On the Benefits of Uncertainty Estimation for Species Classification: A Case Study on Optimizing Expert re-Annotation

Kamenan N'gadi[2], Martin Laviale[1], Cédric Pradalier[3], and Jérémy Fix[2]

[1] Université de Lorraine, CNRS, LIEC, F-57000 Metz, France
[2] Université de Lorraine, CentraleSupélec, CNRS, Loria, F-57000 Metz, France
[3] GeorgiaTech Europe, CNRS IRL 2958, F-57000, Metz, France

Abstract Automated identification of diatoms from microscopic images is a cornerstone of large-scale freshwater biomonitoring, yet the high diversity of species and strong class imbalance pose significant challenges for reliable classification. In this study, we develop a ResNet-18 baseline for a 759-class diatom dataset and investigate how calibrated confidence estimates support selective classification to reduce expert workload without sacrificing taxonomic integrity. While the initial model exhibits overconfidence with an Expected Calibration Error (ECE) of 0.117, post-hoc temperature scaling ($T = 2.12$) dramatically improves reliability, lowering the ECE to 0.055.By leveraging these calibrated probabilities within a triage framework, we demonstrate that a Selective Macro-F1 score of 99.01% can be maintained at 66.2% coverage (rejection threshold $\tau = 0.934$). Qualitative analysis reveals that the system effectively functions as an intelligent filter, consistently deferring rare taxa—such as AAMB and AASI—and ambiguous morphologies for expert review. Our findings, supported by the torch-uncertainty framework, highlight the potential of simple calibration techniques to make automated diatom identification practical and robust, prioritizing the preservation of rare species metrics essential for ecological assessment. To support reproducibility, our code, environment specifications, and calibration results are available anonymously at `https://anonymous.4open.science/r/greenprwksp-B53D/README.md`

Keywords: Diatoms, Bioindication, Deep learning, Uncertainty quantification, Temperature scaling, Calibration, Selective classification, Risk–coverage, Expert-in-the-loop, Species identification

1 Introduction

The over-exploitation of aquatic resources has made robust bio-monitoring tools paramount. Diatoms —micro-algae are ideal bioindicators due to their ubiquity and sensitivity to environmental variables like pH and nutrient levels, as recognized by the EU Water Framework Directive [1]. Standard bio-monitoring protocols rely fundamentally on taxonomic identification at the species level, traditionally performed via microscopy. However, traditional monitoring is hampered

by two bottlenecks: a **taxonomic bottleneck** caused by **high morphological variability**, and a **human resource bottleneck** due to the global **decline in specialized experts**. Consequently, the burden of routine environmental monitoring falls on a shrinking pool of experts, leading to long turnaround times for samples and high costs for regulatory agencies. [1]

However their identification presents significant challenges for automation. First, the MacDonald-Pfitzer rule [2] dictates that cell size and shape change across a species' life cycle, meaning a classifier must map a "trajectory of shapes" rather than a static template. Second, environmental stressors can induce teratologies (structural malformations) which act as out-of-distribution (OOD) samples. It worth to note that for an identification perspective and in automated pipelines, these OOD samples deviate significantly from the "normal" specimens distribution used during model training. Finally, "cryptic diversity" means morphologically identical specimens may belong to different ecological lineages, introducing inherent label noise and inter-expert disagreement [2].
Given the inherent subjectivity in diatom taxonomy, identification is prone to significant inter-expert variability; different specialists may reach conflicting conclusions for the same specimen, especially with degraded samples. If the training labels for an automated system are inconsistent, the model's performance is fundamentally capped. Consequently, a robust monitoring framework must account for this label noise and provide mechanisms to flag cases where expert disagreement is likely.

The effort to **automate diatom identification** has transitioned from early computer vision techniques [1] to modern deep learning architectures. While recent Deep Learning (DL) models approach human-level accuracy on curated datasets [3], they struggle with real-world noise and tend to be "overconfident." Standard Softmax probabilities often mask underlying uncertainty, which is unacceptable in high-stakes ecological diagnostics [4].

To address this, we need systems that not only classify but also **estimate the reliability of their predictive uncertainty**. Selective classification (learning with rejection) [5] provides a formal paradigm: a confidence threshold decides whether a sample is automatically processed or deferred to a human. By trading off coverage (fraction of samples handled automatically) against selective risk (error rate on those samples), such systems can guarantee that expert involvement is reserved for the most ambiguous cases. Effective selective classification requires a nuanced understanding of predictive uncertainty. Uncertainty [6] is not a monolithic concept; rather, it arises from different sources that have different implications for the monitoring process. **Aleatoric uncertainty** —irreducible noise from imaging artifacts, morphological overlap or label ambiguity. Because aleatoric uncertainty is a property of the data, it cannot be reduced by simply collecting more training samples of the same type. Instead, the model must be able to recognize when an image is fundamentally ambiguous and signal this "irreducible noise" to the system. In the other hand, **Epistemic uncertainty**— model ignorance in data-sparse regions of the data-generating process. In diatom

monitoring, epistemic uncertainty is typically triggered by rare species or new imaging conditions.

In this study, we adopt the framework of selective classification for diatom re-annotation by human expert. We equip a deep classifier with a rejection mechanism based on temperature-scaled softmax outputs [7]. By defining a confidence threshold via risk-coverage curves, we ensure that automated identifications stay below a target error rate (e.g., 1%), deferring ambiguous cases to human experts. We view this as a first step towards fully uncertainty-aware diatom monitoring. This lightweight approach provides a practical bridge between high-throughput imaging and reliable ecological assessment and ultimately can meaningfully reduce expert workload.

2 Related Works

Our work lies at the intersection of four research domains: Automated diatom classification, uncertainty quantification (UQ) in deep learning, selective classification, conformal prediction for human-in-the-loop systems, and active learning.

The automation of diatom identification has been pursued for several decades. Early systems such as ADIAC [1] demonstrated that automated identification even for challenging species was feasible, achieving accuracies between 75% and 90% using classical computer vision techniques (e.g., hand-crafted contour and texture features). Since then, deep learning has become the dominant paradigm. Recent studies have achieved impressive classification performance. Kloster *et al.* [3] showed that a VGG16-based CNN reached F1 scores above 97% for species-level classification on virtual slides, demonstrating that domain adaptation to diatoms is viable with limited annotated data. Building on such advances, Venkataramanan *et al.* [8] proposed a method that enforces a Gaussian structure in the latent space of a deep classifier; approximate Gaussian is obtained through a self-supervised procedure that dynamically splits non-Gaussian classes into multiple Gaussian sub-classes. With a well-regularised latent space, the Mahalanobis distance to the corresponding class centroid serves as a principled uncertainty score, enabling both out-of-distribution sample detection and improved probability calibration. Their approach requires minimal architectural modifications and was evaluated on diatoms, making it directly relevant to diatom identification. A complementary direction addresses the scarcity of annotated data: [9] introduced self-supervised pre-training for diatom classification, showing that fine-tuning with only 50 samples per class can recover full-supervised accuracy, reducing annotation requirements by roughly 96%. These works demonstrate that deep learning can match or exceed human-level accuracy on curated benchmarks. However, they focus primarily on maximising classification performance rather than on the reliability of predictions in operational settings. None of these systems provides a principled mechanism for deciding

when a prediction should be trusted versus deferred to an expert.

As discussed in the introduction, the machine learning community distinguishes two sources of uncertainty in neural network predictions: aleatoric (irreducible data noise) and epistemic (model ignorance that can be reduced with more data or better modelling) [4, 6]. This distinction is critical for bioindication—a model trained on typical diatom specimens may exhibit low epistemic uncertainty for common species but high epistemic uncertainty when faced with teratological forms, which are precisely the cases that warrant expert review.

Several techniques have been developed to estimate epistemic uncertainty. Monte Carlo (MC) Dropout [10] treats dropout at test time as a Bayesian approximation: by performing multiple stochastic forward passes and computing the variance of the predictions, one obtains a measure of model uncertainty without any architectural change. Deep Ensembles [11] take an alternative route: several models are trained independently with different random initializations, and their predictions are aggregated; the disagreement among ensemble members provides a reliable epistemic uncertainty estimate. Both approaches, as well as their variants, have been thoroughly surveyed by Gawlikowski *et al.* [12], who cover Bayesian neural networks, ensembles, test-time augmentation, and related calibration aspects. Having methods to quantify uncertainty does not directly prescribe how to use them in decision-making.

The framework of selective classification (learning with rejection) [5] formalises the trade-off between automated throughput and accuracy through risk–coverage curves that characterise model performance as a function of the fraction of examples on which it abstains. Recent extensions include hierarchical selective classification that leverages taxonomic structure to make predictions at coarser ranks when fine-grained confidence is insufficient [13], and extensive benchmarking studies that highlight the dependency of the best method on user objectives [14]. Beyond binary rejection, a more nuanced triage strategy involves providing a list of plausible candidates when a single species cannot be uniquely identified. Conformal prediction [15] is a distribution-free framework that turns any predictor into a set-valued predictor with finite-sample coverage guarantees. Given a pre-trained classifier, a calibration set, and a user-specified error rate α, conformal prediction constructs for each test point x a prediction set $\mathcal{C}(x) \subseteq \mathcal{Y}$ such that

$$\Pr\big[Y_{\text{test}} \in \mathcal{C}(X_{\text{test}})\big] \geq 1 - \alpha \tag{1}$$

holds in the marginal sense over the random draw of the calibration and test samples. The guarantee (1) therefore holds regardless of the underlying data distribution, provided the calibration and test data are exchangeable [15]. EPICSCORE [16] enhances this framework by injecting an explicit estimate of epistemic uncertainty (obtained via MC Dropout or Gaussian Processes) into the nonconformity score, so that prediction sets automatically expand in data-poor regions where the classifier is less confident. This is particularly well-suited to di-

atom monitoring, where rare species and teratological forms naturally correspond to data-sparse regions. Recent work explicitly connects conformal prediction to human-in-the-loop systems: Horton *et al.* [17] proposed Conformal Validation, a deferral policy that uses conformal prediction to identify the most uncertain observations and provides human experts with informative prediction sets.

Our problem also relates to active learning, where the goal is to query a human oracle for labels on the most informative samples in order to improve the model [18]. Active learning aims at building a better classifier over time, whereas our work aims at intelligent *deferral* of decisions in a fixed model, without necessarily retraining. While both involve human experts, the objective differs: active learning seeks better models; we seek better workflows that reduce the expert burden in routine monitoring.

To the best of our knowledge, the combination of selective classification with a calibrated confidence threshold has not yet been systematically explored for diatom identification in operational monitoring scenarios, which motivates the present study.

3 Methodology

3.1 Selective classification with a calibrated threshold

Routine diatom-based bioindication requires processing thousands of specimens, yet taxonomic expertise is scarce and time-consuming. This study therefore seeks to minimise expert involvement by automatically classifying those specimens for which the model is confident, while deferring uncertain cases to a human specialist. Formally, this is cast as selective classification (learning with rejection) [5]: a confidence threshold decides whether a specimen is handled automatically or passed to a taxonomist. Let $\mathcal{X}$ be the input space (e.g., diatom images) and $\mathcal{Y} = \{1, \dots, K\}$ the label set. A training sample $\mathbf{S}_m = \{(x_i, y_i)\}_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y}$ is drawn i.i.d. from an unknown distribution $\mathbf{P}$ over $\mathcal{X} \times \mathcal{Y}$. A classifier $f : \mathcal{X} \to \mathcal{Y}$ has real risk :

$$R(f) \triangleq \mathbf{E}_{(X,Y)\sim\mathbf{P}}\big[\ell(f(X), Y)\big], \tag{2}$$

where ℓ is a loss function (e.g., 0–1 loss). A selective classifier is a pair (f, g), with $g : \mathcal{X} \to \{0, 1\}$ a selection function that qualifies f:

$$(f, g)(x) \triangleq \begin{cases} f(x), & \text{if } g(x) = 1, \\ \text{“abstain”}, & \text{if } g(x) = 0. \end{cases} \tag{3}$$

Thus the model abstains iff $g(x) = 0$. Performance is assessed through coverage and selective risk:

$$\phi(f, g) \triangleq \mathbf{E}_{\mathbf{P}}[g(X)], \tag{4}$$

$$R(f, g) \triangleq \frac{\mathbf{E}_{\mathbf{P}}[\ell(f(X), Y)\, g(X)]}{\phi(f, g)}. \tag{5}$$

From an ecological standpoint, coverage is the fraction of samples processed without human intervention, while selective risk is the expected proportion of misidentified specimens among them—an error that can propagate through multimetric indices and compromise the reliability of an ecological assessment. Given a confidence parameter $\delta > 0$ and a target risk $r^* > 0$, the aim is to build g from $\mathbf{S}_m$ such that

$$\Pr_{\mathbf{S}_m}\big(R(f,g) > r^*\big) < \delta. \tag{6}$$

A standard construction [5] employs a confidence-rate function $\kappa_f : \mathcal{X} \to \mathbb{R}^+$ that ranks predictions: $\kappa_f(x_1) \geq \kappa_f(x_2)$ means the confidence in $f(x_1)$ is at least that in $f(x_2)$. As suggest in [5] κ_f. Here we take the maximum softmax probability as in [5]. The selection function is then parametrised by a threshold $\theta > 0$:

$$g_\theta(x) \triangleq \begin{cases} 1, & \text{if } \kappa_f(x) \geq \theta, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Its empirical selective risk on the sample $\mathbf{S}_m$ is

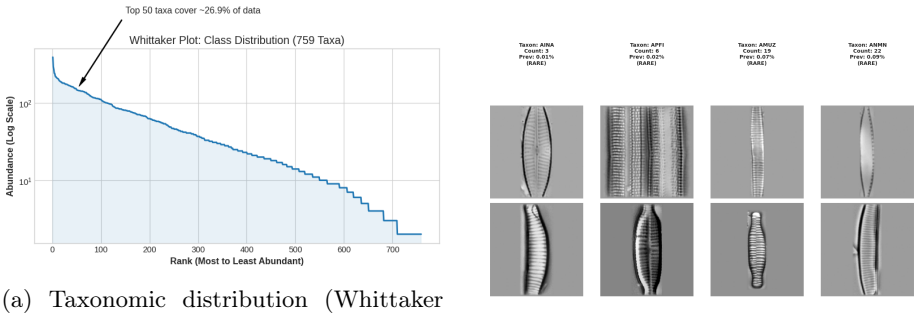
$$\hat{r}(f,g \mid \mathbf{S}_m) \triangleq \frac{\frac{1}{m}\sum_{i=1}^{m} \ell(f(x_i), y_i)\, g(x_i)}{\hat{\phi}(f,g \mid \mathbf{S}_m)}, \qquad \hat{\phi}(f,g \mid \mathbf{S}_m) \triangleq \frac{1}{m}\sum_{i=1}^{m} g(x_i). \tag{8}$$

The optimal θ can be found via the Selection with Guaranteed Risk (SGR) algorithm , which binary-searches for the threshold that respects the risk guarantee [5]. For the sake of simplicity , in this paper, we consider a very simple approach, compared to the full SGR machinery, to compute the threshold.

For the curve to yield a trustworthy threshold, the confidence scores must be well-calibrated: a score p should imply correctness with probability p. Modern neural networks, however, are notoriously overconfident [7], which in the context of biomonitoring would silently increase the fraction of misidentified specimens in the automated set. We therefore apply temperature scaling, a single-parameter extension of Platt scaling [7]. Given a logit vector $\mathbf{z}$, the calibrated confidence for the predicted class is

$$\hat{q} = \max_k \ \sigma_{\mathrm{SM}}\Big(\frac{\mathbf{z}}{T}\Big)^{(k)}, \tag{9}$$

where σ_{SM} is the softmax. The scalar $T > 0$ is learnt on the calibration set by minimising the negative log-likelihood loss. Because rescaling preserves the ranking of the logits, the model's accuracy is unchanged; only the confidence values are adjusted, enabling a more reliable rejection threshold. A well-calibrated model thus provides the ecologist with a trustworthy estimate of when it is safe to rely on automated identifications. The resulting decision rule directly addresses the taxonomic bottleneck in routine bio-monitoring by reserving expert effort for the most challenging specimens while controlling the error rate on the rest.



(a) Taxonomic distribution (Whittaker plot)

(b) Morphological variability examples

Figure 1: **Characterisation of the DIATLAS dataset: taxonomic imbalance and morphological diversity. (a)** The plot illustrates the distribution of 35,447 images across 759 taxa on a logarithmic scale. The initial steep decline represents a small group of dominant species, while the extensive "long tail" contains hundreds of rare taxa with sparse representation (fewer than 10 specimens per class). Notably, the top 50 most abundant taxa account for only 26.9% of the total specimens, indicating that the majority of the dataset consists of low-prevalence classes. **(b)** Representative specimens illustrates a rare taxon ($Prevalence < 0.1\%$) where limited specimen counts—ranging from 3 for AINA to 22 for ANMN—pose a significant challenge for supervised learning. Furthermore, the vertical samples demonstrate high morphological variability, capturing distinct valve and girdle orientations within the same class (e.g., AINA, AMUZ) and variations in frustule preservation and striae visibility (APFI, ANMN). Together, the extreme data scarcity for rare taxa (e.g., CAGL, CIRI) and high visual ambiguity justify the implementation of a selective classification framework to ensure reliable automated bioindication.

3.2 Dataset: DIATLAS

This study uses DIATLAS [19], an open dataset of optical microscopy images of diatoms extracted from French regional atlases and documenting freshwater specimens. Images were acquired using optic or photonic microscopy. Each image is associated with a taxonomic classification following the OMNIDIA nomenclature [20]. The raw corpus comprises 38,719 images. To ensure statistical viability for deep learning, we retained only classes with at least two specimens, yielding a curated set of 35,447 images across 759 taxa. The resulting class distribution exhibits a pronounced **long-tail** shape—an inherent feature of environmental sampling where a few dominant taxa coexist with many rare species—posing a significant challenge for classification (see Figure 1).

For experimentation, the dataset was split into training (25,522 images), calibration (2,835), and test (7,090) subsets via stratified sampling (72/8/20). The calibration set is used as validation set to trigger checkpointing and for temperature scaling and threshold selection. Dataset splits were stratified to preserve class proportions for taxa with sufficient representation; however, for extremely rare species (e.g., those with only one or two instances), this approach naturally resulted in some taxa being present in the test set while remaining absent from the training set or vice-versa.

3.3 Experimental Implementation

Our framework is built on `torch-uncertainty` (TU) [21]. To integrate DIATLAS, we implemented a custom `DiatomDataModule` with a preprocessing pipeline designed to mitigate covariate shift. Images are rescaled to a target physical resolution, vertically aligned via principal axis rotation, and centre-cropped or padded using seamless cloning with a median-colour background to prevent edge artefacts. Grayscale conversion is applied to isolate structural frustule morphology from acquisition-related colour variance.

We trained a ResNet-18 classifier for 60 epochs using Adam (lr $= 10^{-2}$) with a batch size of 32 and 16-bit mixed precision. Training utilised gradient accumulation (factor of 4) and norm-clipping (threshold 1.0).

Post-hoc calibration was performed via temperature scaling [7] as implemented in `torch-uncertainty`. While the architecture is standard, the selective classification mechanism was explicitly implemented to provide direct control over the abstention threshold—enabling precise evaluation of coverage–risk trade-offs and class-wise rejection. Performance is quantified via classification accuracy and Expected Calibration Error (ECE) using the maximum softmax probability (MSP).

All experiments were executed on an NVIDIA RTX 2000 Ada GPU.

4 Results

4.1 Calibration and Selective Triage

The ResNet-18 baseline, implemented via the `torch-uncertainty` framework [21], demonstrates strong classification performance on the test set. However, as illustrated in the reliability diagrams (Figure 2a), the uncalibrated model was initially significantly overconfident (ECE = 0.117). Post-hoc temperature scaling yielded an optimal temperature of $T = 2.12$, aligning the model's confidence closely with the identity diagonal and reducing the calibration error to 0.055. This ensures that the softened softmax outputs are a statistically valid proxy for empirical likelihood, a prerequisite for robust triage.

Beyond point-prediction accuracy, we assess the framework's utility under a rejection policy using metrics that account for taxonomic imbalance (Table 1). The Global Macro-F1 score of 90.35% highlights the challenge of rare species; however, by employing selective classification, we significantly enhance the reliability of the automated subset.

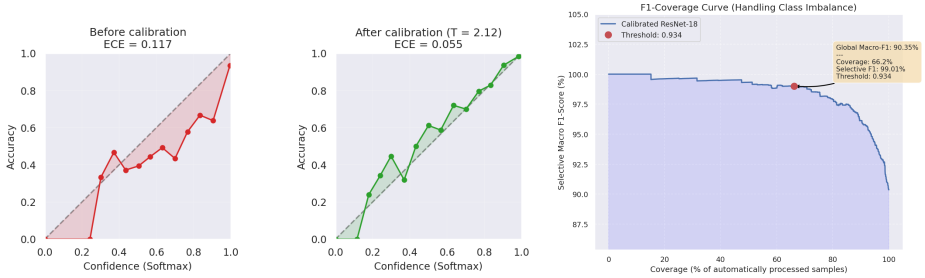
Table 1: Performance metrics for the calibrated ResNet-18 on the test set.

Metric	**Value**
Global Accuracy	95%
Global Macro-F1	90.35%
Expected Calibration Error (ECE)	0.055
Area Under the Risk-Coverage Curve (AURC)	0.0061
Selective Macro-F1 (at target operating point)	99.01%
Coverage (at target operating point)	66.2%
Optimal Temperature (T)	2.12

As shown in the F1-coverage analysis (Figure 2b), we applied an abstention threshold of $\tau = 0.934$ to simulate a high-stakes monitoring deployment. Under this policy, the system automatically processes 66.2% of specimens, achieving a Selective Macro-F1 of 99.01%. The remaining 33.8% of ambiguous or rare cases are deferred to a human specialist. This behaviour fulfils the primary objective of the proposed pipeline: providing high-throughput automation for routine specimens while ensuring that cases prone to misclassification are systematically referred for expert review.

4.2 Qualitative Analysis of Uncertainty and Triage Utility

To assess the practical utility of the selective classification pipeline for taxonomic workflows, we examined the specimens most frequently deferred by the model at the $\tau = 0.934$ operating threshold. As illustrated in Figure 3, the system exhibits a high propensity for rejecting "rare" species—taxa with minimal representation in the training set (e.g., $< 0.5\%$).



(a) Reliability diagrams (Macro-averaged) (b) F1-coverage curve

Figure 2: **Calibration and Selective Triage Performance. (a)** Temperature scaling with $T = 2.12$ reduces the ECE to 0.055, correcting overconfidence. **(b)** The F1-coverage trade-off shows that at a threshold of $\tau = 0.934$, the system maintains a 99.01% Selective Macro-F1 by deferring the most uncertain 33.8% of samples.

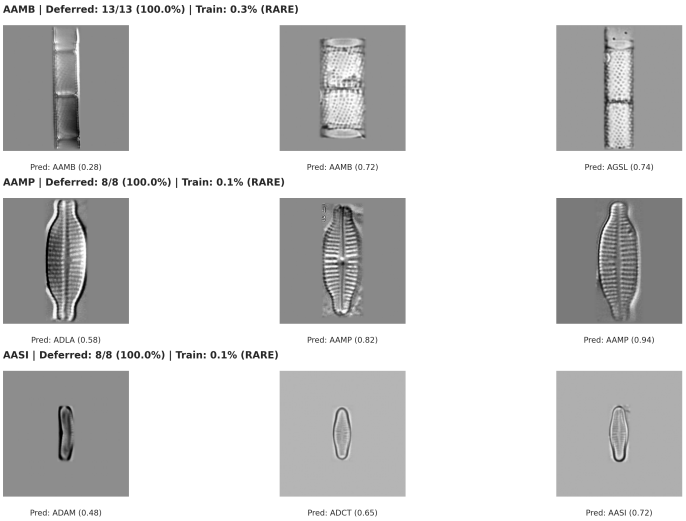


Figure 3: **Qualitative analysis of deferred specimens across rare taxa.** The figure showcases the lowest-confidence specimens for species frequently targeted for human intervention. Taxa such as AAMB, AAMP, and AASI—all representing less than 0.3% of the training data—exhibit 100% deferral rates. The system successfully flags out-of-distribution morphologies (e.g., girdle views or small valves), ensuring these high-stakes specimens are prioritized for expert review.

Specifically, species such as *Aulacoseira ambigua* (AAMB), *Achnanthidium minutissimum* (AAMP), and *Achnanthidium sieminskae* (AASI) were deferred in 100% of test instances. This behavior underscores the model's capacity for epistemic uncertainty detection; rather than forcing a high-confidence prediction for unfamiliar or rare morphologies, the system correctly identifies its own knowledge limits.

From a taxonomist's perspective, this selective mechanism functions as a safeguard for ecological integrity. By isolating ambiguous specimens for human intervention, the pipeline prevents automated misclassification from skewing community indices. This confirms that the calibrated ResNet-18 functions not merely as a rigid classifier, but as an intelligent filter that bridges the gap between high-throughput automated processing and expert domain knowledge.

5 Discussion

The transition from a standard risk-based coverage policy to one governed by the Selective Macro-F1 score represents a fundamental shift toward prioritizing taxonomic integrity over simple data throughput. While global accuracy often masks a model's failure to identify rare bioindicators, the Macro-F1 metric ensures that each taxon, regardless of its frequency in the dataset, is afforded equal weight. Consequently, our calibrated ResNet-18 achieves a Selective Macro-F1 of **99.01%**. This level of reliability directly satisfies the rigorous, high-stakes requirements of ecological monitoring.

A primary observation in our qualitative analysis is that the system's predictive uncertainty aligns closely with genuine taxonomic difficulty. As illustrated by the most frequently deferred specimens (Figure 3), morphological ambiguity—particularly visible in the final row—represents a significant hurdle. Even for seasoned experts, small specimens or those presented in girdle view offer fewer diagnostic features, making confident identification challenging. It is important to clarify that the "rarity" observed here pertains specifically to representation within the current, evolving DIATLAS dataset rather than necessarily reflecting natural abundance within local biodiversity. However, this dataset-driven sparsity directly influences the model's epistemic uncertainty; with fewer examples from which to learn, the model correctly adopts a more conservative, risk-averse posture.

The restrictive nature of optimizing for the Macro-F1 score is inherently more conservative than utilizing a standard accuracy threshold. Because a single error in a rare class catastrophically degrades the macro-average, the resulting rejection threshold ($\tau = 0.934$) is naturally pushed higher. In the context of bioindication, this conservative behavior is highly desirable: the system prefers to defer an ambiguous specimen to a specialist rather than risk a misclassification that could skew sensitive ecological indices. Furthermore, analyzing the subtleties among misclassified cases reveals that many rejected specimens exhibit high visual similarity to overlapping taxa. This suggests that the model successfully identifies regions of the feature space where class boundaries are blurred—a

phenomenon we intend to explore further by presenting side-by-side comparisons of misclassified pairs to explicitly showcase these morphological overlaps.

Ultimately, high-quality biodiversity assessment requires moving beyond global risk optimization toward class-conditional logic. While our temperature scaling has effectively reduced the Expected Calibration Error (ECE) to 0.055—providing a reliable confidence signal for triage—the current deterministic approach still conflates data noise (aleatoric uncertainty) with model ignorance (epistemic uncertainty). Future iterations of this framework will integrate explicit second-order uncertainty estimators to better disentangle these sources of error. By adapting rejection logic to both the data density and the morphological difficulty of specific taxa, we can maximize throughput without compromising the integrity of the ecological signal. Over time, incorporating active learning pipelines will further reduce expert workload as the system matures.

6 Conclusion

This work establishes a baseline for uncertainty-aware diatom identification using a ResNet-18 architecture trained on a taxonomically diverse dataset of 759 species, characterised by extreme class imbalance. While the baseline model achieved a high global accuracy, its raw softmax outputs were significantly overconfident (ECE = 0.117). Post-hoc temperature scaling effectively mitigated this miscalibration, aligning predictive confidence with empirical likelihood and reducing the calibration error to 0.055.

The practical utility of these calibrated confidences was demonstrated through a selective classification framework governed by the Macro-F1 score. By prioritising taxonomic integrity over simple point-accuracy, we established a high-reliability regime: at an operating threshold of $\tau = 0.934$, the system achieves a Selective Macro-F1 of 99.01%. Under this policy, the framework automatically processes 66.2% of specimens while systematically deferring the most ambiguous and rare cases—often those representing less than 0.5% of the training data—for expert review. This balance underscores the feasibility of using automated triage to reduce specialist workload by two-thirds without compromising the ecological signal required for bioindication.

However, several challenges remain for operational deployment. Standard softened softmax outputs, while well-calibrated, do not fully disentangle epistemic model ignorance from aleatoric data noise. Qualitative analysis further revealed that morphological similarities and small specimen sizes continue to pose challenges for both the model and human experts. Future research will focus on benchmarking second-order uncertainty estimators, such as Monte-Carlo Dropout and Conformal Prediction, to provide more nuanced rejection criteria with formal coverage guarantees. By integrating these methods into continuous, human-in-the-loop learning cycles, we move toward a resilient monitoring

framework capable of adapting to the shifting taxonomic distributions inherent in aquatic ecosystems.

Acknowledgements This project work was funded the ANR BIOINDIC-IA ANR-24-CE04-0345.

References

[1] Hans Du Buf. *Automatic Diatom Identification*. Series in Machine Perception and Artificial Intelligence. World Scientific Publishing, 2002.

[2] David G. Mann. 'The species concept in diatoms'. In: *Phycologia* 38.6 (1999), pp. 437–495.

[3] Michael Kloster et al. 'Deep learning-based diatom taxonomy on virtual slides'. In: *Scientific Reports* 10.1 (2020), p. 14416.

[4] Alex Kendall and Yarin Gal. 'What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?' In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017, pp. 5574–5584.

[5] Yonatan Geifman and Ran El-Yaniv. 'Selective Classification for Deep Neural Networks'. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017, pp. 4878–4887.

[6] Eyke Hüllermeier and Willem Waegeman. 'Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods'. In: *Machine Learning* 110.3 (2021), pp. 457–506.

[7] Chuan Guo et al. 'On Calibration of Modern Neural Networks'. In: *Proc. ICML*. 2017, pp. 1321–1330.

[8] A. Venkataramanan et al. 'Gaussian Latent Representations for Uncertainty Estimation using Mahalanobis Distance in Deep Classifiers'. In: *Proc. ICCVW*. 2023, pp. 4490–4499.

[9] Mingkun Tan et al. 'Scaling down annotation needs: The capacity of self-supervised learning on diatom classification'. In: *iScience* 28 (2025).

[10] Yarin Gal and Zoubin Ghahramani. 'Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning'. In: *Proc. ICML*. 2016, pp. 1050–1059.

[11] Balaji Lakshminarayanan, Alexander Pritzel and Charles Blundell. 'Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles'. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[12] J. Gawlikowski et al. 'A survey of uncertainty in deep neural networks'. In: *Artificial Intelligence Review* 56.1 (2023), pp. 1513–1589.

[13] Shani Goren, Ido Galil and Ran El-Yaniv. 'Hierarchical Selective Classification'. In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*. Poster presentation. 2024.

[14] Andrea Pugnana et al. 'Deep Neural Network Benchmarks for Selective Classification'. In: *Journal of Data-centric Machine Learning Research* (2024). Submitted 12/23; Revised 7/24; Published 8/24.

[15] Glenn Shafer and Vladimir Vovk. 'A Tutorial on Conformal Prediction'. In: *Journal of Machine Learning Research* (2008).

[16] Luben M. C. Cabezas et al. 'Epistemic Uncertainty in Conformal Scores: A Unified Approach'. In: *Proceedings of the [X] Conference on Uncertainty in Artificial Intelligence*. Vol. [volume number]. details pending official publication. PMLR, 2025.

[17] Paul Horton, Alexandru Florea and Brandon Stringfield. 'Conformal validation: A deferral policy using uncertainty quantification with a human-in-the-loop for model validation'. In: *Machine Learning with Applications* 22 (2025), p. 100733.

[18] Dongyuan Li et al. 'A Survey on Deep Active Learning: Recent Advances and New Frontiers'. In: *IEEE Transactions on Neural Networks and Learning Systems* 36.4 (2025), pp. 5879–5899.

[19] Corentin Galinier et al. *DIATLAS: The French freshwater DIatom ATLAS image dataset*. `https://doi.org/10.5281/zenodo.16260887`. 2024. DOI: `10.5281/zenodo.16260887`.

[20] Christophe Lecointe. *OMNIDIA software for diatom indices and database management*. `https://omnidia.fr/en/`. Version 6.0. 2024.

[21] Adrien Lafage et al. 'Torch-Uncertainty: A Deep Learning Framework for Uncertainty Quantification'. In: *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*. To appear. 2025.