Sensors and Materials, Vol. 32, No. 11 (2020) 3879–3892
MYU Tokyo

S & M 2382

Comprehensive Review on Application of Machine Learning Algorithms for Water Quality Parameter Estimation Using Remote Sensing Data

Nimisha Wagle,[1] Tri Dev Acharya,[2,3,4] and Dong Ha Lee[2*]

[1]Survey Department, Government of Nepal, Minbhawan, Kathmandu 44600, Nepal
[2]Department of Civil Engineering, Kangwon National University,
1 Kangdaehak-gil, Chuncheon 24341, Republic of Korea
[3]School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture,
No. 15 Yongyuan Road, Daxing District, Beijing 102616, China
[4]Institute of Transportation Studies, University of California Davis, 1605 Tilia Street, Davis, California 95616, USA

(Received May 29, 2020; accepted October 1, 2020)

Keywords: remote sensing, water quality parameters, machine learning, estimation, review

Water is an integral aspect of the world necessary for living creatures to thrive. Owing to unplanned urbanization, rapid industrialization, and uncontrollable human intervention, water quality is gradually degrading. This affects not only marine animals but also humans. Thus, the quality of water should be examined regularly. Water quality parameters should be estimated to monitor water quality. In general, water quality parameters are measured by *in situ* measurements. Although these measurements are accurate, they are costly and do not provide real-time spatial and temporal changes in water quality. To overcome this limitation, water quality parameters can be estimated using machine learning (ML) along with remote sensing (RS) data. A combination of ML and RS data is a powerful approach for the routine assessment of spatial and temporal variations in water quality parameters. In this paper, some articles based on this approach are reviewed. By analyzing the literature, it was found that the integrated use of RS-based geospatial data with ML helps to produce an accurate result. Most of the authors used the regression algorithm in the estimation of the water quality parameters, with a support vector machine (SVM) regression intensively used. The artificial neural network (ANN) algorithm was the most used algorithm of ML in many of the studies. The researchers used multispectral images for their study. By applying ML to RS data, water quality monitoring systems are evolving into real-time artificial intelligence (AI)-enabled models that provide valuable recommendations and insights to support farmers to make decisions and take action in aquaculture.

1. Introduction

Within an aquatic ecosystem, water quality plays an important role in the health of living organisms. In recent years, increasing population and unplanned urbanization have degraded water quality, thus affecting the health of ecosystems.[(1)] If the degradation continues, it will

*Corresponding author: e-mail: geodesy@kangwon.ac.kr
https://doi.org/10.18494/SAM.2020.2953

ISSN 0914-4935
https://myukk.org/

disturb the aquatic ecosystem and even cause the extinction of aquatic organisms, having a great impact on all living organisms including terrestrial ones. Therefore, water quality should be monitored regularly.[2,3]

Many physical, biological, and chemical parameters determine the water quality.[4] Traditionally, *in situ* measurement has mainly been used to estimate and monitor water quality parameters, where water samples are collected and tested in a laboratory. This technique may provide accurate values but is usually uneconomical, time-consuming, and unable to show real-time and spatial changes in water quality.[5]

Over time, there has been a shift from traditional *in situ* measurement to remote sensing (RS) techniques.[6–9] RS technology uses spaceborne or airborne sensors to measure the amount of radiation at various wavelengths reflected from the water's surface and extract information from it.[10] The reflections can be used directly or indirectly to determine different water quality parameters. The spectral characteristics of water and pollutants, which are functions of the hydrological, biological, and chemical characteristics of water, are essential factors in the monitoring and assessment of water quality.[11] The advantages provided by this technique are numerous, the most substantial one being near-real-time water quality mapping over a large spatial extent (e.g., a whole lake) without requiring a time-consuming and expensive field survey for sampling.[12] However, mapping over a large extent comes with a large amount of data. Large-scale RS imagery is difficult to manage and analyze using traditional statistical techniques. Thus, nowadays there is a move towards new technologies such as the incorporation of machine learning (ML) algorithms in geospatial databases. ML has emerged together with big-data technologies and high-performance computing to create new opportunities to unravel, quantify, and understand data-intensive processes for aquatic operational environments.[13] As the best solution, a combination of ML and satellite RS data is a powerful approach for the routine assessment of spatial and temporal variations in water quality parameters and may offer a suitable method to integrate water quality data collected from traditional *in situ* measurements.[8] ML algorithms help to estimate water quality parameters in less time and to provide a real-time measurement.[14] ML algorithms in cooperation with RS imagery reduce the human effort of analyzing big data and are highly cost-effective while producing very accurate results.[15–18]

Considering the literature gap, we present a comprehensive review of the application of ML algorithms for water quality parameter estimation using RS imagery. Sections 2–4 introduce water quality parameters, ML, and RS, respectively. Section 5 reviews the application of ML to water quality parameter estimation using RS imagery. Finally, Sect. 6 discusses the trend in water quality estimation with reference to Sect. 5 and ways forward for near-real-time estimation methods. The presentation of the learning models and algorithms in ML and the water quality parameters are limited to those that have been implemented in the works presented in this review. Figure 1 shows a generalized workflow of water quality parameter estimation using ML algorithms and RS data.

2. Water Quality Parameters

Water quality is measured using different water quality parameters. The commonly studied water quality parameters are given below.

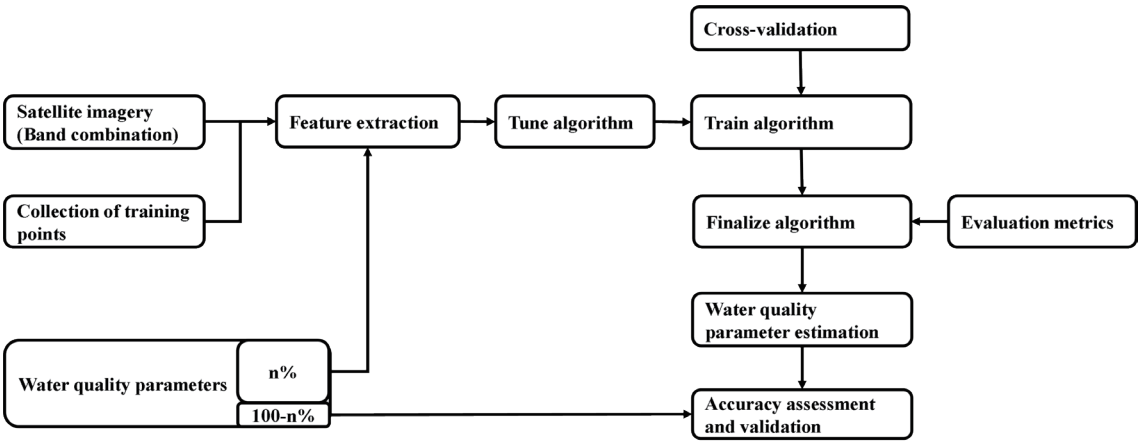

Fig. 1. Generalized workflow of water quality parameter estimation using ML algorithms and RS data.

2.1 Chlorophyll-a (Chl-a)

Chl-a is one of the major indicators of water quality. Eutrophication phenomena that drive algal blooms are related to Chl-a.[6] Eutrophication is the enrichment of water with nutrients. Excessive nutrients in water may harm the living ecosystem of the aquatic region.[19] Thus, Chl-a in aquatic regions should be monitored. Chl-a reflects green wavelengths, so it has high surface reflectance for green wavelengths.

2.2 Chlorides

Chlorides are salt compounds resulting from the combination of gas chlorine and metals. Excessive chlorides are very toxic to the aquatic ecosystem.[20] Therefore, the water used in the fishing industry or processed for any use has a recommended maximum chloride level. Chlorides can contaminate freshwater streams and lakes. Fish and aquatic communities cannot survive in water with high levels of chlorides. Higher chloride levels can affect the health of food sources and pose a risk to the survival, growth, and/or reproduction of aquatic living beings.

2.3 Dissolved oxygen (DO)

DO refers to the level of free, non-compound oxygen present in water or other liquids.[21] It is an important parameter in assessing water quality because it affects the organisms living within a body of water. In limnology (the study of lakes), DO is an essential factor second only to the water itself. A DO level that is too high or too low both can have harmful effects on the aquatic animals.

2.4 pH and total alkalinity

pH is also one of the indicators of water quality. Water with a low pH can damage pond liners and harm aquatic animals and humans. On the other hand, high-pH water can cause scale formation, metal stains, and cloudy water, and reduce the efficiency of chlorine in lakes and rivers. Similarly, total alkalinity is a measurement of the concentration of all alkaline substances dissolved in water.[22] These alkaline substances are primarily carbonates, bicarbonates, and hydroxides, along with a few others. They buffer the pH in water by neutralizing acids. In other words, total alkalinity is a measure of the water's ability to resist changes in pH.

2.5 Temperature

Temperature is also one of the factors determining water quality. Temperature regulates biological, physical, and chemical processes in water. Water with very low and very high temperatures is not suitable for aquatic animals. Water temperature affects other water quality parameters such as DO and solubility. Elevated temperatures and, more importantly, steep temperature gradients, can have direct harmful effects on fish.[23] It is also very important to analyze the temporal variations due to seasonal changes.

2.6 Total phosphorus

Phosphorus is an essential nutrient of plants, animals, and humans. In water, it exists primarily as orthophosphate (PO_4^{3-}) or inorganic compounds.[24] Total phosphorus is defined as the total amount of all phosphorus compounds that exist in various forms. An increased phosphorus concentration leads to the eutrophication of the aquatic environment, causing oxygen deficiency with deadly consequences for fish and other aquatic organisms. This makes it necessary to monitor this parameter. Phosphorus can enter water via wastewater discharge or the drainage of agricultural areas. Also, detergents, such as those used in dishwashers, often contain phosphorus. Their increased usage and disposal have led to increased phosphorus concentrations in wastewater. However, the increasing number of wastewater treatment plants that can remove phosphorus is helping to reduce the pollution that occurs from wastewater discharge.

2.7 Turbidity and total suspended solids

Turbidity is a measure of the ability of light to pass through water, i.e., its murkiness. Suspended solids in water cause the absorption or scattering of light rather than its transmission.[25] Turbidity is measured in nephelometric turbidity units (NTU)[26] and gives an estimate of the number of suspended solids in water. Suspended solids usually enter water as a result of soil erosion from disturbed land or the inflow of effluent from sewage plants or industry.[27] Suspended solids also occur naturally in water from bank and channel erosion; however, this

process has been accelerated by human use of waterways. Suspended residue can also choke sea plants as they settle out in low streams, and clog mouthparts and gills of fish and amphibian macroinvertebrates. In addition to suspended particles, turbidity measurements also consider the algae and plankton present in water. Pollutants such as nutrients and pesticides may bind with suspended solids and settle in bottom sediments, where they may become concentrated. High turbidity affects submerged plants by preventing sufficient light from reaching them for photosynthesis.[4] High turbidity can also significantly increase the water temperature, which needs to remain fairly constant for aquatic fauna to survive.[28] Although high turbidity is often a sign of poor water quality and land management, crystal-clear water does not always guarantee healthy water. Extremely lucid water can indicate very acidic conditions or high levels of salinity, so lucid water is not good for aquatic animals.

3. ML

ML is the science of getting computers to work without being explicitly programmed. In ML, the model is trained automatically using various data, i.e., features and labels, which are later required to obtain new sets of data.[29] The choice of which features to use ("feature learning") for characterizing a data point is very important for the success of the overall ML method. Although features have to be in such forms that they can be computed easily, they still need to contain a sufficient amount of information about the ultimate quantity of interest (the label).[29] The ML model's accuracy is increased by using the right number of parameters and hyperparameters. To calculate the performance of ML models and calculations, different measurable and scientific models are utilized. After the completion of the learning procedure, the prepared model can be utilized to characterize, anticipate, or cluster new models (testing information) using the experience acquired during the preparation procedure, in which the prediction is improved with understanding over the long run. ML has a close connection with statistics (especially nonparametric and computational statistics) and theoretical computer science.

ML tasks are typically classified into different broad categories depending on the learning type (supervised or unsupervised), learning model (classification, regression, clustering, or dimensionality reduction), and the algorithm employed to implement the selected task.[30]

3.1 Learning models

3.1.1 Regression

Regression is a supervised learning model, which aims to predict an output that varies according to the known input variables.[31] Regression algorithms predict the output values based on input features from the data given to the system. The methodology is the algorithm that builds a model on the features of training data and uses this model to predict the value for new data. Most algorithms used in learning models include linear and logistic regressions as well as stepwise regression.[32] Also, more complex regression algorithms, such as ordinary

least-squares regression,[33] multivariate adaptive regression splines, Bayesian regression, nonparametric regression, multiple linear regression, cubist regression, and locally estimated scatterplot smoothing, have been developed.

3.1.2 Classification

Classification is a type of supervised learning. It specifies the class to which data elements belong to and is best used when the output has limited and distinct values. It also predicts a class for an input variable.[34] Classification categorizes a set of data into classes. Its main goal is to classify the data into categorical class labels. The most common classification problems include gaze estimation, text classification, speech recognition, face detection, handwriting recognition, and document classification. Binary and multiclass classification problems exist, and there are many ML algorithms for classification in ML. The algorithms mentioned in this review paper are discussed below.

3.2 ML algorithms

3.2.1 Artificial neural networks (ANNs)

ANNs are the subset of ML that comprises traditional and deep neural networks (NNs). They are computing systems inspired by the biological NNs that constitute animal and human brains. Such systems “learn” to perform tasks by considering examples, generally without being programmed with task-specific rules.[35] The human brain comprises billions of neurons for processing the data obtained from different sensory organs. Likewise, ANN is an improved model of the structure of a natural neural system consisting of interconnected units with a particular topology that automatically trains itself with various sets of training data.[36]

Deep NN or deep learning (DL) uses hidden and deep ANN layers to progressively extract higher-level features from the raw input. One of the fundamental features of DL is that, at times, the feature extraction is performed by the model itself.[37] DL is a variation of ML that is concerned with an unlimited number of levels of limited size, which permits practical application and effect optimization with higher level features from raw input.. In DL, the layers are also permitted to be heterogeneous models for efficiency, trainability, and understandability, from where the "structured" part is obtained. Deep NN is essentially ANN with numerous concealed layers between the information and yield layers and can be either managed, mostly regulated, or even stand-alone. A typical DL model is a convolutional NN, where features are obtained by performing convolutions in images.[38] Other common DL models incorporate profound Boltzmann machines, profound conviction systems, and autoencoders.

3.2.2 Decision tree (DT)

A DT is a tree-like depiction of a decision and its every possible consequence or potential outcome after making that decision.[39] It is one way to display an algorithm that only

contains conditional control statements. Each inner hub of the tree structure represents an alternate pairwise examination on a choice of feature, although each branch is the result of this correlation. Leaf hubs provide an official choice or prediction after following the path from the root to the leaf (communicated as an ordering rule). Currently, the most well-known learning calculations are characterization and relapse trees, the chi-square programmed cooperation finder, and the iterative dichotomiser. DT is used for both classification and regression. Recursive partitioning (REPTree) is a type of binary tree utilized for grouping or regression assignments.[40] It creates a DT that correctly classifies members by splitting it into subpopulations based on several dichotomous independent variables. It is easy to understand and attempt to limit the utilization of all given datasets.[41]

3.2.3 Support vector machines (SVMs)

An SVM is a supervised ML model that uses classification algorithms for two-group classification problems.[42] After giving an SVM model sets of labelled training data for either of two categories, it can categorize new examples. It is intrinsically a binary classifier that constructs a linear separating hyperplane to classify data instances.[43] The classification abilities of SVMs can be significantly improved by changing the first component space into an element space of a higher measurement by utilizing the "kernel trick".[44] SVMs have been utilized for order, relapse, and bunching. SVMs manage overfitting issues, which appear in high-dimensional spaces, making them engage in different applications. Most utilized SVM calculations incorporate the help vector relapse, least-squares bolster vector machine, or progressive projection calculation bolster vector machine.[45] SVM regression (SVR) is commonly used in the water quality parameter estimation.

3.2.4 Gradient boosting algorithms

(a) Gradient boosting machine (GBM)

A GBM is a boosting algorithm utilized when a large amount of information is required to be predicted with high accuracy. Boosting is a type of learning algorithm that consolidates the predictions of a few base estimators to improve accuracy.[46] It consolidates different weak or normal indicators to a solid indicator. The guiding heuristic is that good predictive results can be obtained through increasingly refined approximations.[47]

(b) XGBoost

XGBoost is an advanced optimized distributed gradient boosting library designed to be productive, adaptable, and convenient.[48] It executes artificial intelligence (AI) calculations under the gradient boosting system. XGBoost dominates structured or tabular datasets on classification and regression predictive modeling problems. It resolves numerous issues in information science quickly and accurately.

(c) LightGBM

LightGBM is a gradient boosting framework that uses tree-based learning algorithms. The system provides quick and high-performance gradient boosting dependent on the choice of tree calculations and is utilized for positioning, arrangement, and numerous other AI assignments. It was created as part of the Distributed Machine Learning Toolkit Project of Microsoft.[49]

4. RS

RS is defined as sensing the information of the Earth using satellite or airborne sensors.[10] It is a complete process that captures Earth's surface data using electromagnetic energy and processes and extracts information for the geographic information system (GIS). RS can be categorized into ground-borne, airborne, and spaceborne, and, depending on the energy source, it can be considered as active or passive. The process of obtaining information from satellite images usually requires three steps: preprocessing, image enhancement, and image classification.[50] With advances in space science and the expanding utilization of computer applications and processing control in recent decades, RS technologies make it possible to analyze and study land and water bodies in large areas.[51] The collected remotely sensed data occurs in digital form and is therefore easily readable in computer processing. There are various advantages of the use of RS imagery to estimate water quality parameters over the use of only *in situ* measurements:

a. near-continuous spatial coverage of satellite data over a complete geographic area of a water body,
b. capable of assessing water quality in remote areas,
c. availability of satellite data in all seasons, and
d. efficient analysis of satellite data.

5. Application in Water Quality Parameter Estimation

With advances in technology, many researchers have shifted from traditional water quality parameter estimation techniques to new technologies such as ML and RS. SVR combining *in situ* data and surface reflectance is rapidly replacing linear regression.[18,52–56] This technique provides not only accuracy but also robustness when there are few sample points.[52] Wang *et al.* combined an ML algorithm, the water quality index (WQI), and RS spectral indices (difference, ratio, and normalized difference indices) through fractional derivative methods to establish a model for estimating and assessing the WQI,[57] called the particle swarm optimization (PSO)–SVR model. This model showed high performance with a coefficient of determination R^2 of 0.92, a root mean square error (RMSE) of 58.4, and a slope of the line of best fit of 0.97. They improved the accuracy of the obtained model by using new hyperspectral indices and PSO–SVR.

Kim *et al.* used three ML approaches, random forest (RF), CR, and SVR, to estimate two major water quality indicators, Chl-a and suspended particulate matter (SPM) concentrations, in coastal environments on the west coast of South Korea using the Geostationary Ocean Color

Imager (GOCI) satellite data.[58] They showed that SVR was better than the other techniques. When GOCI-derived radiance data were used, the ratios of band 2 to band 4 and band 6 to band 5 were the most influential input variables in predicting Chl-a and SPM concentrations, respectively. Hafeez *et al*. compared the reflectance data of Landsat 5, 7, and 8 imagery with *in situ* measurement data to evaluate the performance of various ML algorithms.[53] They estimated Chl-a, total suspended solids, and turbidity using various ML algorithms such as ANN, SVR, and CR. They obtained the highest accuracy with the ANN, with 91% accuracy for Chl-a, 92% accuracy for SS, and 85% accuracy for turbidity. It was concluded from their work that NN-based ML techniques provide higher accuracy than other techniques and that ML with satellite imagery has a high potential for future studies in water quality monitoring. Camps-Valls *et al*. evaluated the performance of a relevance vector machine (RVM) for the estimation of Chl-a from RS data.[52] The RVM was used to alleviate the deficiencies of SVR. The RVM was evaluated in terms of the accuracy and bias of the estimations, the sparseness of the solutions, robustness to a low number of training samples, and computational burden. Their study suggested that the RVM produced better results than ANN and SVR. Although the RVM produced a highly accurate result, it was more computationally demanding than SVMs. They suggested that, owing to high levels of uncertainty in both satellite-derived data and *in situ* measurements, robust and stable nonlinear regression models that provide inverse models are desirable. These models can be obtained from an RVM. Maier and Keller focused on the trade-off between the spatial and spectral resolutions of six simulated satellite-based data sets when estimating the Chl-a concentration with supervised ML models.[59] Arias-Rodriguez *et al*. used ML regression and all lifespan MERIS satellite data to estimate water quality parameters.[60] In their study, AI approaches with different complexities were investigated, and the ideal model for SDD and turbidity was resolved. Cross-approval showed that the satellite-based evaluations were consistent with the *in situ* estimations for both SDD and turbidity, with R^2 values of 0.81 to 0.86, an RMSE of 0.15 m, and 0.95 NTU. Chebud *et al*. developed an NN model in which Landsat data was used as a proxy to quantify water quality parameters, namely, Chl-a, turbidity, and phosphorus, before and after ecosystem restoration and during the wet and dry seasons.[61] The NN model was highly correlated with the data with $R^2 > 0.95$. The RMSE values for phosphorus, turbidity, and Chl-a were below 0.03 mg L^{-1}, 0.5 NTU, and 0.17 mg m^{-3}, respectively, in the NN training and validation phases. They determined the usefulness of the NN model for estimating the water quality parameters in a complex ecosystem. The developed NN model reduced the uncertainty resulting from the exclusion of any of the bands and captured both the linear and nonlinear complex relationships. González Vilas *et al*. also developed algorithms based on the NN technique and retrieved the Chl-a concentration in optically complex waters using MERIS data in the Galician Rias region of Spain.[62] They showed that the combination of *in situ* data and the NN algorithm improved the retrieval of Chl-a in water and could be used to obtain more accurate Chl-a maps. Blix and Eltoft presented the concept of an automatic model selection algorithm (AMSA) to find the best model for determining water quality parameters.[63] Their AMSA was designed to estimate oceanic Chl-a for global and optically complex waters by using four ML feature ranking methods and three ML regression models. This was carried out by using various regression

algorithms to retrieve water quality parameters from remotely sensed multispectral data for the given sensor and environment. Wang *et al*. adopted ANNs in RS imagery to improve the monitoring capability of water quality in a reservoir.[64] In their study, the ANN topology retrieved the remotely sensed data to estimate the water quality, with a correlation coefficient of 0.815 at the testing phase. Canziani *et al*. used Landsat bands and an ANN algorithm to determine Chl-a and the turbidity of different shallow Pampean Lakes.[16] The integration of the ANN algorithm and RS data made it possible to retrieve information on shallow lake systems at broad spatial and temporal scales. The result obtained from their study was statistically significant. Liu *et al*. stated that a linear model does not produce a good result for inland and shallow lakes, so nonparametric statistical techniques such as NN analysis should be utilized for water quality parameter estimation.[65] Pu *et al*. used a CNN with a hierarchical structure to determine water quality levels using Landsat-8 imagery.[66] They used CNN to mitigate the problem of estimating water quality parameters, which occurs because of the weak optical characteristics of water and the lack of explicit correlation between RS imagery bands and parameters.

Moser and Serpico used SVMs to calculate the sea surface temperature.[18] Using satellite data and corresponding *in situ* measurements, they found an approximate relation between them, which was subsequently used to estimate unknown surface temperatures from additional satellite data. Even though the proposed technique was experimentally tested in the context of surface temperature estimation, it is not application-specific. Further validation on different regression problems (e.g., estimation of other bio/geophysical parameters of the Earth's surface) will be required to evaluate the effectiveness of the method.

The ANN and SVR are both convenient for nonlinear modeling and produce a better result for water quality parameter estimation than other models. However, both methods need many paired samples (with the inputs and corresponding outputs both known) to construct a reliable and accurate model. In most cases, there are not enough paired samples for modeling since abundant *in situ* measurements are too costly. Wang *et al*. used a new method of semi-supervised SVR with a satellite to deal with the problem of insufficient paired samples and model accuracy.[15] Nascimento Silva and Panella used RS imagery and ANN to determine algal blooms by measuring Chl-a from space.[67] They described empirical algorithms, which incorporate information from the multispectral instrument of the Sentinel-2 satellite, and the obtained result was found to be statistically accurate. Pahlevan *et al*. introduced a new ML model, a mixture density network (MDN), for estimating Chl-a in water using the Sentinel-2 multispectral instrument.[68] It markedly outperformed existing algorithms when applied across different bio-optical regimes in inland and coastal waters. The MDN is a class of NNs, which helps to overcome the non-unique characteristic of the solution to the inverse problem of retrieving Chl-a using likelihoods generated in the training and validation steps.

Jeihouni *et al*. used decision-tree-based data mining to identify high-quality groundwater zones for water supply management.[69] They used different DT methods such as ordinary decision tree (ODT), RF, random tree (RT), chi-square automatic interaction detector (CHAID), and iterative dichotomiser 3 (ID3) to extract key relevant variables affecting water quality (electrical conductivity, pH, hardness, and chloride) in a GIS platform. The RF showed the highest

performance (accuracy of 97.10%) among the methods. Cao *et al*. employed an ML approach called an extreme gradient boosting tree (BST) to develop an algorithm for Chl-a estimation from OLI in turbid lakes.[47] The BST model performed well on a subset of data (N = 102, R^2 = 0.79, root mean squared difference = 7.1 μg L^{-1}, mean absolute percentage error = 24%, mean absolute error = 1.4, and bias = 0.9) and had better Chl-a retrievals than several band-ratio algorithms and the RF approach.

6. Discussion and Conclusion

Most of the studies reviewed in this paper were carried out to evaluate Chl-a using an SVM, whereas very few studies evaluated other water quality parameters. Most of the studies implemented ANN algorithms of ML for water quality parameter estimation. Comparative studies using multispectral RS imagery employed SVR and an ANN as state-of-the-art algorithms for benchmarks. In general, the empirical relationship between the *in situ* data and the surface reflectance has been established through ML-based regression methods. In these reviewed studies, the ANN algorithm had the highest accuracy among the methods. A decision-tree-based method and CNNs were also used by some authors to determine water quality. In those studies, images from satellites such as MERIS, GOCI, and Landsat were used.

The integrated use of ML and RS in water quality parameter estimation is being fostered nowadays. Their integrated use helps to produce a statistically accurate result as well as gives the spatial and temporal water quality in real time. To further improve the results of water quality parameter evaluation, hyperspectral images can be used for the data analysis. Several studies have used hyperspectral images along with ML technologies to retrieve results.[17,70–72] Hyperspectral cameras have a high spectral resolution, enabling them to evaluate water quality parameters when covering the wavelength range from 450 to 950 nm. In general, a hyperspectral camera records the surface reflectance of the water components. Hyperspectral cameras help to see the unseen through the narrow bands.[71] The use of an unmanned aerial vehicle (UAV) as a platform also helps to increase the accuracy as UAV images are captured from a small height and have a high spatial resolution. However, very few works have yet been carried out using UAVs.[73–75] Further adoption of these RS technologies is necessary for these approaches.

The real-time monitoring of water quality is essential in this era of rapid industrialization. Therefore, the concept of smart water quality monitoring should be studied. Geetha and Gouthami presented a low-cost, low-complexity smart water quality monitoring system using a controller with an built-in Wi-Fi module to monitor parameters such as pH, turbidity, and conductivity, enabling the real-time monitoring of water quality.[76] A few other related works have been reported.[77,78] Studies should also be carried out on the use of hyperspectral images to find the real-time status of water quality. Recently, there has been skyrocketing growth in the study and experiments on water quality estimation using ML and RS techniques. ML, being a hot and trending topic for studies, has become the first choice for most researchers.

By incorporating ML with RS data, we can carry out ongoing and timely examinations of water quality with an AI-empowered framework with the ultimate aim of advancing

the fisheries industry. For this purpose, it is expected that the utilization of integrated RS technology with ML algorithms will become increasingly widespread in the future owing to the availability of incorporated and applicable tools. The combined technology will provide valuable recommendations and insights to support decision making and implementation in aquaculture farming.

Acknowledgments

This work was supported by a National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. 2018R1A2B6009363).

References

1 S. Oliver, J. Corburn, and H. Ribeiro: Int. J. Environ. Res. Public Health **16** (2019) 40. https://doi.org/10.3390/ijerph16010040
2 T. D. Acharya, A. Subedi, and D. H. Lee: Sensors **19** (2019) 2769. https://doi.org/10.3390/s19122769
3 T. D. Acharya, A. Subedi, and D. H. Lee: Sensors **18** (2018) 2580. https://doi.org/10.3390/s18082580
4 M. H. Gholizadeh, A. M. Melesse, and L. Reddi: Sensors **16** (2016) 1298. https://doi.org/10.3390/s16081298
5 P. A. Brivio, C. Giardino, and E. Zilioli: Sci. Total Environ. **268** (2001) 3. https://doi.org/10.1016/S0048-9697(00)00693-8
6 C. Zhang and M. Han: E-Proc. 36th IAHR World Congr. (2015) 6.
7 Y. Zhang, J. Pulliainen, S. Koponen, and M. Hallikainen: Boreal Environ. Res. **8** (2003) 251.
8 N. Wagle, R. Pote, R. Shahi, S. Lamsal, S. Thapa, and T. D. Acharya: ISPRS Ann. Photogramm. Remote Sens. Spat. Inf. Sci. (2019) 127. https://doi.org/10.5194/isprs-annals-iv-5-w2-127-2019
9 X. Guan, J. Li, and W. G. Booty: Water Resour. Manag. **25** (2011) 2015. https://doi.org/10.1007/s11269-011-9792-3
10 T. M. Lillesand and R. W. Kiefer: Remote Sens. Image Interpret. 3rd ed. (1994) 763
11 E. Seyhan and A. Dekker: Hydrobiol. Bull. **20** (1986) 41. https://doi.org/10.1007/BF02291149
12 J. R. Jensen: Remote sensing of vegetation: Remote Sensing of the Environment: An Earth Resource Perspective (Pearson Prentice Hall, 2006) 2nd ed.
13 A. N. Ahmed, F. B. Othman, H. A. Afan, R. K. Ibrahim, C. M. Fai, M. S. Hossain, M. Ehteram, and A. Elshafie: J. Hydrol. (2019). https://doi.org/10.1016/j.jhydrol.2019.124084
14 U. Ahmed, R. Mumtaz, H. Anwar, S. Mumtaz, and A. M. Qamar: Water Sci. Technol. Water Supply **20** (2020) 28. https://doi.org/10.2166/ws.2019.144
15 X. Wang, L. Ma, and X. Wang: Int. Geosci. Remote Sens. Symp. (2010) 2757. https://doi.org/10.1109/IGARSS.2010.5653832
16 G. Canziani, R. Ferrati, C. Marinelli, and F. Dukatz: Math. Biosci. Eng. **5** (2008) 691. https://doi.org/10.3934/mbe.2008.5.691
17 S. Keller, P. M. Maier, F. M. Riese, S. Norra, A. Holbach, N. Börsig, A. Wilhelms, C. Moldaenke, A. Zaake, and S. Hinz: Int. J. Environ. Res. Public Health **15** (2018). https://doi.org/10.3390/ijerph15091881
18 G. Moser and S. B. Serpico: Kernel Methods Remote Sens. Data Anal. **47** (2009) 301. https://doi.org/10.1002/9780470748992.ch13
19 L. Han and K. J. Jordan: Int. J. Remote Sens. **26** (2005) 5245. https://doi.org/10.1080/01431160500219182
20 World Health Organization: Chlorides in drinking water, in: Guidel. Drink. Qual. (1996) pp. 520–521.
21 S. A. Rounds, F. D. Wilde, and G. F. Ritz: Dissolved Oxygen (2006) Chap. A6, Sect. 6.2.
22 A. G. Dickson: Deep Sea Res. Part A, Oceanogr. Res. Pap. **28** (1981) 609. https://doi.org/10.1016/0198-0149(81)90121-7
23 J. Sunil, S. Gunwant, and M. Y. P.: Int. J. Eng. Res. Technol. **2** (2013) 2516.
24 T. Reed-Andersen, S. R. Carpenter, and R. C. Lathrop: Ecosystems **3** (2000) 561. https://doi.org/10.1007/s100210000049
25 Md. Serajuddin, Md. A. Chowdhury, Md. M. Haque, and Md. E. Haque: Int. J. Eng. Trends Technol. **67** (2019) 83. https://doi.org/10.14445/22315381/ijett-v67i9p214

26 M. J. Brandt, K. M. Johnson, A. J. Elphinston, and D. D. Ratnayaka: Chapter 12 - Chemical Storage, Dosing and Control: Twort's Water Supply, M. J. Brandt, K. M. Johnson, A. J. Elphinston, and D. D. Ratnayaka, Eds. (7th Ed., Butterworth-Heinemann, Boston, 2017) pp. 513–552.
27 R. B. Susfalk, B. Fitzgerald, and A. M. Knust: Suspended solids in the Upper Carson River, Nevada (Nevada, 2008).
28 D. Jiang, J. Li, Y. Zhou, J. Wang, Y. Chen, and W. Xiao: Water Sci. Technol. Water Supply **18** (2018) 1173. https://doi.org/10.2166/ws.2017.189
29 M. Mohri, A. Rostamizadeh, and A. Talwalkar: Foundations of Machine Learning (The MIT Press, 2012).
30 P. Langley: Mach. Learn. **82** (2011) 275. https://doi.org/10.1007/s10994-011-5242-y
31 D. R. Cox: J. R. Stat. Soc. Ser. B **21** (1959) 238. https://doi.org/10.1111/j.2517-6161.1959.tb00334.x
32 M. Lewis-Beck, A. Bryman, and T. F. Liao: SAGE Encycl. Soc. Sci. Res. Methods (2012) 1. https://doi.org/10.4135/9781412950589.n974
33 L. Leng, T. Zhang, L. Kleinman, and W. Zhu: J. Phys. Conf. Ser. **78** (2007). https://doi.org/10.1088/1742-6596/78/1/012084
34 O. Pentakalos: Introduction to Machine Learning (MIT Press, 2019).
35 Y.-Y. Chen, Y.-H. Lin, C.-C. Kung, M.-H. Chung, and I.-H. Yen: Sensors **19** (2019) 2047. https://doi.org/10.3390/s19092047
36 A. Zell: Simulation of Neural Networks, Ed. Addison-Wesley (1994) 1st ed.
37 Y. Lecun, Y. Bengio, and G. Hinton: Nature **521** (2015) 436. https://doi.org/10.1038/nature14539
38 S. Albawi, T. A. Mohammed, and S. Al-Zawi: Proc. 2017 Int. Conf. Eng. Technol. ICET 2017 (2018) pp. 1–6.
39 J. R. Quinlan: Mach. Learn. **1** (1986) 81. https://doi.org/10.1007/bf00116251
40 C. Strobl, J. Malley, and G. Tutz: Psychol. Methods **14** (2009) 323. https://doi.org/10.1037/a0016973
41 L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone: Classification and Regression Trees (Chapman & Hall, 2017).
42 C. Cortes and V. Vapnik: Mach. Learn. **20** (1995) 273. https://doi.org/10.1007/bf00994018
43 N. Guenther and M. Schonlau: Stata J. **16** (2016) 49. https://doi.org/10.4018/978-1-60960-557-5.ch007
44 D. Decoste and B. Schölkopf: Mach. Learn. **46** (2002) 161. https://doi.org/10.1023/A:1012454411458
45 H. Han and X. Jiang: Cancer Inform. **13s1** (2014) CIN.S13875. https://doi.org/10.4137/cin.s13875
46 G. Biau, B. Cadre, and L. Rouvière: Mach. Learn. **108** (2019) 971. https://doi.org/10.1007/s10994-019-05787-1
47 Z. Cao, R. Ma, H. Duan, N. Pahlevan, J. Melack, M. Shen, and K. Xue: Remote Sens. Environ. **248** (2020) 111974. https://doi.org/https://doi.org/10.1016/j.rse.2020.111974
48 XGBoost Documentation—xgboost 1.2.0-SNAPSHOT documentation, https://xgboost.readthedocs.io/en/latest/
49 Features — LightGBM 2.3.2 documentation, https://lightgbm.readthedocs.io/en/latest/Features.html
50 R. A. Schowengerdt: The Nature of Remote Sensing, R.A.B.T.-R.S. Ed. (3rd Ed., Schowengerdt, Remote Sens., Academic Press, Burlington, 2007).
51 J. B. Campbell: Introduction to Remote Sensing (Guilford Press, 2002) 3rd ed.
52 G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. Vila-Francés, J. Amorós-López, and J. Calpe-Maravilla: Remote Sens. Environ. **105** (2006) 23. https://doi.org/10.1016/j.rse.2006.06.004
53 S. Hafeez, M. Wong, H. Ho, M. Nazeer, J. Nichol, S. Abbas, D. Tang, K. Lee, and L. Pun: Remote Sens. **11** (2019) 617. https://doi.org/10.3390/rs11060617
54 K. P. Singh, N. Basant, and S. Gupta: Anal. Chim. Acta **703** (2011) 152. https://doi.org/10.1016/j.aca.2011.07.027
55 G. Mountrakis, J. Im, and C. Ogole: ISPRS J. Photogramm. Remote Sens. **66** (2011) 247. https://doi.org/10.1016/j.isprsjprs.2010.11.001
56 N. Wagle, T. D. Acharya, and D. H. Lee: Estimating Chlorophyll-a and Dissolved Oxygen Based on Landsat 8 bands using Support Vector Machine and Recursive Partitioning Tree Regressions (2020) p. 6573.
57 X. Wang, F. Zhang, and J. Ding: Sci. Rep. **7** (2017) 1. https://doi.org/10.1038/s41598-017-12853-y
58 Y. H. Kim, J. Im, H. K. Ha, J. K. Choi, and S. Ha: GIScience Remote Sens. **51** (2014) 158. https://doi.org/10.1080/15481603.2014.900983
59 P. M. Maier and S. Keller: Work. Hyperspectral Image Signal Process. Evol. Remote Sens. 2019-Septe (2019). https://doi.org/10.1109/WHISPERS.2019.8921073
60 L. F. Arias-Rodriguez, Z. Duan, R. Sepúlveda, S. I. Martinez-Martinez, and M. Disse: Remote Sens. **12** (2020) 1586. https://doi.org/10.3390/rs12101586
61 Y. Chebud, G. M. Naja, R. G. Rivero, and A. M. Melesse: Water. Air. Soil Pollut. **223** (2012) 4875. https://doi.org/10.1007/s11270-012-1243-0
62 L. González Vilas, E. Spyrakos, and J. M. Torres Palenzuela: Remote Sens. Environ. **115** (2011) 524. https://doi.org/10.1016/j.rse.2010.09.021

63 K. Blix and T. Eltoft: Remote Sens. **10** (2018) 775. https://doi.org/10.3390/rs10050775
64 T. S. Wang, C. H. Tan, L. Chen, and Y. C. Tsai: Proc. 2nd Int. Symp. Intell. Inf. Technol. Appl. IITA 2008 **1** (2008) 540. https://doi.org/10.1109/IITA.2008.279
65 Y. Liu, A. Islam, and J. Gao: Prog. Phys. Geogr. **27** (2003) 24. https://doi.org/10.1191/0309133303pp357ra
66 F. Pu, C. Ding, Z. Chao, Y. Yu, and X. Xu: Remote Sens. **11** (2019) 1674. https://doi.org/10.3390/rs11141674
67 H. A. N. Silva and M. Panella: Prog. Electromagn. Res. Symp. 2018-Augus (2018) 458. https://doi.org/10.23919/PIERS.2018.8597731
68 N. Pahlevan, B. Smith, J. Schalles, C. Binding, Z. Cao, R. Ma, K. Alikas, K. Kangro, D. Gurlin, N. Hà, B. Matsushita, W. Moses, S. Greb, M. K. Lehmann, M. Ondrusek, N. Oppelt, and R. Stumpf: Remote Sens. Environ. **240** (2020) 111604. https://doi.org/https://doi.org/10.1016/j.rse.2019.111604
69 M. Jeihouni, A. Toomanian, and A. Mansourian: Water Resour. Manag. **34** (2020) 139. https://doi.org/10.1007/s11269-019-02447-w
70 M. Mbuh: Intech i (2016) 13. https://doi.org/http://dx.doi.org/10.5772/57353
71 N. A. Shafique, F. Fulk, B. C. Autrey, and J. Flotemersch: J. Geomatics **3** (2009) 9.
72 Y. Zhang, L. Wu, H. Ren, L. Deng, and P. Zhang: Remote Sens. **12** (2020) 1567. https://doi.org/10.3390/rs12101567
73 R. H. Becker, M. Sayers, D. Dehm, R. Shuchman, K. Quintero, K. Bosse, and R. Sawtell: J. Great Lakes Res. **45** (2019) 444. https://doi.org/10.1016/j.jglr.2019.03.006
74 C. Koparan, A. B. Koc, C. V. Privette, and C. B. Sawyer: Water **10** (2018) 264. https://doi.org/10.3390/w10030264
75 F. S. Y. Watanabe, E. Alcântara, T. W. P. Rodrigues, N. N. Imai, C. C. F. Barbosa, and L. H. da S. Rotta: Int. J. Environ. Res. Public Health **12** (2015) 10391. https://doi.org/10.3390/ijerph120910391
76 S. Geetha and S. Gouthami: Smart Water **2** (2016) 1. https://doi.org/10.1186/s40713-017-0005-y
77 K. Spandana and V. R. Seshagiri Rao: Int. J. Eng. Technol. **7** (2018) 259. https://doi.org/10.14419/ijet.v7i3.6.14985
78 N. Vijayakumar and R. Ramya: IEEE Int. Conf. Circuit, Power Comput. Technol. ICCPCT **2015** (2015). https://doi.org/10.1109/ICCPCT.2015.7159459

About the Authors

Nimisha Wagle received her B.E. degree in geomatics from Kathmandu University, Nepal, in 2017. She is a survey officer at the Survey Department, Government of Nepal. Her interests are related to geospatial data preparation, machine/deep learning and mapping for agriculture, land cover, and water quality. (nimisha.wagle@nepal.gov.np)

Tri Dev Acharya received his B.E. degree in geomatics from Kathmandu University, Nepal, in 2011 and his combined M.S. and Ph.D. degrees from Kangwon National University, Korea, in 2018. He is a postdoctoral researcher at Kangwon National University, Korea. His research interests are in the geospatial data preparation, modeling, and simulation of land cover, surface water, natural hazards, and transportation using various machine learning algorithms. (tridevacharya@kangwon.ac.kr)

Dong Ha Lee received his B.E., M.S., and Ph.D. degrees from Sungkyunkwan University, Korea, in 2000, 2003, and 2008, respectively. Since 2015, he has been an associate professor at Kangwon National University, Korea. His research interests are in geodesy, surveying, geospatial information, and natural hazard analysis. (geodesy@kangwon.ac.kr)