PlanktonShift: Cross-Domain Plankton Image Classification with Calibration, Interpretability, and Trait Proxies

Technical Report (PhD CASE Preparation Project)

Samuel Moses Orokpo

MSc Big Data Analytics, Edge Hill University (UK)

January 2026

Abstract

This report documents a successful proof-of-concept machine learning (ML) system for plankton image analytics, developed as preparatory work for a PhD CASE studentship on applying ML to plankton imaging and biodiversity assessment. The work delivers an end-to-end, reproducible pipeline for cross-domain classification under dataset shift, complemented by probability calibration, explainability (Grad-CAM), and lightweight trait proxy extraction (e.g., size/ESD). The system is implemented in TensorFlow and packaged with a Streamlit demonstration application to support rapid inspection, qualitative validation, and stakeholder-facing communication.

1 Project Overview

Project name: PlanktonShift
Repository: https://github.com/freelansire/PlanktonShift

PlanktonShift focuses on: (i) plankton image classification under cross-domain transfer, (ii) uncertainty estimation and calibration for trustworthy deployment, and (iii) trait proxy extraction to bridge ML outputs with ecological interpretation.

2 Datasets

Two open domains were integrated to create a realistic cross-domain benchmark:

- **IFCB-inspired imaging context:** The system is motivated by imaging-in-flow monitoring and the Imaging FlowCytobot concept [1].
- **PlanktonSet 1.0 (NDSB 2015):** Plankton imagery used in the 2015 National Data Science Bowl [2].

To enable rapid cross-domain experiments, folder-level labels were harmonised into a shared coarse label space (e.g., `diatom`, `protist`, `detritus`, `artifact`, `other`). This supports scalable experimentation while preserving ecologically meaningful groupings.

3 System Design and Implementation

3.1 Model and Training

A lightweight CNN backbone (MobileNetV3Small) was used to balance accuracy and computational efficiency. Training was implemented with:

- **Supervised learning** on the source domain.
- **CORAL feature alignment** using an unlabeled target stream to reduce covariance mismatch in latent features under domain shift.
- **Two-stage optimisation:** (1) train classifier head with frozen backbone; (2) fine-tune deeper layers with a reduced learning rate.

3.2 Calibration and Trustworthy Outputs

The project treats probability quality as a first-class concern and includes:

- **Calibration diagnostics:** Expected Calibration Error (ECE) and Brier score.
- **Post-hoc temperature scaling:** learned on a held-out validation subset to improve reliability of predicted probabilities.

3.3 Explainability and Trait Proxies

To support scientific interpretability and stakeholder confidence:

- **Grad-CAM explanations** highlight salient image regions driving predictions.
- **Trait proxies** are estimated using lightweight segmentation and morphology (area, ESD, axis lengths, and a biovolume proxy), enabling early linkage between ML outputs and ecological size-structure summaries.

4 Demonstration Application

The Streamlit application is included to make the research pipeline *inspectable* and *shareable* beyond scripts and logs. It is most useful at three moments: (i) early verification that the trained model behaves sensibly on real images, (ii) qualitative assessment of domain shift, and (iii) communication with supervisors or non-ML collaborators.

When to run it

- **After training (required):** Run the demo once a trained model file exists (e.g., `models/ifcb_to_planktonset_coral.keras`). This enables interactive prediction and Grad-CAM visualisation.
- **After evaluation and calibration (recommended):** Run the demo after generating the JSON artifacts from `src.eval` and `src.calibrate`. This activates the benchmark viewer tab and supports rapid comparison of pre/post calibration metrics.

What it demonstrates

1. **Predict + Explain (Grad-CAM):** Users upload an image and receive top predicted probabilities alongside a Grad-CAM heatmap overlay. This supports qualitative validation (e.g., whether the model attends to biologically plausible structures rather than background artifacts).
2. **Traits (lightweight morphology):** A second tab produces a segmentation overlay and exports trait proxies (e.g., area and equivalent spherical diameter). This provides an early bridge from classification outputs to trait-oriented ecological summaries.
3. **Benchmark Viewer (JSON artifacts):** A third tab reads precomputed evaluation and calibration JSON files and displays key reliability metrics (ECE, Brier) and confusion matrices. This allows rapid reporting without rerunning heavy computations.

Recommended usage in supervisor discussions

In emails or meetings, the Streamlit demo can be used as a concise walkthrough: show (i) a few representative images from each domain to illustrate domain shift, (ii) Grad-CAM outputs for interpretability, and (iii) before/after calibration metrics to demonstrate a focus on trustworthy ML suitable for downstream scientific and policy-facing use. z

5 Experimental Results (Current Baseline)

This section summarises key metrics observed from the completed run and highlights positive outcomes relevant to cross-domain robustness and trustworthy ML.

5.1 Training Stability and Baseline Accuracy

The training procedure converged stably and achieved strong internal performance on the training/validation split in the completed run:

- **Training accuracy:** ≈ 0.975
- **Validation accuracy:** ≈ 0.964

These results indicate effective feature learning and stable optimisation for the coarse-label setting.

5.2 Calibration Improvements Under Domain Shift

Although domain shift can lead to overconfident predictions, temperature scaling substantially improved probabilistic reliability on the PlanktonSet validation subset:

Metric (PlanktonSet validation)	Before	After Temp. Scaling
Negative log-likelihood (NLL)	2.535	0.823
Expected Calibration Error (ECE)	0.921	0.561
Brier score	1.695	0.629
Temperature T	–	10.0

Table 1: Post-hoc calibration improved probability quality substantially in the completed run (lower is better for NLL/ECE/Brier).

Positive takeaway: the pipeline not only detects calibration issues, but also includes a practical corrective method that yields large improvements in NLL/ECE/Brier—an important property for decision support and policy-relevant reporting contexts where confidence must be interpretable.

6 Relevance to a PhD CASE Project

PlanktonShift aligns strongly with PhD CASE priorities in plankton imaging and biodiversity assessment:

- **Cross-instrument translatability:** explicit evaluation under domain shift, with alignment methods (CORAL) to reduce mismatch.
- **Traits for biodiversity indicators:** trait proxies (ESD/biovolume proxy) support size-structure and ecological change summaries.
- **Usable, reproducible ML:** modular codebase, scripted experiments, and a demonstration interface suitable for collaboration across marine and policy partners.

7 Conclusion

PlanktonShift goes beyond a standard image classifier by directly addressing domain shift, uncertainty calibration, interpretability, and trait-based ecological relevance. The combination of rigorous experimentation and a deployable Streamlit demo provides a credible foundation for further research contributions to plankton monitoring and biodiversity assessment.

References

[1] Olson, R. J. and Sosik, H. M. (2007). A submersible imaging-in-flow instrument to analyze nano- and microplankton: Imaging FlowCytobot. *Limnology and Oceanography: Methods*, 5, 195–203. doi:10.4319/lom.2007.5.195. Accessed 19/01/2026.

[2] Cowen, R. K.; Sponaugle, S.; Robinson, K. L.; Luo, J.; Guigand, C. (2015). PlanktonSet 1.0: Plankton imagery data collected from F.G. Walton Smith in Straits of Florida from 2014-06-03 to 2014-06-06 and used in the 2015 National Data Science Bowl (NCEI Accession 0127422). NOAA National Centers for Environmental Information. Dataset. doi:10.7289/v5d21vjd. Accessed 21/01/2026.