# Application of Deep (Machine) Learning for Phytoplankton Identification Using Microscopy Images

Arief Rachman[1,*] Aulia Salsabella Suwarno[2] Susanna Nurdjaman[3]

[1] *Biological Oceanography Research Group, Research Center for Oceanography, National Research and Innovation Agency (RCO-BRIN), Indonesia*
[2] *Faculty of Earth Science and Technology, Bandung Institute of Technology (ITB), Indonesia*
[3] *Department of Oceanography, Faculty of Earth Sciences and Technology, Bandung Institute of Technology (ITB), Indonesia*
[*] *Corresponding author. Email: arief_rachman1987@yahoo.com*

**ABSTRACT**

As a hot spot of marine diversity, between 150 – 400 phytoplankton species have been reported in various Indonesian marine ecosystems. However, phytoplankton identification in Indonesia is mainly made manually by a human expert, which is a time-consuming process with many limitations. Thus, this study aimed to develop automatic phytoplankton identification using Deep Machine Learning algorithms, such as Convolutional Neural Networks (CNNs), to help the identification process of the Indonesian phytoplankton. A pre-trained VGG-16 model was used to build a CNN model to identify phytoplankton up to genus level under five different model scenarios (S) based on curated phytoplankton images from the Plankton Image Database of RCO-BRIN. The cross-entropy loss analysis and confusion matrix showed the simple model (S1) and genus-level model (S4) have the best performance with low classification errors. In the application trial, the S1 model could differentiate diatoms and dinoflagellates group with up to 78% accuracy, while the S4 model could differentiate the target genus of *Ceratium*, *Chaetoceros*, *Coscinodiscus*, *Protoperidinium*, and *Rhizosolenia* up to 79% accuracy. However, the S4 model suffers from forced classification problems due to its inability to identify images of any non-target genus. Unfortunately, the S5 model created to solve the S4 problems has a much lower accuracy at 54% due to highly diverse data stored in the 'Others' category, which confuses the model. Although the CNNs models in this study can automatically identify phytoplankton up to genus level at accuracy >75%, the current limitations in all scenarios need to be solved before the model can be used in a real-world research scenario.

***Keywords:*** *Computer vision, CNNs, Phytoplankton diversity, Taxonomic identification, VGG-16*

## 1. INTRODUCTION

Phytoplankton is the most important foundation of life in the aquatic ecosystems and an essential component in energy transfer in the aquatic food web. Due to its rapid response and high sensitivity to any changes in the aquatic environment, many phytoplankton species have been used as indicators for the health of the environment. The cell shape, growth rate, and biochemical components, such as lipids, pigment, and other secondary metabolites, could signal to many anomalies in the water, such as eutrophication. Furthermore, some phytoplankton tends to grow rapidly under anomalous water condition and cause harmful algal blooms (HABs) events, which could cause a devastating impact on the ecosystem, particularly if the bloomed species was the toxin producers [1]. About 200 taxa of phytoplankton were estimated, which includes dinoflagellates, cyanobacteria, diatoms, raphidophytes, dictyochophytes, pelagophytes, and haptophytes, could produce a harmful toxin that could threaten the health and economy of many human societies in the coastal and inland water areas [2]. HABs events have increased incidence, duration, and frequency over the last few decades from the 1980s. Due to the many combined effects of anthropogenic activities along the world's coastlines and climate change, including ocean warming, acidification, and hypoxia [2,3]. Currently, between 20 to over 80 cases of

Diarrhetic Shellfish Poisoning (DSP) caused by *Dinophysis* species, Amnestic Shellfish Poisoning (ASP) caused by *Pseudo-nitszchia* species, and Paralytic Shellfish Poisoning (PSP) caused by *Alexandrium* species, have been reported for every year around the world [2]. Since phytoplankton species composition could indicate the environmental problem and early warning of the possible emergence of harmful blooms, accurate and fast identification of phytoplankton species has become increasingly important in this Anthropocene era.

Guiry [4] estimated that between 30,000 to 1 million species of phytoplankton are existed in freshwater and marine ecosystems around the world, with diatoms (Phylum: Bacillariophyta) and dinoflagellates (Phylum: Miozoa) as the two major phytoplankton groups. About 8,000 species of diatoms and 2,200 species of dinoflagellates have been named, cataloged, and described [4,5]. Despite being a hot spot for marine diversity, including phytoplankton, there were no official records or estimates of phytoplankton species in Indonesia. More than 150 to over 400 species of phytoplankton have been reported from some studies in oceanic, coastal, and island ecosystems, such as Makassar Strait [6], Lembeh Strait [7], and Seribu Islands [8]. Some species among the known phytoplankton species in Indonesia, such as *Pyrodinium bahamense*, *Margalefdinium* (*Cochlodinium*) *polykrikoides*, and *Noctiluca scintillans*, have also been reported to cause recent harmful blooms, fish kills, and fatal human poisoning in some coastal areas, such as Lampung Bay, Cirebon, and Ambon Bay [9-11]. Currently, phytoplankton identification was made manually under a microscope by human expert or specialist, which were known to have limitations, such as (i) time-consuming process, (ii) highly dependent on the expertise, experience, and skill of the expert, (iii) influenced by physiological and psychological conditions of the expert, and (iv) bias or misidentification caused by morphological variations in some species [12]. Due to the increasing problem of HABs in Indonesia, a rapid and reliable phytoplankton identification tool is required to be used in the monitoring and early warning system of HABs in the country.

The development of computer technology has increased computers' computational power and enabled a much complex artificial intelligence (AI) and machine learning (ML). Machine learning is a type of AI that uses a complex algorithm to learn from the data to improve, describe, and predict on the outcomes of some input data [13]. This study uses Convolutional Neural Network (CNN). This specialized machine learning technique used convolution operation in a neural network to process any data which has a grid-like topology, such as in time-series data (single dimension) or image data (two dimensions) [14,15]. Several studies have used deep learning techniques, particularly CNN, to automate the plankton image acquisition, identification, and enumeration. For example, Schulze *et al.* [16] has developed an open-source automatic microscopy image recognition system, the PlanktoVision, using a neural network to perform segmentation and identification of 10 taxa of phytoplankton. Another study, such as Cheng *et al.* [17], has constructed an enhanced CNN model combined with Support Vector Machine (SVM) to conduct an in-situ automatic zooplankton identification and enumeration. On the other hand, Pedraza *et al.* [18] has used deep learning and CNN model to classify 80 species of diatoms from an extensive dataset of 160,000 brightfield image samples. Similarly, Kloster *et al.* [19] used a deep CNN based on VGG16 framework to perform a taxonomic identification of several diatoms taxa on 'digital/virtual slides' up to species level. Related to the HABs mitigation and management, Henrichs *et al.* [20] have used a CNN to improve an early warning system to blooms of toxic dinoflagellate species, particularly *Karenia brevis*, *Dinophysis ovum*, and *Prorocentrum texanum* in the Texas coastal ecosystem. Furthermore, neural network techniques, such as CNN, FFNN (Feed-Forward Neural Network), RNN (Recurrent Neural Network), and LSTM (Long Short-Term Memory), have also been used to make a prediction or to forecast harmful algal blooms and phytotoxin contamination in the shellfish in the marine ecosystems [21].

In Indonesia, machine learning and neural networks have been used in several ecological and biodiversity studies to help and perform taxonomic identification or taxa classification of plants, fishes, or habitat types. For example, a study by Yunandar *et al*. [22] uses a machine learning technique to classify the inundation typology of peatland to locate key sampling sites and help understand the plankton biodiversity in water bodies of the Paminggir peatland in South Borneo, Indonesia. On the other hand, Böhlen and Sujarwo [23] use five neural network architectures, Alexnet, Squeezenet, Resnet50, Resnext152, and Vanillanet, to automatically identify ethnobotanically important plants in Bali, Indonesia based on an image dataset consisting of 50,000 images from 26 taxa. Another study by Liawatimena *et al.* [24] has explored the use of CNN to classify and identify three commercially important fish in Indonesia, such as *Katsuwonus pelamis* (cakalang), *Euthynnus affinis* (tongkol), and *Coryphaena hippurus* (mahi-mahi). Even so, study on the use or construction of machine learning models, such as CNN to study plankton in Indonesia, was rare, hard to find, or might have not yet been done. Accurate and fast identification of phytoplankton taxa is becoming more important to determine the status of the ecosystem or to detect the emergence of harmful algal blooms in Indonesia quickly. In that case, there is an urgent need to use machine learning to identify automatically and even enumerate the phytoplankton taxa in Indonesia. Thus, this study aims to develop an automatic phytoplankton image identification based on a CNN model specifically trained to classify the phytoplankton species in Indonesia. Furthermore, the study also aims to test the usability of the developed CNN model with a real dataset or scenario, which will be used to determine the strength and limitations of the developed model.

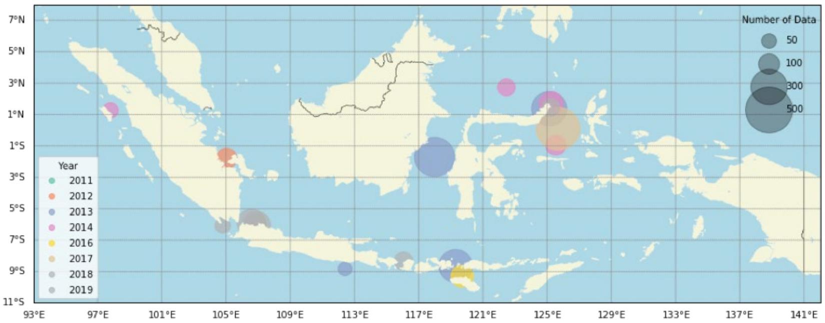


**Figure 1.** Distribution and number of phytoplankton images that were used in this study.

## 2. METHODS

### *2.1. Phytoplankton Data Curation*

This study used CNN models built based on high-resolution microscope images from the Plankton Image Database of the Plankton Laboratory, Research Center for oceanography, National Research and Innovation Agency (RCO-BRIN).

The Plankton Image Database (cPID) contains more than 6000 plankton images in several digital formats. In this study, 2667 images were curated from cPID includes phytoplankton of diatoms, dinoflagellates, and cyanobacteria (Table 1). The images were collected during various research expeditions from 18 different locations across Indonesian waters from 2011 – 2019 (Figure 1, Table 1). The cPID was used to construct the scenario and model of this study. 2.2. Scenario Development. In this study, five different scenarios were constructed to test the ability of the models to identify and classify the phytoplankton images at different taxonomic levels (Table 2). The detailed description for each scenario was as follows:

- Scenario 1 (S1) split the data into binary categories, diatoms (dia) and dinoflagellates (dino), and serve as the simplest model in this study.
- Scenario 2 (S2) further differentiate the diatoms into two different classes, the Centric Diatoms (diaC) and Pennate Diatoms (diaP) based on the general cell morphology of the groups.
- Scenario 3 (S3) was similar to Sc2, but the Centric Diatoms group were further differentiated into Circular Centric Diatoms (diaCC) and Rectangular Centric Diatoms (diaCR) based on the general cell's shape or silhouette in 2D microscope images.
- Scenario 4 (S4) was designed for the model to recognize and classify three genera of diatoms, *Chaetoceros*, *Coscinodiscus*, and *Rhizosolenia*, and two genera of dinoflagellates, *Ceratium* and *Protoperidinium*. As the CNN model requires different images used in the model training, validation, and test, only phytoplankton genus with >100 images were used. The minimum 100 images per class/categories/taxa were set based on the note in the deep learning study of Kloster *et al.* [19] and Pedraza *et al.* [18]. Furthermore, in this scenario, the selected target genus was considered a potentially harmful genus. Some of them, such as *Chaetoceros* and *Ceratium*, have been reported to cause Harmful Algal Blooms events in Indonesian waters [25,26].
- Scenario 5 (S5) was identical to S4, but with an additional class of 'Others', which will serve as a placeholder for any phytoplankton images in the cPID that did not belong to the five targeted phytoplankton genus in this study.

**Table 1.** Detail on the sampled year, location, and the number of phytoplankton images used in this study

| NO | SAMPLED YEAR | LOCATIONS | NUMBER OF IMAGES |
|---|---|---|---|
| 1 | 2011 | Jakarta Bay | 97 |
| 2 | 2012 | West Bangka | 97 |
| 3 | 2013 | Indian Ocean, off southern East Java | 44 |
| 4 | 2013 | Makassar Strait | 365 |
| 5 | 2013 | Komodo Islands | 256 |
| 6 | 2013 | Bitung | 278 |
| 7 | 2014 | Maluku Sea | 92 |
| 8 | 2014 | Sulawesi Sea | 69 |
| 9 | 2014 | Nias | 56 |
| 10 | 2014 | Lembeh Strait & Likupang | 132 |
| 11 | 2016 | Sumba | 113 |
| 12 | 2017 | Eastern ITF Pathway (Maluku Sea) | 430 |
| 13 | 2018 | Ancol | 73 |
| 14 | 2019 | Pari Island | 186 |
| 15 | 2019 | Belimbing Bay, Lampung | 53 |
| 16 | 2019 | Lembeh Strait | 73 |
| 17 | 2019 | Gili Islands | 79 |
| 18 | 2019 | Jakarta Bay | 174 |
|  |  | Total | 2667 |

### *2.2. Image Pre-processing and Augmentation*

The CNN models in this study required the phytoplankton images to be divided into three datasets for each class within each scenario: the training, validation, and test datasets. The numbers of data for each dataset in each scenario are described in Table 3.

In this study, the images in the Original Training dataset undergo a geometric transformation and segmentation process to expand the training dataset's size and enhance the constructed model's ability to recognize the object (phytoplankton cells) in the image. Geometric transformation of the original training images including (i) horizontal flip, (ii) vertical flip, (iii) resize, (iv) rotation, (v) width shift, (vi) height shift, (vii) shear, and (viii) zoom. On the other hand, the segmentation process used in this study was (i) adaptive gaussian thresholding and (ii) Histogram of Oriented Gradients (HOG). Adaptive gaussian thresholding was used to separate the background pixel and target object while dealing with various lighting conditions [27] in the microscope images due to various illumination techniques used to capture the image. On the other hand, HOG was used to quickly and accurately perform edge detection on the object in the images [28], which could improve the ability of the CNN model to recognize the objects in the images.

**Table 2.** Detail on the output (model classes) and availability of image data for each scenario for the model in this study

| Scenario | Output (Model Classes) | Available Image Data* |
|---|---|---|
| 1 | Diatoms (**dia**) | 1419 |
| | Dinoflagellates (**dino**) | 955 |
| 2 | Centric Diatoms (**diaC**) | 1107 |
| | Pennate Diatoms (**diaP**) | 312 |
| | Dinoflagellates (**dino**) | 955 |
| 3 | Circular Centric Diatoms (**diaCC**) | 434 |
| | Rectangular Centric Diatoms (**diaCR**) | 673 |
| | Pennate Diatoms (**diaP**) | 312 |
| | Dinoflagellates (**dino**) | 955 |
| 4 | *Ceratium* (**cer**) | 297 |
| | *Chaetoceros* (**chae**) | 220 |
| | *Coscinodiscus* (**cosc**) | 144 |
| | *Protoperidinium* (**pro**) | 153 |
| | *Rhizosolenia* (**rhizo**) | 116 |
| 5 | *Ceratium* (**cer**) | 297 |
| | *Chaetoceros* (**chae**) | 220 |
| | *Coscinodiscus* (**cosc**) | 144 |
| | *Protoperidinium* (**pro**) | 153 |
| | *Rhizosolenia* (**rhizo**) | 116 |
| | Others | 1444 |

*Total available image data for each scenario = 2667 images. Data was curated and sorted according to the model classes

## *2.3. Model Building*

Python language is used to build CNN models that run on an open-source cloud service, Google Colaboratory, using the KERAS library from the Tensorflow framework. The CNN model used in this study (Figure 2) was built based on a VGG16 CNN model, which was pre-trained with the ImageNet dataset. The VGG16 model was a 16 layers CNN model proposed by Simonyan and Zisserman [29] from the Visual Geometry Group, the University of Oxford, which was designed to perform image classification using a wide range of datasets with high accuracy. In this study, only the VGG16 convolutional part was imported without the 3 fully connected layers at the top of the network. Then, all of the VGG16 trainable layers were frozen, and additional two dense classifiers were added at the end of the layer, one with ReLu activation and softmax activation. The modified VGG16 architecture used in this study can be found in Table 4. For each study scenario, a modified VGG16 model was created and then trained with the Augmented Training dataset for 50 epoch with 16 steps per epoch.

## *2.4. Validation, Evaluation, and Application*

**Table 3.** The number of phytoplankton images in each dataset for each scenario. Original Training dataset consists of the curated images split into different classes/categories according to the scenario design. The Augmented Training dataset resulted from pre-processing and augmentation, which increased the number of available images for model training

| Dataset | Scenario | | | | |
|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 |
| Original Training | 1355 | 891 | 1043 | 248 | 891 |
| Augmented Training | 2048 | 2048 | 2048 | 512 | 512 |
| Validation | 32 | 32 | 32 | 32 | 32 |
| Test | 32 | 32 | 32 | 32 | 32 |

In this study, the trained model for each scenario was validated using the validation dataset to test its accuracy and correctly identify the classes or categories in the model's scenario. Each model was then evaluated using two different metrics: categorical cross-entropy loss and accuracy test using a confusion matrix. The categorical cross-entropy loss was used to quantify the rate of error in predicting the correct categories of the target images [30]. The categorical cross-entropy loss was used in this study as the CNN models in each scenario used a 'Softmax' activation layer. On the other hand, the confusion matrix will show the proportion of success and failure to predict the true classes, which also indicate the complete performance of each classification model constructed in this study.

**Table 4.** The base architecture of the VGG16 model was used to identify the phytoplankton images in this study.

| Layer (type) | | Size |
|---|---|---|
| Input | Image | 128 x 128 x 3 |
| 1 | 2 x Conv_2D | 128 x 128 x 64 |
| | Max Pooling | 64 x 64 x 64 |
| 3 | 2 x Conv_2D | 64 x 64 x 128 |
| | Max Pooling | 32 x 32 x 128 |
| 5 | 2 x Conv_2D | 32 x 32 x 256 |
| | Max Pooling | 16 x 16 x 256 |
| 7 | 3 x Conv_2D | 16 x 16 x 512 |
| | Max Pooling | 8 x 8 x 512 |
| 10 | 3 x Conv_2D | 8 x 8 x 512 |
| | Max Pooling | 4 x 4 x 512 |
| 13 | Fully Connected Layer | 8192 |
| 14 | Fully Connected Layer | 256 |
| Output | Fully Connected Layer | 6 |

This study also incorporated an application trial for each trained CNN model by using a new image dataset obtained from phytoplankton net samples collected from a study in Jakarta Bay in 2019 [31]. The new image dataset consists of 32 high-resolution phytoplankton images that were not included in the curated dataset (Table 1) that were used to construct the CNN models. The application trial was conducted to determine the usability of each CNN model in simulated real-life scenarios and to identify the weakness and strengths of each model.

## 3. RESULTS & DISCUSSIONS

### *3.1. Loss and Accuracy*

Each constructed CNN model with different categorical scenarios in this study did show a different rate of loss and identification accuracy (Figure 3). The detail on the cross-entropy loss value and accuracy of the model for each scenario during model training and validation are shown in Table 5. Based on the research result, the CNN model with the simplest scenario (S1), was the best model with minimal loss and highest accuracy, both in the training and validation test (Figure 3). During the training, the CNN S1 model starts with the lowest loss of 0.95 and ends with a loss of 0.14 after 50 epochs (Figure 3A). On the other hand, the S1 CNN model started with >50% accuracy during the training test and reached 94% accuracy after 50 epoch (Figure 3C). High accuracy and low loss for the S1 CNN model were expected as the model only deals with binary categories, diatoms, or dinoflagellates. On the other hand, it was interesting to find that the S4 CNN model with a complex model and five categories could achieve the second-highest accuracy (>90%) during the training (Figure 3C), while much lower loss compared to other simpler models, such as CNN S3 and S2 (Figure 3A). Generally, the model accuracy would be lower while the loss value would be higher along with the increasingly

**Table 5.** The accuracy and cross-entropy loss values for model training and validation for each scenario in this study.

| Scenario | Cross-Entropy Loss | | | Accuracy (%) | | |
|---|---|---|---|---|---|---|
| | 1st Epoch | 50th Epoch | Difference (Δ) | 1st Epoch | 50th Epoch | Difference (Δ) |
| Training | | | | | | |
| 1 | 0,9543 | 0,1366 | 0,8177 | 55,16 | 93,95 | 38,79 |
| 2 | 1,8196 | 0,4376 | 1,3820 | 34,05 | 81,63 | 47,58 |
| 3 | 1,6747 | 0,4423 | 1,2324 | 25,09 | 82,48 | 57,39 |
| 4 | 1,8429 | 0,2313 | 1,6116 | 21,52 | 91,22 | 69,70 |
| 5 | 1,8917 | 0,4136 | 1,4781 | 23,05 | 83,59 | 60,54 |
| Validation | | | | | | |
| 1 | 0.7088 | 0.4048 | 0.304 | 53.12 | 87.50 | 34.38 |
| 2 | 1.065 | 0.5901 | 0.4749 | 43.75 | 65.62 | 21.87 |
| 3 | 1.3732 | 0.4341 | 0.9391 | 34.38 | 81.25 | 46.87 |
| 4 | 1.5497 | 0.4902 | 1.0595 | 28.12 | 81.25 | 53.13 |
| 5 | 1.6792 | 0.9206 | 0.7586 | 25.00 | 62.50 | 37.50 |

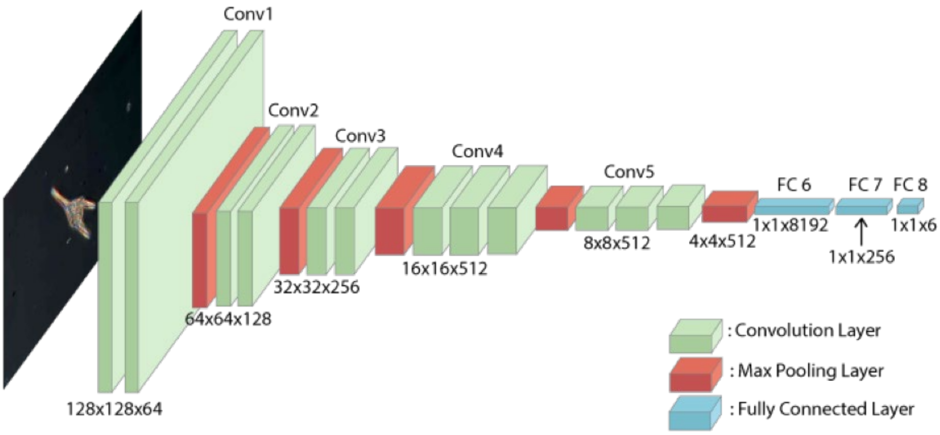


**Figure 2.** Illustration of the base architecture of the VGG16 model used in this study.

complex scenario. But, that was not the case in this study. The most complex model in this study, the CNN S5, perform quite well in training, with an accuracy of 83.5% after 50 epochs (Figure 3C). However, the CNN S5 model was the worst model during the validation test, with accuracy <65% (Figure 3B) and has the highest loss value compared to the other CNN models (Figure 3D). Note that in the validation test, the CNN S1 and CNN S4 models remain as the 1st and 2nd best models based on their loss values and accuracy (Figure 3B & 3D).

On the other hand, erratic patterns in the cross-entropy loss and accuracy graph during validation test (Figure 3B & 3D), along with relatively large differences between those values compared to the model training (Figure 3A & 3C), might indicate a problem of overfitting during the model training. This problem might occur because all CNN models in this study did not use any regularization techniques, such as early stop

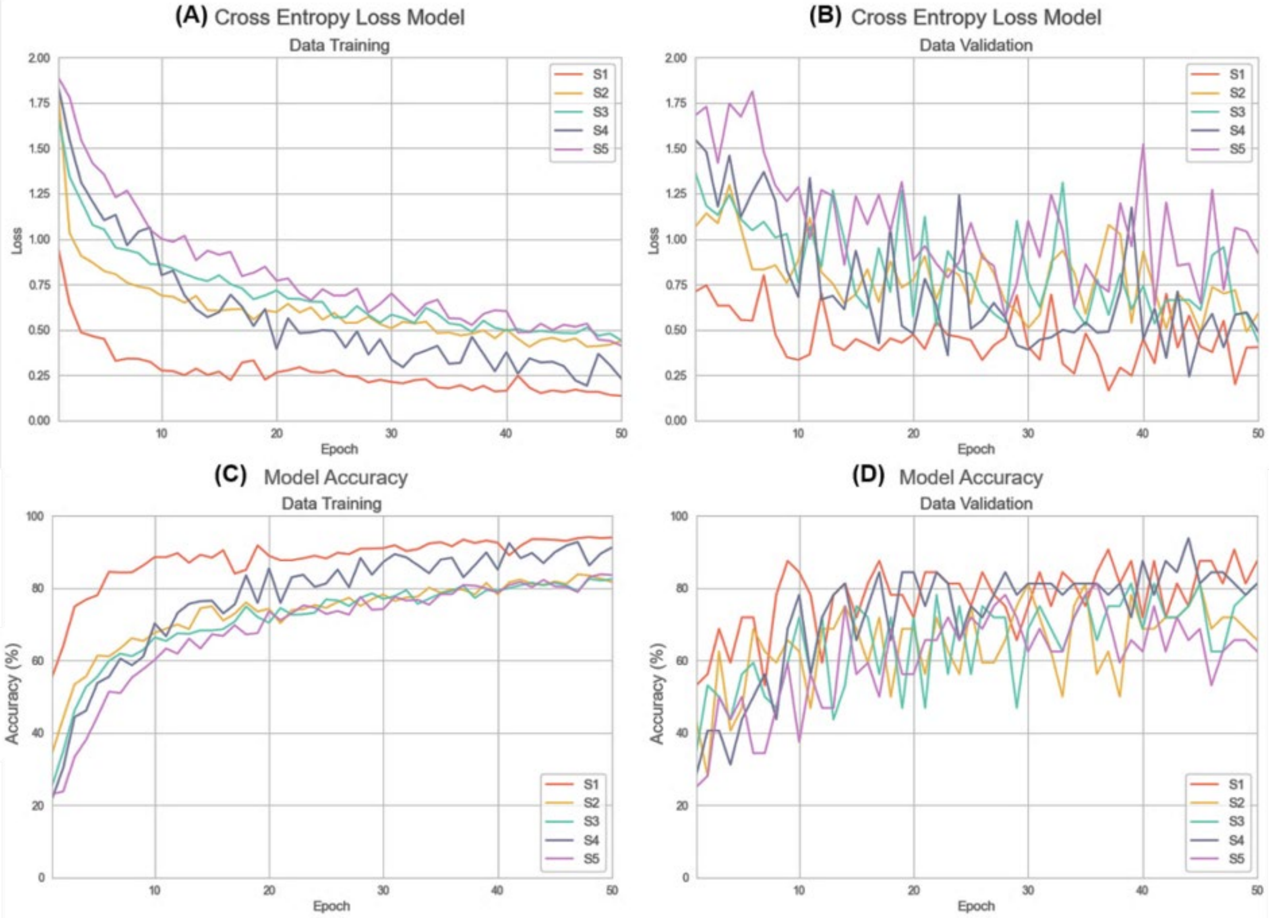


**Figure 3.** Graphic showing the cross-entropy loss of (A) training data and (B) validation data, and model accuracy of (C) training and (D) validation data. Each model run consists of 50 epochs in 16 steps per epoch.

**Table 6.** The accuracy of the CNN model in each scenario during the model test.

| Scenario | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Test Accuracy (%) | 87,50 | 82,29 | 77,34 | 88,75 | 75,00 |

or dropout, which usually was used to avoid overfitting in deep learning models [30].

### *3.2. Test Performance*

The confusion matrix showed the overall model performance during the model test to classify the categories or classes in each scenario of this study (Figure 4). Interestingly, the accuracy of the S1 CNN model was slightly lower than the accuracy of the S4 CNN model during the model testing (Table 6). CNN S1 was 87.5%, while CNN S4 accuracy reached 88.8%. Similar to what has been shown in the training and validation test (Table 5), the most complex model, CNN S5, has the lowest performance (Figure 4) and accuracy compared to the other CNN models (Table 6).

The confusion matrix also shows which categories/classes are less likely to be classified correctly by each CNN model in each scenario (Figure 4). It was unexpected to see the S4 model, with its rather complex classification categories, managed to perform well in identifying most categories. However, it seems to have problems identifying *Ceratium* and *Chaetoceros* (Figure 4). As shown in the confusion matrix, the model predicted the images as *Chaetoceros* six times, although the true label of the image was *Rhizosolenia* (Figure 4). It also failed to predict the *Ceratium* images correctly and classify *Ceratium* images as other categories or genus (Figure 4). It also failed to correctly predict the *Ceratium* images and classify *Ceratium* images as other categories or genus (Figure 4). The problem in the S4 CNN model was carried over and amplified in the S5 CNN model. The error rate in identifying *Ceratium* and *Rhizosolenia* was much higher in the S5 model than the S4 model (Figure 4). The S5 CNN model was only able to correctly classify 10 *Ceratium* images and 12 *Rhizosolenia* images out of 32 images in the test dataset, and it often misclassifies those two categories/classes as Others (Figure 4).

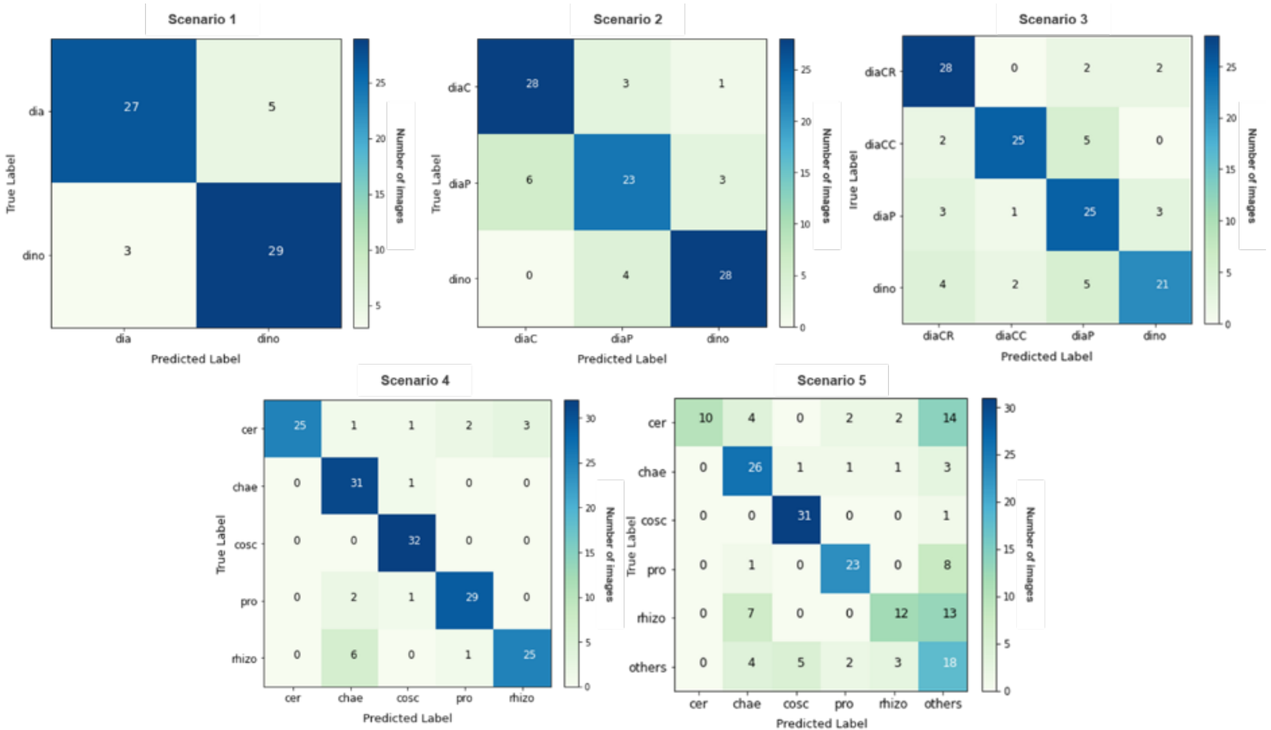


**Figure 4.** Confusion matrix of the CNN model for each scenario during the model testing of this study. Darker colors signify more images that the model correctly or incorrectly predicted either. In the test dataset, each category/class contains unique 32 phytoplankton images different from the one in training and validation datasets. Note: dia = diatoms, dino = dinoflagellates, diaC = centric diatoms, diaP = pennate diatoms, diaCR = rectangular centric diatoms, diaCC = circular centric diatoms, cer = *Ceratium*, chae = *Chaetoceros*, cosc = *Coscinodiscus*, pro = *Protoperidinium*, rhizo = *Rhizosolenia*.

The classification mistake in the CNN model could occur due to several factors, such as (i) image noise due to low light or other non-ideal imaging situations, (ii) feature problem due to unclear or ambiguous internal features, boundaries, or descriptors within the images, (iii) incomplete cell frustule (diatoms) or theca (dinoflagellates) in the image, and (iv) multiple and/or overlapping objects (cells) within an image [17,19]. Two possible explanations for misclassification or low performance in genus-level models could be due to high variation in the cell forms within *Ceratium* genus (Figure 5D & 5F), which often have a similar shape to *Rhizosolenia* (Figure 5Ai & 5E) or *Protoperdinium* (Figure 5G-H). In addition, many images from the cPID contain more than one genus in an image or include multiple genera other than the labeled genus (Figure 5A). The microscope image, which has complex and overlapping cells (Figure 5A) could confuse the model as the current model could not recognize multiple objects in a 2D photo. In this study, that image (Figure 5A) was not used in the CNN model training, validating, or testing, as it could confuse the model and reduce its performance. It was not clear why the S4 and S5 models have difficulties identifying *Rhizosolenia* and often make mistakes in classifying it as *Chaetoceros* (Figure 4). Although both *Rhizosolenia* and *Chaetoceros* belong to the same form group, the centric diatoms, *Chaetoceros* cells (Figure 5B), have distinct morphological characters that could be easily differentiated from *Rhizosolenia* (Figure 5Ai & 5E).

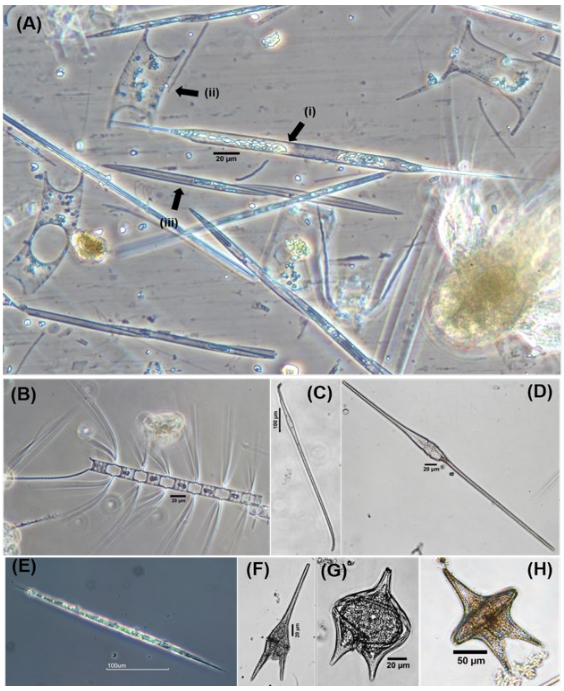


**Figure 5.** Example of phytoplankton images in cPID used to construct and test the CNN models. These examples emphasize similarities in cell forms of several species and complexity in some labeled images in cPID. **(A)** image labeled as *Rhizosolenia setigera* (i) but contain overlapping cells of Eucampia cornuta (ii), *Pseudo-nitzschia* sp (iii) and other hard to identified species, **(B)** *Chaetoceros distans*, **(C)** *Amphisolenia bidentata*, **(D)** *Ceratium fusus*, **(E)** *Rhizosolenia hebetata*, **(F)** *Ceratium furca*, **(G)** *Protoperidinium oceanicum*, and **(H)** *Protoperidinium elegans*.

## *3.3. Application Trial*

In this study, the application trial was conducted using all CNN models to identify an entirely new dataset consisting of random 32 phytoplankton images from the Jakarta Bay 2019 study [31] that were not included in the training or test dataset. The trial was done as a simulation to test how the model behaves under a completely new dataset and apply the model in a pseudo-real-world scenario. As seen in the examples within Figure 6A, all CNN models generally have good performance and managed to identify and classify the phytoplankton image that exists as a category within the model. In this case, the five-target genus in S4 and S5 CNN models, *Ceratium*, *Chaetoceros*, *Rhizosolenia*, *Coscinodiscus*, and *Protoperidinium* (Figure 6A), were used as a benchmark for phytoplankton cell identification of all models. The detailed result of the correct prediction and accuracy was summarized in Table 7.

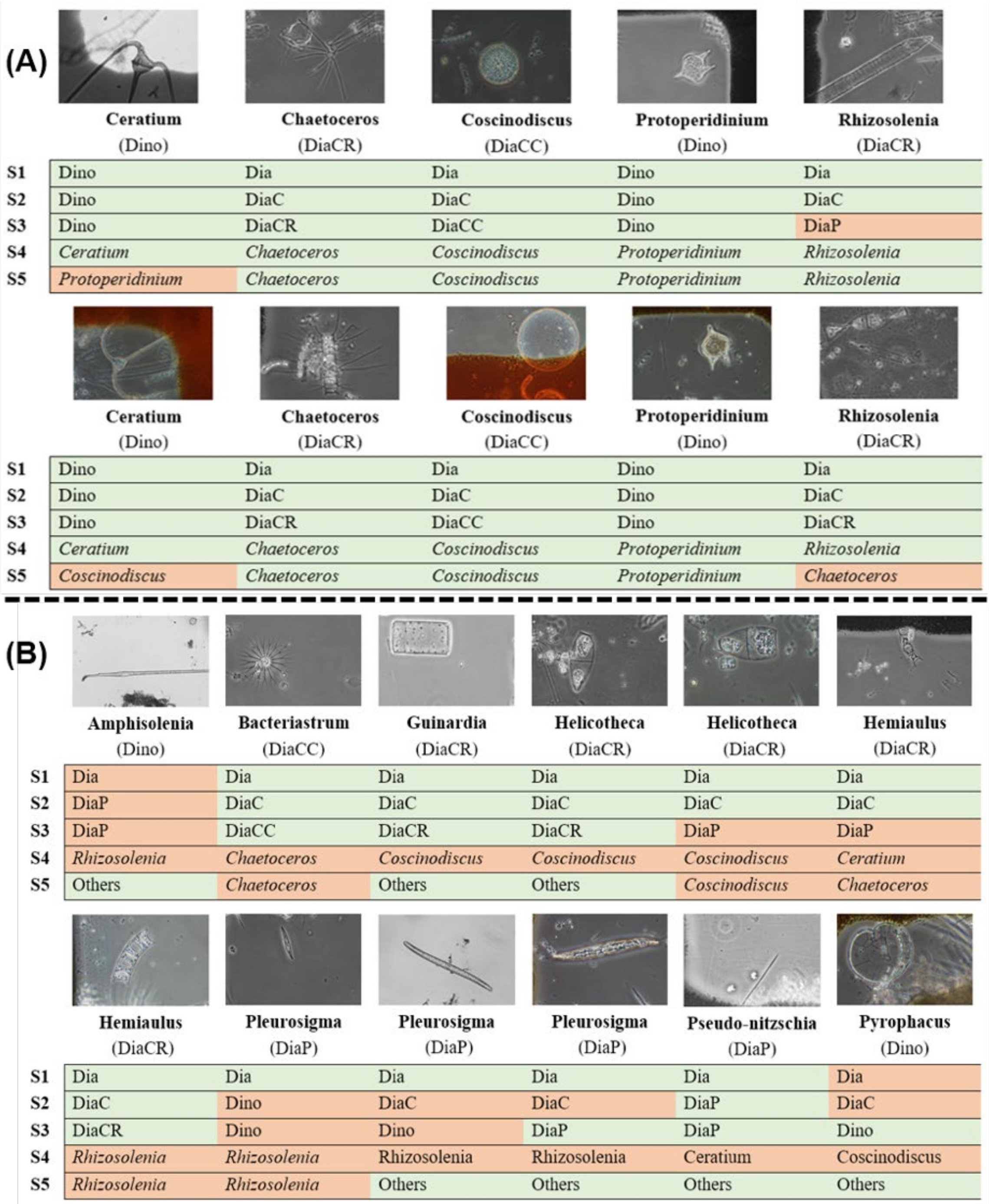
Simpler models, such as S1 and S2, correctly identified the categories in the examples images of the application trial dataset (Figure 6). However, it started to have problems when identifying phytoplankton that might not be included in the training, validation, and test dataset of S1 and S2 CNN models, such as *Amphisolenia* and *Pyrophacus* (Figure 6B). On the other hand, a more complex model, particularly the S4 CNN model, managed to correctly identify all cells in the

**Table 7.** Summary of the number of correct predictions and accuracy for each category in each CNN model. The prediction was made from a new dataset consisting of 32 random phytoplankton images from the Jakarta Bay 2019 study [31], which is different from the images in the training, validation, and test dataset used to build the models.

| Scenario<br>Category | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| *Ceratium* | 7 | 7 | 7 | 11 | 1 |
| *Chaetoceros* | 12 | 12 | 9 | 12 | 10 |
| *Coscinodiscus* | 10 | 11 | 12 | 12 | 12 |
| *Protoperidinium* | 6 | 10 | 10 | 11 | 7 |
| *Rhizosolenia* | 12 | 5 | 2 | 11 | 3 |
| Others | 9 | 6 | 6 | 0 | 6 |
| Total Accuracy (%) | 77,78 | 70,83 | 63,89 | 79,17 | 54,17 |

application trial image examples (Figure 6A) with high accuracy (79%) (Table 7).

However, the S4 CNN model was built only to identify five target genera. Thus, it could not correctly identify any phytoplankton cells outside its built-in categories (genus) (Figure 6A). Despite its best performance, whether in training, validation, test, and application trial (Table 7), the S4 CNN model has a problem of forced classification. Due to its limited ability to classify only five classes/categories, the S4 CNN model will forcefully classify any phytoplankton images outside its categories into one of the categories within the model. This problem would cause



| | **Ceratium** (Dino) | **Chaetoceros** (DiaCR) | **Coscinodiscus** (DiaCC) | **Protoperidinium** (Dino) | **Rhizosolenia** (DiaCR) |
|---|---|---|---|---|---|
| S1 | Dino | Dia | Dia | Dino | Dia |
| S2 | Dino | DiaC | DiaC | Dino | DiaC |
| S3 | Dino | DiaCR | DiaCC | Dino | DiaP |
| S4 | *Ceratium* | *Chaetoceros* | *Coscinodiscus* | *Protoperidinium* | *Rhizosolenia* |
| S5 | *Protoperidinium* | *Chaetoceros* | *Coscinodiscus* | *Protoperidinium* | *Rhizosolenia* |

| | **Ceratium** (Dino) | **Chaetoceros** (DiaCR) | **Coscinodiscus** (DiaCC) | **Protoperidinium** (Dino) | **Rhizosolenia** (DiaCR) |
|---|---|---|---|---|---|
| S1 | Dino | Dia | Dia | Dino | Dia |
| S2 | Dino | DiaC | DiaC | Dino | DiaC |
| S3 | Dino | DiaCR | DiaCC | Dino | DiaCR |
| S4 | *Ceratium* | *Chaetoceros* | *Coscinodiscus* | *Protoperidinium* | *Rhizosolenia* |
| S5 | *Coscinodiscus* | *Chaetoceros* | *Coscinodiscus* | *Protoperidinium* | *Chaetoceros* |

| | **Amphisolenia** (Dino) | **Bacteriastrum** (DiaCC) | **Guinardia** (DiaCR) | **Helicotheca** (DiaCR) | **Helicotheca** (DiaCR) | **Hemiaulus** (DiaCR) |
|---|---|---|---|---|---|---|
| S1 | Dia | Dia | Dia | Dia | Dia | Dia |
| S2 | DiaP | DiaC | DiaC | DiaC | DiaC | DiaC |
| S3 | DiaP | DiaCC | DiaCR | DiaCR | DiaP | DiaP |
| S4 | *Rhizosolenia* | *Chaetoceros* | *Coscinodiscus* | *Coscinodiscus* | *Coscinodiscus* | *Ceratium* |
| S5 | Others | *Chaetoceros* | Others | Others | *Coscinodiscus* | *Chaetoceros* |

| | **Hemiaulus** (DiaCR) | **Pleurosigma** (DiaP) | **Pleurosigma** (DiaP) | **Pleurosigma** (DiaP) | **Pseudo-nitzschia** (DiaP) | **Pyrophacus** (Dino) |
|---|---|---|---|---|---|---|
| S1 | Dia | Dia | Dia | Dia | Dia | Dia |
| S2 | DiaC | Dino | DiaC | DiaC | DiaP | DiaC |
| S3 | DiaCR | Dino | Dino | DiaP | DiaP | Dino |
| S4 | *Rhizosolenia* | *Rhizosolenia* | Rhizosolenia | Rhizosolenia | Ceratium | Coscinodiscus |
| S5 | *Rhizosolenia* | *Rhizosolenia* | Others | Others | Others | Others |

**Figure 6.** Results of application trial using new image dataset from Jakarta Bay 2019 study [31]. The images in this real-world simulation trial were not included in the dataset used to train, validate, and test the CNN models of this study. Result of identification using (A) images that exist as a category or class in the model, and (B) images that do not exist as a category or class in the model, except the S5 model in the 'Others' category. In this test, the result of model prediction or classification was checked and corrected manually by a human expert. The green color signifies correct prediction, while the red color signifies incorrect prediction after a manual taxonomic check

phytoplankton genus, such as the dinoflagellates *Amphisolenia* and *Pyrophacus*, to be classified or identified by the S4 CNN model as *Rhizosolenia* or *Coscinodiscus* (Figure 6B), which would be a classification error according to a manual taxonomic identification by a human expert.

The forced classification problem with the S4 CNN model was the reason to build the S5 CNN model in this study. The S5 model contains the 'Others' class/category as a placeholder for any phytoplankton images that did not belong to the five target genus of the S4 CNN model. That solves the problem of forced classification as any images that were not *Ceratium*, *Chaetoceros*, *Rhizosolenia*, *Coscinodiscus*, and *Protoperidinium*, will now be classified as 'Others' in the S5 CNN model. However, the vastly diverse cell's shapes of many genera in the 'Others' might be the reason for the poor performance of the S5 CNN model in all stages, including the model building stages (training, validating, testing) (Table 5 & Table 6) and the application trial (Table 7). The accuracy of the S5 CNN model was only 54% during the application trial (Table 7). Still, it managed to correctly identify the non-five-target genus, such as *Amphisolenia*, *Pyrophacus*, and *Pleurosigma* as 'Others'. Those non-target genera were unable to be identified correctly by the better performed S4 CNN model (Figure 6B). Even so, the low accuracy of the S5 CNN model of this study still prevents it from being used in a real-world scenario to identify the phytoplankton genus in Indonesia.

The CNN models of this study were capable of automatically identifying phytoplankton images up to genus level at accuracy >75% in some model scenarios. Among the five CNN model scenarios, the simplest model (S1) and the genus-level model (S4) performed well in all testing stages and have low classification errors. The S1 model could differentiate diatoms and dinoflagellates groups with up to 78% accuracy in the application trial. In contrast, S4 model could differentiate the target genus of *Ceratium*, *Chaetoceros*, *Coscinodiscus*, *Protoperidinium*, and *Rhizosolenia* up to 79% accuracy.

However, the S4 model suffers from forced classification problems due to its inability to identify images of any non-target genus. Therefore, any non-target-genus, such as *Amphisolenia* and *Pyrophacus* in the application trial dataset, will be incorrectly classified as one among the five target genus. Unfortunately, the S5 CNN model created to solve the S4 problems has a much lower accuracy at 54% due to highly diverse data stored in the 'Others' category, which often confuses the model. Even so, the S5 CNN model managed to classify non-target-genus, such as *Amphisolenia*, *Pyrophacus*, and *Pleurosigma*, into the 'Others' category.

Despite the success of all CNN's models in this study to automatically identify phytoplankton up to genus level at accuracy > 75%, the current limitations in all scenarios need to be solved before the model can be used in a real-world research scenario. Some of the solutions to solve the problem in the current CNN models including (i) expanding the dataset to include more images, (ii) optimizing the pre-processing and augmentation process, and (iii) fine-tuning the model.

## ACKNOWLEDGMENTS

This study was funded by the National Priority Program of Deputy of Earth Sciences, Indonesian Institute of Sciences (LIPI; now merged in National Research and Innovation Agency or BRIN) 2021 under the *Prioritas Riset Nasional* (National Research Priority) *Mikro- dan Makro-alga Laut Strain Asli Indonesia* (MALSAI) program. This research was a part of the "*Klasifikasi Plankton dengan Menggunakan Microscopic/Underwater Images*" project within the MALSAI program. This research was also a part of the undergraduate thesis (Skripsi) of Aulia Salsabella Suwarno as a prerequisite to finishing her undergraduate study (S1) in the Faculty of Earth Science and Technology, Bandung Institute of Technology (ITB). We want to thank Dr. Esa Prakasa, the project leader of our MALSAI research group, for his input in our research. We would also thank fellow researchers and now-retired research technicians of Plankton Laboratory, RCO-BRIN, for their support and contribution to the image data acquisition for the Plankton Image Database. In this manuscript, authors share equal contributions, with Aulia Salsabella Suwarno as the designer, developer, and tester of the models; Arief Rachman as the Aulia's secondary supervisor, image data provider, manual phytoplankton identifier, and manuscript writer; and Susanna Nurdjaman as the Aulia's primary supervisor, plankton data analyzer, and model's interpreter and appraiser.

## REFERENCES

[1]. I.M. Suthers, A.J. Richardson, D. Rissik, The importance of plankton, in: Plankton: A Guide to Their Ecology and Monitoring for Water Quality. 2019. pp. 1-13.

[2]. G.M. Hallegraeff, D.M. Anderson, C. Belin, M-Y.D Bottein, E. Bresnan, M. Chinain, et al., Perceived global increase in algal blooms is attributable to intensified monitoring and emerging bloom impacts, in : Communications Earth & Environment vol. 2, 2021, pp. 117. DOI: 10.1038/s43247-021-00178-8.

[3]. A.W. Griffith, C.J. Gobler, Harmful algal blooms: A climate change co-stressor in marine and

freshwater ecosystems, Harmful Algae vol. 91, 2020. DOI: 10.1016/j.hal.2019.03.008.

[4]. M.D. Guiry, How many species of algae are there? Journal of phycology vol. 48, 2012 pp. 1057-63. DOI: 10.1111/j.1529-8817.2012.01222.x.

[5]. S.D. Lee, S.M. Yun, J.S. Park, J.H. Lee, Floristic survey of diatom in the three islands (Baeknyeong, Daecheong, Socheong) from Yellow sea of Korea, Journal of Ecology and Environment vol. 38, 2015, pp. 1-36. DOI: 10.5141/ecoenv.2015.000.

[6]. A. Rachman, A. Purwandana, N. Fitriya, Phytoplankton Community Structure of the Makassar Strait, Indonesia, IOP Conference Series: Earth and Environmental Science vol. 789, 2021. DOI: 10.1088/1755-1315/789/1/012006.

[7]. T. Senming, A. Rachman, N. Fitria, H. Thoha, B. Chen, Phytoplankton changes during SE monsoonal period in the Lembeh Strait of North Sulawesi, Indonesia, from 2012 to 2015, Acta Oceanologica Sinica vol. 37, 2018, pp. 9-17. DOI: 10.1007/s13131-018-1283-4.

[8]. A. Rachman, Checklist and estimation of total number of phytoplankton species in Pari, Tidung, and Payung Islands, Indonesia. Biodiversitas Journal of Biological Diversity, vol. 21, 2020. DOI: 10.13057/biodiv/d210616.

[9]. S. Likumahua, Recent blooming of *Pyrodinium bahamens*e var. *compressum* in Ambon Bay, Eastern Indonesia, Marine Research in Indonesia vol. 38, 2015, pp. 31-7. DOI: 10.14203/mri.v38i1.54.

[10]. H. Thoha, Muawanah, M. Bayu Intan, A. Rachman, O.R. Sianturi, T. Sidabutar, *et al.*, Resting cyst distribution and molecular identification of the harmful dinoflagellate *Margalefidinium polykrikoides* (Gymnodiniales, Dinophyceae) in Lampung Bay, Sumatra, Indonesia, Frontiers in microbiology vol. 10, 2019, pp.1-12.
DOI: 10.3389/fmicb.2019.00306.

[11]. A. Nurlina, A.A. Liambo, editors. Kejadian Luar Biasa Paralytic Shellfish Poisoning Pada Konsumsi Kerang Hijau Terkontaminasi Saxitoxin di Kabupaten Cirebon, Indonesia, Desember 2016. Prosiding Seminar Nasional dan Diseminasi Penelitian Kesehatan [Internet]; Tasikmalaya: STIKes Bakti Tunas Husada, 2015. Available from: https://ejurnal.stikes-bth.ac.id/index.php/P3M_PSNDPK/article/view/362.

[12]. P.F. Culverhouse, R. Williams, M. Benfield, P.R. Flood, A.F. Sell, M.G. Mazzocchi, *et al.*, Automatic image analysis of plankton: future perspectives, Marine Ecology Progress Series vol. 312, 2006, pp. 297-309. DOI: 10.3354/meps312297.

[13]. J. Hurwitz, D. Kirsch, Machine learning for dummies. Hoboken, New Jersey, USA, John Wiley & Sons, Inc., 2018, pp. 75.

[14]. I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016. pp. 330-372.

[15]. K.K. Al-jabery, T. Obafemi-Ajayi, G.R. Olbricht, D.C. Wunsch Ii, 4 - Selected approaches to supervised learning. In: KK. Al-jabery, T. Obafemi-Ajayi, G.R. Olbricht, D.C. Wunsch Ii, editors. Computational Learning Approaches to Data Analytics in Biomedical Applications, Academic Press, 2020, pp. 101-23.

[16]. K. Schulze, U.M. Tillich, T. Dandekar, M. Frohme, PlanktoVision - an automated analysis system for the identification of phytoplankton, BMC Bioinformatics volume 14, 2013, pp. 115. DOI: 10.1186/1471-2105-14-115.
[17]. K. Cheng, X. Cheng, Y. Wang, H. Bi, M.C. Benfield, Enhanced convolutional neural network for plankton identification and enumeration, PLOS ONE vol. 14, 2019. DOI: 10.1371/journal.pone.0219570.

[18]. A. Pedraza, G. Bueno, O. Deniz, G. Cristóbal, S. Blanco, M. Borrego-Ramos, Automated Diatom Classification (Part B): A Deep Learning Approach, Applied Sciences vol. 7, 2017, pp. 460. DOI: 10.3390/app7050460

[19]. M. Kloster, D. Langenkämper, M. Zurowietz, D. Beszteri, T.W. Nattkemper, Deep learning-based diatom taxonomy on virtual slides. Scientific Reports vol. 10, 2020, pp. 14416. DOI: 10.1038/s41598-020-71165-w.

[20]. D.W. Henrichs, S. Anglès, C.C. Gaonkar, L. Campbell, Application of a convolutional neural network to improve automated early warning of harmful algal blooms, Environmental Science and Pollution Research vol. 28, 2021, pp. 28544-55. DOI: 10.1007/s11356-021-12471-2.

[21]. R.C. Cruz, P. Reis Costa, S. Vinga, L. Krippahl, M.B. Lopes, A Review of Recent Machine Learning Advances for Forecasting Harmful Algal Blooms and Shellfish Contamination, Journal of Marine Science and Engineering vol. 9, 2021, pp. 283. DOI: 10.3390/jmse9030283.

[22]. D. Yunandar, H. Effendi, Y. Setiawan, Plankton biodiversity in various typologies of inundation in Paminggir peatland, South Kalimantan, Indonesia on dry season, Biodiversitas Journal of Biological Diversity vol. 21, 2020. DOI: 10.13057/biodiv/d210322.

[23]. M. Böhlen, W. Sujarwo, editors, Machine Learning in Ethnobotany, 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2020. DOI: 10.1109/SMC42975.2020.9283069.

[24]. S. Liawatimena, Y. Heryadi, Lukas, A. Trisetyarso, A. Wibowo, B.S. Abbas, *et al.*, editors, A Fish Classification on Images using Transfer Learning and Matlab, 2018 Indonesian Association for Pattern Recognition International Conference (INAPR), 2018. DOI: 10.1109/INAPR.2018.8627007.

[25]. D. Praseno, N. Wiadnyana, HAB organisms in Indonesian waters [Internet]. Canadian Technical Report of Fisheries and Aquatic Sciences; 1996. Available from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.606.1950&rep=rep1&type=pdf#page=93.

[26]. D. Praseno, Y. Fukuyo, R. Widiarti, Sugestiningsih, editors, Red tide occurrences in Indonesian waters and the need to establish a monitoring system [Internet], Proceedings of Workshop on Red tide Monitoring in Asian Coastal Waters, 2003. Available from: https://www.researchgate.net/publication/242468092_RED_TIDE_OCCURRENCES_IN_INDONESIAN_WATERS_AND_THE_NEED_TO_ESTABLISH_A_MONITORING_SYSTEM.

[27]. C.A. Montes, Practical Computer Vision: Theory and Applications [Internet], 2015. Available from: http://www.bcamath.org/documentos_public/courses/course_day2.pdf.

[28]. N. Dalal, B. Triggs, editors. Histograms of oriented gradients for human detection, 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 2005. DOI: 10.1109/CVPR.2005.177.

[29]. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition [Internet]. arXiv preprint arXiv:14091556, 2014. Available from: https://arxiv.org/abs/1409.1556v6.

[30]. N. Ketkar, J. Moolayil, Deep learning with Python: Learn Best Practices of Deep Learning Models with PyTorch, California, APress, 2021, pp. 316.

[31]. H. Thoha, Laporan Akhir Kegiatan Tahun 2019: Status Ekologi Perairan Teluk Jakarta, Jakarta, Pusat Penelitian Oseanografi, Lembaga Ilmu Pengetahuan Indonesia, 2019.