

A Large-Scale Convolutional Neural Network for Fine-Grained Conus Shell Identification (Mollusca: Gastropoda: Conidae)

Document Created: 17 March 2025

Philippe Kerremans  0009-0002-9238-4007

Abstract

Cone snails (genus *Conus*) comprise one of the most diverse groups of marine gastropods, with over 850 recognized species. Despite significant variability in coloration and subtle morphological traits, *Conus* shells often appear similar in overall shape, making their accurate identification a fine-grained image classification challenge. In this study, we present a convolutional neural network (CNN) model trained on a large dataset of 130,373 images spanning 518 *Conus* species. Images were gathered from multiple sources and extensively curated to address issues of inconsistent labeling and background noise. Preprocessing steps included segmentation of individual shells, uniform background replacement, and resizing, thereby standardizing visual inputs for the model. Performance metrics (recall, precision and F1 score) show strong results, with an overall accuracy of 97% and macro-averaged precision and recall around 96–97%. Confidence intervals further support the reliability of these findings, even for classes with fewer validation images. We compare our approach with two previous *Conus* models: one developed at Naturalis Biodiversity Center and another by Qasmi et al., each employing different image-processing and classification strategies. Our results underscore that large-scale species coverage — when coupled with thorough preprocessing — does not necessarily diminish model accuracy. Furthermore, the model's solid performance amid considerable species-level imbalances highlights the viability of CNN-based systems for difficult, fine-grained biodiversity classification tasks. This comprehensive dataset and refined workflow pave the way for future integrative studies that combine museum collections, citizen science, and advanced AI methodologies to enhance *Conus* taxonomy and broader molluscan research.

Introduction

The genus *Conus* (commonly known as cone snails) represents one of the most diverse groups of marine gastropods, with a current total of 853 recognized extant species according to MolluscaBase and WoRMS ([WoRMS](#), [MolluscaBase](#)). The *Conus* genus is one of the largest in the Mollusca phylum. Until a decade ago, its species were split across 89 genera [1], but are now largely consolidated within *Conus* genus ([WoRMS](#), [MolluscaBase](#)). Despite this tremendous species richness, cone snails exhibit a notable uniformity in general shell shape and pattern, which differ only in subtle color variations, banding, or small morphological traits. As a result, discriminating among the many *Conus* species becomes a daunting fine-grained image classification challenge, demanding models that can isolate and interpret minute differences in shell markings.

The *Conus* CNN model is designed to learn features that capture these nuances, allowing it to separate species based on visual cues in shell images. However, the complexity of this task escalates with the number of species (i.e., classes) included: each new class introduces additional inter-class similarities and increases the potential for misclassification [2]. Moreover, adding classes means collecting and processing more images, thus necessitating greater computational resources and longer training times. In total, the dataset comprises 130 373 images across all included *Conus* species. Consequently, scaling a *Conus* CNN model to encompass all 853 species underscores both the difficulty of fine-grained recognition and the importance of robust, efficient training strategies.

The shells of the *Conus* genus have a characteristic morphology, the most important features for species identification are pattern and colouration. Other important features are the form of the spire and the width versus the length.

Table I. The *Conus* genus and the image dataset.

Parameter	Value	Comments
Species in the <i>Conus</i> genus	853	MolluscaBase/WoRMS accessed Jan 2025
Species with images	529	Status Jan. 2025
Species with 25 images or more	518	Status Jan. 2025. 11 species have less than 25 images and were excluded.
Total number of images in the dataset	130 373	Status Jan. 2025
Species with the most images	<i>Conus textile</i> , 2923 images	

Methods

Data Collection

The dataset for the *Conus* CNN model comprises 130,373 shell images representing 518 *Conus* species (see table I). From the 529 species for which images were collected, all species with less than 25 images were removed (see [Minimum number of images needed for each species](#)). A total of 518 species were used. These

images were aggregated from multiple sources, including online databases and museum or field photograph repositories ([Identifying Shells using Convolutional Neural Networks: Data Collection and Model Selection](#)). Additionally, broad community-driven efforts (e.g., citizen science platforms) have contributed to the pool of images – modern biodiversity projects have amassed massive image collections of specimens. The dataset comprises images from the following sources: museum collections (4.3%, 5640 images), online citizen science platforms (18.1%, 23661 images) and commercial shell websites (77.6%, 101071 images). The original Conus image dataset is 12% larger, many images were removed because the image quality is bad, or other objects are visible in the picture (hands, other animals, labels, etc.). Also, images that contains more than one shell and could not be split in images with only 1 shell were eliminated.

Hardware

An HP Omen 30L GT13 was used for training the model. It contains a Intel(R) Core(TM) i9-10850K CPU @ 3.60GHz processor, with 64GB RAM, Nvidia GeForce RTX 3080 10GB.

Image preparation

All images were pre-processed. When an image contained multiple shells, we applied thresholding to binarize the background and then used contour detection to locate each shell's outline, cropping out each detected contour as an individual image. The background was replaced with a uniform black background. A square image was made by padding with a black background. All shells were resized (400 x 400 px). A final visual selection was made before producing the final image dataset. Overall, 10-20% of the images were removed for various reasons (when other objects were visible in the picture such as hands, habitat, text, etc.).

Annotation and Labeling Challenges

Preparing a labeled dataset of 518 species presents significant annotation hurdles. One major challenge is taxonomic ambiguity. Cone snail taxonomy has been in flux – historically cone snails were split in 89 genera [1], but the last decade most species were merged in the genus *Conus* (see MolluscaBase/WoRMS). As a result, the same species might be known by multiple names, or what were once separate species might have been merged. Such inconsistencies across image sources can lead to mislabeling (e.g., an image labeled with an outdated name). Careful curation was needed to reconcile synonyms and ensure each image is tagged with a valid, accepted species name.

Another challenge is the morphological similarity among species: many *Conus* shells differ only in subtle pattern or color variations. Non-experts may confuse one species for another, especially if shell patterns overlap or the specimen is an atypical individual. This means some portion of the images could be erroneously labeled, introducing noise into the training data.

Metrics and confidence intervals

Metrics were calculated using the *sklearn.metrics* module, functions `accuracy_score`, `precision_score`, `recall_score`, `f1_score`, `confusion_matrix`, `classification_report` were used. To calculate the confidence intervals (95%). A cap of 200 images per species was employed when computing performance metrics and confidence intervals, since sampling beyond this limit did not yield any improvement in the statistical estimates. Bootstrapping was used [4]. Bootstrapping, being a non-parametric method, does not rely on the normality assumption. A 1000 runs were performed for each species to calculate the intervals.

Results

The *Conus* image dataset

From these 529 species for which images are available, 518 have more than 25 images which were used to

construct a model (see [Minimum number of images needed for each species](#)). These 518 species and the number of images used are listed in Table II ([Supplementary Material](#)). The distribution of images among species is shown in the next figure.

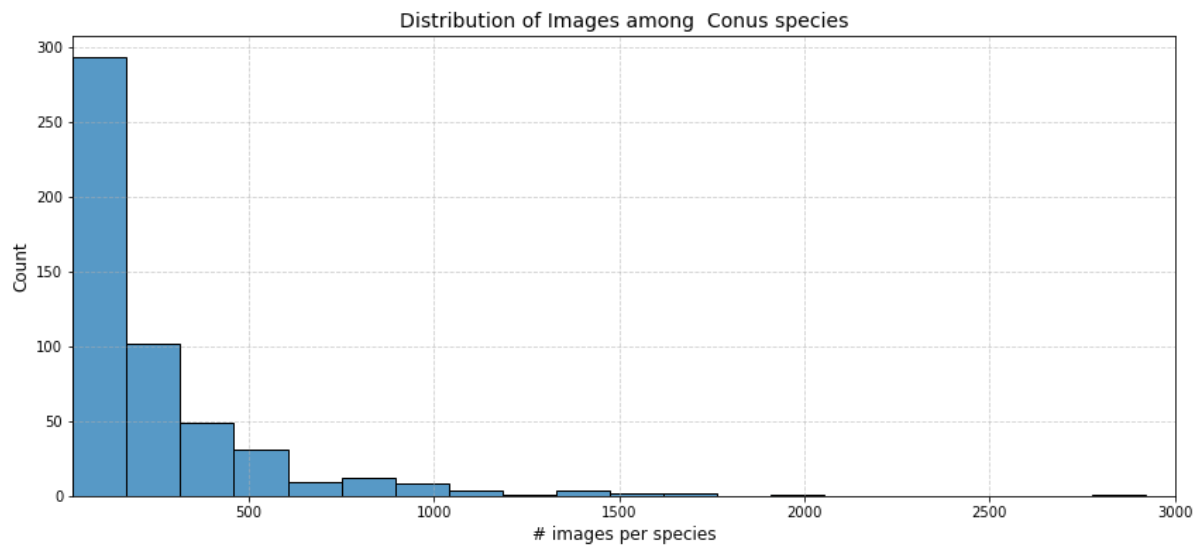


Figure 1: Distribution of images among the 518 *Conus* species available in the dataset

Many species have a low number of images. The first bin (25-170 images) has 293 species which is almost half of all species. There are a few species with a large number of images; *Conus textile* has 2923 images, *Conus furvus* 1976 images and *Conus mercator* 1703 images. There is a considerable imbalance in the dataset. However, preliminary tests with oversampling the minority classes with augmented images (rotate, flip, change brightness and contrast) did not improve the results significantly (data not shown).

Image Preparation

The dataset of 130,373 images was split into 80% training and 20% testing data. This means that the species with the least images (total of 25 images) has 20 images in the training dataset. There may be more than 20 images if the original image shows several shells (or views of the same shell) because separate images were made for each shell in the original images.

Model creation

The *Conus* model was created as described in [Identifying Shells using Convolutional Neural Networks: Data Collection and Model Selection](#). The hyperparameters used are provided in table III.

Table III. Hyperparameters

Hyperparameter	Value	Comments
Batch Size	64	The batch size determines the number of samples processed in each iteration.
Epochs	100	The number of epochs determines how many times the entire training dataset is passed through the model. Because early-stopping is used, less than 100 epochs were needed. Fine-tuning usually requires fewer epochs compared to training from scratch.

Hyperparameter	Value	Comments
----------------	-------	----------

Optimizer	Adam	The optimizer determines the algorithm used to update model weights during training.
Learning rate	0.0002	
Fine-tuning	top 3 layers unfrozen	
Top layer dropout	0.25	
Regularization	0.0001	

Some limited parameter tuning was performed, however the initial hyperparameters gave already good results (data not shown). The learning rate was decreased from initial 0.0005 to 0.0002, and the top layer dropout increased from 0.2 to 0.25 (see also [Identifying Shells using Convolutional Neural Networks: Data Collection and Model Selection](#)). This limited hyperparameter tuning was done iteratively. The final training was run for 73 epochs using early stopping. Inference was performed on the validation set and analyzed using *sklearn.metrics.classification_report*. The summary statistics are provided in table IV.

Table IV. Summary statistics using *sklearn.metrics*

Statistic	Value
Categorical Accuracy	0.97
Macro Average Recall	0.96
Macro Average Precision	0.97
Macro Average F1	0.96
Weighted Average Recall	0.97
Weighted Average Precision	0.97
Weighted Average F1	0.97

Because a small number of validation images were used to calculate the metrics for a significant proportion of

the species (180 species have less than 20 validation images), the confidence intervals were calculated for the F1 score using bootstrapping. The metrics and confidence intervals are given in table V ([Supplementary Material](#)). Only 6 species had low metrics (and large confidence intervals): *Conus adenensis* 0.540 (CI: 0.193-0.645), *Conus auricomus* 0.652 (CI: 0.476-0.792), *Conus compressus* 0.696 (CI: 0.428-0.880), *Conus conspersus* 0.640 (CI: 0.363-0.827), *Conus gilvus* 0.666 (CI: 0.399-0.864), *Conus turritinus* 0.588 (CI: 0.200-0.823), and *Conus vezzaroi* 0.741 (CI: 0.500-0.903). The figure below shows the distribution of the F1 score for the validation set.

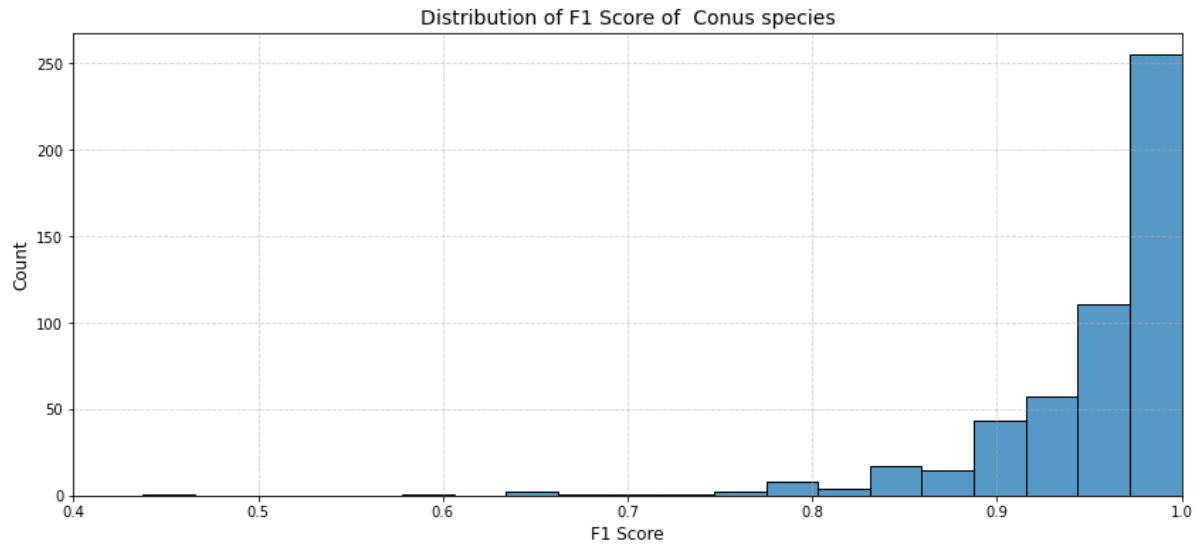


Figure 2: Distribution of the F1 score for the validation set

The figure below shows a scatterplot where the number of images in the validation set is plotted against the F1 score. Classes with a low F1 score (points on the left) consistently have a low number of validation images. There are no classes shown with many validation images that result in a poor F1 score. As the F1 score increases towards the right, the number of validation images per class varies widely. The dense, almost vertical cluster on the far right indicates that many classes achieved a high F1 score, regardless of whether they had a few or many validation images. A low number of validation images (and training images) is strongly correlated with poor model performance. While having more images doesn't guarantee a perfect score, a lack of images appears to be a primary factor for the classes where the model fails.

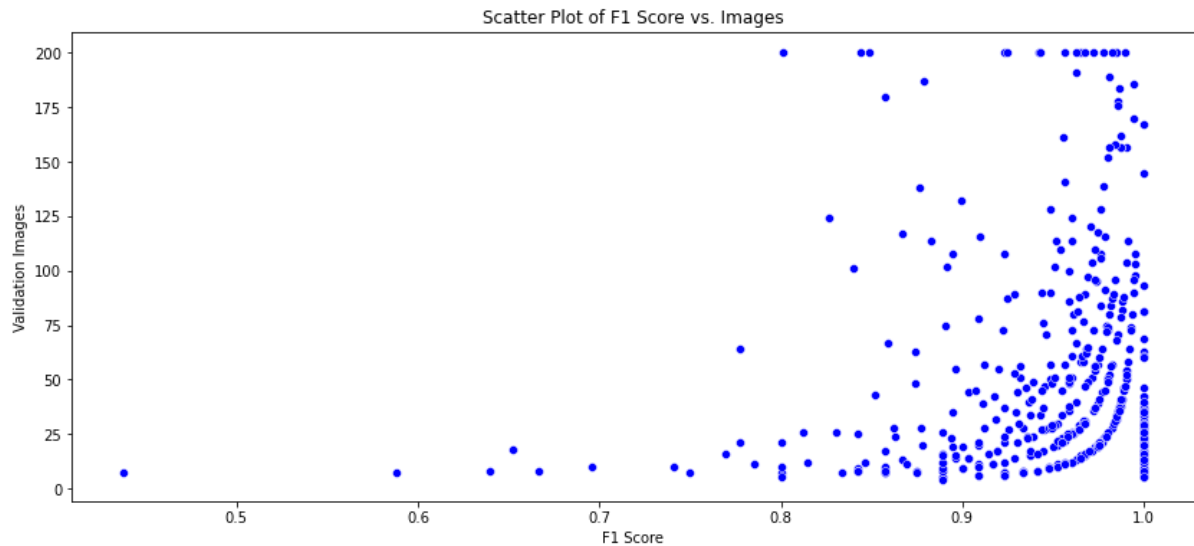


Figure 3: Scatterplot of the F1 score versus number of images

When comparing the F1 score based on the validation set with the F1 Score where training set images were added (to a max. of 200 images), we see for the majority of the species no large difference (see figure below). Only for a few species, those that have a low "Validation" F1 Score, we see a large difference with the F1 Score that includes training images (the species *Conus compressus*, *Conus gloriamaris*, *Conus turritinus* and *Conus vezzaroii*). This is expected, since testing on training data provides an easy performance boost that has the largest impact on the lowest initial scores.

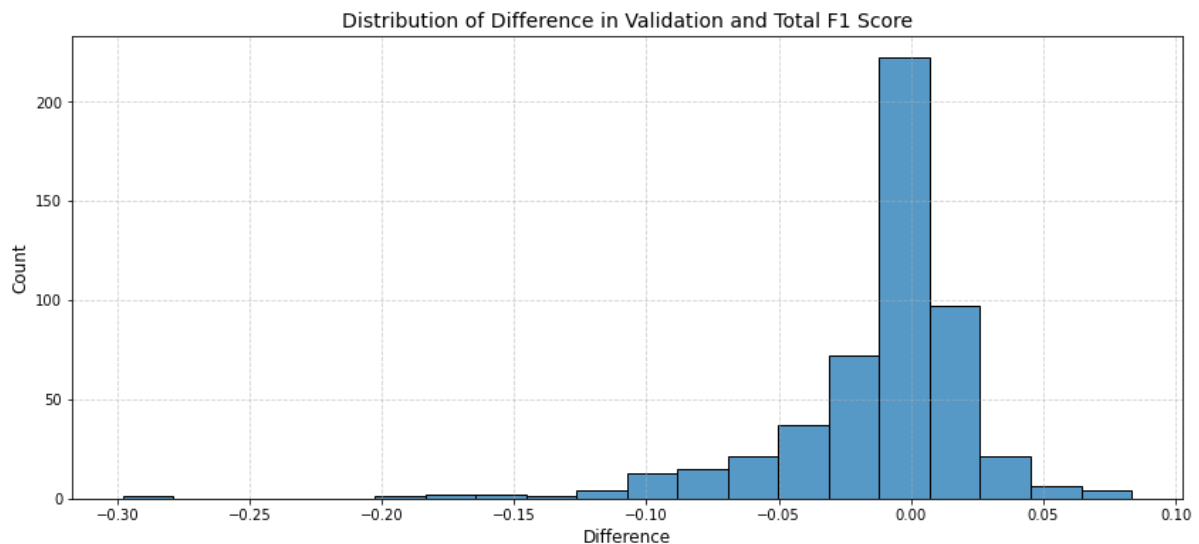


Figure 4: F1 score of the validation set vs. val. + train. set.







Confusion Matrix

The confusion matrix provides a class-by-class breakdown of the CNN model's performance. Overall, the strong main diagonal indicates a high degree of accuracy across most species. However, the off-diagonal values reveal that significant misclassification was confined to a few specific pairs. The most frequent confusion occurred between *Conus ardisiaceus* and *Conus aemulus*, likely due to their remarkable similarity in

shell morphology and color patterns. Other notable problem pairs included *Conus asiaticus* with *Conus alabaster*, and *Conus angasi* with *Conus adenensis*. Beyond these specific cases, misclassifications for other species were sparse and not concentrated in any particular group, suggesting the model's errors are systematic to visually ambiguous species rather than random.

Visual inspection shows there are similarities between these species:

Table VI. Images of the most confused *Conus* species

Conus ardisiacus	Conus aemulus	Conus asiaticus	Conus alabaster	Conus adenensis	Conus angasi
					

Significant misclassification was confined to these species. In contrast, misclassified individuals of other species were not concentrated in any particular group.

Additional tests

Additional images were collected after model creation. This anecdotal test for several species confirms the performance of the model (Table VII).

Table VII. Additional tests of the performance of the *Conus* model.

Species	Recall (Conf. interval)	New images	Correct prediction	Wrong prediction	Recall (for test images)	Avg. (softmax) probability
<i>Conus ammiralis</i>	0.98 (0.96-0.995)	85	85	0	1.00	0.99
<i>Conus striatus</i>	0.965 (0.935-0.986)	43	41	2	0.95	0.93
<i>Conus mustelinus</i>	0.994 (0.979-1.0)	53	52	1	0.94	0.98
<i>Conus merletti</i>	0.973 (0.903-1.0)	51	46	5	0.89	0.90
<i>Conus amadis</i>	0.975 (0.941-1.0)	24	23	1	0.90	0.96

Table VII presents the recall and associated confidence intervals calculated from the validation sets for each tested species, alongside the classification outcomes for newly tested images. Two images of *Conus striatus* were misclassified, although with relatively low prediction probabilities (0.58 and 0.55). More notably, four images of *Conus merletti* were consistently misclassified. These images depict a single specimen photographed from different angles and under slightly varying contrast conditions. All were incorrectly identified as *Conus moluccensis*, a species visually similar to *Conus merletti*. Potential explanations for this misclassification include the possibility that the specimen is genuinely *Conus moluccensis* rather than *Conus merletti*, or that there are inaccuracies or overlaps within the training dataset—for example, *Conus moluccensis* specimens mistakenly included within the *Conus merletti* class. Another interpretation is that these two species might actually represent varieties of the same species.

Discussion

Developing a CNN-based classification model for the *Conus* genus involves a variety of unique challenges stemming from both the taxonomic complexity of this group and the subtle characteristics that distinguish its species. One of the most significant hurdles is that *Conus* species are primarily differentiated by their color patterns, which can be very similar across species. Even minor variations in lighting, shell wear, and image quality can obscure these differences, making it difficult for a CNN to accurately distinguish one species from another.

Compounding this challenge is the sheer size of the *Conus* genus—currently recognized to include 853 species—making it one of the largest genera within the Mollusca phylum ([WoRMS](#), [MolluscaBase](#)). Handling such a large number of classes naturally demands more computational resources, including significant memory capacity. As a result, training a *Conus* CNN model can be both computationally expensive and time-consuming, often taking over four hours on our infrastructure.

Another important factor is that *Conus* shells are highly sought after by collectors, leading to a relatively large pool of publicly available images. On the one hand, this abundance of data provides a rich resource for model training. On the other hand, it requires careful data management to account for variations in image quality, resolution, and lighting conditions, as well as potential imbalances in how frequently each species is photographed.

Altogether, these considerations — high species diversity, subtle morphological and color distinctions, and a sizable but inconsistently curated dataset — highlight the complexity of creating a robust *Conus* CNN model. Building such a model requires thorough data curation, meticulous preprocessing, and a well-designed computational infrastructure capable of supporting prolonged training periods. Our CNN model achieved an accuracy of 97%, utilizing 130 373 cone snail shell images.

Before model training, extensive pre-processing is performed. All images were analyzed to detect the number of shells in the image and a separate image was made for each shell. If possible, the background was changed to black. A fixed input size of 400x400 pixels was used. Images were made square if needed. A final manual step was included to select images that clearly show shell features that help in species identification.

Metrics calculated for each species (recall, precision and F1 score) shows that a large proportion of all species can be identified reliably. Calculation of the confidence intervals support this conclusion.

Two other models of the *Conus* genus were created before, both with good performance [5, 6]. The team at Naturalis, Leiden created several models for several topics, including also a *Conus* model. The model was trained on 797 *Conus* species, 15 877 images. Performance of the model was not communicated, but limited tests (data not shown) show good performance. This suggests that even with a moderately sized dataset (15

877 images), accurate species-level classification can still be achieved if the training images are curated carefully and taxonomic labels are standardized (e.g., via WoRMS). N. Qasmi et al. have also made a Conus AI model, based on 47 600 images on 119 Conus species. Their model has 95% accuracy using a combination of Random Forest (RF), XGBoost (XGB) methods and feature extraction using a CNN.

The recently reported model by Qasmi et al. [6] employed a combined approach of deep learning (VGG16 for feature extraction) and ensemble supervised learning (Random Forest and XGBoost). Notably, their workflow involved explicit feature-engineering steps—such as color moments, local binary patterns, and Haralick textures — before applying ensemble classifiers. This pipeline effectively demonstrated how hybrid methods (deep feature extraction plus machine-learning classifiers) can yield strong performance in a challenging fine-grained domain.

Compared to these two models, the CNN model described in this study substantially broadens the species coverage to 518 species, incorporating 130,373 images — a dataset volume almost three times as large as that of Qasmi et al. and well above the Naturalis pilot. Nonetheless, it attains a similarly strong performance: an accuracy of 97% with a macro-average F1 score around 0.96. This outcome underscores two important points:

1. **Broader Taxonomic Scope vs. High Accuracy:** Expanding classification from 119 to 518 Conus species introduces further inter-class similarity, increasing the risk of misclassification. Despite this, the final accuracy remains comparable to previous efforts, implying that large-scale coverage does not necessarily diminish model precision—provided the dataset is well curated and sufficient computational resources are available.
2. **End-to-End CNN Training vs. Hybrid Feature Extraction:** Unlike Qasmi et al.'s approach, which used a CNN (VGG16) mainly for feature extraction before applying classical ensemble methods, this model employs a fine-tuned convolutional neural network pipeline. Both methods illustrate valid strategies for biodiversity image classification. Ensemble approaches may be easier to interpret or to integrate with domain-specific features, whereas fine-tuning a CNN can leverage the model's internal feature hierarchy, especially when the training set is extensive.

Another distinguishing factor is the volume and diversity of images. In the Naturalis pilot, images came predominantly from a handful of museum collections plus some private collections (15,877 total) [5]. Here, over 130,000 images were aggregated from a wide array of sources, including community-driven repositories, potentially bringing greater variance in lighting conditions, viewpoints, and shell morphologies. While this diversity strengthens generalizability, it also escalates demands on data preprocessing, standardization, and computational power. In particular, the workflow included automated image segmentation (one shell per image) and uniform background replacement—steps that appear to significantly streamline model training. Both the Naturalis pilot and Qasmi et al. [6] used similarly rigorous approaches for data cleaning, but with smaller datasets and fewer species, the effect of image variation may have been comparatively lower.

At last, a notable distinction in the current study is the use of transfer learning and fine-tuning, particularly leveraging an EfficientNet architecture pretrained on ImageNet. Although ImageNet does not contain seashell images, the extensive and diverse features learned from over a million labeled images still confer a significant advantage when training on the Conus dataset—or any other seashell dataset. By pretraining on ImageNet and then fine-tuning on domain-specific images, the model inherits rich, general-purpose visual representations that aid in discerning even subtle morphological details of shells. This approach is particularly effective for fine-grained biodiversity classification, where expert-labeled data are often scarce and species distinctions can be minute. [7].

Although direct performance comparisons can be confounded by differences in taxonomy, image sources, or evaluation protocols, these concurrent findings strongly support the viability of AI-based classification for large,

visually diverse mollusk genera. Future work may involve combining the strengths of these approaches: unifying data from multiple sources, benchmarking different architectures or ensemble methods, and assessing the impact of refined taxonomic standards on model reliability.

One of the major challenges in building a robust *Conus* classification model is the imbalance in species representation, where some species have thousands of images while others have only a few. This imbalance is a common issue in biological datasets, where rare or newly discovered species often have limited available data [1, 2]. Few-shot learning techniques based on meta-learning and contrastive learning have been successfully applied in biodiversity classification to address data scarcity [8]. In recent studies, prototypical networks and metric-based learning have enabled models to recognize species with only a few labeled images by learning generalized feature spaces that capture taxonomic similarities [9, 10]. Similarly, contrastive learning, which pretrains models using large unlabeled datasets, has demonstrated superior transferability for species recognition tasks [11, 12]. Future work could explore such approaches to enhance classification performance for *Conus* species with very few training images, reducing the impact of dataset imbalance. By incorporating these advanced transfer-learning methods, AI models could better support biodiversity research, particularly for rare and underrepresented species.

This *Conus* AI model is a node of the hierarchical CNN model available at Identifyshell.org.

Conclusion

In summary, this work demonstrates the feasibility and accuracy of a large-scale CNN-based classification model for *Conus* shells — one of the most diverse and taxonomically challenging groups within the Mollusca. By assembling a dataset of over 130 000 images representing 518 species, we highlight the key hurdles inherent to fine-grained shell identification, including taxonomic ambiguity, limited or imbalanced species-specific data, and subtle morphological differences. Careful data curation, background standardization, and strategic model fine-tuning were crucial in achieving consistent performance across hundreds of species, as evidenced by high macro-averaged metrics and reliable confidence intervals.

The findings underscore that high coverage of *Conus* species need not compromise classification accuracy, provided the dataset is sufficiently robust and preprocessing steps are meticulously executed. Comparing our results to earlier *Conus* AI models further illustrates how diverse computational strategies—ranging from end-to-end CNN training to hybrid feature extraction—can yield strong results in challenging biodiversity contexts. These approaches collectively validate the viability of automated shell recognition on a scale that can significantly accelerate research and improve collection management for museums, citizen science platforms, and other stakeholders interested in marine biodiversity.

Supplementary Material

Tables II (Images per Species) and Table V (Performance Metrics per Species) are available at [DOI: 10.5281/zenodo.16013529](https://doi.org/10.5281/zenodo.16013529).

References

- [1] N. Puillandre et al. *One, four or 100 genera? A new classification of the cone snails* [Journal of Molluscan Studies](#) 81: 1–23 (2015)
- [2] Van Horn, G., et al. *The iNaturalist Species Classification and Detection Dataset* [Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition \(CVPR\)](#), 8769-8778. (2018)
- [3] Brown, L. D. et al. *Interval Estimation for a Binomial Proportion* [Statistical Science](#), 16(2), 101–133 (2001)
- [4] A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application* [Cambridge University Press](#) (1997)
- [5] A. Hendriksen *Automatische beeldherkenning als instrument voor museumcollecties. Proces en resultaten - lessons learned* [Symposium Museumcollecties & AI](#), 16 mei 2022 (2022)
- [6] N. Qasmi et al. *Recognition of Conus species using a combined approach of supervised learning and deep learning-based feature extraction* . [PLoS ONE](#) 19(12):e0313329 (2024)
- [7] Yin Cui et al. *Large Scale Fine-Grained Categorization and Domain-Specific Transfer Learning* [Computer Vision and Pattern Recognition](#) (2018)
- [8] Chen, C. et al. *Few-shot learning for wildlife recognition: A case study in Senegal*. [AI](#), 4(3), 574-597 (2023)
- [9] Snell, J. et al. *Prototypical networks for few-shot learning* [Advances in Neural Information Processing Systems](#), 30, 4077–4087 (2017)
- [10] Wang, Y. et al. *Generalizing from a few examples: A survey on few-shot learning*. [ACM Computing Surveys](#), 53(3), 1–34
- [11] Jaiswal, A. et al. *A Survey on Contrastive Self-Supervised Learning* [Applied Intelligence](#), 51, 3498–3514. (2021)
- [12] Grill, J.-B. et al. *Bootstrap your own latent: A new approach to self-supervised learning* [Advances in Neural Information Processing Systems](#), 33, 21271–21284 (2020)