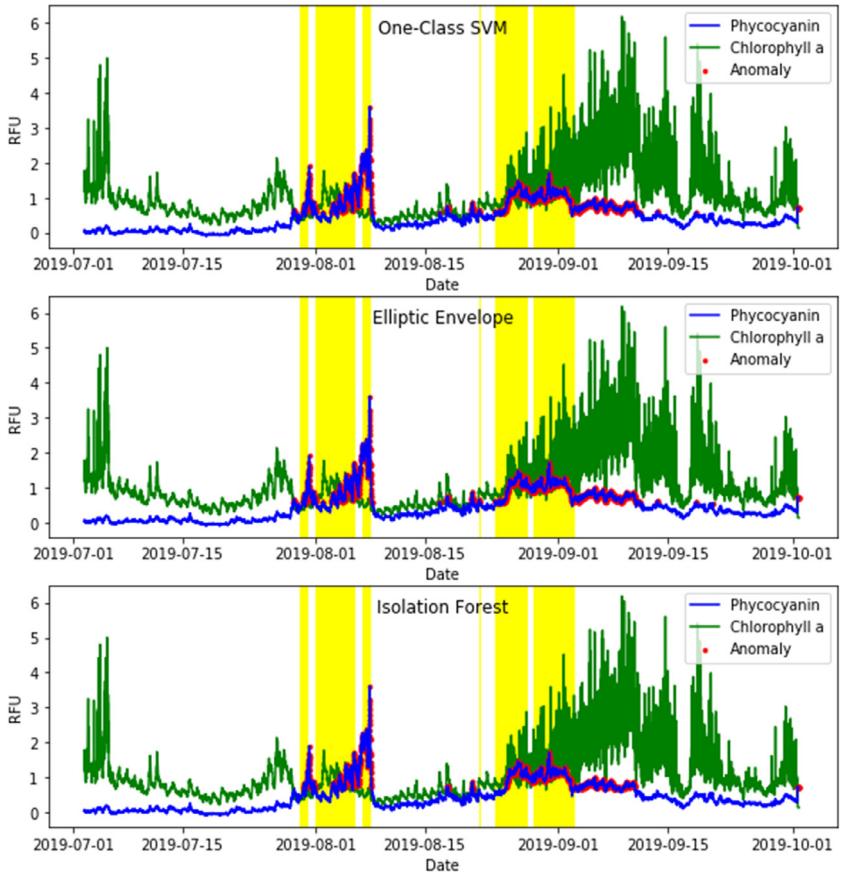

Fig. 7. The One-Class SVM, elliptic envelope, and iForest algorithms achieved maximum F1 scores of only 0.67, 0.66, and 0.64, respectively, for the WE13 dataset. Even when optimized, a significant number of anomalies continued to be identified following the second major cyanobacterial bloom that ended on September 3, 2019. The yellow shaded regions indicate periods where cyanobacterial activity was detected by the HAB Tracker. It is likely that the elevated chlorophyll a fluorescence interfered with the phycocyanin sensor and caused elevated phycocyanin to be measured and anomalies to be identified despite no cyanobacterial activity detected by remote sensing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

terference is systematic for the YSI EXO2 probe, correction factors can effectively eliminate the bias from non-cyanobacterial green algae (Choo et al., 2019). A similar effect may be achieved using supervised machine learning such that the effects of certain variables on the model output are learned (Gomez-Alvarez and Revetta, 2020). Future work should explore these possibilities as well as other algorithms with deep learning and data streaming capabilities for assessing multivariate data spanning many years and for detecting anomalies in real time, respectively.

5. Conclusions

The conclusions drawn from this study are:

- Elevated cyanobacterial activity can be reliably detected from phycocyanin fluorescence data using unsupervised machine learning algorithms.
- Standardization of the training data is an important preprocessing step for this approach, especially where multiple datasets are used, because it improves the consistency of an algorithm's performance across datasets.
- Several promising algorithms are identified that may be readily implemented by drinking water utilities or adopted by fluorometer manufacturers in their monitoring software: One-Class SVM with a contamination rate of 0.3 or 035 and elliptic envelope with a contamination rate of 0.15. Similar performance was also achieved by the iForest algorithm with a contamination rate of 0.15.
- Training and testing the algorithms on only phycocyanin fluorescence was sufficient to accurately identify anomalies except in one dataset where interference from chlorophyll a affected the sensor. Nonetheless, it may be important for some utilities to identify periods of elevated chlorophyll a, and this should be considered in future studies.

Therefore, these and other unsupervised machine learning models have the potential to be applied to phycocyanin fluorescence data to identify anomalies indicating cyanobacterial activity and may eventually be included in monitoring software to be continually optimized to trigger alarms in real time.

Funding sources

This work was supported by the Natural Sciences and Engineering Research Council of Canada Project CRDPJ 482052–15.

Declaration of Competing Interest

None.

Acknowledgments

The authors acknowledge funding from the City of Toronto, the City of Hamilton, the Regional Municipality of Niagara, the Regional Municipality of York, the Regional Municipality of Durham, and the Union Water Supply System. The authors thank Parsa Torabian for his insight and guidance on machine learning.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.watres.2021.117073.

References

Abou-Moustafa, K.T., Schuurmans, D., 2015. Generalization in unsupervised learning. In: Efficient Learning Machines. Apress, Berkeley, CA, pp. 300–317. doi:10.1007/978-3-319-23528-8_19.

Alameddine, I., Kenney, M.A., Gosnell, R.J., Reckhow, K.H., 2010. Robust multivariate outlier detection methods for environmental data. J. Environ. Eng. 136, 1299–1304. doi:10.1061/(ASCE)EE.1943-7870.0000271.

Alla, S., Adari, S.K., 2019. Beginning Anomaly Detection Using Python-Based Deep Learning. Apress, Berkeley, CA doi:10.1007/978-1-4842-5177-5.

Almuhtaram, H., Cui, Y., Zamyadi, A., Hofmann, R., 2018. Cyanotoxins and cyanobacteria cell accumulations in drinking water treatment plants with a low risk of bloom formation at the source. Toxins (Basel) 10, 430. doi:10.3390/toxins10110430.

Aragão, M.C., dos Reis, K.C., Souza, A.C., Rocha, M.A.M., Capelo Neto, J., 2020. Modeling total microcystin production by microcystis aeruginosa using multiple regression. J. Water Supply Res. Technol. 69, 415–426. doi:10.2166/aqua.2020.128.

Azimi, S., Azhdary Moghaddam, M., Hashemi Monfared, S.A., 2018. Anomaly detection and reliability analysis of groundwater by crude Monte Carlo and importance sampling approaches. Water Resour. Manag. 32, 4447–4467. doi:10.1007/s11269-018-2029-y.

Bastien, C., Cardin, R., Veilleux, É., Deblois, C., Warren, A., Laurion, I., 2011. Performance evaluation of phycocyanin probes for the monitoring of cyanobacteria. J. Environ. Monit. 13, 110–118. doi:10.1039/C0EM00366B.

Bertone, E., Burford, M.A., Hamilton, D.P., 2018. Fluorescence probes for real-time remote cyanobacteria monitoring: a review of challenges and opportunities. Water Res. 141, 152–162. doi:10.1016/j.watres.2018.05.001.

Braei, M., Wagner, S., 2020. Anomaly detection in univariate time-series: a survey on the State-of-the-Art. arXiv.

Breunig, M.M., Kriegel, H.-P., Ng, R.T., Sander, J., 2000. LOF: identifying density-based local outliers. In: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data - SIGMOD '00, New York, New York, USA. ACM Press, pp. 93–104. doi:10.1145/342009.335388.

Brient, L., Lengronne, M., Bertrand, E., Rolland, D., Sipel, A., Steinmann, D., Baudin, I., Legeas, M., Le Rouzic, B., Bormans, M., 2008. A phycocyanin probe as a tool for monitoring cyanobacteria in freshwater bodies. J. Environ. Monit. 10, 248–255. doi:10.1039/B714238B.

Celebi, E.M., Aydin, K., 2016. Unsupervised Learning Algorithms. Springer International Publishing, Cham doi:10.1007/978-3-319-24211-8.

Chalapathy, R., Chawla, S., 2019. Deep learning for anomaly detection: a survey. arXiv 1–50.

Chang, D.W., Hobson, P., Burch, M., Lin, T.F., 2012. Measurement of cyanobacteria using in-vivo fluoroscopy - Effect of cyanobacterial species, pigments, and colonies. Water Res. 46, 5037–5048. doi:10.1016/j.watres.2012.06.050.

Cheng, T., Dairi, A., Harrou, F., Sun, Y., Leiknes, T., 2019. Monitoring influent conditions of wastewater treatment plants by nonlinear data-based techniques. IEEE Access 7, 108827–108837. doi:10.1109/ACCESS.2019.2933616.

Choo, F., Zamyadi, A., Newton, K., Newcombe, G., Bowling, L., Stuetz, R., Henderson, R.K., 2018. Performance evaluation of in situ fluorometers for real-time cyanobacterial monitoring. H2Open J. 1, 26–46. doi:10.2166/h2oj.2018.009.

Choo, F., Zamyadi, A., Stuetz, R.M., Newcombe, G., Newton, K., Henderson, R.K., 2019. Enhanced real-time cyanobacterial fluorescence monitoring through chlorophyll-a interference compensation corrections. Water Res. 148, 86–96. doi:10.1016/j.watres.2018.10.034.

Chorus, I., 2012. Current Approaches to Cyanotoxin Risk Assessment, Risk Management and Regulations in Different Countries.

Chorus, I., Bartram, J., 1999. Toxic Cyanobacteria in Water: A guide to Their Public Health Consequences, Monitoring and Management, Retrieved March doi:10.1046/j.1365-2427.2003.01107.x.

Chorus, I., Welker, M., 2021. Toxic Cyanobacteria in Water. CRC Press, Second edition. Boca Rataon : CRC Press, an imprint of Informa, 2021. 10.1201/9781003081449

Dogo, E.M., Nwulu, N.I., Twala, B., Aigbavboa, C., 2019. A survey of machine learning methods applied to anomaly detection on drinking-water quality data. Urban Water J. 16, 235–248. doi:10.1080/1573062X.2019.1637002.

EPA Office of Water, 2015. Recommendations for public water systems to manage cyanotoxins in drinking water.

Erickson, J.S., Hashemi, N., Sullivan, J.M., Weidemann, A.D., Ligler, F.S., 2012. In situ phytoplankton analysis: theres plenty of room at the bottom. Anal. Chem. 84, 839–850. doi:10.1021/ac201623k.

Fernández, C., Estrada, V., Parodi, E.R., 2015. Factors triggering cyanobacteria dominance and succession during blooms in a hypereutrophic drinking water supply reservoir. 10.1007/s11270-014-2290-5

Francy, D.S., Brady, A.M.G., Ecker, C.D., Graham, J.L., Stelzer, E.A., Struffolino, P., Dwyer, D.F., Loftin, K.A., 2016. Estimating microcystin levels at recreational sites in western Lake Erie and Ohio. Harmful Algae 58, 23–34. doi:10.1016/j.hal.2016.07.003.

Gao, G., Xiao, K., Chen, M., 2019. An intelligent IoT-based control and traceability system to forecast and maintain water quality in freshwater fish farms. Comput. Electron. Agric. 166, 105013. doi:10.1016/j.compag.2019.105013.

Goldstein, M., Uchida, S., 2016. A Comparative evaluation of unsupervised anomaly detection algorithms for multivariate data 1–31. 10.1371/journal.pone.0152173

Gomez-Alvarez, V., Revetta, R.P., 2020. Monitoring of nitrification in chloraminated drinking water distribution systems with microbiome bioindicators using supervised machine learning. Front. Microbiol. 11, 1–13. doi:10.3389/fmicb.2020.571009.

Gong, Z., Zhong, P., Hu, W., 2019. Diversity in machine learning. IEEE Access 7, 64323–64350. doi:10.1109/ACCESS.2019.2917620.

Harke, M.J., Davis, T.W., Watson, S.B., Gobler, C.J., 2016. Nutrient-controlled niche differentiation of western lake erie cyanobacterial populations revealed via metatranscriptomic surveys. Environ. Sci. Technol. 50, 604–615. doi:10.1021/acs.est.5b03931.

Harrou, F., Dairi, A., Sun, Y., Senouci, M., 2018. Statistical monitoring of a wastewater treatment plant: a case study. J. Environ. Manage. 223, 807–814. doi:10.1016/j.jenvman.2018.06.087.

Health Canada, 2016. Cyanobacterial Toxins in Drinking Water.

Hill, D.J., Minsker, B.S., 2010. Anomaly detection in streaming environmental sensor data: a data-driven modeling approach. Environ. Model. Softw. 25, 1014–1022. doi:10.1016/j.envsoft.2009.08.010.

Hodge, V., Austin, J., 2004. a survey of outlier detection methodologies. Artif. Intell. Rev. 22, 85–126. doi:10.1023/B:AIRE.0000045502.10941.a9.

Hou, D., He, H., Huang, P., Zhang, G., Loaiciga, H., 2013. Detection of water-quality contamination events based on multi-sensor fusion using an extented Dempster–Shafer method. Meas. Sci. Technol. 24, 055801. doi:10.1088/0957-0233/24/5/055801.

Hoyle, B., Rau, M.M., Paech, K., Bonnett, C., Seitz, S., Weller, J., 2015. Anomaly detection for machine learning redshifts applied to SDSS galaxies. Mon. Not. R. Astron. Soc. 452, 4183–4194. doi:10.1093/mnras/stv1551.

Izydorczyk, K., Carpentier, C., Mrówczyński, J., Wagenvoort, A., Jurczak, T., Tarczyńska, M., 2009. Establishment of an alert level framework for cyanobacteria in drinking water resources by using the Algae online analyser for monitoring cyanobacterial chlorophylla. Water Res. 43, 989–996. doi:10.1016/j.watres.2008.11.048.

Jeong, J., Park, E., Han, W.S., Kim, K.-Y., 2017. A subagging regression method for estimating the qualitative and quantitative state of groundwater. Hydrogeol. J. 25, 1491–1500. doi:10.1007/s10040-017-1561-9.

Jin, C., Mesquita, M.M.F.F., Deglint, J.L., Emelko, M.B., Wong, A., 2018. Quantification of cyanobacterial cells via a novel imaging-driven technique with an integrated fluorescence signature. Sci. Rep. 8, 9055. doi:10.1038/s41598-018-27406-0.

Jin, T., Cai, S., Jiang, D., Liu, J., 2019. A data-driven model for real-time water quality prediction and early warning by an integration method. Environ. Sci. Pollut. Res. 26, 30374–30385. doi:10.1007/s11356-019-06049-2.

Kommineni, S., Amante, K., Karnik, B., Sommerfeld, M., Dempster, T., Area, S., Quality, W., 2009. Strategies for controlling and mitigating algal growth within water treatment plants.

Kyan, M., Muneesawang, P., Jarrah, K., Guan, L., 2014. Unsupervised Learning. John Wiley & Sons, Inc., Hoboken, NJ, USA doi:10.1002/9781118875568.

Liu, F.T., Ting, K.M., Zhou, Z.-H., 2008. Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining. IEEE, pp. 413–422. doi:10.1109/ICDM.2008.17.

Liu, J., Wang, P., Jiang, D., Nan, J., Zhu, W., 2020. An integrated data-driven framework for surface water quality anomaly detection and early warning. J. Clean. Prod. 251, 119145. doi:10.1016/j.jclepro.2019.119145.

Loisa, O., Kääriä, J., Laaksonlaita, J., Niemi, J., Sarvala, J., Saario, J., 2015. From phycocyanin fluorescence to absolute cyanobacteria biomass: an application using in-situ fluorometer probes in the monitoring of potentially harmful cyanobacteria blooms. Water Pract. Technol. 10, 695–698. doi:10.2166/wpt.2015.083.

Macário, I.P.E., Castro, B.B., Nunes, M.I.S., Antunes, S.C., Pizarro, C., Coelho, C., Gonçalves, F., de Figueiredo, D.R., 2015. New insights towards the establishment of phycocyanin concentration thresholds considering species-specific variability of bloom-forming cyanobacteria. Hydrobiologia 757, 155–165. doi:10.1007/s10750-015-2248-7.

Mehrotra, K.G., Mohan, C.K., Huang, H., 2017. Anomaly Detection Principles and Algorithms, Terrorism, Security, and Computation. Springer International Publishing, Cham doi:10.1007/978-3-319-67526-8.

Meyer, K.A., Davis, T.W., Watson, S.B., Denef, V.J., Berry, M.A., Dick, G.J., 2017. Genome sequences of lower Great Lakes Microcystis sp. reveal strain-specific genes that are present and expressed in western Lake Erie blooms. PLoS ONE 12, e0183859. doi:10.1371/journal.pone.0183859.

Mladenov, V., Koprinkova-Hristova, P., Palm, G., Villa, A.E.P., Appollini, B., Kasabov, N., 2013. Artificial neural networks and machine learning – ICANN 2013. 23rd International Conference on Artificial Neural Networks doi:10.1007/978-3-642-40728-4.

Muharemi, F., Logofătu, D., Leon, F., 2019. Machine learning approaches for anomaly detection of water quality on a real-world data set. J. Inf. Telecommun. 3, 294–307. doi:10.1080/24751839.2019.1565653.

Namuduri, S., Narayanan, B.N., Davuluru, V.S.P., Burton, L., Bhansali, S., 2020. Review—deep learning methods for sensor based predictive maintenance and future perspectives for electrochemical sensors. J. Electrochem. Soc. 167, 037552. doi:10.1149/1945-7111/ab67a8.

NOAA/GLERL, 2020. Experimental Lake Erie Harmful Algal Bloom (HAB) Tracker [WWW Document] URL https://www.glerl.noaa.gov/res/HABs_and_Hypoxia/habTracker.html .

Pacheco, A., Guedes, I., Azevedo, S., 2016. Is qPCR a reliable indicator of cyanotoxin risk in freshwater? Toxins (Basel) 8, 172. doi:10.3390/toxins8060172.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res. 12, 2825–2830.

Rousseeuw, P., Driessen, K., 1999. A Fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat. Mach. Intell. 1, 206–215. doi:10.1038/s42256-019-0048-x.

Samuelsson, O., Zambrano, J., Björk, A., Carlsson, B., 2019. Automated active fault detection in fouled dissolved oxygen sensors. Water Res. 166, 115029. doi:10.1016/j.watres.2019.115029.

Schölkopf, B., Platt, J.C., Shawe-Taylor, J., Smola, A.J., Williamson, R.C., 2001. Estimating the support of a high-dimensional distribution. Neural. Comput. 13, 1443–1471. doi:10.1162/089976601750264965.

Shi, B., Wang, P., Jiang, J., Liu, R., 2018. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. Sci. Total Environ. 610–611, 1390–1399. doi:10.1016/j.scitotenv.2017.08.232.

Srivastava, A., Singh, S., Ahn, C.-Y., Oh, H.-M., Asthana, R.K., 2013. Monitoring approaches for a toxic cyanobacterial bloom. Environ. Sci. Technol. 47, 8999–9013. doi:10.1021/es401245k.

Symes, E., van Ogtrop, F., 2016. Determining the efficacy of a submersible in situ fluorometric device for cyanobacteria monitoring coalesced with total suspended solids characteristic of lowland reservoirs. River Res. Appl. 32, 1632–1641. doi:10.1002/rra.2993.

Tan, F.H.S., Park, J.R., Jung, K., Lee, J.S., Kang, D.-K., 2020. Cascade of One Class Classifiers for Water Level Anomaly Detection. Electronics 9, 1012. doi:10.3390/electronics9061012.

Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. Neural. Comput. 8, 1341–1390. doi:10.1162/neco.1996.8.7.1341.

Xu, Z., Kakde, D., Chaudhuri, A., 2019. Automatic hyperparameter tuning method for local outlier factor, with applications to anomaly detection. In: 2019 IEEE International Conference on Big Data (Big Data. IEEE, pp. 4201–4207. doi:10.1109/BigData47090.2019.9006151.

Zamyadi, A., Choo, F., Newcombe, G., Stuetz, R., Henderson, R.K., 2016a. A review of monitoring technologies for real-time management of cyanobacteria: recent advances and future direction. TrAC Trends Anal. Chem. 85, 83–96. doi:10.1016/j.trac.2016.06.023.

Zamyadi, A., Dorner, S., Sauvé, S., Ellis, D., Bolduc, A., Bastien, C., Prévost, M., 2013. Species-dependence of cyanobacteria removal efficiency by different drinking water treatment processes. Water Res. 47, 2689–2700. doi:10.1016/j.watres.2013.02.040.

Zamyadi, A., Henderson, R.K., Stuetz, R., Newcombe, G., Newtown, K., Gladman, B., 2016b. Cyanobacterial management in full-scale water treatment and recycling processes: reactive dosing following intensive monitoring. Environ. Sci. Water Res. Technol. 2, 362–375. doi:10.1039/C5EW00269A.

Zamyadi, A., McQuaid, N., Prévost, M., Dorner, S., 2012. Monitoring of potentially toxic cyanobacteria using an online multi-probe in drinking water sources. J. Environ. Monit. 14, 579–588. doi:10.1039/c1em10819k.