applied sciences

Article

Machine Learning Powered Microalgae Classification by Use of Polarized Light Scattering Data

Zepeng Zhuo [1,2], Hongjian Wang [2], Ran Liao [2,*] and Hui Ma [1,2]

[1] Department of Physics, Tsinghua University, Beijing 100084, China; zzp17@mails.tsinghua.edu.cn (Z.Z.); mahui@tsinghua.edu.cn (H.M.)

[2] Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China; whj20@mails.tsinghua.com

* Correspondence: liao.ran@sz.tsinghua.edu.cn

Abstract: Microalgae are widely distributed in the ocean, which greatly affects the ocean environment. In this work, a dataset is presented, including the polarized light scattering data of 35 categories of marine microalgae. To analyze the dataset, several machine learning algorithms are applied and compared, such as linear discrimination analysis (LDA) and two types of support vector machine (SVM). Results show that non-linear SVM performs the best among these algorithms. Then, two data preparation approaches for non-linear SVM are compared. Subsequently, more than 10 categories of microalgae out of the dataset can be identified with an accuracy greater than 0.80. The basis of the dataset is shown by finding the categories independent to each other. The discussions about the performance of different incident polarization of light gives some clues to design the optimal incident polarization of light for future instrumentation. With this proposed technique and the dataset, these microalgae can be well differentiated by polarized light scattering data.

Keywords: polarized light scattering; suspended particles; machine learning

Citation: Zhuo, Z.; Wang, H.; Liao, R.; Ma, H. Machine Learning Powered Microalgae Classification by Use of Polarized Light Scattering Data. *Appl. Sci.* **2022**, *12*, 3422. https://doi.org/10.3390/app12073422

Academic Editors: Santiago Royo and Alexander N. Pisarchik

Received: 31 December 2021
Accepted: 21 March 2022
Published: 28 March 2022

1. Introduction

Microalgae are widely distributed in the ocean, which greatly affects the ocean environment [1]. Monitoring the categories and growing states of microalgae is important, as it can help explain and forecast the change of marine ecological environment and reduce loss from toxic blooms [2,3]. The common methods to observe and identify microalgae are based on optical microscopy, which is labor-intensive and needs specialized knowledge [4].

PCR is a vital tool to identify the categories of microalgae, but the preparation of the sample is time-consuming [5]. A spectrophotometer can measure the absorption spectrum of microalgae and they can be used to analyze the density and biomass of microalgae [6]. In addition, the signal of acoustic backscattering shows a good correlation with the abundance of the microalgae under certain concentration range [7]. However, these methods are based on the analysis of bulk volume, which limits their application in further detailed classification. Recently, there have been some tools developed to assist automatic phytoplankton taxonomy. Li et al. [8] introduced an imaging system to monitor marine organisms with sizes ranging from 200 µm to 40 mm. Göröcs et al. [9] reconstructed the holographic diffraction images to analyze natural water samples. However, these imaging methods are limited by the speed, resolution, and visual field, and meet their bottleneck when facing with micron-sized algae.

The scattering measurement has the advantage of characterizing the physical microstructure of different microparticles. Katherine et al. [10] classified different suspensions by scattering intensities features at multiple angles. Ye et al. [11] measured overall microparticle size using the scattering spectrum. Polarization is an inherent property of light [12]. Polarized light scattering, as an emerging tool, has been applied to characterize different states of microalgae [13,14], cancerous tissues [15], atmospheric microparticles [16], and microplastics [17].

Chami et al. [18] demonstrated the potential of using the polarized signal to analyze biogenic and highly refractive particles in coastal waters. Koestner et al. [19] used polarized light scattering measurements to characterize particle size and composition of natural assemblages of marine particles. Based on the pre-trained model, the states and categories of particles can be recognized from the mixture. Chen et al. [20] quantitatively studied the flocculation process with polarized light scattering. Wang et al. [21] applied polarization parameters to recognize different states of *Microcystis aeruginosa*, and gave an early warning strategy. In short, these works show the significance and value of a polarized light scattering dataset. However, there is still no such dataset of the diverse microalgae, which limits the optical polarization tools' applications in monitoring the microalgae in water.

In this work, a dataset by polarized light scattering measurement is presented, including the information of polarization parameters of 35 categories of marine microalgae. For each category, 10 states of polarization (SOP) of incident light are applied to respectively illuminate the samples, and for each SOP, there are more than 1000 records of the particles. To analyze the dataset, several machine learning algorithms are applied and compared to build the classifier which is used to identify different categories. This work compares linear discrimination analysis (LDA) and different types of support vector machine (SVM). Results showcase that non-linear SVM performs the best among these algorithms. Then, two data preparation approaches for non-linear SVM are compared. Subsequently, we show that more than 10 categories of microalgae out of the dataset can be identified with an accuracy greater than 0.80. With this proposed technique and the dataset, these microalgae can be well differentiated by polarized light scattering.

2. Materials and Methods

2.1. Samples

With the characteristics of morphology of the microalgae and types of intracellular pigments, microalgae can be classified into different phyla, such as Bacillariophyta, Chlorophyta, Pyrrophyta, etc. [22]. The samples in this work include 35 categories of marine microalgae, and the detailed information can be referred to in Table 1. These samples were provided by (Shanghai Guangyu Biological Technology Co. Ltd., Shanghai, China). During the experiments, all of these microalgae were sampled from the original sample, and they were added into the filtered seawater in the sample pool.

2.2. Experimental Setup

The experimental setup is shown in Figure 1, and it was designed to measure the polarized light scattering of individual microalgae. The light source (S) emitted linearly polarized light with a wavelength of 532 nm. Then, the linearly polarized light could be modulated into different states of polarization (SOP) with a polarized state generator (PSG), which consisted of the half-wave plate (HW1) and the quarter-wave plate (QW1). With lens (L1), the modulated light was focused into the sample pool (P) full of filtered seawater. In the sample pool, the microalgae were suspended with a stirrer rotating at the speed of 200 rounds per minute.

Table 1. Detailed information of samples used in this work.

No.	Name	Phylum	No.	Name	Phylum
1	*Skeletonema costatum*	Bacillariophyta	19	*Cyclotella meneghiniana*	Bacillariophyta
2	*Platymonas subcordiformis*	Chlorophyta	20	*Tetraselmis*	Chlorophyta
3	*Thalassiosira pseudonana*	Bacillariophyta	21	*Isochrysis zhangjiangensis*	Chrysophyta
4	*Chattonella marina*	Chrysophyta	22	*Chaetoceros muelleri*	Bacillariophyta
5	*Amphora*	Bacillariophyta	23	*Phaeocystis globosa Scherffel*	Chrysophyta
6	*Thalassiosira rotula Meunier*	Bacillariophyta	24	*Isochrysis CCMP3180*	Chrysophyta
7	*Isochrysis galbana*	Chrysophyta	25	*Karenia mikimotoi*	Pyrrophyta
8	*Thalassiosira weissflogii*	Bacillariophyta	26	*Porphyridaceae*	Rhodophyta
9	*Cyclotella meneghiniana*	Bacillariophyta	27	*Chaetoceros gracilis*	Bacillariophyta
10	*Isochrysis*	Chrysophyta	28	*Leptocylindrus*	Bacillariophyta
11	*Nitzschia closterium f.minutissima*	Bacillariophyta	29	*Zooxanthella*	Pyrrophyta
12	*Pavlova viridis*	Chrysophyta	30	*Heterosigma akashiwo (Hada)*	Xanthophyta
13	*Amphidinium carterae Hulburt*	Pyrrophyta	31	*Scrippsiella trochoidea*	Pyrrophyta
14	*Nannochloropsis*	Chlorophyta	32	*Cyclotellacryptica*	Bacillariophyta
15	*Chaetoceros curvisetus*	Bacillariophyta	33	*Cryptomonas*	Cryptophyta
16	*Dunaliella bardawil*	Chlorophyta	34	*Platymonas helgolandica tsingtaoensis*	Chlorophyta
17	*Trichodesmium*	Cyanophyta	35	*Coccolithophorids*	Chrysophyta
18	*Chaetoceros debilis Cleve*	Bacillariophyta			

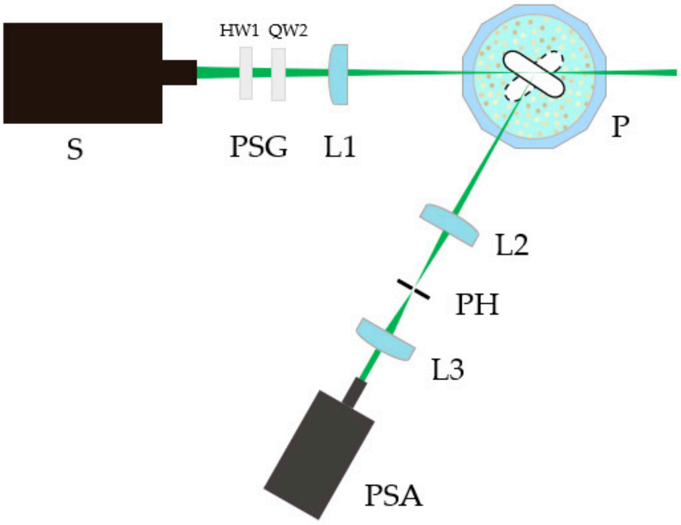

Figure 1. The schematic diagram of the experiment setup.

Scattering happens once microalgae pass the scattering volume which is the intersection volume of the optical systems of the illuminating path and receiving path [18]. Then, the scattered light is received at the backward 120°, and this angle has been proven to be sensitive to the microstructure of microalgae [8]. Subsequently, the scattered light passes through the lenses (L2 and L3), and the pinhole (PH) of 100 µm, so as to collimate the light and confine the size of the scattering volume to 100 µm. Then, the measurement of individual microalgae can be realized if the concentration is lower than 10^5 cells per mL. The collimated light after L3 can be analyzed by the polarization state analyzer (PSA), to finally get the Stokes vector S [23], as Equation (1),

$$S = \begin{bmatrix} I \\ Q \\ U \\ V \end{bmatrix} \tag{1}$$

where I is the light intensity, and Q, U, V are the residual intensity of 0° linear polarization, 45° linear polarization, and right-handed circular polarization, respectively. The three normalized polarization parameters, q, u, v, are defined as Equation (2),

$$q = \frac{Q}{I}, \; u = \frac{U}{I}, \; v = \frac{V}{I} \tag{2}$$

2.3. Machine Learning Algorithms

2.3.1. Linear Discrimination Analysis

Linear discrimination analysis (LDA) is a machine learning algorithm aiming to reduce the dimension of the original features and realize the classification between different categories [24]. The optimal goal of LDA is to find a projection axis by maximizing the between-class difference $|\mu_1 - \mu_2|^2$ and minimizing the within-class difference $\left(\delta_1^2 + \delta_2^2\right)$, which is equivalent to maximizing the value L, defined as Equation (3)

$$L = \frac{|\mu_1 - \mu_2|^2}{\left(\delta_1^2 + \delta_2^2\right)} \tag{3}$$

2.3.2. Support Vector Machine

Support vector machine (SVM) is a machine learning algorithm that maximizes the margin between different categories of data and classifies different categories with a hyperplane [25]. SVM is designed for two-class classification tasks; one-versus-one (OVO) and one-versus-rest (OVR) are two methods applied in the classification of the multi-class task. Regarding the classification of different categories, SVM with the linear kernel can separate the data of different categories by a linear hyperplane, while the non-linear SVM model applies the Gaussian kernel to map the original features into a higher-dimensional space and realize the classification by a non-linear hyperplane.

2.3.3. Performance Evaluation

In this work, we use accuracy to evaluate the performance of the built classifiers, and the accuracy is defined as Equation (4),

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{4}$$

where TP is the true positive (the actual and predicted labels are positive), TN is the true negative (the actual and predicted labels are negative), FP is the false positive (the actual label is negative and the predicted label is positive), and FN is the false negative (the actual label is positive and the predicted label is negative).

3. Results

3.1. Algorithm Comparison and Selection

During the measurement, 10 SOPs of incident light were modulated into 7 differently oriented linear SOPs, i.e., 120°, 150°, 30°, 60°, 90°, −45°, and 45°; one elliptical SOP, E; the left-handed circular SOP, L; and the right-handed circular SOP, R. For each SOP, each category of microalgal cells was continuously measured to obtain more than 1000 records. Note that the time durations for different categories of microalgae were quite different according to the concentration of the particles. To obtain 1000 records for 10 polarization states of each sample, the average time duration was about 15 min. Here, each record corresponded to an individual microalgal cell and it had four feature parameters, [I, q, u, v]. Therefore, for each category of microalgae in Table 1, the measured dataset included 10 SOPs and more than 1000 records for each SOP. Then, the input dataset was fed into the machine learning algorithms to build the models and train different classifiers, and the algorithm comparisons were conducted for different categories of microalgae in a same SOP of the incident light.

The sampling methodology for the machine learning process is shown in Table 2. For each sample, there were in total 1000 records for one polarization state, where 800 records were used for training process and the remaining 200 were used for validation. Note that there were 10 illuminating polarization states, the sampling number for these 35 categories of samples was equally balanced for OVO and OVR.

Table 2. The sampling methodology for the machine learning.

	Training Subset	Validating Subset	Total
Each sample	800	200	1000
Sampling ratio	0.8	0.2	1

Machine learning algorithms were tested and compared in this work, including LDA, linear SVM, and non-linear SVM. The programing language in this work was Python 3.8, the package of data processing was mostly NumPy, and the package used for algorithm training was sklearn. During the training process, two categories of microalgae were randomly selected out of the dataset, and then the data were separated into training and validating subsets. After the classifier was trained by the training subset, the classifier was tested in the validating subset, and the accuracy was recorded as the accuracy of classifier. In this section, we go through all categories of microalgae, and for each pair of target microalgae, the accuracy is recorded for further comparison.

To compare the performance of LDA with linear SVM, a binary array (A_1, A_2) was obtained for each pair of microalgae, where A_1 is the accuracy of LDA and A_2 is the accuracy of linear SVM. We applied the LDA to project the polarization parameters, q, u, v, into one value. During the training process, the solver of LDA was the singular value decomposition, the threshold for a singular value was 0.5, and the penalty of SVM was the hinge loss. After going through all the categories of microalgae in the dataset by permutation and combination, a series of arrays were recorded and they are shown in Figure 2a.

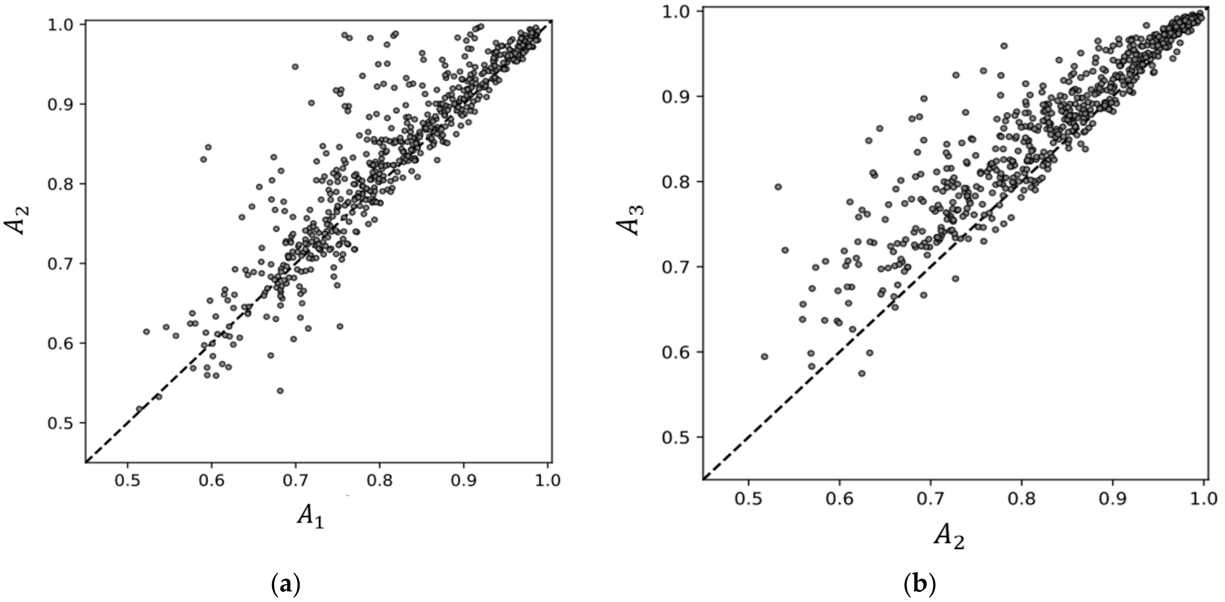

Figure 2. Comparison between the performances of different algorithms; (**a**) the comparison results of LDA and linear SVM; (**b**) the comparison results of linear SVM and non-linear SVM.

Results show that A_1 and A_2 were mostly lined up on the diagonal line, while linear SVM was slightly better compared with LDA. Similarly, in the comparison of linear and non-linear SVM, the result shown in Figure 2b indicates that the accuracy of the non-linear

SVM, A_3, was generally higher than that of the linear SVM, A_2. Thereafter, non-linear SVM was applied to train and classify different categories of microalgae.

3.2. Classification Result of the Dataset

Since SVM is designed for two-class tasks, the dataset was firstly separated by OVR or OVO into two categories. For OVR, every target class was compared with the remaining classes, and 35 classifiers could be built with the non-linear SVM with OVR. For an unknown validating sample, 35 classifiers were applied to predict the measured data and output 35 probability scores, then the highest score corresponded to the identified category.

Different levels of recognition accuracy were set in order to quantitatively analyze the classification performance with the built classifier. Moreover, it could provide some information about the microalgae, which is important for understanding the physical mechanism of the discrimination or instructing the future prototype deployments in water. Empirically, we selected four levels to evaluate the results. The performance of the classifier on the validating dataset is presented in Table 3, which shows that *Isochrysis* has the highest recognition accuracy greater than 0.95, which implies that *Isochrysis* has distinct scattering property compared with the other microalgae in Table 1. Referring to the result in Table 3, both *Chattonella marina* and *Cryptomonas* achieved a recognition accuracy greater than 0.9 but smaller than 0.95, and *Chaetoceros curvisetus* and *Zooxanthella* achieved an accuracy falling in the range of 0.85–0.90. In addition, the other five categories could be identified with an accuracy within 0.80–0.85. Based on the OVR approach, there were in total 10 categories of microalgae that could be identified in the dataset with an accuracy greater than 0.80. The classifiers derived from OVR were able to recognize the certain targeted category of microalgae, and the trained model could be used to evaluate whether the measured sample was similar to the targeted category.

Table 3. Classification results using SVM by one-versus-rest (OVR).

Recognition Accuracy	Categories of Microalgae
>0.95	*Isochrysis*
0.90–0.95	*Chattonella marina; Cryptomonas*
0.85–0.90	*Chaetoceros curvisetus; Zooxanthella*
0.80–0.85	*Porphyridaceae; Trichodesmium; Amphidinium carterae Hulburt; Amphora; Skeletonema costatum*

For OVO, every class was respectively compared with the remaining $n-1$ classes, and then $n \times (n-1)/2$ classifiers could be built in total. During the validating process, the identified category was voted among these classifiers. Different from OVR, we set an accuracy threshold to find the largest subset out of the original dataset, and all of the microalgae in the subset could be recognized above the accuracy threshold. Subsequently, different accuracy thresholds were compared, including 0.95, 0.90, 0.85, and 0.80. With these thresholds, the results of the largest subset are shown in Table 4.

The results in Table 4 show that four categories ($n = 4$) of microalgae could be extracted from the original dataset, and each category in the subset could be retrieved with an accuracy of more than 0.95. Eight categories of microalgae could be extracted out of the dataset, and they could be recognized with an accuracy above 0.9, and 11 and 15 categories of microalgae could be extracted if the accuracy threshold was 0.85 and 0.8, respectively.

Comparing these two approaches, the OVO performed better than the OVR. However, it is notable that the complexity of OVO is $O(n^2)$ which is much higher than that of OVR. Different from OVR, the OVO for SVM in this work tried to find the subset of the microalgae where different categories of microalgae could be well classified from others. Then, these categories of microalgae in the subset could be used as the skeleton categories which are independent of each other and are the basis of the dataset, and the other categories are similar to them or their combinations.

Table 4. Classification results using SVM by one versus one (OVO).

Accuracy Threshold	Categories of Microalgae	*n*
0.95	*Chattonella marina; Isochrysis; Chaetoceros debilis Cleve; Porphyridaceae*	4
0.90	*Skeletonema costatum; Chattonella marina; Isochrysis; Dunaliella bardawil; Isochrysis zhangjiangensis; Isochrysis CCMP3180; Scrippsiella trochoidea; Cryptomonas*	8
0.85	*Skeletonema costatum; Chattonella marina; Thalassiosira weissflogii; Isochrysis; Chaetoceros curvisetus; Dunaliella bardawil; Isochrysis zhangjiangensis; Isochrysis CCMP3180; Porphyridaceae; Leptocylindrus; Cryptomonas*	11
0.80	*Skeletonema costatum; Chattonella marina; Isochrysis galbana; Isochrysis; Chaetoceros curvisetus; Cyclotella meneghiniana; Isochrysis zhangjiangensis; Isochrysis CCMP3180; Karenia mikimotoi; Porphyridaceae; Leptocylindrus; Zooxanthella; Scrippsiella trochoidea; Cryptomonas; Platymonas helgolandica tsingtaoensis*	15

4. Discussion

The four polarization parameters of each record are [I, q, u, v], and they are derived from the Stokes vectors of scattered light, which are basically related to the incident SOP. To find the best incident SOP and reduce the detection complexity, the performance of different incident SOP and different combinations of features are discussed and the averaged classification accuracy of all categories in Table 1 based on non-linear SVM and OVO is used to evaluate the performances.

The results shown in Figure 3 collect the accuracy among all the incident SOP and different combinations of polarization parameters. When the input was [I, q, u, v], all the SOPs achieved an accuracy of about 0.85, while the 150° linear SOP was slightly better than other SOPs, and the 120° linear SOP was the worst. Compared with the classification accuracy derived from I, it is notable that the original classification accuracy was greatly improved with the addition of the polarization features.

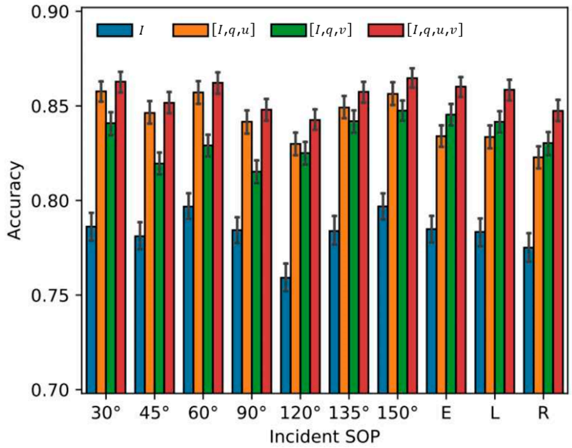

Figure 3. The performance comparison of different incident SOPs and different combinations of features.

However, the relative contribution of the polarization parameters u and v was not same. For the SOP of E, L, and R, the addition of the parameter v resulted in a higher accuracy than u. However, for the other linear SOPs, the parameter u brought more effective information than the parameter v.

The Mueller matrix is a 4×4 matrix, $\boldsymbol{M}$, describes the polarization property of the particle, and combines the incident SOP, S_i, and the scattered SOP, S_s, that is, $S_s = \boldsymbol{M} \times S_i$.

The 16 elements of $\boldsymbol{M}$ contain different aspects of physical information of particles. Usually, we have to change S_i 4 times and measure the respective S_s to calculate $\boldsymbol{M}$. We can always simultaneously get the Stoke vector in a single shot but it is hard to get $\boldsymbol{M}$ in that way, especially for the suspended particles. Therefore, in the dataset, we can only provide the S_s of the individual microalgae in a given S_i, as the data used in the classifications in Figure 2, and at most we can obtain the statistical $\boldsymbol{M}$ but not the individual $\boldsymbol{M}$ of the microalgal cells [26].

Considering that different S_i will bring different information regarding the elements of the $\boldsymbol{M}$, it is important to discuss the contribution of the elements based on the dataset. Moreover, the quantitative contributions of the elements will guide us to improve the detecting speed of the system by reducing the number of modulated S_i, and find the optimal SOP to better characterize and classify these categories of microalgae, which is important for fast field probing applications.

The 150° linear SOP and right-handed circular SOP are further discussed with the Mueller matrix theorem. For simplicity, we discuss the spherical particles. Theoretically, for spherical particles, the top right and bottom left elements of $\boldsymbol{M}$ are zero [27]. When S_i is $[I_i, Q_i, U_i, V_i]^T$, S_s can be calculated by Equation (5),

$$S_s = \begin{bmatrix} M_{11} & M_{12} & 0 & 0 \\ M_{12} & M_{22} & 0 & 0 \\ 0 & 0 & M_{33} & M_{34} \\ 0 & 0 & -M_{34} & M_{44} \end{bmatrix} \begin{bmatrix} I_i \\ Q_i \\ U_i \\ V_i \end{bmatrix} = \begin{bmatrix} M_{11}I_i + M_{12}Q_i \\ M_{12}I_i + M_{22}Q_i \\ M_{33}U_i + M_{34}V_i \\ -M_{34}U_i + M_{44}V_i \end{bmatrix} = I * \begin{bmatrix} 1 \\ q \\ u \\ v \end{bmatrix} \tag{5}$$

where I is the scattered intensity and q, u, v are the normalized polarization parameters of S_s.

Usually, the standard SOP of incident light can be represented as $\left[1, 1/2, -\sqrt{3}/2,\ 0\right]^T$ for 150° and $[1,\ 0,\ 0,\ 1]^T$ for R. Then, the Stokes vector of the scattered light is calculated as $[M_{11} + M_{12}/2,\ M_{12} + M_{22}/2, -\sqrt{3}M_{33}/2,\ \sqrt{3}M_{34}/2]^T$ for 150° linear SOP and the Stokes vector of the scattered light is $[M_{11},\ M_{12},\ M_{34},\ M_{44}]^T$ when the incident SOP is R.

The classification performance of the incident SOP of 150° is obviously better than the performance of the incident SOP of R. Compared with these two derived Stokes vectors of scattered light, the Stokes vector of 150° linear SOP has the distinctive information of the elements M_{22} and M_{34}, and the Stokes vector of R has the distinctive information of M_{44}. Since the incident SOP of 150° has a better performance of the incident SOP of R, it seems that M_{22} and M_{34} may contain more important information than M_{44} for the classification task.

However, for the contribution of the polarization parameters u and v, the parameter u contributes less information compared with the parameter v for linear SOP, while the parameter v contributes less information compared with the parameter u for circular SOP. Thus, it seems that M_{34} is less useful compared with M_{33} or M_{44}.

The result in Figure 2 shows that non-linear SVM displays the best performance to classify these categories of microalgae based on Stokes vectors. However, this result does not explicitly claim the contributions of the specific polarization parameters, since it learns the polarization parameters of Stokes vectors in a hyper feature space. To verify the above analysis with Mueller matrix, we select two categories of microalgae to quantitatively evaluate the contribution of the polarization parameter v and the symmetry of the polarization states. Both the *Isochrysis* and *Chattonella marina* are easily distinguishable from other categories of microalgae. Their biological features are investigated in a previous book [28,29]. The result is shown in Table 5; note that the relative difference is the difference between these two cases over the retrieved accuracy with $[I, q, u]$. Moreover, the precision information of these two categories of microalgae can be referred to in Tables 6 and 7, which shows the classification performance of the trained model on these two categories. The two circular SOPs of the incident light, R and L, have theoretical Stokes vectors of $[M_{11},\ M_{12},\ M_{34},\ M_{44}]$ and $[M_{11},\ M_{12},\ M_{34}, -M_{44}]$. The classification performances with $[I, q, u, v]$ and $[I, q, u]$ were compared and the result is shown in Table 5, which indicates that the SOP of R and L can retrieve a close classification accuracy and the contributions of the feature, v or M_{44}, are about 0.12 in both cases. This analysis indicates that the SOP of incident light L and R has a similar performance in the classification task of the dataset, and this result corresponds to the theoretical explanation with the Mueller matrix.

Table 5. Classification *Isochrysis* and *Chattonella marina*.

SOP	Accuracy ([I, q, u, v])	Accuracy ([I, q, u])	Relative Difference
L	96.98%	86.47%	0.12
R	95.42%	85.08%	0.12

Table 6. Precision of *Chattonella marina*.

SOP	Precision ([I, q, u, v])	Precision ([I, q, u])	Relative Difference
L	98.52%	89.08%	10.60%
R	97.89%	90.16%	8.57%

Table 7. Precision of *Isochrysis*.

SOP	Precision ([I, q, u, v])	Precision ([I, q, u])	Relative Difference
L	95.50%	83.73%	14.06%
R	93.22%	81.36%	14.58%

With the analysis above, the information of M_{44} is suggested to be contained during the polarized light scattering measurement and the circular SOP of incident light is suggested to be included in the further measurement, as an addition to the measurement with linear SOP of the incident light, such as 150° linear SOP. There may only be one incident SOP allowed when the conceptual prototype is deployed to classify the suspended microalgae in the aquatic field, so an optimal incident SOP should be necessarily designed based on these considerations.

Note that most microalgal cells are not uniform spheres in the morphology and structure; previous literature demonstrates that the top right and bottom left elements of their Mueller matrices are approximately zero [30]. Therefore, the discussion based on Equation (5) is still valid. During the experiments in this work, we measured 10 polarization states, which is time-consuming. The discussion gives clues to reducing the modulated polarization states; 150° linear SOP and the circular SOP are suggested to be included in future applications.

The results in this work indicate that, due to the diversity and complexity, polarization data equipped by the machine learning algorithm are a feasible way to effectively classify marine microalgae, and more comprehensive methods and an abundant number of data would achieve a better classification performance. The fast-developing machine learning method will provide tools for us [31], and the Mueller matrix polarimetry of the individual microalgae may be expected in near future, and they both promote classification ability. Moreover, we notice that the optical microscopy is still the standard method to identify the species of the microalgae [4] and the morphological information and the internal structure play vital roles in the classification of diverse microalgae to which polarization parameters are sensitive [32,33]. In addition, an in situ prototype based on polarized light scattering was easily built and demonstrated to be powerful in classifying particles in seawater [34]. As such, the combination of the microscope and polarized light scattering is a promising tendency and may provide a way to accurately and rapidly classify the microalgae in water.

5. Conclusions

In this work, we presented a dataset including the polarized light scattering data of 35 categories of marine microalgae and explored a feasible way to classify the microalgae using the polarization data. The dataset included the diversity and complexity of the marine microalgae, covering their different physical properties such as size, shape, structure, etc. After comparing the three machine learning algorithms, we showed that the classifier based on non-linear SVM could classify them with an accuracy above 80%. Subsequently, two comparison approaches, one-versus-one and one-versus-rest, were applied to separate the dataset. Results showed that 15 categories of microalgae could be identified with an accuracy greater than 80%, and they could be treated as the basis of the dataset, since these categories of microalgae are independent from each other. Moreover, the polarization data with the full Stokes vector obtained the best classification accuracy, which shows the advantage and the necessity of circular polarization in both illumination and detection. The above results indicate that, due to the diversity and complexity, marine microalgae need comprehensive machine learning algorithms and abundant polarization data to achieve the best classification performance. The results in this work give hints to understand the physical mechanism of the classification and further instruct the future prototype deployments in water.

Author Contributions: Z.Z. and H.W.: writing—original draft preparation; R.L.: writing—review and editing; H.M.: resources. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Key-Area Research and Development Program of Guangdong Province (2020B1111040001); the National Key Research and Development Program of China (2018YFC1406600); and the National Natural Science Foundation of China (NSFC) (41527901, 61975088, 62075042).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Tan, J.S.; Tan, J.S.; Chew, K.W.; Lam, M.K.; Lim, J.W.; Ho, S.H.; Show, P.L. A review on microalgae cultivation and harvesting, and their biomass extraction processing using ionic liquids. *Bioengineered* **2020**, *11*, 116–129. [CrossRef]
2. Huisman, J.; Codd, G.A.; Paerl, H.W.; Ibelings, B.W.; Verspagen, J.M.H.; Visser, P.M. Cyanobacterial blooms. *Nat. Rev. Genet.* **2018**, *16*, 471–483. [CrossRef] [PubMed]
3. Rekully, C.M.; Faulkner, S.T.; Lachenmyer, E.M.; Cunningham, B.R.; Shaw, T.J.; Richardson, T.L.; Myrick, M.L. Fluorescence Excitation Spectroscopy for Phytoplankton Species Classification Using an All-Pairs Method: Characterization of a System with Unexpectedly Low Rank. *Appl. Spectrosc.* **2018**, *72*, 442–462. [CrossRef] [PubMed]
4. Coltelli, P.; Barsanti, L.; Evangelista, V.; Frassanito, A.M.; Gualtieri, P. Water monitoring: Automated and real time identification and classification of algae using digital microscopy. *Environ. Sci.-Proc. Imp.* **2014**, *16*, 2656–2665. [CrossRef] [PubMed]
5. Radha, S.; Fatima, A.; Sellamuthu, I.; Mohandss, R. Direct colony PCR for rapid identification of varied microalgae from freshwater environment. *J. Appl. Phycol.* **2013**, *25*, 609–613. [CrossRef]
6. Johan, F.; Jafri, M.; Lim, H.; Omar, W. Laboratory measurement: Chlorophyll-a concentration measurement with acetone method using spectrophotometer. In Proceedings of the 1014 IEEE International Conference on Industrial Engineering and Engineering Management, Selangor, Malaysia, 9–12 December 2014; pp. 744–748. Available online: http://www.ieem2014.org/public.asp?page=home.htm (accessed on 30 December 2021).
7. Kim, H.; Kang, D.; Jung, S.W.; Kim, M. High-frequency acoustic backscattering characteristics for acoustic detection of the red tide species Akashiwo sanguinea and Alexandrium affine. *J. Ocean. Limnol.* **2019**, *37*, 1268–1276. [CrossRef]
8. Li, J.; Chen, T.; Yang, Z.; Chen, L.; Liu, P.; Zhang, Y.; Yu, G.; Chen, J.; Li, H.; Sun, X. Development of a Buoy-Borne Underwater Imaging System for In Situ Mesoplankton Monitoring of Coastal Waters. *IEEE J. Ocean. Eng.* **2021**, *47*, 88–110. [CrossRef]
9. Göröcs, Z.; Tamamitsu, M.; Bianco, V.; Wolf, P.; Roy, S.; Shindo, K.; Yanny, K.; Wu, Y.; Koydemir, H.C.; Rivenson, Y.; et al. A deep learning-enabled portable imaging flow cytometer for cost-effective, high-throughput, and label-free analysis of natural water samples. *Light-Sci. Appl.* **2018**, *7*, 66. [CrossRef]
10. Klug, K.E.; Jennings, C.M.; Lytal, N.; An, L.; Yoon, J.Y. Mie scattering and microparticle-based characterization of heavy metal ions and classification by statistical inference methods. *R. Soc. Open Sci.* **2019**, *6*, 190001. [CrossRef]
11. Ye, M.; Wang, S.; Lu, Y.; Hu, T.; Zhu, Z.; Xu, Y. Inversion of particle-size distribution from angular light-scattering data with genetic algorithms. *Appl. Opt.* **1998**, *38*, 2677–2685. [CrossRef]
12. Chami, M. Importance of the polarization in the retrieval of oceanic constituents from the remote sensing reflectance. *J. Geophys. Res. Space Phys.* **2007**, *112*, 05026. [CrossRef]
13. Wang, Y.; Liao, R.; Dai, J.; Liu, Z.; Xiong, Z.; Zhang, T.; Chen, H.; Ma, H. Differentiation of suspended particles by polarized light scattering at 120°. *Opt. Express* **2018**, *17*, 22419. [CrossRef] [PubMed]

14. Wang, H.; Liao, R.; Xiong, Z.; Wang, Z.; Li, J.; Zhou, Q.; Tao, Y.; Ma, H. Simultaneously Acquiring Optical and Acoustic Properties of Individual Microalgae Cells Suspended in Water. *Biosensors* **2022**, *12*, 176. [CrossRef]
15. He, C.; He, H.; Chang, J.; Chen, B.; Ma, H.; Booth, M.J. Polarisation optics for biomedical and clinical applications: A review. *Light-Sci. Appl.* **2021**, *10*, 194. [CrossRef]
16. Xu, Q.; Zeng, N.; Guo, W.; Guo, J.; He, Y.; Ma, H. Real time and online aerosol identification based on deep learning of multi-angle synchronous polarization scattering indexes. *Opt. Express* **2021**, *29*, 18540–18564. [CrossRef] [PubMed]
17. Yu, S.; Dai, J.; Liao, R.; Chen, L.; Zhong, W.; Wang, H.; Jiang, Y.; Li, J.; Ma, H. Probing the nanoplastics adsorbed by microalgae in water using polarized light scattering. *Optik* **2021**, *231*, 166407. [CrossRef]
18. Chami, M.; Platel, M.D. Sensitivity of the retrieval of the inherent optical properties of marine particles in coastal waters to the directional variations and the polarization of the reflectance. *J. Geophys. Res.* **2017**, *112*, C05037. [CrossRef]
19. Koestner, D.; Stramski, D.; Reynolds, R.A. Polarized light scattering measurements as a means to characterize particle size and composition of natural assemblages of marine particles. *Appl. Opt.* **2020**, *59*, 8314–8334. [CrossRef]
20. Chen, Y.; Liao, R.; Li, J.; Zhou, H.; Wang, H.; Zhuo, Z.; Wang, Q.; Yan, C.; Ma, H. Monitoring particulate composition changes during the flocculation process using polarized light scattering. *Appl. Opt.* **2021**, *60*, 10264. [CrossRef]
21. Wang, H.; Li, J.; Liao, R.; Tao, Y.; Peng, L.; Li, H.; Deng, H.; Ma, H. Early warning of cyanobacterial blooms based on polarized light scattering powered by machine learning. *Measurement* **2021**, *184*, 109902. [CrossRef]
22. Boddy, L.; Morris, C.W.; Wilkins, M.F.; Al-Haddad, L.; Tarran, G.A.; Jonker, R.R.; Burkill, P.H. Identification of 72 phytoplankton species by radial basis function neural network analysis of flow cytometric data. *Mar. Ecol. Prog. Ser.* **2000**, *195*, 47–59. [CrossRef]
23. Bohren, C.F.; Huffman, D.R. *Absoption and Scattering of Light by Small Particles*; John Wiley and Sons: New York, NY, USA, 1983.
24. Yan, C.; Chang, X.; Luo, M.; Zheng, Q.; Zhang, X.; Li, Z.; Nie, F. Self-weighted Robust LDA for Multiclass Classification with Edge Classes. *ACM Trans. Intell. Syst. Technol.* **2021**, *12*, 4. [CrossRef]
25. Chang, C.C.; Lin, C.J. Libsvm: A Library for Support Vector Machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27. [CrossRef]
26. Li, J.; Wang, H.; Liao, R.; Wang, Y.; Liu, Z.; Zhuo, Z.; Guo, Z.; Ma, H. Statistical Mueller matrix driven discrimination of suspended particles. *Opt. Lett.* **2021**, *46*, 3645–3648. [CrossRef] [PubMed]
27. Svensen, Ø.; Stamnes, J.J.; Kildemo, M.; Aas, L.M.S.; Erga, S.R.; Frette, Ø. Mueller matrix measurements of algae with different shape and size distributions. *Appl. Opt.* **2011**, *50*, 5149–5157. [CrossRef] [PubMed]
28. Mishchenko, M.I.; Liu, L. Weak localization of electromagnetic waves by densely packed many-particle groups: Exact 3D results. *J. Quant. Spectrosc. Radiat. Transf.* **2007**, *106*, 616–621. [CrossRef]
29. Carmelo, R.T. *Identifying Marine Phytoplankton*; Academic Press: San Diego, CA, USA, 1997.
30. Volten, H.; de Haan, J.F.; Hovenier, J.W.; Schreurs, R.; Vassen, W.; Dekker, A.G.; Hoogenboom, H.J.; Charlton, F.; Wouts, R. Laboratory measurements of angular distributions of light scattered by phytoplankton and silt. *Limnol. Oceanogr.* **1998**, *43*, 1180–1197. [CrossRef]
31. Marx, V. The big challenges of big data. *Nature* **2013**, *498*, 255–260. [CrossRef]
32. Brosnahan, M.L.; Velo-Suárez, L.; Ralston, D.K.; Fox, S.E.; Sehein, T.R.; Shalapyonok, A.; Sosik, H.M.; Olson, R.J.; Anderson, D.M. Rapid growth and concerted sexual transitions by a bloom of the harmful dinoflagellate Alexandrium fundyense (Dinophyceae). *Limnol. Oceanogr.* **2015**, *60*, 2059–2078. [CrossRef]
33. Wang, Y.; Dai, J.; Liao, R.; Zhou, J.; Meng, F.; Yao, Y.; Chen, H.; Tao, Y.; Ma, H. Characterization of physiological states of the suspended marine microalgae using polarized light scattering. *Appl. Opt.* **2020**, *59*, 1307–1312. [CrossRef]
34. Liao, R.; Li, Q.; Mao, X. A prototype for detection of particles in sea water by using polarize-light scattering. In Proceedings of the OCEANS 2019, Marseille, France, 17–20 June 2019.