Optics and Lasers in Engineering 166 (2023) 107567

Contents lists available at ScienceDirect

Optics and Lasers in Engineering

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

Polarization fingerprint for microalgae classification

Jiajin Li[a,b], Jinfu Wei[a], Hongyuan Liu[a], Jiachen Wan[b], Tongyu Huang[b], Hongjian Wang[a], Ran Liao[a,b,*], Meng Yan[c], Hui Ma[a,b]

[a] Shenzhen Key Laboratory of Marine IntelliSensing and Computation, Institute for Ocean Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[b] Guangdong Research Center of Polarization Imaging and Measurement Engineering Technology, Institute of Biopharmaceutical and Health Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[c] State Key Laboratory of Marine Pollution, City University of Hong Kong, Hong Kong, China

ARTICLE INFO

Keywords:
Polarized light scattering
Polarization fingerprint
Microalgae
Mueller matrix
Classification
Physical features

ABSTRACT

Microalgae classification is important for aquatic environment monitoring, but it is difficult in some in-situ cases. The Mueller matrix is known to encode the physical properties including the sophisticated microstructure of microalgae cells. In this paper, the concept of polarization fingerprint based on the Mueller matrix is presented to extract the parameters to classify microalgae. An experimental setup based on the particulate Mueller matrix polarimetry (PMP) is used to measure 25 kinds of microalgae covering seven phyla to form a Mueller matrix dataset. The proposed sixteen polarization parameters with explicit physical meanings and associations with the structural properties are selected to compose the polarization fingerprint. As a result, different microalgae can be effectively classified by the polarization fingerprint and machine learning algorithms. The contribution of the parameters in the polarization fingerprint is analyzed to provide the specific features for microalgae classification, which gives insights into microalgae's exclusive physical properties. The extensibility and flexibility of the polarization fingerprint are discussed. This work shows the polarization fingerprint's power to classify microalgae, which can help retrieve in-situ information on the community structures of microalgae, and further monitor the aquatic environment.

1. Introduction

Microalgae cells are the key component of the Earth's biosphere and aquatic food webs. They are the sensitive indicators of the ecological environment and can help explain and forecast the changes in water ecosystems [1–2]. Microalgae cells in water ecosystems vary in their concentration and composition with location and time, whose diversity and dynamics can directly affect climate, fisheries, and human livelihoods [3]. Also, microalgae cells can regulate the biological carbon pump locally and globally, and they are widely involved in the assessment of risk and development of environmental regulations [4]. Besides, the kinds and concentrations of microalgae have a great impact on the inherent and apparent optical properties of the water body, which contributes to the physical basis of modern water color remote sensing [5]. Hitherto, the in-situ accurate classification of different microalgae is still quite hard for the research community, because it indeed needs multimodal measurement with high spatial and temporal resolution [6].

Recently, the microalgae classification always depends on their physical, biological, and chemical information, such as size, shape, microstructure, pigment, and intracellular molecules [6–7]. The optical microscope is a common tool to classify different species of microalgae by observing their cellular shapes and internal pigment fluorescence [8]. Meanwhile, optical microscopes are subjected to the diffraction limit and fail to obtain the microstructure information of the microalgae cell, which can be especially important in the classification when the microalgae cells have similar shapes. The electron microscope, such as scanning electron microscope (SEM) and transmission electron microscopy (TEM), is also usually adopted to provide the detailed intracellular or extracellular microstructure of microalgae cells [9–10]. The observation of microalgae using the electron microscope is restricted by high costs, high labor intensity, and time consumption [7]. Moreover, there are intrinsic difficulties when using these microscopes for in-situ classification of microalgae, due to their complex structure and rigorous sample preparation.

The flow cytometer is also widely used for a variety of tasks to analyze the compositions and physiological states of the microalgae community. It needs strict hydrodynamic focusing to successfully pass cells through the detection volume one by one and its characterization generally depends on the intensity of scattering light and the fluorescence. Most flow cytometers need to stain the specific dyes to label the intra-

* Corresponding author.
E-mail address: liao.ran@sz.tsinghua.edu.cn (R. Liao).

https://doi.org/10.1016/j.optlaseng.2023.107567
Received 25 November 2022; Received in revised form 9 February 2023; Accepted 3 March 2023
Available online 11 March 2023

cellular molecule which is hard for in-situ detection [7]. In addition, some submersible flow cytometers have equipped the ability to get the cell's image with limited resolution [11]. Moreover, the side scattering intensity is used to characterize the microstructure of microalgae cells, which is too insufficient to accurately classify various microalgae cells. At present, there still lacks a potential tool to obtain the abundant microstructure information for in-situ microalgae classification.

Polarization is an inherent property of light. Polarization measurements can provide abundant information about the microstructural properties of samples. Usually, the polarization states of light can be described by a 4-dimensional vector, called Stokes vector S. The 4×4 Mueller matrix M is a transformation matrix of polarization states of light [12], i.e. $S_{out} = M S_{in}$, as shown in Eq. (1). The Mueller matrix elements represent the capability of the sample to transform the polarization states of light before and after the interaction. The Mueller matrix elements are affected by many different factors, including the geometrical properties (size, shape, microstructure, orientation, and alignment), the optical properties (refractive index, diattenuation, absorption, and retardance), and so on [13]. Generally, the Mueller matrix elements can be normalized by M_{11}, that is, $m_{ij} = M_{ij}/M_{11}$, where i, j =1, 2, 3, 4.

$$S_{in} = \begin{bmatrix} I_{in} \\ Q_{in} \\ U_{in} \\ V_{in} \end{bmatrix}, \; S_{out} = \begin{bmatrix} I_{out} \\ Q_{out} \\ U_{out} \\ V_{out} \end{bmatrix}, \; M = \begin{bmatrix} M_{11} & M_{12} & M_{13} & M_{14} \\ M_{21} & M_{22} & M_{23} & M_{24} \\ M_{31} & M_{32} & M_{33} & M_{34} \\ M_{41} & M_{42} & M_{43} & M_{44} \end{bmatrix} \quad (1)$$

Theoretically, the Mueller matrix can characterize the comprehensive polarization properties of samples [12]. However, individual Mueller matrix elements often fail to reveal the apparent relationship with specific microstructural features [14]. Many polarization parameters derived by the Mueller matrix decomposition (MMD) [15] and Mueller matrix transformation (MMT) [16], are used to quantitatively describe the characteristic features of samples. These have been proven to be sensitive to the micro- or macro-structural features and optical or other physical properties of biomedical samples [17]. Due to the great diversity of microalgae, each of these polarization parameters alone may be not adequate in the descriptions and classifications of microalgae cells. So, it still needs to find a way to combine these polarization parameters with explicit physical meanings together to classify microalgae cells, which is beneficial for the microalgae research and resource utilization.

The Mueller matrix of individual suspended particles can be measured by particulate Mueller matrix polarimetry (PMP) [18]. In previous work, different suspended particles are successfully classified by the Mueller matrix elements powered by convolutional neural network. However, it is hard to physically explain why different suspended particles can be classified and to explicitly find which parameters dominate the classification. Meanwhile, previous researches have reported the featured parameters from the Mueller matrix images to characterize the tissue's microstructure [16]. These encourage us to further find the way to exploit the potentialities of PMP on the classification of microalgae.

In this work, we propose the concept of polarization fingerprint consisting of these polarization parameters to effectively classify the microalgae, with the aid of machine learning algorithms. 25 kinds of microalgae in seven phyla are chosen as the samples with the broad representation. Sixteen polarization parameters with explicit physical meanings and associations with certain structural properties are selected to form the polarization fingerprint. Different microalgae are effectively classified by the polarization fingerprint and the machine learning algorithms. Then, the combination of polarization parameters and their contribution to the microalgae classification are analyzed, which reveals the inclusiveness and exclusiveness of the polarization fingerprint. After that, the classification of all kinds of microalgae is carried out and the correlation between parameters in the polarization fingerprint is analyzed, which discusses the extensibility and flexibility of the polarization fingerprint.

2. Methods

2.1. Samples

In this work, 25 kinds of common microalgae are collected as the measured samples, and their microscopic photos are shown in Figure 1. These samples are provided by Shanghai Guangyu Biological Technology Co. Ltd., Shanghai, China. The detailed information on these microalgae is shown in Table 1 including seven phyla, and these microalgae are wide-ranging and representative in shape and size. During the experiments, all of these microalgae are sampled from the original suspensions, and they are added into the sample pool containing filtered seawater, and then measured by PMP.

2.2. Experimental setup and data

PMP has been designed to measure the Mueller matrix of individual microalgae cells [18], and its schematic diagram is shown in Figure 2. The illumination arm consists of a 532 nm laser (S) and a polarization state generator which consists of a polarizer (P) and a pair of electro-optic modulators (E1 and E2) to modulate the illuminating light. The modulated illuminating light is divided into two parts by a 10:90 non-polarizing beam splitter (N). Herein, 10 percent of the modulated illuminating light enters into a monitoring polarization state analyzer (M-PSA) to monitor S_{in}, and the rest of the light is then focused by lens 1 (L1) to illuminate the suspended microalgae cell.

The curved surface of the glass beaker (GB) can act as a cylindrical lens for both the illuminating light and scattered light. To remove these effects, in the sample pool, the GB is placed at the center of a glass dodecagon cuvette (DC) filled with distilled water. The measured samples are contained in the GB and stirred by a magnetic stirrer at a speed of 200 rounds per minute. The 120° backscattering detection arm consists of lens 2 (L2), lens 3 (L3), a 100 μm pinhole (PH), and a detection polarization state analyzer (D-PSA), where the scattered Stokes vector S_{out} of microalgae cell can be measured. Finally, we can use the measured S_{in} and S_{out} to calculate the Mueller matrix M of individual microalgae cells in the GB.

The intersection volume of illuminating light and the optical path of the detection arm is the scattering volume. In Figure 2, the PH and the scattering volume are the object-image relationships by L2. The scattering volume is limited to less than 0.01 μL which is determined by the focused spot of the illuminating light and the size of PH. If the concentration of the suspended particle is controlled to be less than 10^5 particles / mL, there is statistically only one particle in the scattering volume. Therefore, the measurement of the individual microalgae cell can be realized [18].

2.3. Polarimetry basis parameters

Herein, sixteen polarization parameters, which are the functions of Mueller matrix elements, are adopted to quantitatively analyze the experimental data of microalgae [19], as shown in Table 2. Parameter M_{11} is related to the transmittance or reflectivity of samples [20], and generally can be used to evaluate the intensity of scattered light. M_2 is the sum of squares of normalized Mueller matrix elements, which can describe the degree of depolarization, and $M_2 = 4$ is the necessary and sufficient condition for the non-depolarizing normalized Mueller matrix [21]. M_D may be related to the size or depolarization of the scatterer [22–23]. Generally, the multiple scattering between the complexities of cellular organelles can depolarize the illuminating polarized light, causing a low degree of polarization [9]. t represents the overall degree of linear anisotropy, which can represent the linear anisotropy degree of fibrous structures [24]. β is usually used to evaluate the optical rotation [24–25], and CD is related to the circular dichroism anisotropy [26]. D_s shows the summation of differences in the nonzero and symmetrical

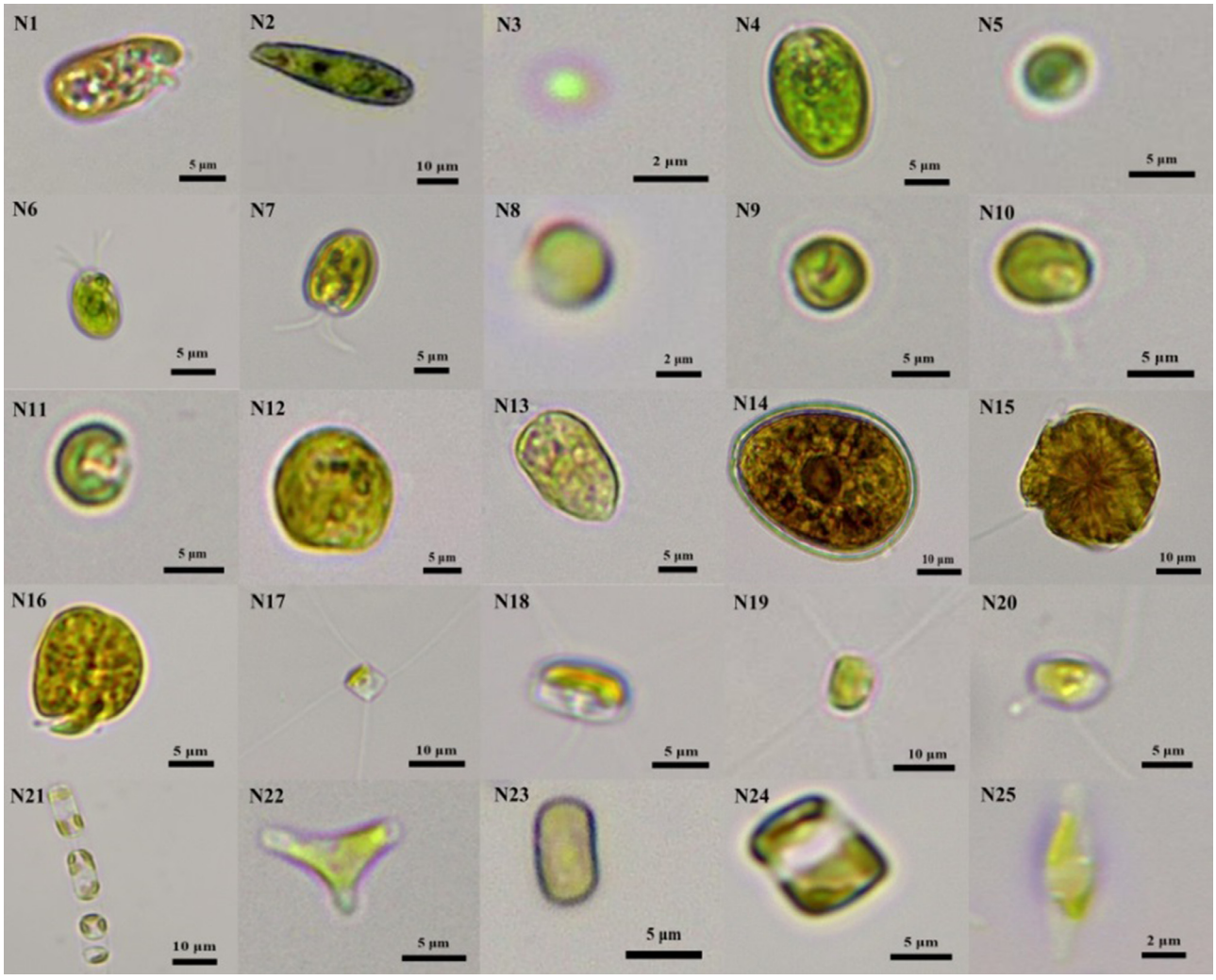

Fig. 1. Optical microscope images of 25 kinds of microalgae.

Table 1
Detailed information on the selected microalgae.

No.	Name	Phylum	Shape and Size
N1	*Cryptomonas ovata*	Cryptophyta	Ovoid, 8-12 μm × 3-8 μm
N2	*Euglena gracilis*	Euglenophyta	Spindly, 40-50 μm × 5-15 μm
N3	*Synechococcus* WH7803	Cyanophyta	Spherical, 0.5-2 μm
N4	*Dunaliella salina*	Chlorophyta	Ovoid, 18-28 μm × 10-14 μm
N5	*Chlorella marina*	Chlorophyta	Spherical, 2.5-6 μm
N6	*Platymonas subcordiformis*	Chlorophyta	Ovoid, 11-14 μm × 7-9 μm
N7	*Platymonas helgolandica*	Chlorophyta	Ovoid, 20-24 μm × 12-15 μm
N8	*Emiliania huxleyi*	Chrysophyta	Spherical, 1-8 μm
N9	*Isochrysis galbana*	Chrysophyta	Ovoid, 4-7 μm × 2-4 μm
N10	*Isochrysis zhangjiangensis*	Chrysophyta	Ovoid, 6-7 μm × 5-6 μm
N11	*Phaeocystis globsa*	Chrysophyta	Spherical, 2-7 μm
N12	*Prorocentrum minimum*	Pyrroptata	Ovoid, 15-23 μm × 13-17 μm
N13	*Prorocentrum donghaiense*	Pyrroptata	Ovoid, 15-25 μm × 10-15 μm
N14	*Prorocentrum lima*	Pyrroptata	Ovoid, 33-44 μm × 23-33 μm
N15	*Alexandrium tamarense*	Pyrroptata	Spherical, 20-40 μm
N16	*Amphidinium carterae* Hulburt	Pyrroptata	Biconical, 11-24 μm × 6-17 μm
N17	*Chaetoceros debilis* Cleve	Bacillariophyta	Cylindrical, 5-10 μm
N18	*Chaetoceros gracilis*	Bacillariophyta	Cylindrical, 5-10 μm
N19	*Chaetoceros curvisetus* Cleve	Bacillariophyta	Cylindrical, 5-10 μm
N20	*Chaetoceros meülleri*	Bacillariophyta	Cylindrical, 5-10 μm
N21	*Skeletonema costatum*	Bacillariophyta	Cylindrical, 6-7 μm
N22	*Phaeodactylum tricornutum*	Bacillariophyta	Spindly, 8-12 μm × 2-5 μm
N23	*Thalassiosira weissflogii*	Bacillariophyta	Cylindrical, 5-10 μm
N24	*Cyclotella cryptica*	Bacillariophyta	Cylindrical, 5-10 μm
N25	*Nitzschia closterium f.minutissima*	Bacillariophyta	Spindly, 12-20 μm × 2-5 μm

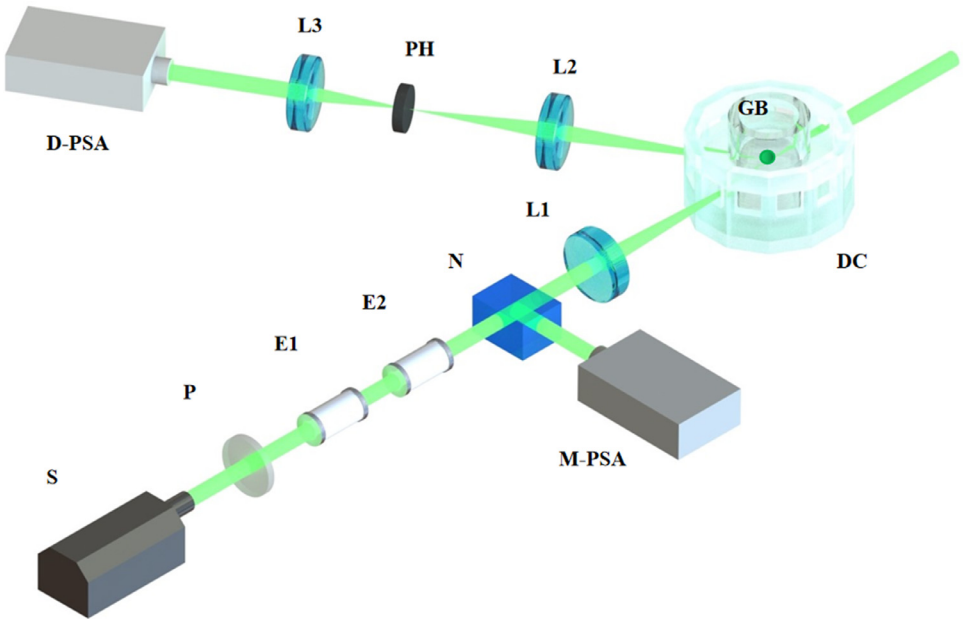

Fig. 2. Schematic diagram of particulate Mueller matrix polarimetry.

Table 2
Detailed information on polarimetry basis parameters.

Parameter	Formula	Physical property
M_{11}	M_{11}	Intensity of scattered light [20]
M_2	Sum of squares of normalized Mueller matrix elements	Degree of depolarization [21]
M_D	Determinant of normalized Mueller matrix	Related to depolarization or size [22,23]
t	$\frac{1}{2}\sqrt{(m_{22}-m_{33})^2+(m_{23}+m_{32})^2}$	Degree of overall linear anisotropy [24]
β	$\frac{1}{2}(m_{23}-m_{32})$	Circular retardance, optical rotation [25]
CD	$m_{14}+m_{41}$	Circular dichroism anisotropy [26]
D_s	$\lvert(m_{12}-m_{21})\rvert+\lvert(m_{33}-m_{44})\rvert+\lvert(m_{34}+m_{43})\rvert$	Difference with ideal sphere [12]
q_L	$\sqrt{m_{42}^2+m_{43}^2}$	Capability of transforming circular polarization to linear polarization [27]
r_L	$\sqrt{m_{24}^2+m_{34}^2}$	Capability of transforming linear polarization to circular polarization [27]
D_L	$\sqrt{m_{12}^2+m_{13}^2}$	Linear diattenuation, dichroism [24]
P_L	$\sqrt{m_{21}^2+m_{31}^2}$	Linear polarization [24]
PD	$\sqrt{(m_{12}-m_{21})^2+(m_{13}-m_{31})^2}$	Breaking of transpose symmetry [19]
rq	$\sqrt{(m_{24}+m_{42})^2+(m_{34}+m_{43})^2}$	Breaking of transpose symmetry [19]
α_{DP}	$\frac{1}{2}(\mathrm{atan2}(m_{13},m_{12})-atan2(m_{31},m_{21}))$	Breaking of transpose symmetry [19]
α_{rq}	$\frac{1}{2}(\mathrm{atan2}(-m_{24},m_{34})-atan2(m_{42},-m_{43}))$	Breaking of transpose symmetry [19]
α_1	$\frac{1}{4}\mathrm{atan2}(m_{23}+m_{32},m_{22}-m_{33})$	Azimuth orientation of anisotropy [28]

elements of the Mueller matrix of an ideal solid sphere, which may indicate the shapes or symmetries [12]. q_L, r_L, D_L and P_L are related to the anisotropic properties from different origins [24,27], which may reflect the factors causing the anisotropy. PD, rq, α_{DP}, and α_{rq}, are related to the breaking of transpose symmetry, which may be caused by many physical reasons [19]: (a) overlapping of sample; (b) nontrivial zenith angle for cylindrical scatterer or sphere-birefringence sample; (c) absorption. α_1 is related to the azimuth orientation of the anisotropy [28], which may be sensitive to microalgae's shape, configuration, or motion trail. Herein, $\mathrm{atan2}(y,x)$ returns the four-quadrant inverse tangent ($\tan^{-1}$) of y and x.

Those sixteen polarization parameters belong to the group consisting of the so-called polarimetry basis parameter (PBP). In the works of literature, there are more than 60 PBPs with explicitly physical meanings, and each of them connects the measured optical quantities, i.e., Mueller matrix elements, with one aspect of the sample's physical properties [19]. Those PBPs have been proven in research of the forward scattering or backscattering of media, whose feasibility in the present 120° backscattering setup would be tested. In this work, powered by those PBPs and machine learning algorithms, new polarimetry feature parameters (PFPs) or polarimetry classification models (PCMs) can be obtained and estimated by the performance of microalgae classification. These sixteen PBPs in Table 2 are selected since their PFPs and PCMs perform well to classify the microalgae cells, and we call them the polarization fingerprint of these microalgae.

2.4. Machine learning algorithm

2.4.1. Linear discriminant analysis

The linear discriminant analysis (LDA) algorithm aims to project the original d-dimensional data into the m-dimensional data ($m<d$) using an optimal linear transformation matrix W [29], which is a simple

and effective supervised feature extraction algorithm to characterize or distinguish different samples. After transformation, the variance of the same sample is the smallest, and the distance of different samples is the largest. Particularly, the classification performance can be easy to visually observe when the *d*-dimensional data are projected into the subspace whose dimensionality is less than or equal to three. Moreover, the new PFP can be obtained to effectively classify the special microalgae using PBPs and LDA algorithm, which is the linear combination of PBPs and is also called LDA-PFP. Note that the coefficient of each PBP in the linear combinations represents its importance or contribution to the classification for the normalized dataset. The degree of separation DoS is defined as Eq. (2), which can be used to quantitatively evaluate the classification ability of LDA-PFP. In other words, the larger DoS represents the better classification ability for two different microalgae.

$$DoS = \frac{(\mu_1 - \mu_2)^2}{(\sigma_1 + \sigma_2)} \tag{2}$$

where μ_1 and μ_2 are the arithmetic means of two data sets after projection, and σ_1 and σ_2 are the variances of two data sets after projection.

2.4.2. Random forest

The random forest (RF) algorithm is another commonly-used machine learning algorithm to handle both classification and regression problems. The RF algorithm is made up of a collection of decision trees, and each tree in the ensemble is comprised of a data sample drawn from a training set with replacement. Generally, the RF algorithm combines the output of multiple decision trees to reach a single result (prediction), and its simplicity and flexibility has fueled its adoption in many fields [30]. The number of decision trees is set as 100 in all models in this work.

Each tree in random forests is trained by a random subset of the original whole dataset. The randomness includes features randomness and samples randomness. Usually, n samples are selected randomly and repeatedly from all n samples and $\sqrt{m}$ features are selected randomly from all m features. The purpose of the randomness is to decrease the variance of the estimator and reduce the overfitting problem.

When a random subset is determined, an individual decision tree is also built. The Gini impurity is used to measure the uncertainty of data labels in the RF algorithm, as shown in Eq. (3). When the decision trees in the RF algorithm are being trained, the splitting strategy on every node is based on how much each feature contributes to decreasing the weighted Gini impurity. Further, the Gini gain of every feature can be calculated by Eq. (4) to select the feature with the maximum Gini gain to split the dataset into two subsets. The above operations are repeated until all classes are classified in each decision tree.

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2 \tag{3}$$

where k is the number of classes and the p_i is the probability of class i at given dataset D.

$$\Delta Gini(A) = Gini(D) - Gini(D|A) \tag{4}$$

where A is the selected feature and $Gini(D|A)$ is shown as Eq. (5).

$$Gini(D|A) = \sum_{i=1}^{2} \frac{n_i}{n} Gini\left(D_i\right) \tag{5}$$

where the D_i with size n_i is the i-th subset of D and n is the size of D.

With many decision trees built by different subsets, we can make decisions by averaging the probabilistic prediction of these decision trees. The RF algorithm has some key advantages, such as strong robustness, and a high degree of accuracy. Besides, the most amazing quality of RF is that we can easily observe the importance of each feature in predictions. In addition, the adjustable parameters in the RF algorithm are fewer, which can be easier to build stable and practical classification models than neural networks. In this work, the RF algorithm is also used to build special PCMs based on PBPs to classify different microalgae with a high degree of accuracy, which is quite different from the LDA algorithm.

For the RF algorithm, two types of variable importance measures (VIM) have been defined: the Gini importance, and the permutation importance, and their calculation details can be found elsewhere [31]. Herein, the Gini importance is used. The Gini importance measures the (normalized) mean decrease of Gini impurity over decision trees. In this work, 70% of the dataset is used as the training set, and 30% of the dataset set is used as the test set. The larger VIM represents the larger contribution. In each case, the summation of all VIM of features is 1.

3. Results

3.1. PBPs of different microalgae

The PBPs of 25 kinds of microalgae are calculated based on their Mueller matrix elements. Note that each PBP is linearly normalized by the maximum and minimum in the whole dataset into the range of 0 to 1. For simplicity, the mean values and standard variations of Mueller matrix elements are calculated to observe the overall situation, as shown in Figure 3 where microalgae in different phyla are plotted by different colors. Notably, different microalgae show multifarious differences in PBPs. N1 and N2 have similarities in some PBPs, such as M_{11} and PD, but have obvious differences in M_D, t, D_L, and P_L. N4 and N14 show great differences compared with other microalgae in many PBPs, such as M_2, t, D_L, and P_L. Besides, N3 also shows various differences in PBPs such as r_L, rq, and α_{DP}. The M_D of N8, N9, N10, and N11 are at a similar level and relatively larger than others, which may imply that M_D may be an important feature to characterize the Chrysophyta. Besides, the results of N12, N13, and N14 point out that the polarization fingerprints can be quite different from each other in various species of the same genus. N14 has an obvious pyrenoid in the cell center, and many valve pores and marginal pores, which are the outstanding characteristics that may cause the unusual PBPs.

On one hand, there are many potential differences in the PBPs, which reflect the physical features of different microalgae. However, for most microalgae, it is not enough for each PBP to classify one kind of microalgae from the others. Therefore, machine learning methods are necessarily used to combine the tiny difference in PBPs together and finally find the new PFPs to classify various microalgae. On the other hand, from Figure 3, it is obvious that the difference of the microalgae distributes unevenly in the PBPs, which means there are some PBPs that are the key contributors, and the weight of these PBPs may be the important feature of each microalgae cell. In this work, these key PBPs and their weight are the main part when using polarization fingerprint to classify the microalgae.

By acquiring the data of samples for enough time, the sample numbers of each class of microalgae are ensured to be more than 5000. Note that we measure the individual microalgae cells; even for the same kind of cells, the variance of size, shape, or intracellular structure of the cells which are subjected to their own response to the environmental factors, would lead to the difference of PBPs. In addition, some individual cells would be detected in different view since they are suspended and may pass the detection volume in random routes. For these reasons, the standard deviations of the PBP values are large. Also, for each PBP, the difference between the microalgae cells is not significant, so it seems hard to discriminate the different species only by one PBP. This pushes us to combine the tiny difference together to find a better way to classify these microalgae.

3.2. Classification of N1 and each kind of microalgae by LDA algorithm

For simplicity, we use the LDA algorithm to investigate the classification between N1 and the 25 kinds of microalgae. Subsequently, we find the key PBPs and their weights for the specified classification. Based on

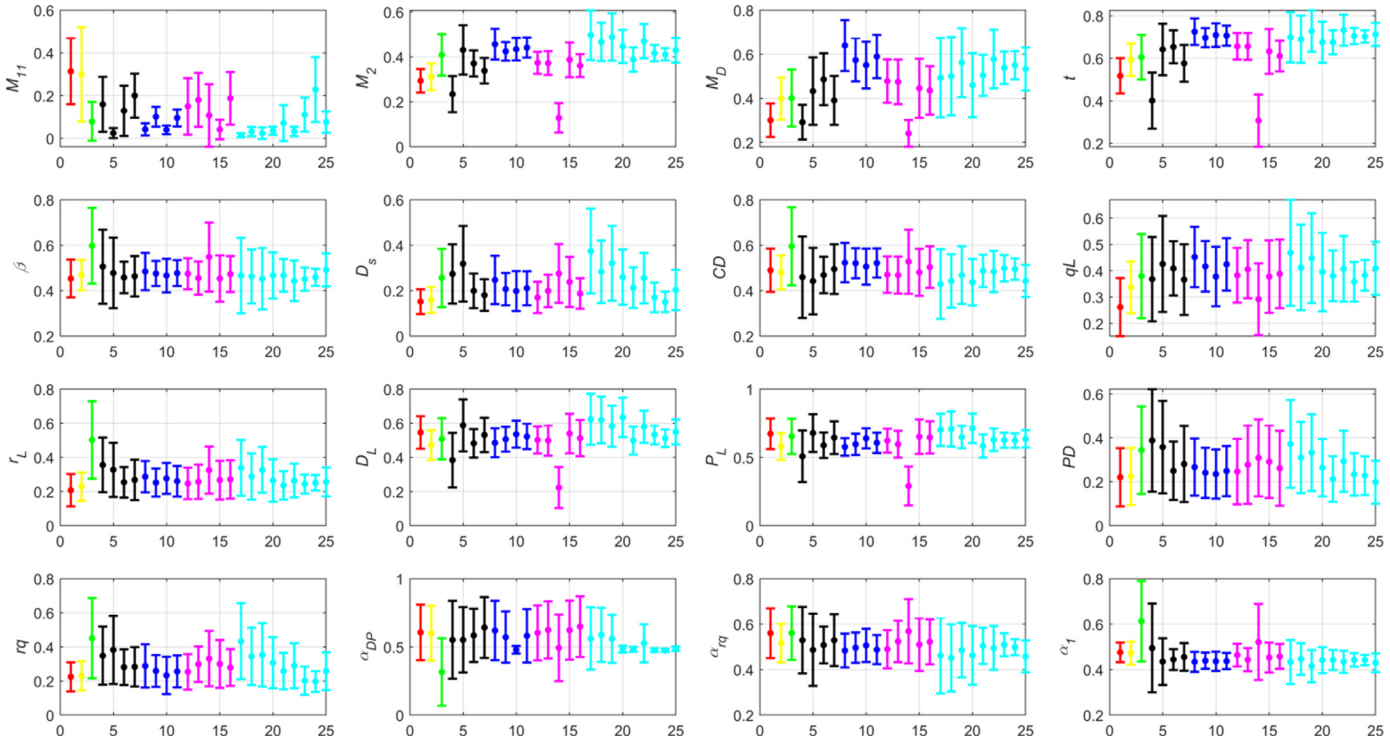

Fig. 3. PBPs of different microalgae. The x-axis is the number of mentioned microalgae.

the PBPs and LDA algorithm, LDA-PFPs can be found to better classify different microalgae than each PBP. These LDA-PFPs are combinations of PBPs that are sensitive to the special features between two different microalgae. Herein, N1 is used to compare with other each kind of microalgae in pairs using the LDA algorithm, and the inherent physical features of N1 are availably extracted. For convenience, the original 16-dimensional data is projected into the 1-dimensional subspace, so that each classification case of N1 and other microalgae has one corresponding LDA-PFP and DoS.

Further, we can calculate the normalized distributions of LDA-PFPs that can easily classify the two different microalgae, as shown in Figure 4. If N1 is compared with itself, the value of DoS is 0. Note that most microalgae can be separated with N1 after the simple linear projection, and most of the values of DoS are larger than 2. However, some microalgae may be confused with N1, such as N2 and N7, and the values of DoS are less than 1.5. It can be seen that the full width at half maximum of N1 in different projections will also have the high-variability. It implies that there are some important features lost when transforming from 16-dimensional PBP data to LDA-PFP. And it can be imagined that if the original 16-dimensional data is projected into the higher dimensional subspace, we will get better classification ability than LDA-PFPs.

Particularly, LDA-PFP has an explicit expression which is the linear combination of PBPs. Therefore, the absolute value of the coefficient of each PBP can be regarded as the weight of contribution to the special classification case. Thus, the weight of the LDA-PFP of N1 and other kind of microalgae can be sorted from the largest to the smallest to observe and compare the contribution of PBPs, which is collected for all cases and shown in Figure 5. Note that for each case, the weight is normalized such that the sum of all weights is 1.

From Figure 5, for different classification cases, the contribution of each PBP will be quite different. In some cases, like N1 vs. N3, N1 vs. N4, and N1 vs. N6, the highest weight is no more than 0.2, which indicates that many PBPs contribute to the classification. However, like N1 vs. N2, N1 vs. N5, N1 vs. N11, N1 vs. N20, and N1 vs. N22, there is an obvious key PBP whose weight is close to or larger than 0.3, and it indicates that the key PBP contributes much to the classification. In addition, although there is a key PBP in the case N1 vs. N2, the value of DoS is quite low, which is quite different from other cases. It indicates that the LDA algorithm may fail in this case. These results in Figure 5 would be further investigated in subsection 3.4.

3.3. Classification of N1 and each kind of microalgae by RF algorithm

The above result of LDA analysis reveals that the PFPs provided by LDA can effectively classify N1 from the 25 kinds of microalgae, and there are the key PBPs with larger weight in PFPs than other PBPs, which is encouraging for the search of key PBPs in the polarization fingerprint. Meantime, due to the intrinsic limitation of the LDA algorithm to find the non-linear relationship of PBPs, RF with more powerful classification ability than LDA, is adopted to analyze the dataset, including the possible complex cases, such as N1 vs. N2 and N1 vs. N7 in Figure 4. Note that the RF algorithm fails to easily obtain the PFPs with explicit analytic formula, but it can find the more useful PCMs for microalgae classification.

The RF algorithm is used to build the special PCMs for classifying N1 with another kind of microalgae. The classification accuracy of N1 and each of the other microalgae is calculated, and the corresponding VIM of PBPs can be sorted from highest to lowest. All classification cases are collected and shown in Figure 6. We use the classification accuracy to evaluate the classification performance which is shown in the title of each subplot in Figure 6. Notably, the classification accuracies of all cases are all higher than 98%. Even in the hard cases, such as N1 vs. N2 and N1 vs. N7, the PCMs based on RF can easily classify N1 and the other two respectively. It is noticeable that all PBPs contribute much to the classification between N1 and N2. Comparing the results in Figure 4 with Figure 6, the RF algorithm explores more potential differentiated features between N1 and N2 than the LDA algorithm, to ensure the excellent classification ability of N1 and N2. These results in Figure 6 indicate that the RF algorithm has the advantage to extract the characteristics of data and classify different microalgae cells.

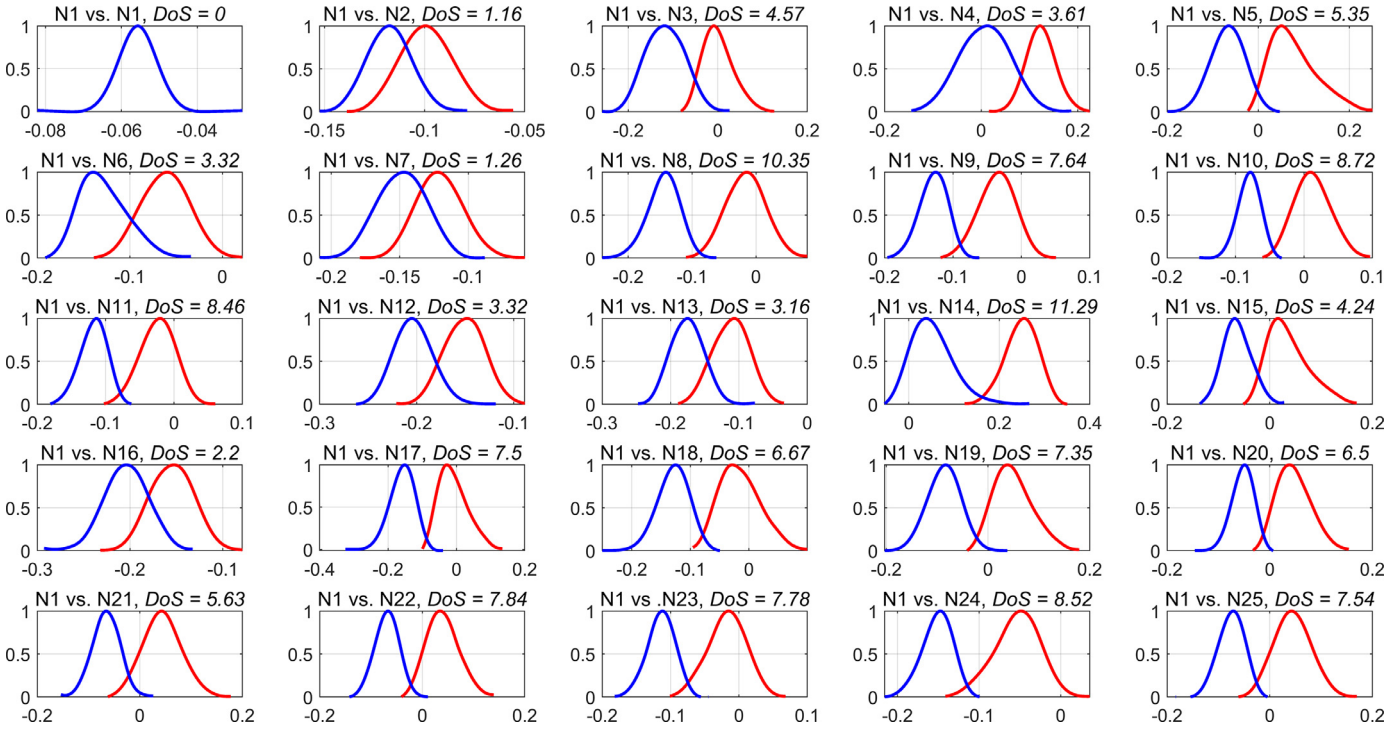

Fig. 4. LDA-PFP distributions of N1 (red line) and other microalgae (blue line). For each subplot, the horizontal axis is LDA-PFP, and the vertical axis is the normalized distribution.

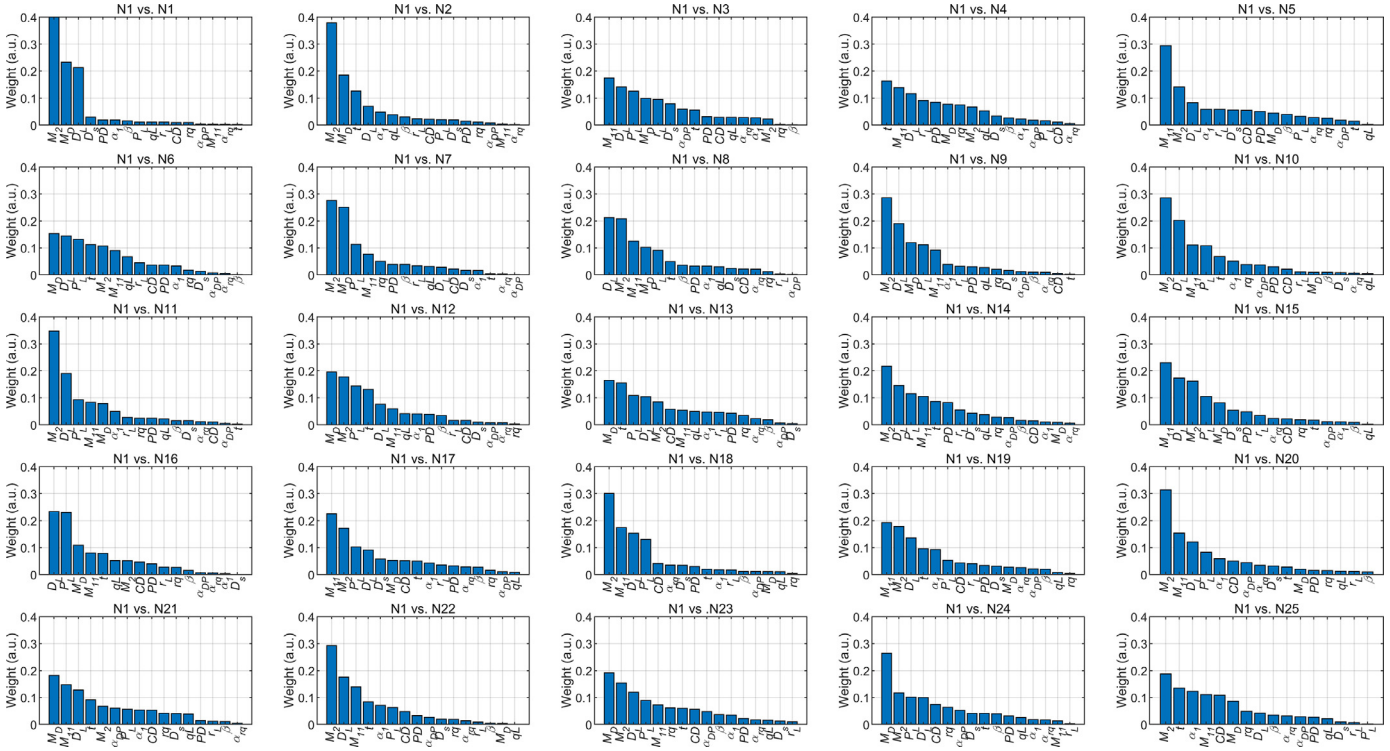

Fig. 5. Weights of PBPs in the LDA classification.

3.4. Feature PBPs in the classification of microalgae

In Figure 5, for each classification case, we consider PBPs ranking in the top five as the key PBPs of N1. Then, we can calculate the cumulative frequency of the key PBPs in Figure 5 for all classification cases and rank them in descending order, and further select the key PBPs whose frequencies are not less than 5. Particularly, we name these selected PBPs as the feature PBPs (f-PBPs) of N1. Similarly, we can get the f-PBPs by using the RF algorithm. The f-PBPs of N1 by LDA and RF algorithms are both collected as the second line in Table 3, where the numbers under the f-PBPs are the corresponding cumulative frequencies in descending order. Got by the LDA and RF algorithms, there are

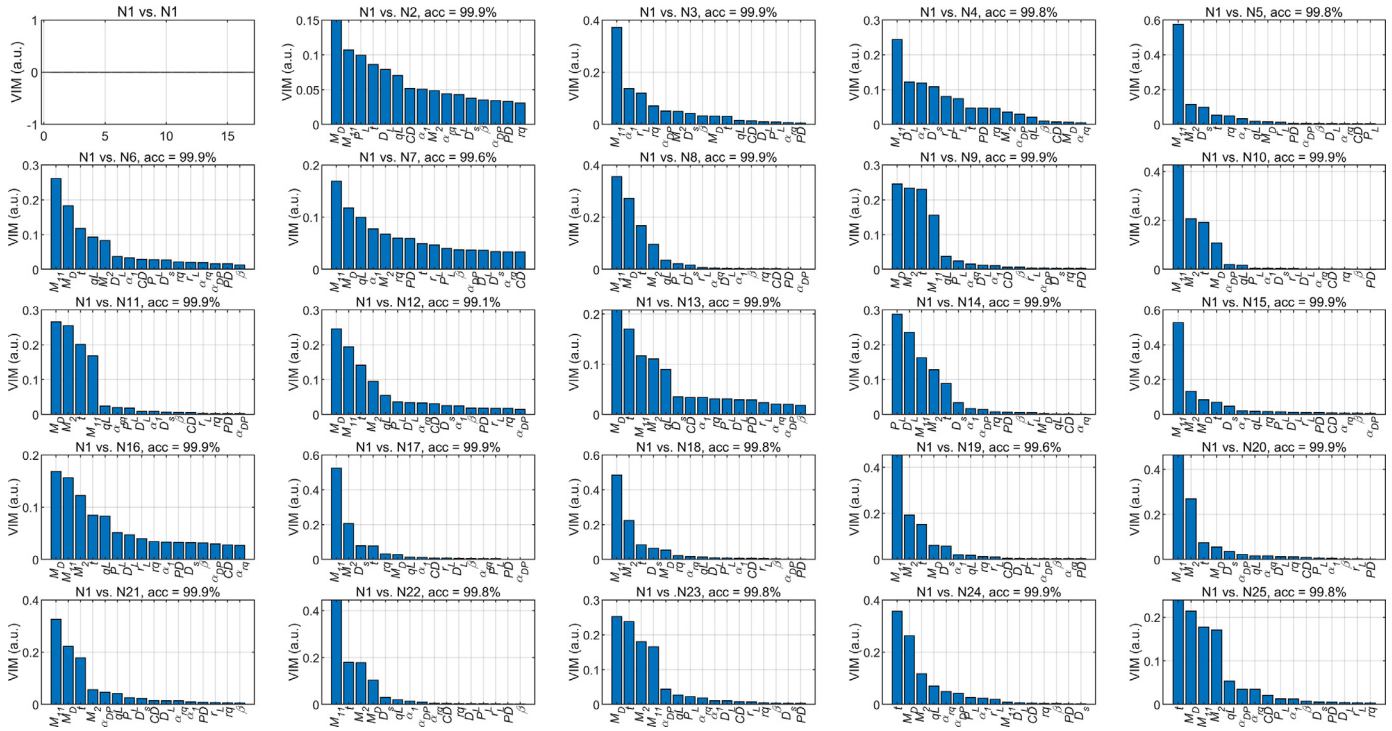

Fig. 6. VIM of PBPs in the RF classification.

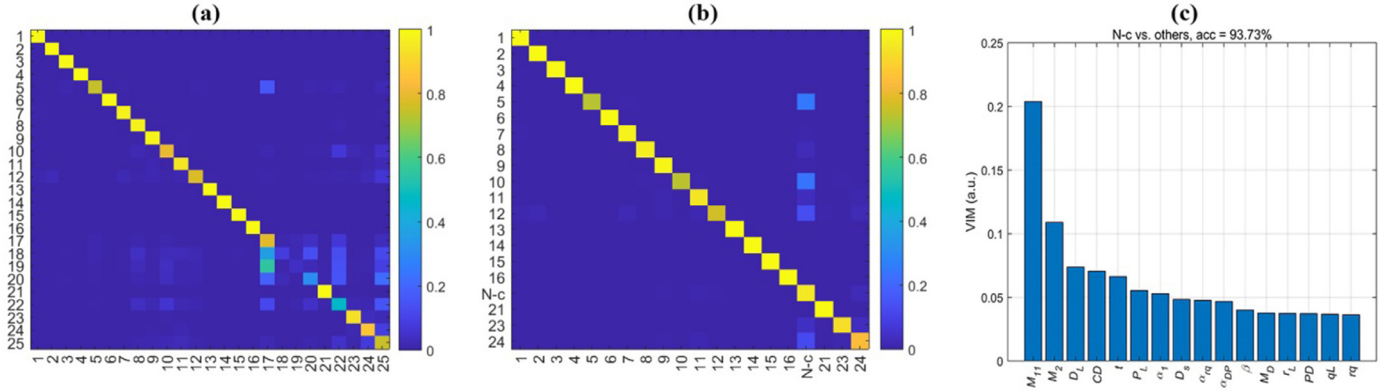

Fig. 7. (a) Recall of confusion matrix of 25 kinds of microalgae; (b) Recall of confusion matrix of 20 kinds of microalgae; (c) VIM of PBPs between N-c and the rest of the microalgae.

four f-PBPs (M_{11}, M_2, M_D, *and* t) that are the same although their frequencies are quite different. This indicates that these four f-PBPs are the intrinsic features that can classify N1 and the other microalgae. Recalling Figure 1 and Table 1, it can be understood that N1's optical features, such as the ability to scattered light, depolarization, and overall linear anisotropy, play a vital role in classifying the other microalgae. These features are believed to be related to the N1's physical properties, such as size, intracellular organelle, and microstructure. Considering the high classification accuracy in Figure 6, from the second line in Table 3, one can see that the f-PBPs can accurately classify N1 from the 25 kinds of microalgae. Also, we notice that the order and frequency of the f-PBPs are different and there are two other different PBPs, which could be explained by the different abilities of LDA and RF algorithms.

Similarly, we can get f-PBPs of the other kinds of microalgae by both LDA and RF algorithms, which are collected in Table 3. For each kind of microalgae, the f-PBPs by each algorithm can be used to classify this kind from the 25 kinds of microalgae. Also, we can find the f-PBPs regardless of the machine learning algorithms, and they are related to the intrinsic and exclusive physical properties of this kind of microalgae. Moreover, we find that the f-PBPs for different kinds of microalgae are quite different. In other words, the f-PBPs characterize the microalgae although they are got from the classification between the microalgae.

Also, we find that the 16 PBPs are all involved but diversely distributed in the f-PBPs of the 25 kinds of microalgae, which indicates that the differences of microalgae are diverse. Considering the high classification accuracy when we use the machine learning algorithms to get these f-PBPs, we can believe that these 16 PBPs are basically enough to be the polarization fingerprint of 25 kinds of microalgae. The polarization fingerprint is a metaphor for the human fingerprint in that it is inclusive of all features of microalgae but is exclusive to each kind of microalgae.

4. Discussions

4.1. Classification of 25 kinds of microalgae based on the RF algorithm

Hereinbefore, the PFPs and PCMs obtained by N1 and other kind of microalgae are introduced in detail. Together with the other pair-

Table 3
Cumulative frequency of f-PBPs in the pairwise classification.

No.	LDA	RF
N1	D_L, M_2, M_{11}, P_L, M_D, t, α_1 (23, 22, 19, 17, 15 12, 6)	M_{11}, M_2, t, M_D, q_L, D_s (23, 22, 21, 20, 11 7)
N2	M_2, P_L, α_1, t, D_L, M_{11}, M_D, rq (23, 19, 16, 14, 14, 10,10, 5)	M_{11}, M_2, t, M_D, α_1, D_s, D_L, P_L, rq, α_{DP} (21, 20, 16, 12, 9, 8, 8, 7, 7, 7)
N3	M_D, r_L, D_L, M_{11}, M_2, t, D_s, P_L, CD (19, 19, 18, 15, 10, 10, 8, 8, 6)	α_1, r_L, α_{DP}, M_{11}, t, M_D, β, rq, M_2 (23, 22,19, 18, 9, 8, 7, 6, 5)
N4	t, M_D, D_L, M_2, M_{11}, P_L, r_L, D_s, rq (24, 21, 18, 15, 13, 13, 6, 5, 5)	M_2, t, M_D, M_{11}, D_L, P_L, α_1 (22, 22, 19, 14, 14, 11, 11)
N5	M_2, M_{11}, M_D, t, D_L, P_L, CD, r_L, α_1 (25, 24, 20, 18, 11, 8, 5, 5, 5)	M_{11}, t, M_2, α_1, M_D, rq, P_L, D_s, D_L, α_{DP} (22, 14, 12, 12, 11, 11, 10, 8, 8, 8)
N6	P_L, M_2, t, D_L, M_D, M_{11}, CD, β, q_L, α_1 (21, 16, 15, 14, 13, 12, 7, 6, 6, 6)	M_{11}, M_2, t, P_L, D_L, α_{DP}, M_D, D_s (20, 17, 15, 14, 12, 10, 9, 5)
N7	M_2, P_L, M_{11}, t, D_L, M_D, CD, rq, α_1 (19, 18, 17, 16, 16, 15, 6, 6, 5)	M_{11}, t, M_2, M_D, α_{DP}, D_L, P_L, D_s, rq, α_1 (23, 21, 18, 15, 9, 7, 6, 5, 5, 5)
N8	M_{11}, M_D, M_2, D_L, P_L, CD, β, t, α_1 (19, 17, 14, 14, 14, 9, 8, 7, 6)	M_D, M_{11}, P_L, M_2, D_L, t, α_{DP}, CD (21, 17, 15, 13, 13, 11, 9, 5)
N9	M_2, M_D, D_L, P_L, M_{11}, CD, t, β, r_L (19, 17, 14, 14, 13, 12, 10, 8, 7)	M_{11}, M_2, M_D, t, α_1, D_L, P_L, rq, α_{DP}, CD (22, 17, 15, 11, 9, 8, 7, 7, 7, 6)
N10	M_{11}, t, D_L, M_2, P_L, M_D, CD, α_{DP} (21, 17, 17, 16, 16, 11, 7, 7)	M_{11}, α_{DP}, M_2, t, M_D, P_L, rq, α_{rq}, D_L (20, 20, 15, 12, 11, 9, 7, 6, 5)
N11	M_2, M_D, D_L, M_{11}, P_L, β, t, CD, r_L (21, 20, 16, 14, 12, 10, 9, 9, 6)	M_{11}, M_2, M_D, t, P_L, α_{DP}, α_1, D_L, rq (22, 17, 17, 12, 10, 8, 8, 6, 5)
N12	M_2, t, M_D, α_1, P_L, M_{11}, D_L, CD (23, 19, 16, 14, 12, 11, 10, 5)	M_{11}, M_2, t, D_s, α_1, α_{DP}, M_D, D_L, P_L (21, 18, 14, 13, 13, 9, 8, 7, 7)
N13	M_2, P_L, t, M_{11}, M_D, D_L, CD, α_{rq} (18, 18, 17, 16, 15, 12, 6, 5)	M_{11}, M_2, t, M_D, α_{DP}, P_L, rq, D_L, α_{rq}, D_s, α_1 (18, 18, 15, 10, 10, 8, 8, 7, 7, 6, 6)
N14	M_D, P_L, D_L, M_2, t, PD, M_{11} (24, 24, 23, 22, 14, 9, 5)	M_2, D_L, P_L, t, M_D, M_{11} (24, 24, 24, 23, 20, 5)
N15	M_2, M_{11}, M_D, D_L, t, P_L, CD (22, 20, 18, 18, 14, 8, 7)	M_{11}, M_2, t, M_D, P_L, α_{DP}, D_s, D_L, α_1 (21, 18, 17, 12, 12, 10, 9, 8, 6)
N16	t, M_2, D_L, M_{11}, P_L, M_D, CD, rq, α_1 (20, 18, 16, 15, 13, 10, 10, 6, 5)	t, M_{11}, M_2, M_D, α_{DP}, P_L, α_1, D_s, rq, CD (21, 19, 18, 13, 10, 8, 8, 6, 6, 5)
N17	M_2, M_{11}, P_L, t, M_D, D_s, D_L, CD (25, 24, 18, 11, 10, 9, 9, 8)	M_{11}, D_s, M_2, P_L, rq, D_L, t, α_1, M_D, CD, α_{DP} (23, 17, 13, 12, 12, 9, 7, 7, 6, 6, 5)
N18	M_2, M_{11}, P_L, D_L, M_D, t, CD, q_L (24, 22, 22, 13, 12, 10, 9, 5)	P_L, M_{11}, M_2, D_L, M_D, α_{DP}, t, α_{rq}, D_s, α_1, rq (20, 19, 13, 13, 11, 9, 8, 8, 7, 6, 5)
N19	M_2, M_{11}, α_1, M_D, t, r_L, P_L, D_L (23, 22, 20, 11, 10, 10, 9, 8)	M_{11}, M_2, D_s, α_1, M_D, t, P_L, rq, α_{DP}, D_L (22, 16, 13, 13, 12, 12, 8, 7, 7, 5)
N20	M_2, M_{11}, D_L, P_L, α_{DP}, CD, M_D, t, α_1, q_L (22, 21, 15, 12, 11, 10, 9, 9, 6, 5)	M_{11}, D_L, P_L, α_{DP}, M_2, M_D, t, α_{rq}, α_1 (21, 20, 17, 15, 13, 11, 5, 5, 5)
N21	P_L, M_2, M_{11}, M_D, D_L, t, CD, α_{DP} (17, 16, 15, 15, 15, 12, 9, 8)	M_{11}, M_2, P_L, M_D, t, α_{DP}, D_L (19, 17, 17, 15, 14, 14, 6)
N22	M_{11}, M_2, D_L, t, M_D, CD, P_L, q_L, α_1 (20, 20, 16, 14, 11, 9, 9, 7, 6)	M_{11}, M_2, t, M_D, α_{DP}, D_L, P_L, α_1, rq, α_{rq} (20, 16, 16, 12, 12, 8, 7, 6, 5, 5)
N23	M_2, t, rq, M_D, P_L, M_{11}, D_L, β, α_{DP}, α_{rq} (20, 16, 14, 12, 11, 10, 10, 6, 5,5)	M_{11}, α_{DP}, t, rq, M_2, M_D, P_L, D_L, q_L, r_L, α_{rq} (17, 14, 13, 13, 11, 11, 8, 7, 6, 6, 6)
N24	M_2, t, M_D, P_L, rq, M_{11}, D_L, q_L, α_{DP} (21, 17, 16, 12, 12, 11, 8, 6, 5)	M_{11}, rq, M_2, t, D_s, α_{DP}, M_D, α_{rq}, α_1 (17, 17, 15, 13, 12, 11, 9, 9, 6)
N25	D_L, t, CD, M_2, M_{11}, α_1, M_D, P_L, PD, β (18, 15, 15, 14, 13, 13, 9, 7, 7, 5)	M_{11}, α_{DP}, t, P_L, M_2, CD, M_D, α_{rq}, D_L, PD, α_1 (19, 18, 13, 12, 11, 11, 9, 9, 6, 6, 5)

wise classification of the microalgae, they show the advantages of the polarization fingerprint in microalgae classification. The f-PBPs shown in Table 3 reveal the unique or exclusive physical properties of each kind of microalgae. However, it would be a concern whether the polarization fingerprint would work or not if putting these data of the 25 kinds of microalgae together. It is an ideal case to classify them at once, which is much more difficult than pairwise classification. Apparently, the machine learning algorithms and the settings play a vital role in the multiple microalgae classifications.

For an easier comparison, a new PCM based on the RF algorithm is built to classify all 25 kinds of microalgae together using 16 PBPs, and each kind has similar samplings. The recall of the confusion matrix is shown in Figure 7(a), which is generally considered a performance measurement for the machine learning classification [32]. The average accuracy of the confusion matrix is 83.13%, and we can find that most of the microalgae can be classified from each other. Most kinds of microalgae can be accurately classified. In addition, this PCM accurately distinguishes those microalgae that are very similar in shape and size, such as N9, N10, and N11.

However, some kinds of microalgae (N17 to N20, denoted as N17-20) can't be well recognized, and the classification accuracy of N19 in the PCM is lower than 10%. Note that, the microalgae in N17-20 are different species but the same genus, whose microscopic photos in Figure 1 look similar in both size and shape. Recalling that the average accuracy of pairwise classification of N17-20 using the RF algorithm can be higher than 73%, the bad performance may be mainly attributed to the machine learning algorithm. And in other words, the current polarization fingerprint may need more additional PBPs to better the PCM.

Compared with Figure 1, we can find that the M_2 and D_L of N17-20 are relatively higher, which may be caused by the relatively higher polarization-maintaining and linear diattenuation ability. In addition, the α_{rq} of N17-20 are related relatively lower, which may result from their multiple and long flagella. However, the current PBPs don't enlarge the difference between N17-20, which may lead to the bad performance of the present PCM.

Except for the kinds from N17-20, the classification accuracies of N22 and N25 (Bacillariophyta) are also relatively low. We combine them as one category (kind) and name them N-c (N17-20, N22, and N25).

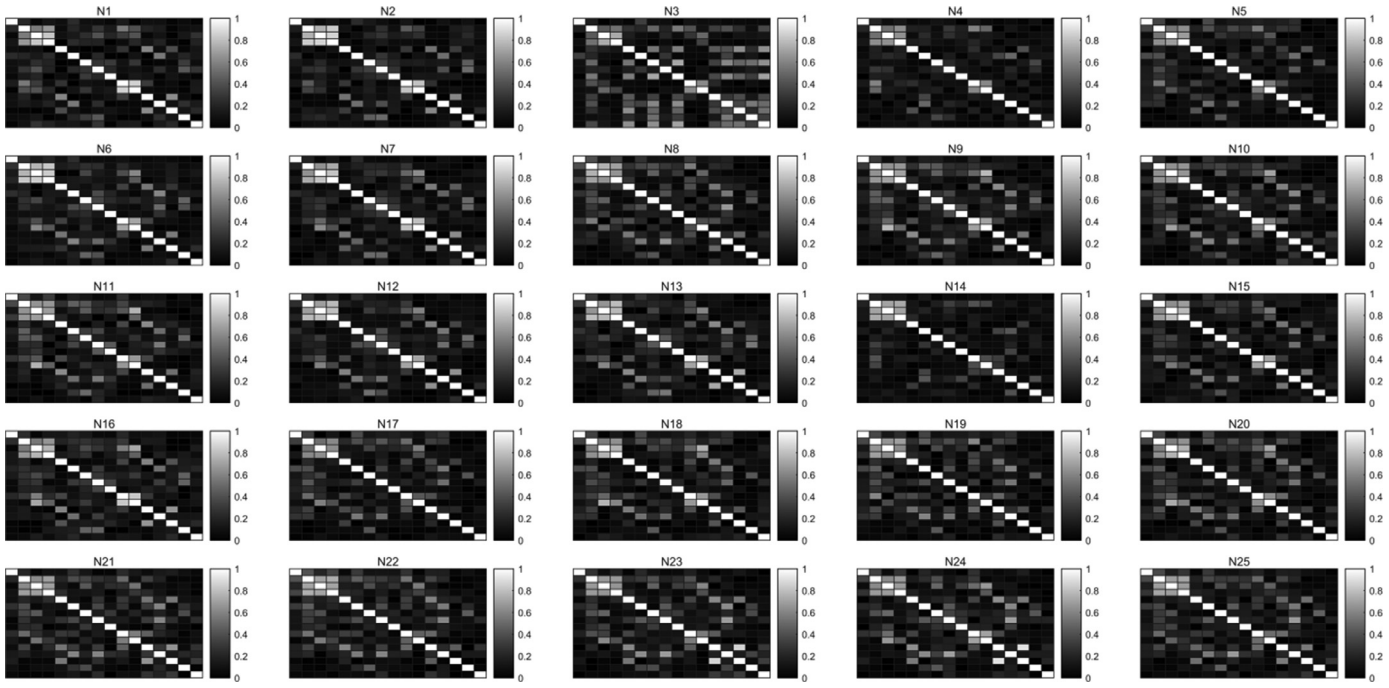

Fig. 8. Correlation maps for microalgae. The same order of PBPs, from top to bottom and from left to right: M_{11}, M_2, M_D, t, β, D_s, CD, q_L, r_L, D_L, P_L, DP, rq, α_{DP}, α_{rq}, α_1.

Then, the PCM is renewed using the RF algorithm and the current PBPs, and the confusion matrix is shown in Figure 7(b). Surprisingly, the notable confused area in Figure 7(a) disappears, and the PCM's average accuracy grows up to 94.59%, which is much higher than the previous one.

Further, we regard N-c as the first category and the rest of the microalgae as the second category. We also build a new PCM using all PBPs of the polarization fingerprint and RF algorithm to classify those two categories and the average accuracy is 93.73%. This result looks wired but reveals the necessity of fine classification to find the full information included in the sample group, which depends on the inclusiveness and exclusiveness of the polarization fingerprint. Further, we obtain the VIM of PBPs, as shown in Figure 7(c). Notably, all the sixteen PBPs of the polarization fingerprint contribute to the classification, which means that all potential features have been fully extracted by the PCM to achieve the classification, and also implies that the numbers of PBPs in the current polarization fingerprint need to be increased.

4.2. Linear correlation between different PBPs

The current polarization fingerprint consists of sixteen PBPs. One may wonder whether they are independent of each other or somehow correlated. To investigate this, the correlation coefficient (r) defined as Eq. (6) is usually used to measure the linear correlation between two variables. Then r is calculated to study the degree of linear correlation between different PBPs, and the results are shown in Figure 8. Generally, the larger r of two variables represents the higher linear correlation.

$$r(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}} \tag{6}$$

where $\mathrm{Cov}(X,Y)$ is the covariance of X and Y, $\mathrm{Var}[X]$ is the variance of X, and $\mathrm{Var}[Y]$ is the variance of Y.

In Figure 8, for each kind of microalgae, we calculate r for each PBPs pair and form a correlation map, for which PBPs from the left to right are M_{11}, M_2, M_D, t, β, D_s, CD, q_L, r_L, D_L, P_L, DP, rq, α_{DP}, α_{rq}, α_1, and those from the top to bottom are in the same order. In each map, white grids are with a strong correlation, and dark grids are those with a weak correlation. The correlation map is symmetric due to the definition of r as Eq. (6). Most PBPs have a low correlation to each other; meanwhile, some PBPs, such as M_2, M_D, and t, have a relatively higher correlation between them than others. Also, there are still some gray grids here or there, which indicates the two related PBPs have some kinds of correlation.

The correlation maps change with the different microalgae, except for M_2, M_D, and t. It means that these three in principle are related to the similar physical properties of the microalgae, regardless of the kinds of the microalgae. For example, from Table 2, both M_2 and M_D are related to the depolarization which could be caused the complexity of intracellular microstructure. Meanwhile, most microalgae have different correlation maps from each other. The D_L and P_L of N16 have a strong linear correlation (r >0.8), while D_L and P_L of N17 are weakly correlated (r <0.5). For the correlation map of N14, most r except the diagonal pixels are lower than 0.4. Contrastively, the correlation map of N3 seems complicated since bright pixels distribute dispersedly. These results mean that the polarization fingerprint can effectively characterize the different features of microalgae.

Further, to evaluate the overall correlation degree of each PBP with the other PBPs in the polarization fingerprint, the variance inflation factor (VIF) is calculated using Eq. (7) [33]. For each given PBP, the large VIF represents the great possibility of collinearity between it and the other PBPs. And, if the VIF of PBP is larger than 1, we consider that the PBP has a strong correlation with the other PBPs.

$$\mathrm{VIF} = \log_{10}\left(\frac{1}{1-R_i^2}\right) \tag{7}$$

where R_i^2 is the determination coefficient calculated by the square of the multi-correlation coefficient of regression analysis between i-th variable and other variables.

The VIF of each PBP for all microalgae can be observed in Figure 9. Notably, M_2, M_D, and D_L seem much easier to be collinear than other PBPs, since the VIF larger than 1 of M_2, M_D, and D_L are respectively for 21, 19, and 13 kinds of microalgae. It can be explained by the correlation map in Figure 8. The M_2, M_D, and t have a relatively large correlation to each other, so their VIF are large; D_L and P_L are correlated to each

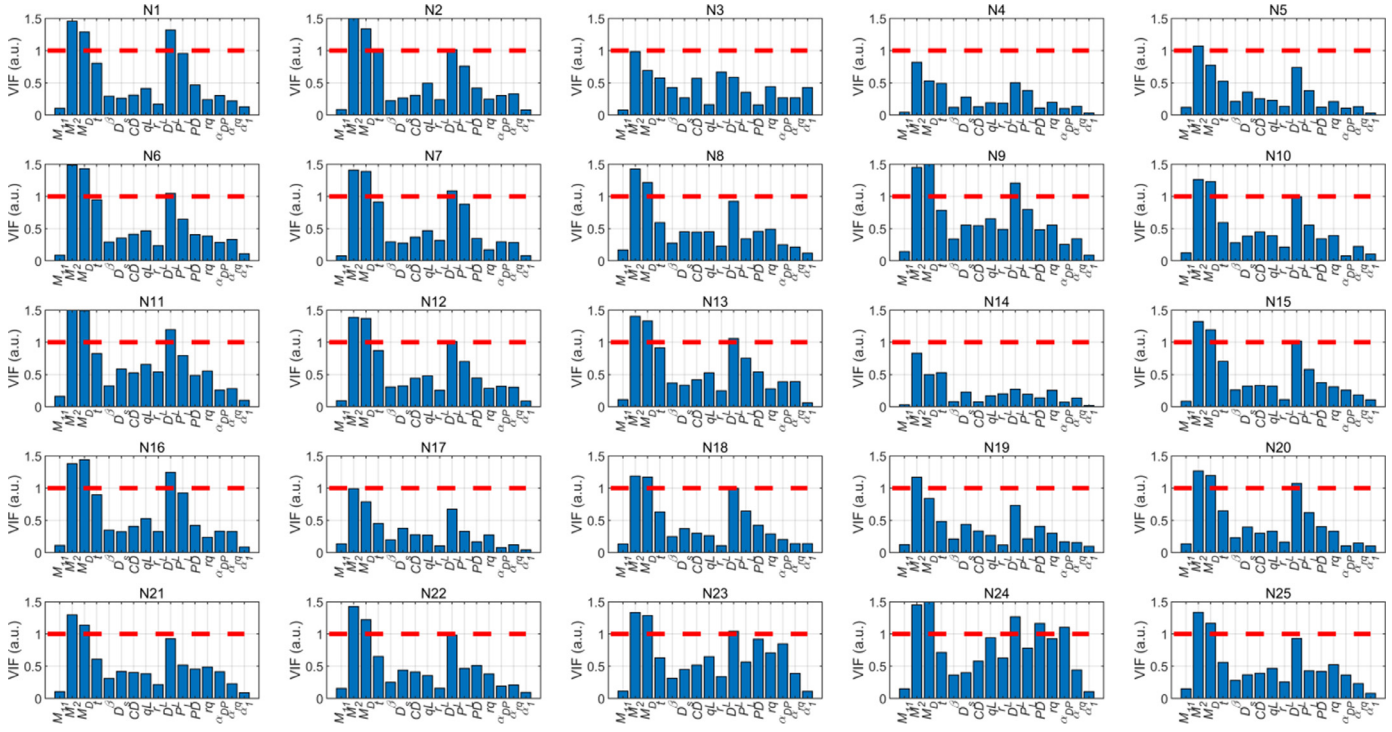

Fig. 9. VIF of each PBP of different microalgae.

other for some microalgae, so the VIF of D_L is large. However, the VIF of t and P_L are relatively lower than the other three, although they are still larger than the others. Generally, the PBPs with large VIF are not good for the polarization fingerprint, since they may be redundant and would weaken the polarization fingerprint's exclusiveness. However, combining the physical meaning of PBPs in Table 2 and Figure 9, the VIF of each PBP can change for different microalgae. Even among M_2, M_D, and D_L, they also change relatively for different microalgae, which can indicate that these PBPs also have their specific contributions to microalgae classification.

Also, the microalgae have different characteristics in VIF. Some microalgae have three PBPs whose VIF are much larger than 1, such as N1, N9, N11, and N16, while the number for N24 is five. On the contrary, some microalgae have no parameters whose VIF are larger than 1, such as N4 and N14. Combined with the microscopic photos in Figure 1, the more PBPs with the large VIF (>1), the simpler or more specific the microalgae's microstructure. For N24, there is an obvious intracellular lipid droplet (white) in the cell; and for N9, there is a clear (dark or brown) block between the chloroplast in the cell. In these cases, the polarization fingerprint is somehow redundant since their unique microstructures dominate their classification with others. And on the contrary, for N14, the intracellular microstructure is quite complex, besides the pyrenoid located in the center volume. Those interrelationships are potentially important information to classify different microalgae.

Each microalgae cell in the aquatic environment has various connections with the surrounding media, such as physical interaction, and chemical or biological exchanges. And its physical properties, such as size, shape, and intracellular microstructure, are firstly determined by its DNA, but also are subjected to the media, such as the flow, pressure, temperature, salinity, light illumination, nutrients, and so on. Therefore, the cell itself is quite complex in physical properties, which leads to the difficulty of characterization of the individual cell by optical quantities until today. In this work, we firstly take the advantage of the Mueller matrix that has sixteen elements which are sensitive to the cell's physical properties, especially the microstructure in the cell. Meanwhile, the concept of polarization fingerprint allows us to collect multiple PBPs (as the function or derivation of the Mueller matrix elements) together to characterize the microalgae cells. The results indicate that the polarization fingerprint as a whole works well to classify the microalgae.

Honestly, PBPs or parameters selected for the fingerprint are not limited to the current version, whose numbers of parameters are not minimal; and adding new ones or replacing some are encouraged if necessary. Note that the current polarization fingerprint is proven to work for the classification of the current 25 kinds of microalgae. If the total microalgae numbers or the members change, the polarization fingerprint would be adjusted on the basis of the current version.

In this work, we secondly take the advantage of the machine learning algorithms which can explore the potential of the polarization fingerprint as possible to accurately classify different microalgae. Obviously, different algorithms show their different ability, advantages, and disadvantages. It is easy to imagine that the more powerful and sophisticated algorithms would be more helpful to select the PBPs to the polarization fingerprint and also be more accurate to classify the microalgae, depending on the time efficiency or cost constraint.

Summarily, the concept of the polarization fingerprint enabled by the particulate Mueller matrix polarimetry is proven as a powerful tool to accurately classify the microalgae, with the aid of machine learning algorithms. These give an insight into microalgae classification and helps pave the way to apply PMP in aquatic environment monitoring. In the future, we believe that the polarization fingerprint would be improved to be much more stable, powerful, and feasible for popular utilization. On the other hand, the polarization fingerprint can be used to promote the rapid, accurate, and in-situ microalgae classification in environmental engineering, which is desired now in aquatic environment monitoring and microalgae resource exploitation.

4.3. *The detected angles of polarized light scattering*

The measurement of polarized light scattering at 120° has been proven that can classify many kinds of suspended particle [18], and can quantitatively characterize the microstructures changes of particles [9–10]. Therefore, the presented results in this paper are performed for the polarized light scattering at 120° to focus on the exploration of the novel concept. According to some previous simulations and experimen-

tal results [34–35], the different detected angles of polarized scattering light can have the advantages in the description of suspended particles. The concept of polarization fingerprint is presented, and its classification ability of microalgae has been shown in Figure 7. It is promising that we can simultaneously measure the Mueller matrix at multiple angles of the scattered light in the future [36]. Further, we can obtain the abundant high-dimensional data of the individual particle, which enables us to accurately classify and analyze suspended particles.

Conclusions

The accurate classification of microalgae is important to aquatic environment monitoring but is still challenging for the scientific community. In this work, a concept of the polarization fingerprint is proposed and proven to be a powerful tool to classify different microalgae, with the aid of machine learning algorithms. The Mueller matrix elements of 25 kinds of microalgae are measured by the particulate Mueller matrix polarimetry and their PBPs are calculated. Sixteen selected PBPs with explicit physical meanings and associations to the structural properties are selected to form the polarization fingerprint. The analysis and experiment results show that the polarization fingerprint includes the most useful features of microalgae to classify microalgae effectively. Besides the inclusiveness and exclusiveness, the polarization fingerprint is also extensive and flexible by following the selection method given in this paper. Due to the power of polarization fingerprint to classify the microalgae, in the future, the polarization fingerprint would promote the in-situ monitoring of microalgae in the aquatic environments.

Funding

This research was funded by Guangdong Development Project of Science and Technology (2020B1111040001); National Key Research and Development Program of China (2018YFC1406600); National Natural Science Foundation of China (NSFC) (41527901, 61975088); Shenzhen-Hong Kong Joint Project (SGDX20201103095403017); Shenzhen Key Laboratory of Marine IntelliSense and Computation under Contract (ZDSYS20200811142605016).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Jiajin Li: Investigation, Writing – original draft, Formal analysis, Methodology, Validation. **Jinfu Wei:** Formal analysis, Validation. **Hongyuan Liu:** Investigation, Validation. **Jiachen Wan:** Investigation, Validation. **Tongyu Huang:** Investigation, Validation. **Hongjian Wang:** Investigation, Validation. **Ran Liao:** Supervision, Resources, Methodology, Writing – review & editing. **Hui Ma:** Supervision, Resources.

Data availability

Data will be made available on request.

References

[1] Yan Z, Shen T, Li W, Cheng W, Wang X, Zhu M, et al. Contribution of microalgae to carbon sequestration in a natural karst wetland aquatic ecosystem: An in-situ mesocosm study. Sci Total Environ 2021;768(10):144387.
[2] Wang H, Li J, Liao R, Tao Y, Peng L, Li H, et al. Early warning of cyanobacterial blooms based on polarized light scattering powered by machine learning. Measurement 2021;184:109902.
[3] Apostolopoulou NG, Smeti E, Lamorgese M, Varkitzi I, Whitfield P, Regnault C, et al. Microalgae show a range of responses to exometabolites of foreign species. Algal Res 2022;62:102627.
[4] Jin D, Hoagland P, Buesseler KO. The value of scientific research on the ocean's biological carbon pump. Sci Total Environ 2020;749(20):141357.
[5] Hu L, Zhang X, Xiong Y. Contribution of submicron particles to the unpolarized and linearly polarized angular scattering. Front Remote Sens 2022;3:925654.
[6] Liu F, Zhang C, Wang Y, Chen G. A review of the current and emerging detection methods of marine harmful microalgae. Sci Total Environ 2022;815:152913.
[7] Arora M, Sahoo D. Concepts and techniques for the study of algae. In: Sahoo D, Seckbach J, editors. The Algae World. Netherlands, Dordrecht: Springer; 2015. p. 519–53.
[8] Barsanti L, Birindelli L, Gualtieri P. Water monitoring by means of digital microscopy identification and classification of microalgae. Environ Sci: Processes Impacts 2021;23:1443–57.
[9] Li J, Zou C, Liao R, Peng L, Wang H, Guo Z, et al. Characterization of intracellular structure changes of *Microcystis* under sonication treatment by polarized light scattering. Biosens - Basel 2021;11(8):279.
[10] Li J, Liao R, Tao Y, Zhuo Z, Liu Z, Deng H, et al. Probing the cyanobacterial *Microcystis* gas vesicles after static pressure treatment: a potential in situ rapid method. Sens -Basel 2020;20(15):4170.
[11] Otálora P, Guzmán JL, Acién FG, Berenguel M, Reul A. Microalgae classification based on machine learning techniques. Algal Res 2021;55:102256.
[12] Bohren CF, Huffman DR. Absorption and Scattering of Light by Small Particles. New York: Wiley; 1983.
[13] He C, He H, Chang J, Chen B, Ma H, Booth MJ. Polarisation optics for biomedical and clinical applications: a review. Light Sci Appl 2021;10:194.
[14] Liu Z, Liao R, Ma H, Li J, Leung PTY, Yan M, et al. Classification of marine microalgae using low-resolution Mueller matrix images and convolutional neural network. Appl Optics 2020;59(31):9698–709.
[15] Lu SY, Chipman RA. Interpretation of Mueller matrices based on polar decomposition. J Opt Soc Am A-Opt Image Sci Vis 1996;13(5):1106–13.
[16] He H, Liao R, Zeng N, Li P, Chen Z, Liu X, et al. Mueller matrix polarimetry—an emerging new tool for characterizing the microstructural feature of complex biological specimen. J Lightwave Technol 2019;37(11):2534–48.
[17] Zhu Y, Dong Y, Yao Y, Si L, Liu Y, He H, et al. Probing layered structures by multi-color backscattering polarimetry and machine learning. Biomed Opt Express 2021;12(7):4324–39.
[18] Li J, Liao R, Guan C, Wang H, Zhuo Z, Zeng Y, et al. Particulate Mueller matrix polarimetry. Opt Laser Technol 2023;158:108780.
[19] Li P, Dong Y, Wan J, He H, Aziz T, Ma H. Polaromics: deriving polarization parameters from a Mueller matrix for quantitative characterization of biomedical specimen. J Phys D-Appl Phys 2022;55(3):034002.
[20] Gil JJ. Invariant quantities of a Mueller matrix under rotation and retarder transformations. J Opt Soc Am A 2016;33(1):52–8.
[21] Gil JJ, Bernabeu E. A depolarization criterion in mueller matrices. Opt Acta: Int J Opt 1985;32(3):259–61.
[22] Qi J, He H, Ma H, Elson DS. Extended polar decomposition method of Mueller matrices for turbid media in reflection geometry. 2017;42(20):4048-51.
[23] Ossikovski R, Vizet J. Polar decompositions of negative-determinant Mueller matrices featuring nondiagonal depolarizers. Appl Optics 2017;56(30):8446–51.
[24] Li P, Lv D, He H, Ma H. Separating azimuthal orientation dependence in polarization measurements of anisotropic media. Opt Express 2018;26(4):3791–800.
[25] Phan Q, Lo Y. Differential Mueller matrix polarimetry technique for non-invasive measurement of glucose concentration on human fingertip. 2017;25(13):15179-87.
[26] Arteaga O, Garcia-Caurel E, Ossikovski R. Anisotropy coefficients of a Mueller matrix. J Opt Soc Am A 2011;28(4):548–53.
[27] Yun T, Zeng N, Li W, Li D, Jiang X, Ma H. Monte Carlo simulation of polarized photon scattering in anisotropic media. Opt Express 2009;17(19):16590–602.
[28] He H, Sun M, Zeng N, Du E, Liu S, Guo Y, et al. Mapping local orientation of aligned fibrous scatterers for cancerous tissues using backscattering Mueller matrix imaging. J Biomed Opt 2014;19(10):106007.
[29] Martinez AM, Kak AC. PCA versus LDA. IEEE Trans Pattern Anal Mach Intell 2002;23(2):228–33.
[30] Goldstein BA, Polley EC, Briggs FB. Random forests for genetic association studies. Stat Appl Genet Mol Biol 2011;10(1):32.
[31] Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources, and a solution. BMC Bioinformat 2007;8:25.
[32] Orenstein EC, Ayata S, Maps F, Becker EC, Benedetti F, Biard T, et al. Machine learning techniques to characterize functional traits of plankton from image data. Limnol Oceanogr 2022;67(8):1647–69.
[33] García CB, García J, López Martín MM, Salmerón R. Collinearity: revisiting the variance inflation factor in ridge regression. J Appl Stat 2015;42(3):648–61.
[34] Guo W, Zeng N, Liao R, Xu Q, Guo J, He Y, et al. Simultaneous retrieval of aerosol size and composition by multi-angle polarization scattering measurements. Opt Laser Eng 2022;149:106799.
[35] Yuan X, Song J, Zeng N, Guo J, Ma H. Correlation analysis and application investigation of multi-angle simultaneous polarization measurement data and concentration of suspended particulate matter in the atmosphere. Front Environ Sci 2022;10:1031863.
[36] Chen Y, Wang H, Liao R, Li H, Wang Y, Zhou H, et al. Rapidly measuring scattered polarization parameters of the individual suspended particle with continuously large angular range. Biosens -Basel 2022;12:321.