Can Foreign Datasets Help Improve Plankton Classification Performance for Local Data?

Malini Ramberran[1(✉)], Adhir Soechit[2], and Patrick Hosein[1]

[1] The University of the West Indies, St. Augustine Campus, St. Augustine, Trinidad
malini.ramberran@my.uwi.edu, patrick.hosein@uwi.edu
[2] University of Liverpool, Liverpool L69 7ZX, UK
a.k.soechit@liverpool.ac.uk

Abstract. Plankton play a critical role in aquatic ecosystems, contributing to oxygen production, nutrient cycling, and the regulation of global carbon dynamics. Effective monitoring of plankton populations is essential for understanding environmental change and ecosystem health. However, traditional plankton classification relies on manual image annotation by taxonomic experts, a process that is labor-intensive and difficult to scale. These challenges are particularly acute in data-sparse regions such as the Caribbean, where labeled datasets are scarce. This study presents a plankton classification approach for a microscopy image dataset collected from the coastal waters of Trinidad and Tobago by the Department of Life Science at the University of the West Indies (UWI). Due to the limited size of this regional dataset, transfer learning was applied using the National Data Science Bowl (NDSB) dataset, which includes over 30,000 images across 121 plankton classes. Convolutional neural networks (CNNs) using the NDSB data are fine-tuned to adapt to the local Caribbean samples. The proposed method significantly improves classification performance, especially in the context of limited data and class imbalance. The results demonstrate the viability of using large, publicly available datasets to enhance local ecological monitoring efforts, offering a scalable and efficient alternative to manual annotation in underrepresented regions.

Keywords: Plankton classification · Deep learning · Convolutional Neural Networks (CNNs) · Transfer learning · Caribbean biodiversity

1 Introduction

Plankton occupy a fundamental position in aquatic food chains, serving as the primary producers and initial consumers that support higher trophic levels in marine and freshwater systems [1]. Phytoplankton, through photosynthesis, contribute nearly half of the world's oxygen supply, while zooplankton regulate population dynamics and nutrient cycling by feeding on phytoplankton and providing a food source for larger organisms such as fish and whales [2]. Beyond

K. Arai (Ed.): FTC 2025, LNNS 1676, pp. 19–37, 2026.
https://doi.org/10.1007/978-3-032-07989-3_2

their ecological roles, plankton mediate carbon transport from surface to deep ocean layers, making them essential agents in climate regulation. As sentinels of ocean health, shifts in their abundance and diversity are often early indicators of environmental change, underlining their importance in long-term marine monitoring programs [3].

Despite their significance, classifying plankton remains a complex task. Manual annotation of microscope imagery requires trained specialists, making the process both time-intensive and difficult to scale. This challenge is exacerbated by the high throughput of modern imaging technologies such as the Imaging FlowCytobot (IFCB) and the In Situ Ichthyoplankton Imaging System (ISIIS), which can generate millions of images during a single deployment [4]. In regions like the Caribbean, where biological data collection is less developed, the absence of automated classification systems and a shortage of taxonomic expertise can hinder the ability to conduct large-scale and continuous ecological assessments. In particular, the absence of publicly available datasets from Caribbean waters restricts the development of machine learning tools tailored to this biodiverse region.

This research utilizes a dataset collected in the coastal waters of Trinidad and Tobago, provided by the Department of Life Sciences at UWI. The data collection pipeline involved extracting plankton samples, imaging them via microscopy, isolating frames containing individual organisms, and segmenting and labeling those organisms for training purposes. The labor-intensive nature of this workflow, along with the limited size of the resulting dataset, underscores the need for scalable and automated classification methods. Furthermore, the lack of regional datasets presents a barrier to training locally effective models from scratch.

To overcome these limitations, this study employs transfer learning from a well-established external dataset: the NDSB dataset released on Kaggle in 2015. The NDSB dataset, made publicly available during a large-scale data science competition, consists of over 30,000 grayscale images distributed across 121 plankton classes [5]. These images were captured using the ISIIS system in temperate and subtropical marine regions. Although geographic and species differences exist, taxonomic groups appear in both the NDSB dataset and Caribbean plankton samples, making transfer learning a viable strategy for knowledge reuse.

While transfer learning has been widely adopted in fields like medical imaging and object recognition, its application to marine plankton classification particularly in ecologically underrepresented regions remains limited. Most existing studies focus on well resourced regions in North America, Europe, or East Asia, where large annotated datasets are available and domain characteristics are more uniform. In contrast, regions such as the Caribbean face challenges including limited data, diverse and unique species distributions, and a lack of pretrained models suited to their specific ecosystems. These gaps present a critical need to evaluate whether models trained on large foreign datasets can effectively support classification tasks on sparse, local datasets from data-poor environments.

Advancements in deep learning techniques, for the most part convolutional neural networks (CNNs), have significantly enhanced the automated identifi-

cation of plankton from image data. CNN-based systems have been shown to perform well even with limited data when pretrained on larger collections. Nevertheless, challenges remain in ensuring robustness when transferring models across datasets with variations in species makeup, image quality, and environmental settings [6]. The intention of this research is to strengthen the classification accuracy on the University dataset by utilizing a CNN trained on a merged dataset that combines both the NDSB Kaggle and the Department of Life Sciences UWI datasets. Model evaluation is focused specifically on the University test partition to assess how effectively knowledge from the external dataset transfers to the regional context. The resulting system offers a practical and scalable approach to plankton monitoring in resource-limited regions, while also demonstrating the broader utility of leveraging global datasets to support regional ecological research.

This research begins by detailing the data acquisition and preprocessing steps applied to both local and external plankton image datasets. It then presents the convolutional neural network architecture and describes the transfer learning strategies used to assess cross-domain performance. The results section provides a comparative evaluation of classification accuracy across different dataset configurations. This is accompanied by a discussion of the broader implications for automated ecological monitoring in data-constrained regions. The research concludes by highlighting its main contributions and outlining potential directions for advancing plankton classification research.

2 Related Work

Yuan et al. [7] developed a lightweight convolutional neural network (CNN) tailored for in-situ plankton classification, targeting embedded devices for real-time ecological monitoring. The study utilized a custom dataset comprising 6,674 plankton images from 12 categories, augmented with environmental metadata such as sampling depth and time. The authors modified the MobileNetV2 architecture by integrating coordinate attention modules and streamlining convolutional blocks to optimize for low-latency environments. Experimental results showed that the improved model achieved an accuracy of 95.46% and recall of 94.48%, with an average inference time of only 6.15 ms per image. When deployed on the Huawei Ascend Atlas board, the model enabled real-time plankton abundance profiling, underscoring its applicability for autonomous marine sensors.

Hassan et al. [8] proposed a fused deep learning framework for accurate and interpretable plankton classification using the WHOI benchmark dataset. Their architecture combines feature maps from InceptionResNetV2 and a custom network, DeepPlanktonNet. Feature fusion is optimized through the Whale Optimization Algorithm (WOA), while the LIME (Local Interpretable Model-Agnostic Explanations) tool is employed to enhance model interpretability. The proposed method achieved 98.79% accuracy, outperforming prior models on the WHOI dataset. The use of WOA allowed the model to reduce feature dimensionality, thereby improving both generalization and computational efficiency.

LIME visualizations confirmed that key plankton structures were the dominant drivers of model predictions, adding a layer of trust and transparency.

Eerola et al. [9] conducted a comprehensive survey of plankton image classification techniques, analyzing over two decades of progress from classical feature extraction to state-of-the-art deep learning models. The review focuses on multiple datasets including WHOI, ZooScan, and laboratory-captured samples, addressing the specific challenges faced in aquatic imaging such as class imbalance, domain shift, and annotation scarcity. The authors evaluated various model families such as CNNs, transformer-based architectures, ensemble models, and semi-supervised approaches. They emphasized the importance of lightweight models for edge deployment and called for more interpretable AI tools in marine science. This survey provides a useful roadmap for future research at the intersection of computer vision and aquatic ecology.

Shi et al. [10] introduced a hybrid quantum-classical convolutional neural network for the classification of phytoplankton imagery. Motivated by the potential of quantum computing for high-dimensional data processing, their architecture combines classical CNN layers with quantum variational circuits. A small, controlled laboratory dataset of phytoplankton species served as the testbed for evaluating performance. The hybrid model showed faster convergence and achieved accuracy levels comparable to classical CNN baselines. While the work remains experimental, it opens the door to leveraging quantum acceleration in ecological monitoring and highlights the viability of quantum-classical models for complex image classification tasks.

Liu et al. [11] presented DeepLOKI, a robust plankton classification system built upon the ResNet-18 architecture, enhanced through self-supervised learning and cross-device domain adaptation. The dataset consisted of high-resolution zooplankton images captured using multiple imaging systems, annotated into taxonomic categories. Pretraining was performed using unlabeled data, followed by supervised fine-tuning on labeled subsets. The model achieved an overall accuracy of 83.9% and was able to generalize effectively across imaging platforms. Its capacity to recognize rare and morphologically ambiguous taxa demonstrates the benefits of self-supervised learning in overcoming dataset imbalance and annotation limitations.

3 Datasets

Two distinct plankton image datasets were utilized in this study: a localized dataset collected by the Department of Life Science at the UWI, from the coastal waters of Trinidad and Tobago, referred to as the University dataset, and a larger, publicly available dataset from the NDSB hosted on the Kaggle platform referred to as the Kaggle dataset.

3.1 University Dataset

The University dataset contains 44 plankton images captured from the nearshore waters of Trinidad and Tobago. The data represents a limited, region-specific col-

lection, reflecting the challenges of data acquisition. These samples were prepared through a detailed laboratory pipeline involving sample extraction, microscopy-based imaging, and the manual segmentation and labeling of individual organisms, as illustrated in Fig. 1.

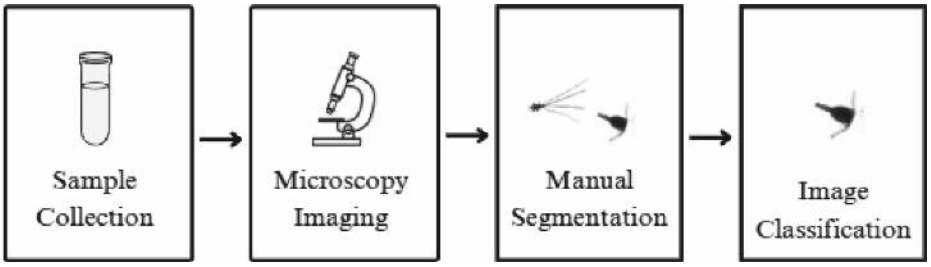


Fig. 1. Manual laboratory process of plankton image classification.

The dataset captures seven distinct plankton classes: *Bacteriastrum*, *Chaetoceros*, *Diatoms*, *Pyrophacus*, *Tintinnopsis*, *Zooplankton*, and *Dinoflagellates*. As shown in Table 1, the distribution of images across classes is imbalanced, reflecting the natural prevalence of these organisms in the sampled marine environment. The limited size of the dataset makes it difficult to train deep learning models effectively from the ground up, thereby supporting the need for transfer learning.

Table 1. Class distribution in the University dataset

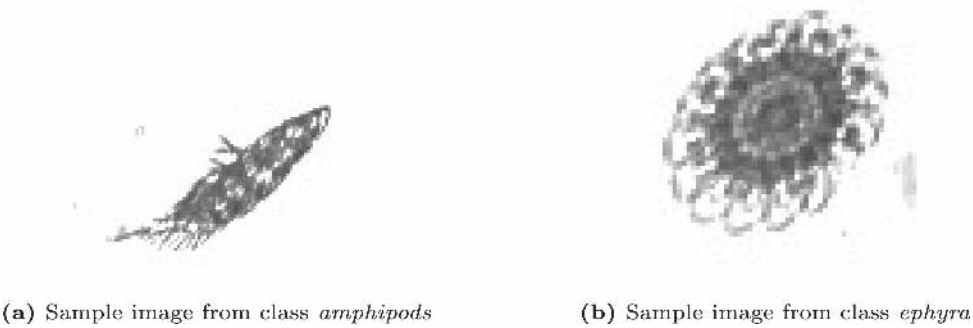
Class Name	Image Count
Bacteriastrum	2
Chaetoceros	4
Diatoms	7
Pyrophacus	5
Tintinnopsis	4
Zooplankton	17
Dinoflagellates	5
Total	44

3.2 Kaggle Dataset

The Kaggle dataset is a large-scale collection comprising over 30,000 grayscale plankton images, categorized into 121 distinct classes. These images were acquired using the ISIIS, which captures high-resolution underwater imagery

in natural environments. The dataset covers various categories of plankton, such as phytoplankton and zooplankton, captured under different imaging scenarios. This variability provides a broad and representative base of visual features, making the dataset highly suitable for training deep learning models.

In contrast to the smaller and region-specific University dataset, the Kaggle dataset offers both scale and class diversity. However, it is not uniformly balanced since some plankton categories contain thousands of images, while others are represented by only a few samples. Despite this imbalance, the size and diversity of the dataset support the development of models that are capable of learning generalized features applicable across different marine environments. As a widely used benchmark in the field of plankton classification, the Kaggle dataset serves as a valuable global reference. In this study, it is leveraged as a pretraining source to enhance classification performance on the smaller, localized dataset from Trinidad and Tobago through transfer learning. Samples from the Kaggle plankton dataset are shown in Fig. 2.



(a) Sample image from class *amphipods* **(b)** Sample image from class *ephyra*

Fig. 2. Representative samples from the Kaggle plankton dataset.

4 Methodology

4.1 Overview

This section describes the experimental setup and methods used to assess how well deep learning performs in classifying plankton images, especially when working with a small amount of locally sourced data. It outlines the datasets used, the pre-processing steps applied, the CNN architecture employed, and the distinct training and evaluation scenarios designed to assess the impact of a larger external dataset on localized classification performance. Three main experimental configurations were explored, which are one using only the University dataset for both training and testing, another using just the Kaggle dataset, and a third combining both datasets for training, while testing was conducted on the University portion. These scenarios were developed to quantify the benefit of transfer learning from a large, diverse external dataset to a smaller, domain-specific

dataset. Cross-validation methods were integrated to ensure robust and generalizable performance metrics across these scenarios. This framework enables a comprehensive analysis of how external data can enhance plankton classification in resource-limited ecological settings.

4.2 Image Loading and Resizing

All images from both the University and Kaggle datasets were loaded using the Open CV library (cv2.imread). During loading,images were uniformly resized to a dimension of 128×128 pixels.This standardization is necessary to ensure all inputs to the CNN model have consistent spatial dimensions, regardless of their original resolution. Images were read in their native color format (3 channels).

4.3 Pixel Normalization

Image pixel intensities, usually spanning 0 to 255, were scaled to a range between 0.0 and 1.0 by converting them to the float32 format and dividing by 255. This normalization step promotes training stability by placing all input features on a consistent scale, which can help speed up the optimization process.

4.4 Label Encoding

Original class labels, which were stored as folder name strings, were first translated into numerical values. These values were then encoded using the one-hot method, where each label is represented by a binary array with 1 marking the active class and 0 for all others. This encoding approach is commonly applied in multi-class classification tasks, particularly when using categorical cross-entropy as the loss function.

4.5 Data Augmentation

To increase dataset diversity and model robustness, a suite of data augmentation techniques was applied using Keras' ImageDataGenerator. These included:

- **Rotation:** Applying random rotations to images within a specified angle range to mimic different viewing angles.
- **Width and Height Shifts:** Moving images along the horizontal or vertical axis by a certain proportion (e.g., up to 20%) to introduce spatial variation.
- **Shearing and Zooming:** Distorting images using shear transformations and scaling up or down by upto 20
- **Horizontal Flipping:** Flipping images along the vertical axis to capture mirror-image instances.
- **Fill Mode:** Filling in newly created pixels after transformations using the nearest method.

These transformations were applied only during training to avoid information leakage into the validation or test sets.

4.6 Train-Test Validation Splits

The University dataset, despite its limited size, was rigorously split to ensure robust evaluation. Initially, the entire University dataset ($X_{university}$, $y_{\text{uni-encoded}}$) underwent a single train-test split. A proportion of 20% of the data was reserved as a dedicated, unseen test set ($X_{test-university}$, $y_{\text{test-uni}}$) for final model evaluation. The remaining 80% formed the University training set ($X_{train-university}$, $y_{\text{train-uni}}$). This split was performed with shuffle to a fixed random sate to ensure reproducibility and a representative distribution of classes across partitions.

The Kaggle dataset was divided using stratified sampling, with 80% of the images allocated for training and the remaining 20% reserved for testing. This stratified train-test split preserved the relative class proportions across both subsets, ensuring that all well-represented classes were adequately sampled during training and evaluation. A fixed random seed was used to guarantee reproducibility of the splits, and this partition formed the basis for the baseline performance scenario using only the Kaggle data.

For the combined training scenario, the Kaggle dataset was used in its entirety (after re-mapping to align with University classes). The University training set ($X_{train-university}$, $y_{\text{train-uni}}$) was then combined with (X_{kaggle}, $y_{\text{kaggle-encoded}}$) to form the comprehensive training dataset ($X_{train-combined}$, $y_{\text{train-combined}}$).

4.7 CNN Architecture

A sequential CNN architecture was designed for plankton image classification. The model's design balances computational efficiency with sufficient capacity to learn complex visual features from the plankton images. The architecture comprises three convolutional blocks, followed by flattening and dense layers. The model structure is as follows:

- **Input Layer:** This is where the model first looks at the image. It takes in color images that are 128×128 pixels in size. The model starts by scanning small patches (3×3) of the image using 32 filters and uses a method called ReLU to help it learn better.
- **Pooling Layer 1:** MaxPooling2D$\big((2,2)\big)$ Here, the image size is reduced by half. This makes the model faster and helps it concentrate on the most essential key components of the image.
- **Convolutional Layer 2:** In this step, halving the image dimensions increases the model's processing speed and sharpens its attention on the most crucial visual elements.
- **Pooling Layer 2:** MaxPooling2D$\big((2,2)\big)$ Here, the image size is reduced by half. This makes the model faster and helps it focus on the most important parts of the image.
- **Convolutional Layer 3:** Here, the image size is reduced by half. This makes the model faster and helps it focus on the most important parts of the image.

- **Pooling Layer 3:** MaxPooling2D$\big((2,2)\big)$ This layer further condenses the image data, enabling the model to concentrate exclusively on the most critical information.
- **Flatten Layer:** All the data from the previous steps, which is in a 3D format, is flattened into a straight line of numbers so it can be used for decision-making.
- **Dense Layer 1:** This is a fully connected layer with 512 units. It gathers and processes all the information learned so far to help the model make better decisions.
- **Dropout Layer:** Dropout(0.5) This is a fully connected layer with 512 units. It gathers and processes all the information learned so far to help the model make better decisions.
- **Output Layer:** Dense$\big($num_classes, activation = softmax$\big)$ In the final step, the model decides which class the image belongs to. It does this by giving each possible class a score and choosing the one with the highest value, using something called a softmax function.

The model utilized the Adam optimizer during compilation, benefiting from its adaptive learning rate that supports efficient and reliable training in deep learning applications. For loss calculation, categorical cross-entropy was employed, making it well-suited for handling multi-class classification tasks with one-hot encoded targets. Model performance was monitored using accuracy as the primary metric.

4.8 Evaluation Metrics

A diverse set of evaluation metrics was utilized to thoroughly analyze how well the model performed. Each metric provided specific information that helped assess different dimensions of classification accuracy and reliability.

- **Accuracy:** Accuracy tells us how many predictions the model got right overall. It adds up the number of correct positive and negative predictions and compares them to the total number of predictions. As shown in Eq. (1), accuracy is computed as the ratio of correctly predicted samples (true positives and true negatives) to all predictions made. However, it may give a false sense of performance when the dataset has uneven class sizes.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Loss:** Loss is a value that shows how far the model's predicted answers are from the actual ones. A smaller loss means the model is making better predictions. The specific method used here, called categorical cross-entropy, is designed for problems where there are multiple possible output categories. Equation (2) defines this loss as the negative log-likelihood of the true class probabilities.

$$\text{Loss} = -\sum_{i} y_i \log(\hat{y}_i) \tag{2}$$

- **Precision:** Precision measures how often the model is correct when it predicts something as positive. High precision means the model does not make many false positive errors. It's especially important when predicting something wrongly as positive can cause problems. As shown in Eq. (3), precision is the ratio of true positives to the total predicted positives.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

- **Recall (Sensitivity):** Recall looks at how well the model finds all the actual positive cases. It's important in situations where missing a true positive (like a medical diagnosis) can be harmful. High recall means the model finds most of what it should. Equation (4) defines recall as the ratio of true positives to all actual positives.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

- **F1-Score:** The F1-score combines precision and recall into one number. It balances both, which is helpful when you need to consider both missed positives and incorrect positives at the same time. Equation (5) shows that the F1-score is the harmonic mean of precision and recall.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

Confusion matrices provide a detailed view of class-level performance by organizing predictions in a tabular format. They display the relationship between the predicted labels and the true labels across all classes. In this matrix, rows indicate the actual classes, while columns represent the predicted classes. Correct predictions appear along the diagonal, while incorrect ones are shown in the off-diagonal positions. This format helps to pinpoint recurring classification errors and identify which classes are often misclassified. This visualization allows for the identification of systematic errors, such as specific classes being consistently confused with others. Analyzing confusion matrices is particularly useful in multi-class classification tasks to assess per-class strengths and weaknesses of the model.

4.9 Implementation

All experiments were conducted using Google Colab Pro, which provided access to cloud-based GPU acceleration. Utilizing Colab Pro's high random access memory (RAM) environment and Graphics Processing Units (GPUs) significantly reduced training and inference times, making it feasible to experiment with multiple configurations efficiently. All codes were implemented in Python using TensorFlow and OpenCV libraries.

4.10 University Dataset Baseline

This setup served as the benchmark for evaluating CNN performance using only locally sourced data. The CNN model was both trained and evaluated on the 44-image University dataset. The dataset was divided into 80% for training and 20% for testing, with stratification by class to preserve label distribution. Training was carried out over 20 epochs, incorporating early stopping based on validation loss. To address class imbalance, techniques such as data augmentation and class weighting were used.

4.11 Kaggle Dataset Baseline

This scenario aimed to establish a performance baseline for the CNN when trained solely on the large, global Kaggle dataset that had over 30,000 images. The dataset was carefully selected to improve training effectiveness by focusing on categories that exhibit clear and consistent visual characteristics. The model was trained and evaluated using an 80/20 train-test split. Class weighting and data augmentation were applied, and the CNN was trained for 20 epochs.

4.12 Combined Training with University-Focused Baseline

This is the core experimental scenario designed to investigate the primary research question on whether incorporating the Kaggle dataset improves classification performance on the limited University dataset. It involved combining the University training subset with the Kaggle dataset to create a diverse and enriched training set. A total of over 30,000 Kaggle images were incorporated to leverage their extensive feature variability. To maintain label consistency, a unified label map was generated and the Kaggle data was reindexed accordingly. The combined dataset was shuffled with a fixed seed prior to training to ensure a randomized yet reproducible sample distribution. Evaluation was conducted exclusively on the University test partition to isolate the effects of external data on local classification performance.

5 Results

This section summarizes the results obtained from three experiments aimed at assessing CNN-driven plankton classification across different levels of data availability. Each scenario progressively explores the impact of training data scale and source:

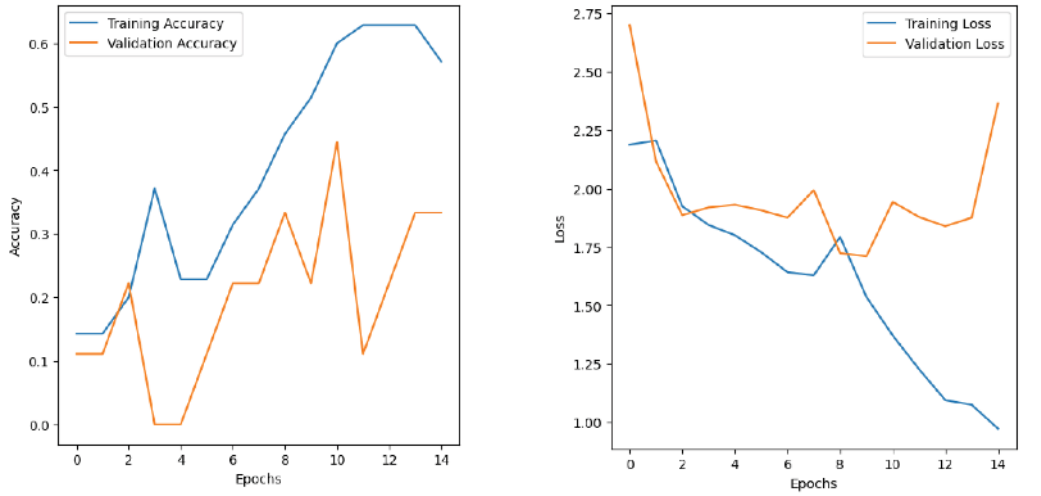
1. Training exclusively on the small, localized University dataset.
2. Training on a large, diverse global dataset from Kaggle.
3. Combining both datasets to assess the benefits of transfer learning on performance over the local test set.

Evaluation metrics include training and test accuracy, loss, confusion matrices, and class-wise F1-scores, providing a comprehensive view of model effectiveness across different training regimes.

5.1 University Dataset

Training on the small University dataset of 44 images achieved limited success. The CNN model achieved a validation accuracy of 33.3%, with a training accuracy of 62.7% and a validation loss of 1.875. These results are illustrated in Fig. 3, which shows the accuracy and loss trends over training epochs. The confusion matrix revealed that only a subset of classes was predicted correctly, with many predictions biased toward the majority class. Precision and recall were low across most classes, indicating overfitting and limited generalization. The test accuracy was 22.2%.

The confusion matrix highlighted a strong prediction bias toward dominant classes such as *Zoo plankton*, while minority classes like *Bacteriastrum* and *Tintinnopsis* were consistently misclassified.



(a) Training and validation accuracy over epochs for the University dataset

(b) Training and validation loss over epochs for the University dataset

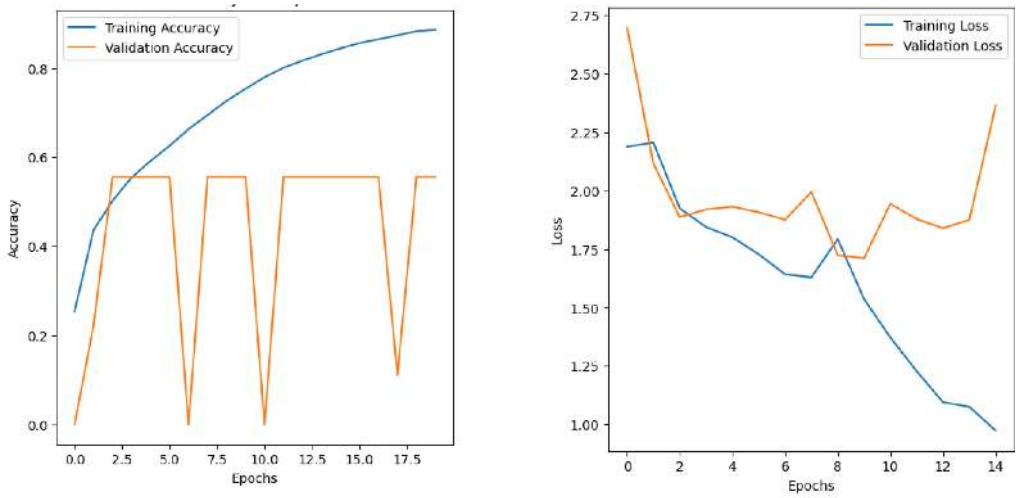
Fig. 3. CNN Performance metrics for training on the University dataset.

The low performance in this scenario can be attributed to the extreme scarcity of training data, which severely limited the model's ability to learn discriminative features. The significant overfitting gap between training and validation accuracy suggests that the model memorized the training examples rather than learning generalizable patterns. The confusion matrix reveals that predictions were disproportionately skewed toward the majority class (Zooplankton), which is expected in highly imbalanced datasets without strong regularization or prior knowledge. These findings emphasize the limitations of training deep models from scratch in data-constrained regions.

5.2 Combined Training with University-Focused Evaluation

This scenario demonstrated significant improvement. The model trained on over 30,000 images achieved a test accuracy of 55.6% on the University dataset, more

than double the performance of Scenario 1. The training accuracy reached 89%, and validation accuracy peaked at 55.6%, with reduced validation loss (1.41). These results are illustrated in Fig. 4, which shows the accuracy and loss trends over training epochs. The confusion matrix indicated better coverage across classes, though some misclassification remained. Notably, the model learned to recognize classes like Diatoms and Chaetoceros more accurately, indicating successful knowledge transfer from the Kaggle dataset to the local domain. Combining the Kaggle dataset with the University training set significantly improved model performance.



(a) Training and validation accuracy over epochs for the combined training with University-focused evaluation

(b) Training and validation loss over epochs for the combined training with University-focused evaluation

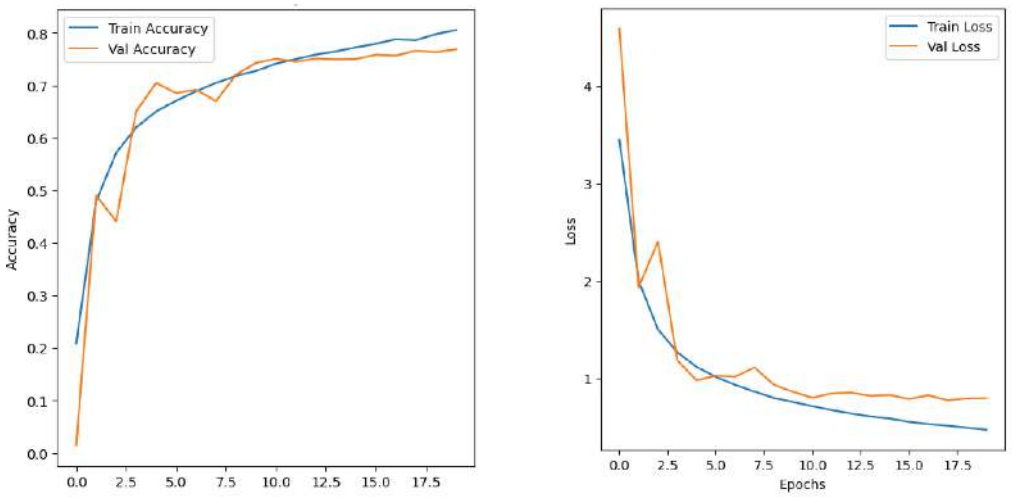
Fig. 4. Performance metrics during training of the CNN on the combined training with University-focused evaluation.

The noticeable performance gain in this scenario illustrates the impact of transfer learning, where the CNN pretrained on diverse Kaggle data was able to extract general visual features transferable to the University dataset. Interestingly, certain classes like Diatoms and Chaetoceros showed marked improvement, likely because their morphological traits are well-represented in the Kaggle dataset. However, residual misclassifications indicate that domain shift effects, such as different imaging conditions or subtle species-level differences, still present challenges. This suggests that while transfer learning improves performance, domain-specific fine-tuning remains critical for peak accuracy.

5.3 Kaggle Dataset

Training solely on the filtered Kaggle dataset yielded the highest performance, with a final test accuracy of 76.6% and a test loss of 0.77. The classification report showed strong precision, F_1, and recall scores across the classes.

Top-performing classes included trichodesmium_puff, chaetognath_other, and copepod_cyclopoid_oithona_eggs. In contrast, underperforming classes such as unknown_unclassified and detritus_blob highlighted potential labeling or feature ambiguity. The learning curves (Fig. 5) exhibited stable convergence with minimal overfitting, underscoring the effectiveness of the data augmentation strategies and class balancing techniques applied during training.



(a) Training and validation accuracy over epochs for the Kaggle dataset

(b) Training and validation loss over epochs for the Kaggle dataset

Fig. 5. Performance metrics during training of the CNN on the Kaggle dataset.

The high accuracy achieved when training and testing solely on the Kaggle dataset reflects the internal consistency and large volume of data available in that benchmark. The low test loss and high precision across most classes confirm that the model performs well when data is abundant and class distributions are broad.

5.4 Results Comparison

To holistically compare model performance across the three experimental scenarios, Table 2 shows the test accuracy achieved under each setting: training solely on the University dataset, exclusively on the Kaggle dataset, and on a combined dataset evaluated on the University test set.

The progression across these scenarios highlights a core insight, that is deep learning models depend heavily on the availability and diversity of training data. While the Kaggle-only scenario offers strong baseline performance, the University-only setup fails due to data insufficiency. Incorporating a large-scale external dataset such as Kaggle significantly improves performance. The combined approach achieves more than double the accuracy of the University-only baseline. It effectively balances generalization and local adaptation, proving the

Table 2. Summary of performance across scenarios

Training Scenario	Test Accuracy (%)	Test Loss
University	22.2	1.875
Combined (Kaggle + Univ)	55.6	1.412
Kaggle	76.6	0.7729

utility of transfer learning for ecological applications in resource-limited regions. Notably, this improvement comes despite the model being evaluated solely on University data, affirming the hypothesis that diverse external datasets can enhance generalization in localized, resource-constrained ecological settings.

6 Discussion

This section interprets the results presented previously, explaining their significance and situating them within the broader context of plankton classification and deep learning. It moves from the specific findings to their wider implications, offering insights into the effectiveness of the proposed methodology and its relevance to ecological monitoring in data-limited regions.

6.1 Key Findings and Hypothesis Validation

This study primarily investigated the efficacy of leveraging a large, publicly available dataset via transfer learning to enhance plankton classification performance on a significantly smaller, regionally specific dataset. The results demonstrate that combining the University dataset's training partition with the extensive Kaggle dataset during CNN training yielded superior classification performance on the unseen University dataset, compared to training on the University dataset alone. This substantiates the central hypothesis that is augmenting limited local data with a large, relevant external dataset through transfer learning significantly improves model generalization for regional ecological monitoring.

Notably, the University and Kaggle datasets did not share any explicitly labeled classes in common. However, qualitative inspection revealed that several classes within the University dataset, such as *Diatoms* and *Dinoflagellates*, encapsulated a variety of subcategories and morphological variants. These subclasses are well represented in the Kaggle dataset, as evident from the WHOI taxonomy [12], suggesting that the model was able to recognize intra-class diversity beyond the scope of human labeling. This further supports the premise that the CNN, through transfer learning, not only generalized well but in some instances demonstrated greater discriminatory power than manual labeling.

6.2 Interpretation of Performance Improvements

The performance enhancement observed with the Kaggle dataset can be attributed to the foundational principles of transfer learning. The vast volume

and diversity of images within the Kaggle dataset enabled the CNN to learn a rich hierarchy of generalizable visual features such as edges, textures, and morphological structures common across various plankton taxa. When fine-tuned with the limited University data, these pre-learned features served as a robust starting point, allowing the model to adapt efficiently to the nuances of Caribbean plankton imagery without requiring extensive local examples for initial feature extraction.

This approach effectively mitigated the challenges posed by the small size and class imbalance of the University dataset. Training a CNN from scratch under such constraints would likely result in poor generalization due to overfitting and insufficient exposure to varied data points. Instead, transfer learning provided a mechanism for knowledge reuse, where globally learned patterns were refined for local specificity.

6.3 Strengths of the Approach

A major strength of the proposed method is its ability to scale deep learning to environments with limited labeled data. The use of the well-curated global Kaggle dataset allowed for effective feature learning, while data augmentation, class reindexing, and label harmonization ensured that model training was not adversely impacted by dataset heterogeneity. The experimental framework also incorporated robust evaluation practices, including stratified train-test splits, confusion matrices, and F1-score analysis to ensure fair assessment of generalization performance.

6.4 Limitations

Although the outcomes are encouraging, the study is not without limitations. The main drawback lies in the limited size of the University dataset, which contains only 44 images. While transfer learning and data augmentation significantly improved performance, the limited diversity of actual local examples still poses a fundamental barrier to achieving high classification performance for all specific subcategories or rare species within the Caribbean context.

While newer plankton collections have appeared in recent years, none match the open availability, single-organism vignette format, and class breadth of the NDSB dataset from Kaggle. More recent options like the WHOI IFCB system, while extensive with over 3 million continuous flow-cytometry images across 70 classes, lack pre-segmented, one-cell vignettes and typically require institutional access agreements [13]. Similarly, the ZooScan system generates large mixed-sample scans demanding offline segmentation and manual validation, which would introduce significant additional computational and manual overhead to the rapid, transfer learning based classification pipeline employed in this study. Since no open-access release since 2015 offers this unique combination of scale, taxonomic diversity, and single-organism segmentation, the NDSB Kaggle collection remains the sole viable public benchmark aligning with the objective of

enhancing local data performance through readily available global knowledge transfer.

Moreover, although class remapping was performed, the inherent class imbalance within both the Kaggle and University datasets likely influenced the per-class performance, leading to lower recall for minority classes, even with class weighting. The Kaggle dataset itself, though extensive, may not fully capture the morphological diversity or environmental conditions specific to the Caribbean region. As a result, some domain mismatch effects persisted.

Another challenge stems from the visual complexity and small physical size of some plankton taxa, which are often difficult to distinguish even by expert annotators. While the CNN model exhibited strong generalization, accurate identification of minute or visually ambiguous classes remains an ongoing hurdle.

Furthermore, by conducting experiments on Google Colab Pro, the study leveraged accessible cloud infrastructure to accelerate training. However, this platform introduced variability in performance due to shared resource allocation. This fluctuation in compute availability occasionally affected model convergence consistency and training time, presenting both a technical and financial consideration for researchers dependent on such environments.

6.5 Future Work

Future studies should aim to increase both the quantity and taxonomic variety of Caribbean plankton datasets, enabling a more comprehensive analysis of the region's biodiversity. This would provide a more comprehensive local training base, reducing reliance on external datasets for effective fine-tuning. Advanced transfer learning strategies, such as domain adaptation, few-shot learning, or meta-learning, could further enhance classification performance for rare or under-represented classes.

Exploring alternate CNN architectures, including lightweight models optimized for mobile or edge deployment, would be valuable for real-time monitoring systems. Ultimately, integrating these classification systems into in-situ plankton monitoring pipelines could support continuous ecological assessments and improve responsiveness to environmental changes in under-monitored regions.

7 Conclusion

This study demonstrates the feasibility and effectiveness of using transfer learning from a large, public plankton image dataset (Kaggle and NDSB) to enhance classification performance on a small, localized dataset from Trinidad and Tobago. The experimental results showed that the combined training scenario leveraging both global and local data achieved a test accuracy of 55.6%, more than double the accuracy obtained when training solely on local data (22.2%). This highlights the substantial performance gap between models trained solely on limited local data and those augmented with globally sourced data. The combined training scenario improved test accuracy by over 30% compared to the

local baseline, validating the hypothesis that incorporating external datasets can substantially improve classification accuracy in data-constrained environments.

In particular, the successful transfer of learned features across geographically and taxonomically diverse datasets underscores the robustness of convolutional neural networks for marine image classification tasks. Despite the absence of exact label alignment between datasets, the model effectively generalized to local plankton categories, benefiting from the rich morphological representations present in the external data.

This study provides a practical framework for regions with limited annotated datasets to deploy deep learning based classification systems without requiring extensive local labeling efforts. It demonstrates a scalable and cost-effective strategy for biodiversity monitoring, with implications for advancing automated ecological assessments in underrepresented marine regions.

Future directions include expanding the local dataset for better coverage of Caribbean taxa, integrating domain adaptation techniques to address dataset discrepancies more robustly, and exploring edge-deployable models for real-time environmental monitoring. Ultimately, this study contributes to the broader goal of democratizing access to intelligent environmental sensing systems by aligning global resources with regional conservation needs.

References

1. Lombard, F., et al.: Globally consistent quantitative observations of planktonic ecosystems. Front. Marine Sci. **6**, 196 (2019). https://doi.org/10.3389/fmars.2019.00196
2. Falkowski, P.G., Barber, R.T., Smetacek, V.: Biogeochemical controls and feedbacks on ocean primary production. Science **281**(5374), 200–206 (1998). https://doi.org/10.1126/science.281.5374.200
3. Irisson, J.-O., Beamish, P., Culverhouse, P.F., et al.: Automatic plankton identification using computer vision and machine learning. Ann. Rev. Marine Sci. **13**, 485–512 (2021). https://doi.org/10.1146/annurev-marine-121219-081634
4. Cowen, R.K., Guigand, C.M.: Advancing the frontiers of in situ plankton imaging. Ann. Rev. Marine Sci. **6**, 43–65 (2014). https://doi.org/10.1146/annurev-marine-120710-100610
5. Cowan, R.K., Guigand, C.M.: National Data Science Bowl Plankton Dataset. Kaggle (2015). https://www.kaggle.com/competitions/datasciencebowl
6. Zheng, H., Wang, R., Yu, Z., Wang, N., Gu, Z., Zheng, B.: Automatic plankton image classification combining multiple view features via multiple kernel learning. BMC Bioinf. **18**(16), 570 (2017). https://doi.org/10.1186/s12859-017-1954-8
7. Yuan, C., et al.: Research on in situ observation method of plankton based on convolutional neural network. J. Marine Sci. Eng. **12**(10), 1702 (2024). https://doi.org/10.3390/jmse12101702
8. Hassan, M., Salbitani, G., Carfagna, S., Khan, J.A.: Deep learning meets marine biology: optimized fused features and LIME-driven insights for automated plankton classification. Comput. Biol. Med. **192**, 110273 (2025). https://doi.org/10.1016/j.compbiomed.2025.110273

9. Eerola, T., et al.: Survey of automatic plankton image recognition: challenges, existing solutions and future perspectives. Artif. Intell. Rev. (2023). https://doi.org/10.1007/s10462-024-10745-y
10. Shi, S., et al.: Hybrid quantum-classical convolutional neural network for phytoplankton classification. Front. Marine Sci. **10**, 1158548 (2023). https://doi.org/10.3389/fmars.2023.1158548
11. Oldenburg, E., Kronberg, R.M., Niehoff, B., Ebenhöh, O., Popa, O.: DeepLOKI: a deep learning based approach to identify zooplankton taxa on high-resolution images from the optical plankton recorder LOKI. Front. Marine Sci. **10**, 1280510 (2023). https://doi.org/10.3389/fmars.2023.1280510
12. Woods Hole Oceanographic Institution: WHOI Plankton Classification Index (2024). https://whoigit.github.io/whoi-plankton/index.html
13. Orenstein, E.C., Beijbom, O., Peacock, E.E., Sosik, H.M.: WHOI-plankton – a large scale fine grained visual recognition benchmark dataset for plankton classification. arXiv preprint arXiv:1510.00745 (2015)