Automated Taxonomic Identification of Marine Plankton Using a Large-Language-Model and Retrieval-Augmented-Generation Framework

Jaronchai Dilokkalayakul*†[1], Akane Kitamura[2], and Takeshi Obayashi‡[1,2]

[1]Graduate School of Information Sciences, Tohoku University – Japan
[2]Advanced Institute for Marine Ecosystem Change (WPI-AIMEC), Tohoku University – Japan

Abstract

Marine plankton are essential to ocean ecosystems and serve as indicators of environmental change and biodiversity patterns. Monitoring plankton communities at scale is increasingly important for understanding ecological dynamics and the impacts of climate change. However, traditional methods for plankton identification rely on manual classification, which are time-consuming and difficult to scale.
While convolutional neural networks (CNNs) have been applied to automate image classification, they typically require large, labeled datasets. Furthermore, CNNs struggle to incorporate contextual information such as environmental metadata or morphological variability, factors that are considered by human experts during manual classification.

Our framework addresses both challenges using large language models (LLMs) and retrieval-augmented generation (RAG). LLMs offer greater flexibility than CNNs by reasoning over multimodal inputs-such as image and texts-allowing for more context-aware classification. RAG enhances this process by retrieving semantically similar examples from a curated plankton image database and injecting them into the model's input as context. This mechanism supplies the model's internal knowledge with relevant examples at inference time, reducing the system's reliance on large labeled training datasets. Preliminary results using PlanktoScope video data suggest that this approach yields promising performance in classifying plankton.
The system processes plankton samples to detect individual plankton, then pairs each instance with a classification prompt. Retrieved examples supply context, and the LLM produces structured JSON output that includes family and genus classification metadata. The modular architecture supports scalable deployment and potential integration with long-term monitoring platforms. It also has applications in microbiome annotation, while reducing manual workload.

Keywords: Planktons, Marine Plankton, Large Language Models (LLM), Retrieval Augmented Generation (RAG), Vector Embedding, Semantic Search, Marine Biodiversity Monitoring, Image Based Classification, PlanktoScope, Automated Taxonomic Classification

*Speaker
†Corresponding author: dilokkalayakul.jaronchai.p8@dc.tohoku.ac.jp
‡Corresponding author: takeshi.obayashi.a6@tohoku.ac.jp