

Article

Remote Sensing Identification of Harmful Algae in Ulansuhai Lake with Machine Learning

Jianglong Cui, Xiaodie Zhang, Caili Du * and Guowen Li *

State Key Laboratory of Environmental Criteria and Risk Assessment, Chinese Research Academy of Environmental Sciences, Beijing 100012, China; cuijianglong1990@sina.com (J.C.); zhangxiaodie1206@163.com (X.Z.)

* Correspondence: aquaducl@163.com (C.D.); iliguowen@163.com (G.L.)

Abstract: Frequent algal blooms in lakes pose a serious threat to aquatic ecosystems. It is of great significance to quickly and accurately monitor the distribution of algae in lakes for the regulation of algal blooms. While remote sensing techniques and machine learning methods can be used in combination to identify algae and analyze their spatial and temporal distribution, these methods still face challenges in practical applications due to uncertainties in lake boundaries and imbalances between algae and non-algae. In order to overcome these difficulties, we studied the dynamic open water range of Ulansuhai Lake and used a non-equilibrium data processing method to identify its algae. We also performed a spatiotemporal analysis of the algal range over a long time series. The results show that (1) the spectral characteristics of Landsat 8 images are very suitable for algal identification based on remote sensing, especially in the random forest method, where the fourth band plays an important role. (2) Among various machine learning methods, the accuracy of the random forest method on the training set and validation set is more than 90%. This indicates that the random forest method is suitable for the long-term monitoring of algal blooms. This study provides scientific and technical support for the management of Ulansuhai Lake, which will be helpful in guiding future management and control work.

Keywords: metaphytic blooms; Ulansuhai Lake; remote sensing; machine learning



Academic Editor: Eric D. Stein

Received: 8 November 2024

Revised: 15 December 2024

Accepted: 26 December 2024

Published: 28 December 2024

Citation: Cui, J.; Zhang, X.; Du, C.; Li, G. Remote Sensing Identification of Harmful Algae in Ulansuhai Lake with Machine Learning. *Water* **2025**, *17*, 50. <https://doi.org/10.3390/w17010050>

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Lakes worldwide have faced serious environmental problems, such as water eutrophication and algal blooms, over the past several decades due to increases in human activities and their associated pollution as well as climate change. The excessive growth of algae has seriously affected the aquatic environment of lakes, resulting in a reduction in ecosystem service functions and serious threats to the ecological and environmental security of river basins [1–4]. Ulansuhai Lake is the largest freshwater lake in the Yellow River Basin and the largest natural wetland at the same latitude, and it plays an important role in maintaining the ecological balance of the basin, regulating the regional climate, and storing water for flood control [5]. However, in recent years, it has been plagued by an outbreak of “Huangtai algae,” which are mainly composed of three filamentous algae: *Spirogyra*, *Zygnema*, and *Mougeotia*. In nutrient-rich water bodies, these algae often grow in large quantities in the summer, die, decompose, and consume more oxygen, resulting in the death of other organisms from hypoxia, thereby destroying the water quality and directly harming the ecology of the Ulansuhai Lake water body [6–9]. Therefore, methods for quickly and accurately identifying Huangtai algae are required. Moreover, mastering the spatial and temporal changes in Huangtai algae is the initial prerequisite for effectively controlling and

preventing the outbreak of Huangtai algae, which are also an important basis for ensuring the ecological environment security of Ulansuhai Lake.

The drift route of algal outbreaks changes greatly; therefore, traditional monitoring methods generally conduct ship-based monitoring surveys along certain routes, which is not only time-consuming and arduous but also increases the difficulty of comprehensively and accurately understanding the time, scope, and drift routes of algae outbreaks. Satellite remote sensing technology, however, can be operated over a wide range, produces rapid results at a low cost, and increases the convenience of continuous observations. Temporal and spatial information on lake water environmental factors can be quickly obtained via remote sensing, which has thus become an important technical method for effectively monitoring algal growth and providing early warnings of bloom outbreaks [10–12].

Early image classification mainly relies on a large number of experts with professional field knowledge and practical experience to design image features, such as color, shape, texture, and spectral information. In general, the remote sensing index threshold method has been adopted to monitor floating algae, and it mainly utilizes the normalized vegetation index (NDVI), enhanced vegetation index (EVI), floating algae index (FAI), alternative floating algae index (AFAI), and red index (RI) [13]. With the development of remote sensing technology, the resolution of the obtained imagery has increased; however, these images present a large number of details, which increases the difficulty of fully expressing the target object via image features alone, and this issue is exacerbated for more complex images. Machine learning algorithms provide a number of feasible methods for classifying remote sensing images. Classification techniques for extracting aquatic vegetation information from remote sensing images are generally divided into two categories: classification algorithms and clustering algorithms.

Clustering algorithms divide the dataset into different groups (clusters) so that the data points in the same cluster have a high similarity, and there is no pre-defined category label in the clustering process. Dogan et al. (2009) used an unsupervised classification method to identify submerged vegetation in shallow lakes based on QuickBird images, and the classification accuracy of various submerged vegetation categories exceeded 70% [14]. Shi et al. (2013) carried out the remote estimation of chlorophyll-a in inland waters based on cluster classification [15]. Zhang et al. (2015) evaluated the chlorophyll in the turbid water of Lake Taihu by means of remote sensing and the K-means method [16]. Wei et al. (2020) proposed an improved unsupervised representation learning model (multilayer feature fusion) for the scene classification of remote sensing images, and it realized richer data expansion and better classification performance [17].

Classification algorithms are methods for formulating decision rules according to sample categories that have been confirmed a priori and identifying the sample data of other unknown categories according to the decision rules. In classification algorithms, training samples of the research area must be established according to prior information. Then, the computer can learn according to the training samples and finally identify unknown categories. Based on APEX images, Bolpagni et al. (2014) used the Mahalanobis distance supervised classification method to divide aquatic vegetation within lakes in Mantua into upstanding vegetation, submerged vegetation, and floating vegetation [18]. Hao et al. (2016) used historical yearly CDL data collected in Kansas, USA, to extract hypothetical samples that were then screened by an artificial antibody network (ABNet) method to obtain a “training sample” for current-season crop classification, and they achieved 90% accuracy [19]. Zhang et al. (2018) used Honghu Lake as their research object, applied classification and regression tree methods to extract the spectral characteristics and other characteristic variables of aquatic plants in Honghu Lake, and established a decision tree model for wetland information extraction in the study area [20]. Bareuther et al. (2020)

conducted a supervised maximum likelihood classification of 62 true-color satellite images of Bellandur and Varthur Lakes in Bangalore, a large city in South India, to distinguish large plants, algae, and free water surfaces and observe changes in the lake cover over many years [21]. Classification algorithms provide more data than clustering algorithms. Comparisons of algal recognition results in previous studies have revealed that the accuracy of algal recognition with classification algorithms is generally better than that with clustering algorithms.

Historically, analyses of long time series of remote sensing images have often ignored temporal variations in the water body boundary and performed algae identification based on a fixed water body boundary, which is contrary to reality. Although scholars have added the land category to the classification category, the probability of false identification increases as the number of categories increases. This study improves upon previous methods of classifying remote sensing images of water bodies in the following three aspects: First, compared with the fixed water body boundary in previous studies, this study first used the MNDWI index to separate land from water in remote sensing images, extracting a dynamic water body boundary. The identification of Huangtai algae considering the dynamic water body boundary can improve the precision of model training. Second, only a rough boundary of the Huangtai algae could be delineated via UAV real shooting monitoring images. Because of the previous setting of the fixed water body boundary, all delineated Huangtai algae were within the fixed water body boundary, which could easily cause a blurring of the Huangtai algae boundaries. Therefore, this study fully considered the dynamic water body boundary and distinguished Huangtai algae areas from non-Huangtai algae areas (water bodies) by superimposing Huangtai algae real shooting images. Compared with multi-level classification models, the construction of a two-level classification model effectively reduced the probability of incorrect identifications. Third, during the training of the Huangtai algae recognition model, the data in the training set for the Huangtai algae and non-Huangtai algae labels were unbalanced. Since most machine learning algorithms for classification are designed based on the assumption that the number of each class is equal, this imbalance will lead to a poor prediction performance of the model for a few classes. Therefore, in this study, the imbalance between the Huangtai and non-Huangtai algae data was fully considered during model training to improve the prediction accuracy of the model. Subsequently, this study compared a variety of machine learning classification models based on remote sensing images of Ulansuhai Lake and aerial monitoring data of Huangtai algae and selected the classification model with the highest accuracy using a 10-fold cross-verification method. After identifying the Huangtai algae, the area coverage degree of the Huangtai algae during the growth cycle was explored, and the spatial and temporal changes in the Huangtai algae were investigated. This study provides scientific and technological support and guidance for the control of algal blooms in Ulansuhai Lake.

2. Materials and Methods

2.1. Study Area

Ulansuhai Lake, located in Bayannur City, Inner Mongolia Autonomous Region, China, is a river-track lake formed by the diversion of the Yellow River and represents the largest lake wetland in the Yellow River Basin (Figure 1). Moreover, this type of multifunctional lake is rarely observed in arid grassland and desert areas worldwide. At present, Ulansuhai Lake has a stable area of 293 km² and performs important functions in terms of Yellow River water regulation, water purification, and flood prevention. Ulansuhai Lake is rich in biodiversity, with 282 species of birds and fish. Moreover, the lake hosts the most important production bases of 24 species of fish for the Yellow River Basin. Since the 1990s, the

natural water supply of Ulansuhai Lake has continuously decreased, resulting in a sharp reduction in the area of the lake. In addition, with a significant increase in the discharge of municipal sewage and industrial wastewater, its ecological function has severely degraded, the eutrophication of the lake water is serious, and Huangtai algae have proliferated in large quantities. Since then, frequent outbreaks of Huangtai algae have become a chronic issue affecting the ecological environment of Ulansuhai Lake.

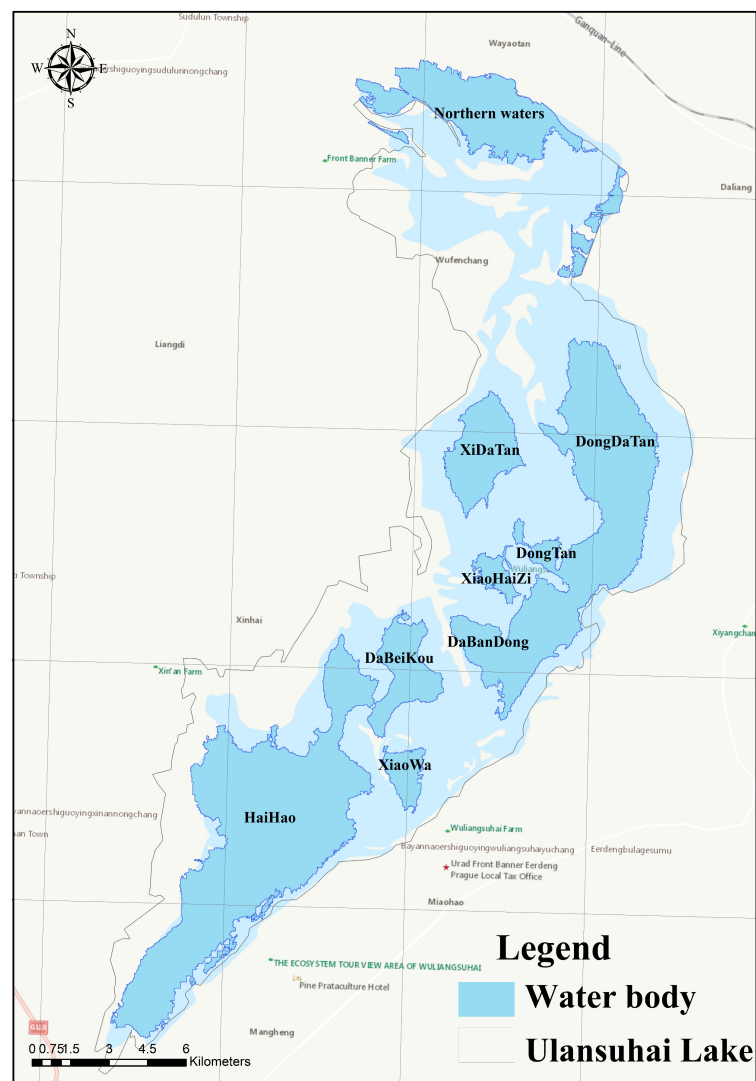


Figure 1. Location of the study area.

2.2. Data

2.2.1. Remote Sensing Image Data

Landsat 8 is the best satellite for detecting harmful algae in small- to medium-sized waterbodies [22]. The Landsat 8 remote sensing image data used in this study were obtained from the remote sensing data service system of the Institute of Aerospace Information Innovation of Chinese Academy of Sciences (<http://eds.ceode.ac.cn/nuds/businessdataquery> (accessed on 24 June 2024)). The Landsat-series satellites launched by NASA provide abundant and high-quality data resources for earth resource exploration, earth resource management, and land habitat monitoring. Landsat 8 was successfully launched in February 2013 and provides global coverage every 16 days. Landsat 8 carries two main payloads, Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS), both of which have significant improvements over previous Landsat 1–7 satellites. ETM + has eight bands: B1 is the blue-green band, and it has a certain ability to penetrate water and can be used to

detect soil and vegetation; B2 is the green band, and it can be used to detect vegetation types; B3 is the red band, and it is located in the chlorophyll absorption region and thus be used to distinguish bare soil, roads, and vegetation; B4 is in the near-infrared band, and it can be used to distinguish between water and vegetation because vegetation usually shows high reflectance in the near-infrared region while water presents very low reflectance; B5 is the mid-infrared band, and it can be used to monitor the water content of vegetation; B6 is the thermal infrared band, and it can be used to detect the thermal radiation of ground objects; B7 is the mid-infrared band, and it that can be used for rock and mineral classification; and B8 is the panchromatic band, and it can be used to improve the image resolution. According to the spatial location of Ulansuhai Lake, this study selected strip number 128, row number 32, longitude range 108.68–108.98, and latitude range 40.75–41.15 for Landsat 8. Based on the growth and extinction cycle of the Huangtai algae in Ulansuhai Lake, remote sensing images from May to September for the period from 2013 to 2020 were selected. Because cloud cover blocks the ground scene and changes the spectral and texture information of the image to a certain extent, 43 remote sensing images with a cloud cover of less than 10% were selected for this study. Considering the uniformity of the spectral resolution, bands 1–7 were selected as the characteristic variables.

2.2.2. Huangtai Algae Aerial Data

The sample collection HR image interpretation method is widely accepted [23]. Aerial surveillance data of Huangtai algae via UAV were collected on 17 June 2022 by the Urat National Reserve Administration in Bayannaoer City (Figure 2). Specifically, by projecting field and UAV-based samples onto HR images, we obtained prior knowledge of the Huangtai algae. By synthesizing the images captured by the UAV at low altitude, a complete JPEG photo of Ulansuhai Lake was obtained. On this basis, a vectorization operation was carried out to construct a Huangtai algae label for the model training data (Figure 3).

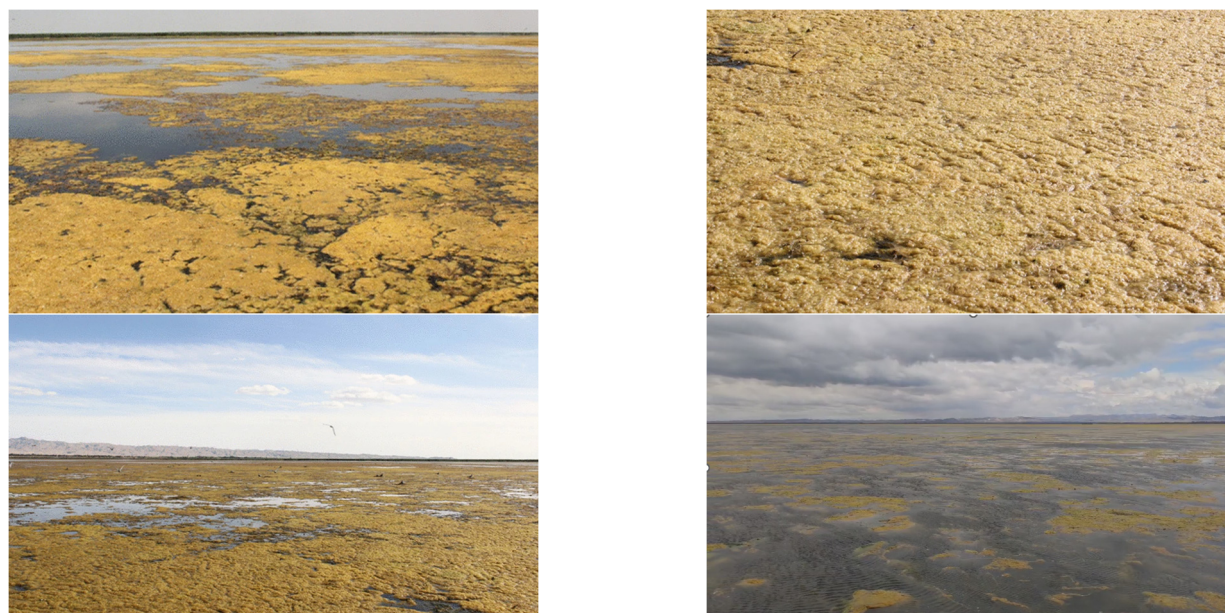


Figure 2. Huangtai algae field collection photos.

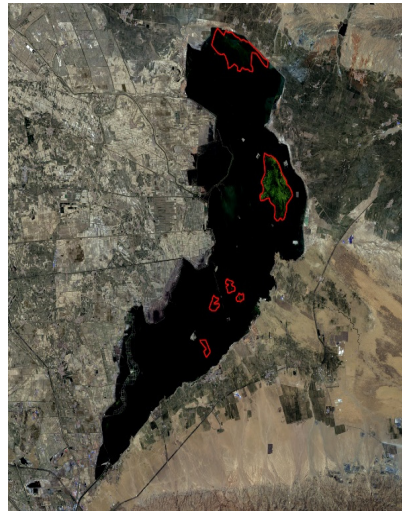


Figure 3. Huangtai algae label after visual interpretation.

2.3. Model

2.3.1. Workflow

The tasks in this study were carried out according to the following technical workflow (Figure 4). First, the UAV image of the Huangtai algae was vectorized by visual interpretation to obtain the Huangtai algae label. The satellite remote sensing images closest to the UAV aerial shooting time were selected for radiometric calibration, atmospheric correction, and water body extraction. The remote sensing band information of the water range was used as the model feature variable and combined with the Huangtai algae label of the UAV to form the model training dataset. To fully consider the imbalanced amount of sample data for the Huangtai algae and non-Huangtai algae labels, the model with the highest accuracy was selected from a variety of machine learning models and the remote sensing image was expanded based on other time periods. Finally, temporal and spatial variations in the growth and extinction periods of the Huangtai algae in Ulansuhai Lake were comprehensively analyzed.

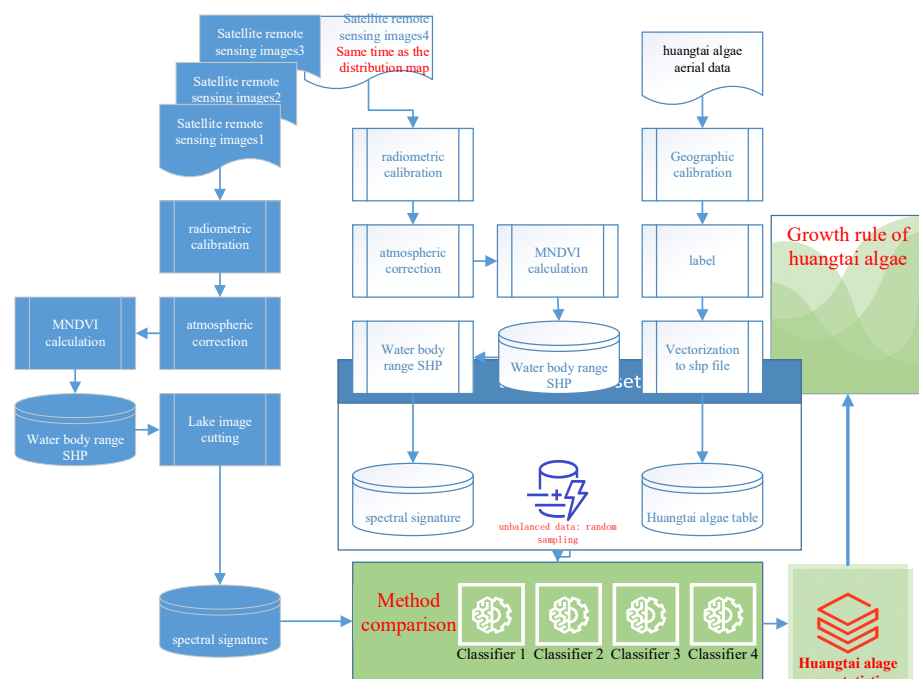


Figure 4. Workflow.

2.3.2. Preprocessing of Remote Sensing Data

As the Landsat 8 data product is at Level 1T, it is a geometric correction product that has been topographed relative to other Landsat data series; thus, these data can be used directly without geometric correction.

(1) Radiometric calibration

The *DN* value is the gray level of the pixel value in the image and represents the digital signal value recorded by the sensor. Because the gray level is relative, the *DN* values can only be compared within the same image, which does not meet the requirements of remote sensing for long time series, multiple regions, and multiple sources. Therefore, data for different simultaneous phases, regions, and sensors can only be analyzed by converting the *DN* values of the image into the actual electromagnetic radiation intensity (radiance) of each pixel. This transformation process is known as radiometric calibration.

This process is calculated as follows:

$$Y = DN \times F\_scales + F\_offsets \quad (1)$$

where *Y* is the reflectance value of the pixel, *DN* is the *DN* value of each original image element, *F_scales* denotes the linear coefficient of the reflection value of the original image, and *F_offsets* is the intercept of the reflectance value of the original image.

Calibration was performed using the Radiometric Calibration module provided by ENVI. The calibration type was set to Radiance, the image output interleave was set to BIL, and the scale factor was set to 0.1. The output storage sequence was BIL, and the unit radiation brightness conversion coefficient (scale factor) was set to 0.1.

(2) Atmospheric correction

All types of radiation energy used in remote sensing interact with the Earth's atmosphere (scattering or absorption), causing energy attenuation and changes in the spectral distribution. The attenuation effect of the atmosphere differs in remote sensing images because of differences in the wavelength of light, length of the atmospheric path, and imaging time. Thus, remote sensing signals are processed via atmospheric correction to remove this attenuation. Presently, the models and methods of atmospheric correction mainly include the image feature model, ground linear regression empirical model, and atmospheric radiation transmission.

In this study, the FLAASH model provided by ENVI was used for atmospheric correction. Based on the MODTRAN transmission theory, FLAASH is a highly accurate atmospheric radiation correction model. The radiometric calibration results and atmospheric correction parameters required by the model were input into the correction module, wherein the sensor height was 705 km and the ground elevation was set to 1.023 km. The atmospheric models were based on a season–latitude surface temperature model, and the selected models are shown in Table 1. Rural was selected for the aerosol model, and the spectral response function of the sensor in the multispectrum setting panel was selected. Aerosol retrieval was set to 2-Band (K-T). For Kaufman–Tanre aerosol retrieval, defeat values were assigned based on the retrieval condition defaults over land according to the retrieval standard (660:2100 nm).

Table 1. Atmospheric models selected according to season and latitude.

Latitude (° N)	Jan.	March	May	July	Sept.	Nov.
40	SAS	SAS	SAS	MLS	MLS	SAS

(3) MNDWI was used to separate water from land and extract water bodies

The most important step was to classify the water and land areas in each remote sensing image. Based on their remarkable spectral differences, the following two types of water indices are commonly used:

$$NDWI = (Green - NIR) / (Green + NIR) \quad (2)$$

$$MNDWI = (Green - MIR) / (Green + MIR) \quad (3)$$

where *Green* is the green band, *NIR* is the near-infrared band, and *MIR* is the mid-infrared band. The threshold for the water index was [0, 1]. For lakes with less algae coverage, the difference between the two water body indices was not obvious, resulting in an NDWI extraction error mainly located in the algal coverage area. MNDWI errors were mainly located in water areas near the shore. In lakes with large algae coverage, the accuracy of the NDWI is very low; however, the MNDWI can achieve a better water boundary extraction effect and thus is suitable for water boundary extraction in eutrophic lakes [24–26].

The MNDWI can be calculated directly using the Spectral Indices tool in ENVI. After calculation, outliers need to be processed using the formula $[-1 > b1 < 1]$, which is entered using the Band Math tool. In the toolbox, Classification→Decision Tree→Build New Decision Tree was selected for land and water separation [b1 gt 0.2].

In the Vector tool in ENVI, Classification→Post Classification→Classification was selected to convert the separated water body part into a base map in shp format, as shown on the left side of Figure 5. Then, the isolated water shp base map was used to cut the remote sensing image after atmospheric correction, and the Edit ENVI Header tool was used to remove the black area around the cut remote sensing image. The cutting process results are shown on the right of Figure 5.

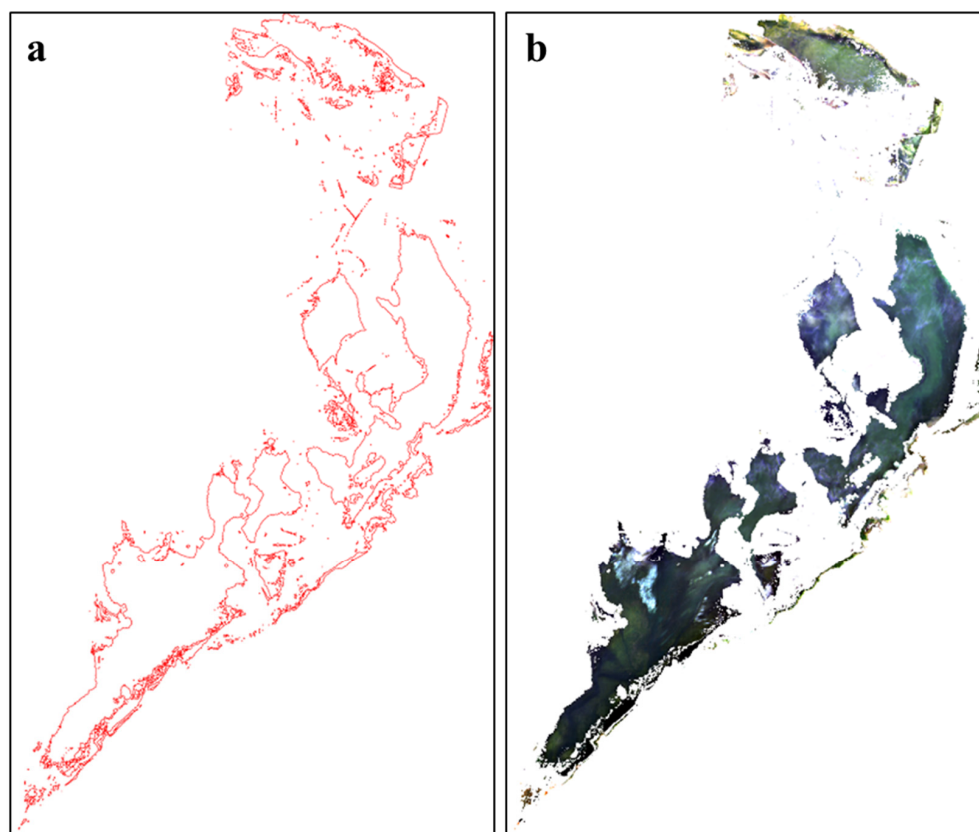


Figure 5. (a) Ulansuhai Lake waterbody and (b) remote sensing image after cutting.

(4) UAV monitoring image registration

Image registration is the process of matching and superimposing two or more images obtained at different times, with different sensors (imaging equipment), or under different conditions (e.g., weather, illumination, camera position, and angle). The Image Registration Workflow in ENVI was used to register two images with different geometric positions. This tool is an automatic, accurate, and fast image registration workflow that integrates complex parameter setting steps into a unified panel. Moreover, it can quickly and accurately realize automatic registration between images with little or no human intervention. Using this tool, the UAV image was geographically registered based on the nearest remote sensing image.

(5) Huangtai algae vectorization

Using the identification tool in ArcGIS 10.4.1, the geometric intersection of the input and identification elements was determined. Input elements (or a portion of the input elements) that overlap with identity elements acquire the attributes of these identity elements. By identifying the water and Huangtai algae shapefiles, a shapefile base map with both water and Huangtai algae categories was formed, as shown in Figure 6, where the highlighted area is the Huangtai algae area. In this study, the non-Huangtai algae area was treated as a class, which could reduce the identification error caused by an increase in the classes [27].

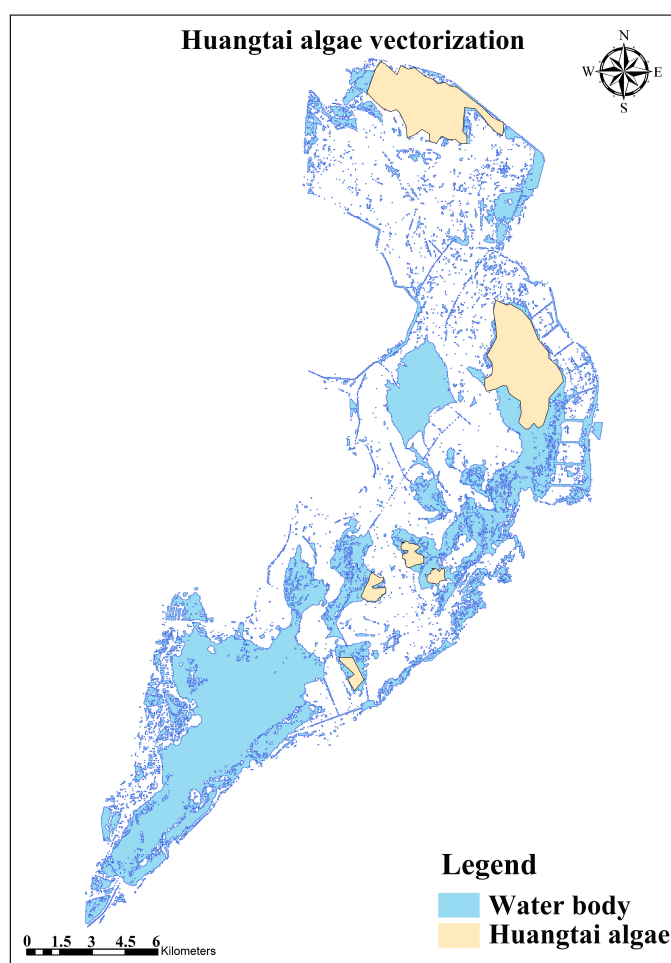


Figure 6. Huangtai algae vectorization results.

2.3.3. Research Methods

After the above analysis, a sample dataset for training the machine learning classifier was constructed. The data mainly included two parts: the remote sensing image of the

cut water body and the shapefile file after identifying the category of Huangtai algae. Considering that the accuracy of the two-classification model will improve more than that of the multi-classification model, the label variables were dichotomized into Huangtai algae and non-Huangtai algae (water bodies) during the classification model training. In addition, the sample dataset faced the problem of non-equilibrium. Because of the large difference between the amount of data in the Huangtai algae and non-Huangtai algae categories, the data belonging to the non-Huangtai algae category were highly accurate while those belonging to the Huangtai algae category were unacceptable. In this study, the sample distribution was balanced by changing the proportion of samples into categories of large numbers. After 10 iterations of undersampling the model training, the most suitable model was selected from a series of common machine learning classifiers according to their accuracy rate [28–30]: random forest (information partitioning), decision tree (information partitioning), KNN (distance partitioning), support vector machine (space partitioning), naive Bayes (conditional probability formula), and linear discriminant analysis (LDA6). The extrapolation was performed according to the optimal classifier.

First, the series models are briefly introduced and the parameters are set.

(1) Random forest

Random forest is an integrated classification algorithm based on decision trees proposed by the American scholar Leo Breiman. In the random forest model, each decision tree has a random vector composed of trained parameters and a bagging algorithm is used to build each decision tree into an independent set of training samples. Subsequently, certain features are selected from the feature set to construct the decision tree.

The random forest algorithm is a parameterless classification and regression algorithm that requires no prior information and is easy to operate. A classifier based on a decision tree can ensure high accuracy. Based on bagging ensemble learning, this method is tolerant to noise and outliers and can process high-dimensional and large-scale data in parallel. Thus, it is a very effective machine learning method, and additional subtrees can improve the performance of the model. Here, we set the number of subtrees to 100 and set the pattern of the maximum number of features to be considered in random forest division as “sqrt.” In addition, “bootstrap” was set to true, which means that bootstrap samples were used when building trees.

(2) Decision tree model

A decision tree is a basic classification and regression method. The decision tree model considers the dataset as the root node and uses the threshold value of the feature parameter to divide it step-by-step. While continuing the division, each sub-dataset is of the same type to achieve classification. In the taxonomy of remote sensing, a decision tree is used to classify images step-by-step using the eigenvalues of pixels and setting appropriate boundary values on the nodes.

By defining rules for the image spectrum, color, space, and other information starting from the central node, various information values of the images are compared and new branches are obtained. New decision trees are obtained by updating the rules until the classification requirements are met, and the final node is the classification result. The feature selection criterion of the decision tree was set as Gini, the maximum number of features was considered when the classification was divided, and the depth of the subtree was not set when the decision tree was established.

(3) K Nearest Neighbor

The KNN classification algorithm is based on sample learning. The basic idea is to first find the *k* nearest neighbor points (usually known as the Euclidean distance) in the

classified data in the training set and then judge the type of the classified object according to the types of most of these nearest neighbor points. The k value is primarily determined by the number and dispersion of various samples, and the corresponding k value is selected for different situations. Here, the number of neighbors in the model was set to five.

(4) Support vector machine

Support vector machine (SVM) was proposed in 1964. It is a generalized linear classifier that classifies data binaries based on supervised learning. The decision boundary is the maximum-margin hyperplane used to solve the learning sample. The SVM uses a hinge loss function to calculate the empirical risk and adds a regularization term to solve the system to optimize the structural risk. A SVM is a sparse and robust classifier that introduces the concept of a kernel function to solve optimization problems in a high-dimensional feature space, and then searches for the optimal classification hyperplane to solve complex data classification problems. Since the 1990s, it has developed rapidly and has derived a series of improved and extended algorithms that have been applied to pattern recognition problems, such as portrait recognition and text classification.

In this model, the penalty parameter C was set to 1, the kernel function type was set to the Gaussian kernel function, and the decision function type was 'ovo', which means one vs. one.

(5) Naive Bayes

The Bayesian method is based on the Bayesian principle and uses knowledge of probability statistics to classify the sample dataset. The Bayes method is characterized by a combination of prior probability and posterior probability; that is, it avoids the subjective bias of using only prior probability and also avoids the overfitting phenomenon of using sample information alone. Owing to its solid mathematical foundation, the error rate of the Bayesian classification algorithm is very low. The Bayes classification algorithm exhibits high accuracy for large datasets and the algorithm itself is relatively simple.

(6) Linear discriminant analysis

LDA is a classic classification model widely used in pattern recognition, machine learning, and data mining. The basic idea of LDA is to map samples from high-dimensional space to low-dimensional space by projection and to project all samples onto a straight line so that similar samples are clustered as much as possible and heterogeneous samples are separated as much as possible. The core idea of LDA is to make the sample easier to classify. Specifically, the goal of LDA is to find a projection direction such that the eigenvectors of the same type of sample are as close as possible after the projection and the dissimilar samples are as far away as possible. This projection direction is the linear direction vector that we are seeking, and the classification of samples becomes simple and easy to implement after the projection.

3. Result and Discussion

3.1. Prediction Accuracy and Model Selection

The UAV aerial photography data from 17 June 2022, along with satellite remote sensing data of proximate dates, were employed to build a training verification set that contained 1,042,524 pixels. Given the imbalance characteristic between the Huangtai algae and non-Huangtai algae categories, this research adopts the undersampling method to balance the dataset by reducing the number of samples in the majority class. A total of 10 groups of data were designed. The decision tree, random forest, KNN, SVM, naive Bayes, and LDA models were used to predict the constructed training and validation sets with the 10-fold cross-validation method. Each method was experimented on each group

of the training and validation set. As an imbalance of samples were processed, the accuracy rate of the index, which is the most intuitive method for evaluating the performance of the classification model, could be adopted. The accuracy rate [31] refers to the ratio of the number of samples predicted by the classification model to the total number of samples. The prediction accuracy of each method is detailed in Table 2.

Table 2. Prediction accuracy of various machine learning methods using the training and validation sets.

Method	Dataset	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group 7	Group 8	Group 9	Group10	Mean
Random Forest	Training	0.999	1.000	1.000	0.999	0.999	1.000	1.000	1.000	1.000	0.999	0.999
	Validation	0.927	0.926	0.929	0.925	0.928	0.930	0.930	0.930	0.928	0.925	0.928
Decision Tree	Training	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	Validation	0.894	0.896	0.896	0.897	0.888	0.892	0.899	0.895	0.896	0.891	0.894
KNN	Training	0.942	0.941	0.941	0.941	0.941	0.942	0.942	0.941	0.940	0.940	0.941
	Validation	0.917	0.913	0.917	0.916	0.915	0.919	0.919	0.919	0.917	0.913	0.916
SVM	Training	0.884	0.885	0.885	0.884	0.884	0.885	0.884	0.883	0.883	0.884	0.884
	Validation	0.877	0.877	0.877	0.877	0.880	0.879	0.882	0.879	0.876	0.874	0.878
Bayes	Training	0.783	0.772	0.777	0.779	0.780	0.777	0.778	0.775	0.776	0.781	0.778
	Validation	0.776	0.761	0.765	0.767	0.769	0.772	0.772	0.771	0.774	0.768	0.769
LDA	Training	0.795	0.797	0.796	0.796	0.795	0.800	0.798	0.799	0.798	0.797	0.797
	Validation	0.790	0.788	0.785	0.789	0.789	0.792	0.791	0.793	0.791	0.788	0.790

As shown in Table 2, the decision tree, random forest, and KNN methods all performed well on the training set, and the prediction accuracy reached more than 90%. In the validation set, only the prediction accuracies of the decision tree and random forest reached 90%, and the random forest model was better than the decision tree model. The results indicate that the above methods (random forest, decision tree, KNN, SVM, naive Bayes, and LDA) were superior to the FAI method with an accuracy of 84.49% [13], SVM method with an accuracy of 68% [32], and LSP method with an accuracy of 84.8% [23]. These differences can be explained in part by the fact that we extracted the area of water at different times before identifying the Huangtai algae in this study, which improved the accuracy of the model for identifying Huangtai algae. In summary, random forest is the preferred method for identifying Huangtai algae in the waters of Ulansuhai Lake, and the following research results are based on the application of the random forest classification prediction method.

3.2. Feature Importance Analysis

The random forest method enables feature importance assessment based on a very simple idea: assess how much each feature contributes to each tree in the random forest. According to the results of the random forest model in Table 3, B4 is the most important, followed by B3 and B5, which further confirms the role of each band in the Landsat 8 spectrum from the side. B4 can be used to distinguish between water bodies and vegetation; B3 can be used to distinguish bare soil, roads, and vegetation; and B5 can be used to monitor the water content of vegetation. Landsat 8 has advantages in the identification of Huangtai algae in remote sensing images because it can detect highly concentrated algal blooms [13]. The analysis of remote sensing images has become an important technical means for people to quickly monitor the outbreak of Huangtai algae and can provide early warnings of algal outbreaks in water bodies.

Table 3. Band importance analysis.

Band	B1	B2	B3	B4	B5	B6	B7
Importance	0.0841201	0.1250335	0.1660145	0.227836	0.155908	0.1448167	0.0962712

The band differences between the Huangtai algae and non-Huangtai algae after remote sensing image preprocessing are shown in Figure 7. The bands with a high degree of differentiation between the Huangtai and non-Huangtai algae were B3, B4, and B5, which also confirms the importance of random forest features.

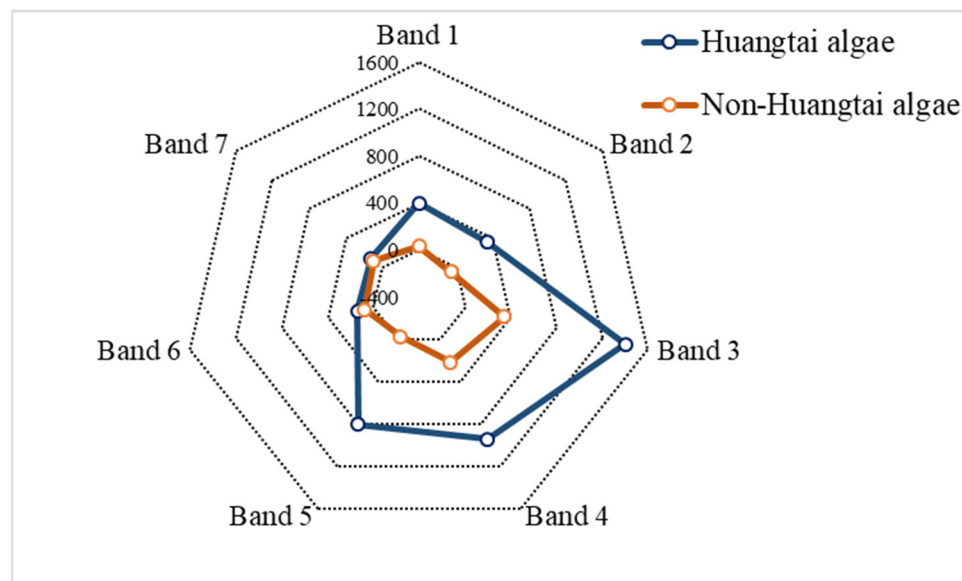


Figure 7. Spectral dissimilarity between the taxonomic classes.

3.3. Classification

3.3.1. Verification of the Open Water Area

According to the satellite remote sensing image identification, the open water area of Ulansuhai Lake showed little fluctuation in other years apart from a peak in 2014 and 2015 (see Figure 8). These relationships may be partly explained by the fact that Ulansuhai Lake is a grass-type lake, and Huangtai algae is mainly distributed in the open water area of Ulansuhai Lake. This observation may support the hypothesis that identifying the open water area of Ulansuhai Lake is important for improving the classification accuracy of Huangtai algae. The average open water area of Ulansuhai Lake from 2013 to 2022 was 140.56 km², which is consistent with the open water area of approximately 200,000 mu published by the Chinese government network (https://www.gov.cn/jrzq/2009-08/23/content_1399286.htm (accessed on 24 June 2024)).

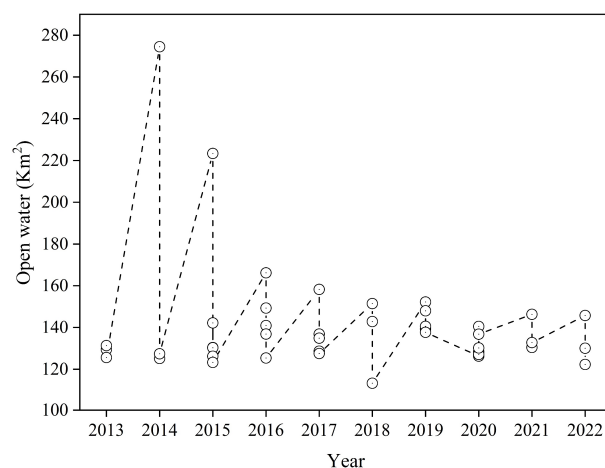


Figure 8. Variation trend chart of the open water area of Ulansuhai Lake.

3.3.2. Spatial and Temporal Distribution of Huangtai Algae

The machine learning results indicated that the life cycle of Huangtai algae includes growth, bloom, decay, and extinction periods from May to September (Figure 9). Typically, water temperature has a significant impact on the growth of aquatic vegetation. Huangtai algae were sensitive to short-term hot temperature conditions [33]. The area coverage degree of algae refers to the proportion of the area occupied by algae on the water surface or within a specific water area. It is a crucial indicator for measuring the degree of algal bloom and can directly reflect the distribution range and density of algae in the water body. According to the analysis of the area proportion of Huangtai algae in Ulansuhai Lake during the life cycle over the past ten years (2013–2022), Huangtai algae grow abundantly in July–August, which is a period of rapidly increasing water temperatures in the lake.

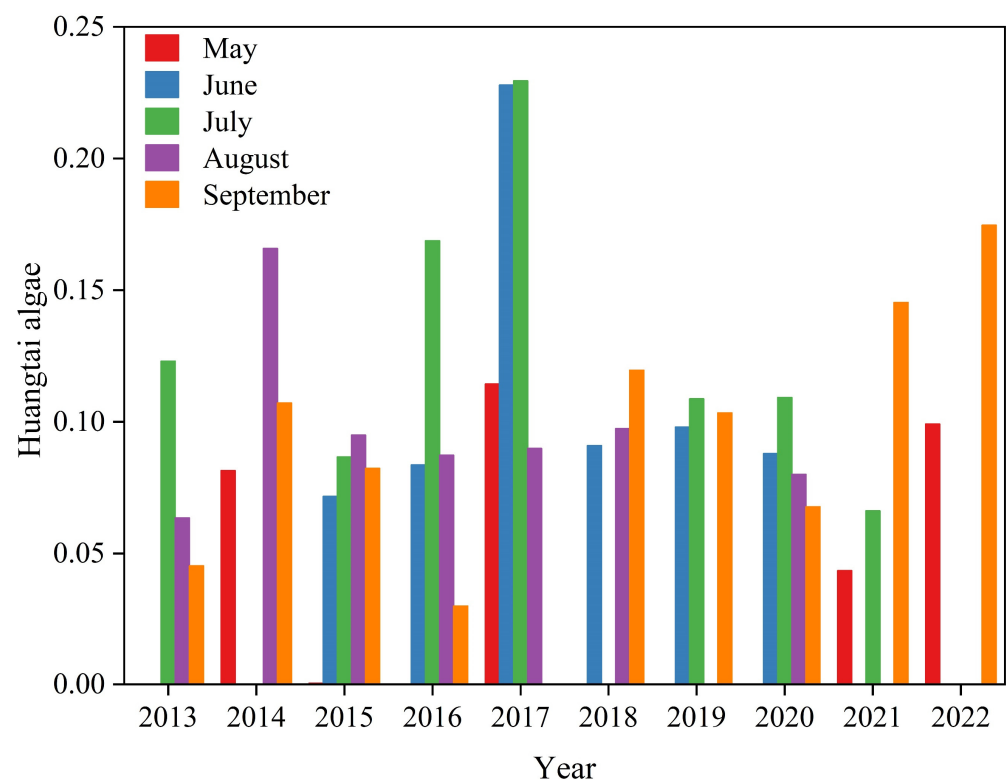


Figure 9. Area proportion of Huangtai algae per month.

The random forest classification method was used to identify Huangtai algae in the waters of Ulansuhai Lake, and some of the recognition results are shown in Figure 10. Consistent with the literature, this study found that participants who also reported using Qing reported that the growth range of Huangtai algae was mainly concentrated in the northern part of Ulansuhai Lake [33].

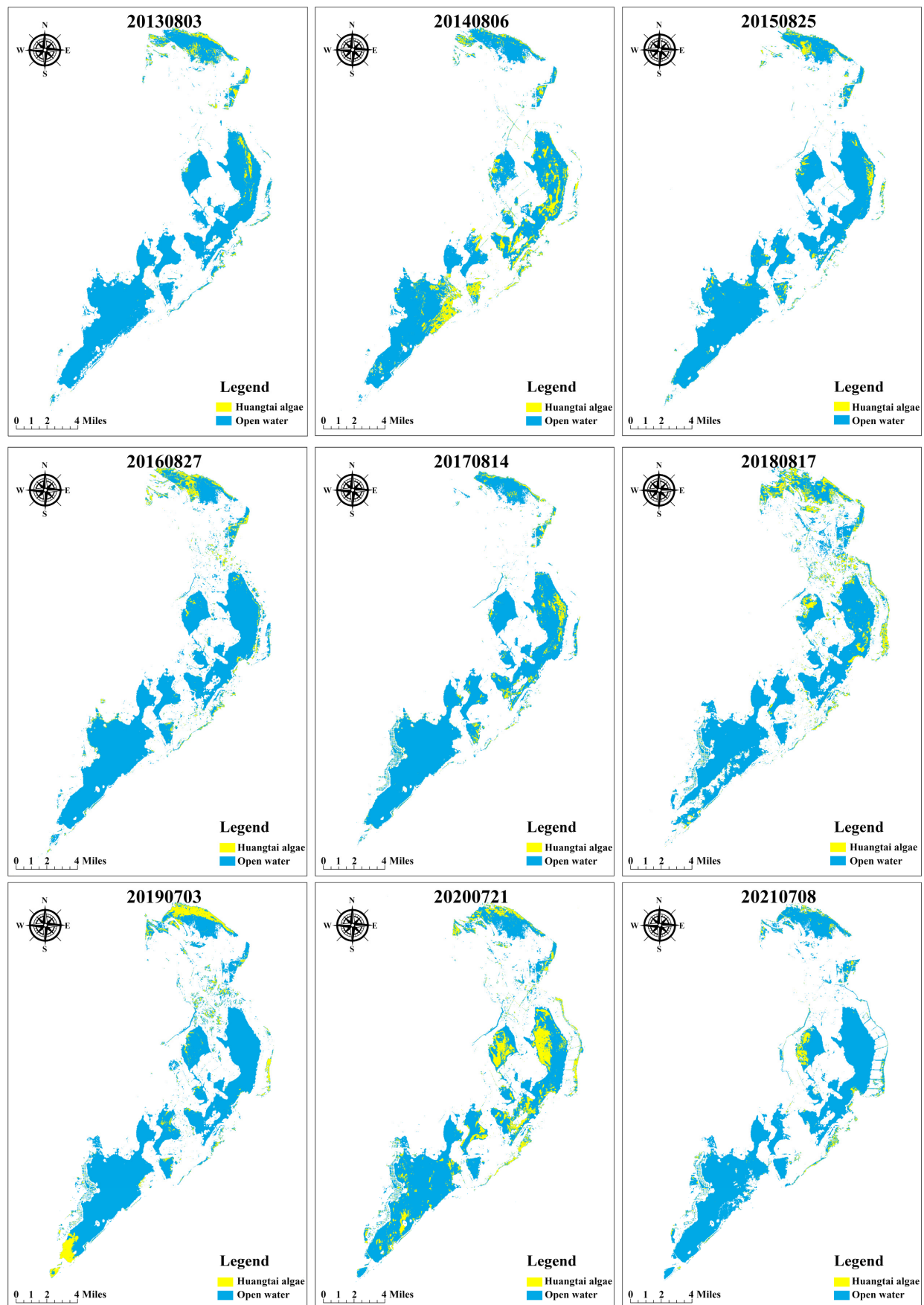


Figure 10. Portion of the recognition results.

4. Conclusions

In this study, a highly accurate system for identifying metaphytic blooms was developed using remote sensing images and machine learning methods. This method has many practical applications owing to its low cost and high accuracy. Moreover, our study has important practical guiding significance for monitoring water ecology and providing early warnings of ecological problems, such as water blooms. Through the identification and analysis of metaphytic blooms in the remote sensing images of Ulansuhai Lake, the main conclusions of this study are as follows.

1. Landsat 8 is very suitable for identifying aquatic plants because of its band characteristics. Before performing remote sensing image recognition, the MNDVI was used to dynamically identify the water boundaries, which can effectively reduce classification errors.
2. Machine learning models can effectively classify water bodies and metaphytic blooms. Among the several commonly used machine learning models, random forest exhibited good performance for both the training and validation sets. However, due to differences in the number of categories, processing data imbalances should be performed during model construction to improve the accuracy of the classification model.
3. Areas of metaphytic blooms show certain spatial and temporal distribution characteristics. The identification of metaphytic blooms in the remote sensing images of Ulansuhai Lake can effectively assist in the ecological supervision of Ulansuhai Lake.

Based on data acquisition limitations and other factors that affected this research, future studies should focus on optimizing the results presented herein based on the following aspects:

First, high-precision remote sensing images should be used to construct feature variables. Currently, submeter-level remote sensing images are available. High-precision remote sensing images contain more spectral information, which could better train the model and improve recognition accuracy.

Second, remote sensing images with a high acquisition frequency and short return period should be used to ensure that the time granularity is fine enough to facilitate epitaxial predictions after Huangtai algae identification.

Third, higher-resolution drone aerial photography should be performed to obtain aerial images that can generate more accurate classification category labels.

Fourth, owing to the large amount of remote sensing image data collected over a long period of time, deep learning model studies should focus on improving distributed computing to better meet immediate computing needs.

Author Contributions: Conceptualization, J.C.; Methodology, X.Z.; Software, X.Z.; Investigation, C.D.; Data curation, C.D.; Writing—original draft, J.C.; Writing—review & editing, C.D.; Visualization, G.L.; Supervision, G.L.; Funding acquisition, G.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Inner Mongolia Autonomous Region Science and Technology Plan Project (No. 2023KJHZ0026).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

References

1. Ahn, Y.-H.; Shanmugam, P. Detecting the red tide algal blooms from satellite ocean color observations in optically complex Northeast-Asia Coastal waters. *Remote Sens. Environ.* **2006**, *103*, 419–437. [[CrossRef](#)]

2. Beck, R.; Zhan, S.; Liu, H.; Tong, S.; Yang, B.; Xu, M.; Ye, Z.; Huang, Y.; Shu, S.; Wu, Q.; et al. Comparison of satellite reflectance algorithms for estimating chlorophyll-a in a temperate reservoir using coincident hyperspectral aircraft imagery and dense coincident surface observations. *Remote Sens. Environ.* **2016**, *178*, 15–30. [[CrossRef](#)]
3. Ho, J.C.; Michalak, A.M.; Pahlevan, N. Widespread global increase in intense lake phytoplankton blooms since the 1980s. *Nature* **2019**, *574*, 667–670. [[CrossRef](#)]
4. Legleiter, C.J.; King, T.V.; Carpenter, K.D.; Hall, N.C.; Mumford, A.C.; Slonecker, T.; Graham, J.L.; Stengel, V.G.; Simon, N.; Rosen, B.H. Spectral mixture analysis for surveillance of harmful algal blooms (SMASH): A field-, laboratory-, and satellite-based approach to identifying cyanobacteria genera from remotely sensed data. *Remote Sens. Environ.* **2022**, *279*, 113089. [[CrossRef](#)]
5. Wang, Z.; Mei, B. Current Status and Challenges of the Ecological Environment of Wuliangsuhai Basin in China. *IOP Conf. Ser. Earth Environ. Sci.* **2021**, *829*, 012012. [[CrossRef](#)]
6. Smayda, T.J. Harmful algal blooms: Their ecophysiology and general relevance to phytoplankton blooms in the sea. *Limnol. Oceanogr.* **1997**, *42*, 1137–1153. [[CrossRef](#)]
7. Yu, H.; Shi, X.; Zhao, S.; Sun, B.; Liu, Y.; Arvola, L.; Li, G.; Wang, Y.; Pan, X.; Wu, R.; et al. Primary productivity of phytoplankton and its influencing factors in cold and arid regions: A case study of Wuliangsuhai Lake, China. *Ecol. Indic.* **2022**, *144*, 109545. [[CrossRef](#)]
8. Sun, L.; Zhang, Z.; Li, Y.; Zhang, L.; Chen, Q.; Yu, R.; Hao, Y.; Lu, C. A new method based on additive vegetation index for mapping Huangtai algae coverage in Lake Ulansuhai. *Environ. Sci. Pollut. Res.* **2023**, *30*, 24590–24605. [[CrossRef](#)] [[PubMed](#)]
9. Du, C.; Cui, J.; Wang, D.; Li, G.; Lu, H.; Tian, Z.; Zhao, C.; Li, M.; Zhang, L. Prediction of aquatic vegetation growth under ecological recharge based on machine learning and remote sensing. *J. Clean. Prod.* **2024**, *452*, 142054. [[CrossRef](#)]
10. Barale, V.; Jaquet, J.-M.; Ndiaye, M. Algal blooming patterns and anomalies in the Mediterranean Sea as derived from the SeaWiFS data set (1998–2003). *Remote Sens. Environ.* **2008**, *112*, 3300–3313. [[CrossRef](#)]
11. Wei, G.; Tang, D.; Wang, S. Distribution of chlorophyll and harmful algal blooms (HABs): A review on space based studies in the coastal environments of Chinese marginal seas. *Adv. Space Res.* **2008**, *41*, 12–19. [[CrossRef](#)]
12. Kudela, R.M.; Palacios, S.L.; Austerberry, D.C.; Accorsi, E.K.; Guild, L.S.; Torres-Perez, J. Application of hyperspectral remote sensing to cyanobacterial blooms in inland waters. *Remote Sens. Environ.* **2015**, *167*, 196–205. [[CrossRef](#)]
13. Luo, J.; Ni, G.; Zhang, Y.; Wang, K.; Shen, M.; Cao, Z.; Qi, T.; Xiao, Q.; Qiu, Y.; Cai, Y.; et al. A new technique for quantifying algal bloom, floating/emergent and submerged vegetation in eutrophic shallow lakes using Landsat imagery. *Remote Sens. Environ.* **2023**, *287*, 113480. [[CrossRef](#)]
14. Dogan, O.K.; Akyurek, Z.; Beklioglu, M. Identification and mapping of submerged plants in a shallow lake using quickbird satellite data. *J. Environ. Manag.* **2009**, *90*, 2138–2143. [[CrossRef](#)] [[PubMed](#)]
15. Shi, K.; Li, Y.; Li, L.; Lu, H.; Song, K.; Liu, Z.; Xu, Y.; Li, Z. Remote chlorophyll-a estimates for inland waters based on a cluster-based classification. *Sci. Total Environ.* **2013**, *444*, 1–15. [[CrossRef](#)] [[PubMed](#)]
16. Zhang, F.; Li, J.; Shen, Q.; Zhang, B.; Wu, C.; Wu, Y.; Wang, G.; Wang, S.; Lu, Z. Algorithms and Schemes for Chlorophyll a Estimation by Remote Sensing and Optical Classification for Turbid Lake Taihu, China. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2015**, *8*, 350–364. [[CrossRef](#)]
17. Wei, Y.; Luo, X.; Hu, L.; Peng, Y.; Feng, J. An improved unsupervised representation learning generative adversarial network for remote sensing image scene classification. *Remote Sens. Lett.* **2020**, *11*, 598–607. [[CrossRef](#)]
18. Bolpagni, R.; Bresciani, M.; Laini, A.; Pinardi, M.; Matta, E.; Ampe, E.M.; Giardino, C.; Viaroli, P.; Bartoli, M. Remote sensing of phytoplankton-macrophyte coexistence in shallow hypereutrophic fluvial lakes. *Hydrobiologia* **2014**, *737*, 67–76. [[CrossRef](#)]
19. Hao, P.; Wang, L.; Zhan, Y.; Wang, C.; Niu, Z.; Wu, M. Crop classification using crop knowledge of the previous-year: Case study in Southwest Kansas, USA. *Eur. J. Remote Sens.* **2016**, *49*, 1061–1077. [[CrossRef](#)]
20. Zhang, Y.; Cai, X.; Song, X.; Suo, J.; Wang, Z.; Li, E.; Wang, X. Remote sensing information extraction of hydrophyte in Honghu Lake based on decision tree. *Wetl. Sci.* **2018**, *16*, 213–222. [[CrossRef](#)]
21. Bareuther, M.; Klinge, M.; Buerkert, A. Spatio-Temporal Dynamics of Algae and Macrophyte Cover in Urban Lakes: A Remote Sensing Analysis of Bellandur and Varthur Wetlands in Bengaluru, India. *Remote Sens.* **2020**, *12*, 3843. [[CrossRef](#)]
22. Liu, S.; Glamore, W.; Tamburic, B.; Morrow, A.; Johnson, F. Remote sensing to detect harmful algal blooms in inland waterbodies. *Sci. Total Environ.* **2022**, *851*, 158096. [[CrossRef](#)] [[PubMed](#)]
23. Sun, C.; Li, J.; Liu, Y.; Zhao, S.; Zheng, J.; Zhang, S. Tracking annual changes in the distribution and composition of saltmarsh vegetation on the Jiangsu coast of China using Landsat time series-based phenological parameters. *Remote Sens. Environ.* **2023**, *284*, 113370. [[CrossRef](#)]
24. Xu, H. Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *Int. J. Remote Sens.* **2006**, *27*, 3025–3033. [[CrossRef](#)]
25. Wang, Q.; Liu, D.; Yang, H.; Yin, J.; Bin, Z.; Zhu, S.; Zhang, Y. Comparative study on the water index of MNDWI and NDWI for water boundary extraction in eutrophic lakes. *Adv. Geosci* **2017**, *7*, 732–738. [[CrossRef](#)]

26. Haibo, Y.; Zongmin, W.; Hongling, Z.; Yu, G. Water Body Extraction Methods Study Based on RS and GIS. *Procedia Environ. Sci.* **2011**, *10*, 2619–2624. [[CrossRef](#)]
27. Feng, L.; Dai, Y.; Hou, X.; Xu, Y.; Liu, J.; Zheng, C. Concerns about phytoplankton bloom trends in global lakes. *Nature* **2021**, *590*, E35–E47. [[CrossRef](#)] [[PubMed](#)]
28. Lai, L.; Zhang, Y.; Cao, Z.; Liu, Z.; Yang, Q. Algal biomass mapping of eutrophic lakes using a machine learning approach with MODIS images. *Sci. Total Environ.* **2023**, *880*, 163357. [[CrossRef](#)] [[PubMed](#)]
29. Ghatkar, J.G.; Singh, R.K.; Shanmugam, P. Classification of algal bloom species from remote sensing data using an extreme gradient boosted decision tree model. *Int. J. Remote Sens.* **2019**, *40*, 9412–9438. [[CrossRef](#)]
30. Yang, C.; Tan, Z.; Li, Y.; Shen, M.; Duan, H. A Comparative Analysis of Machine Learning Methods for Algal Bloom Detection Using Remote Sensing Images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *16*, 7953–7967. [[CrossRef](#)]
31. Ali, S.; Smith, K.A. On learning algorithm selection for classification. *Appl. Soft Comput.* **2006**, *6*, 119–138. [[CrossRef](#)]
32. Raczko, E.; Zagajewski, B. Comparison of support vector machine, random forest and neural network classifiers for tree species classification on airborne hyperspectral APEX images. *Eur. J. Remote Sens.* **2017**, *50*, 144–154. [[CrossRef](#)]
33. Qing, S.; A, R.; Shun, B.; Zhao, W.; Bao, Y.; Hao, Y. Distinguishing and mapping of aquatic vegetations and yellow algae bloom with Landsat satellite data in a complex shallow Lake, China during 1986–2018. *Ecol. Indic.* **2020**, *112*, 106073. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.