
FORAMDEEPSLICE: A HIGH-ACCURACY DEEP LEARNING FRAMEWORK FOR FORAMINIFERA SPECIES CLASSIFICATION FROM 2D MICRO-CT SLICES

A PREPRINT

 **Abdelghafour Halimi**

Visualization Core Lab
Thuwal, 23955, Saudi Arabia
abdelghafour.halimi@kaust.edu.sa

 **Ali Alibrahim**

Physical Sciences and Engineering Division
Thuwal, 23955, Saudi Arabia
ali.alibrahim@kaust.edu.sa

 **Didier Barradas-Bautista**

Visualization Core Lab
Thuwal, 23955, Saudi Arabia
didier.barradas@kaust.edu.sa

 **Ronell Sicat**

Visualization Core Lab
Thuwal, 23955, Saudi Arabia
ronell.sicat@kaust.edu.sa

 **Abdulkader M. Afifi**

Physical Sciences and Engineering Division
Thuwal, 23955, Saudi Arabia
abdulkader.alafifi@kaust.edu.sa

December 2, 2025

ABSTRACT

This study presents a comprehensive deep learning pipeline for the automated classification of 12 foraminifera species using 2D micro-CT slices derived from 3D scans. We curated a scientifically rigorous dataset comprising 97 micro-CT scanned specimens across 27 species, selecting 12 species with sufficient representation for robust machine learning. To ensure methodological integrity and prevent data leakage, we employed specimen-level data splitting, resulting in 109,617 high-quality 2D slices (44,103 for training, 14,046 for validation, and 51,468 for testing). We evaluated seven state-of-the-art 2D convolutional neural network (CNN) architectures using transfer learning. Our final ensemble model, combining ConvNeXt-Large and EfficientNetV2-Small, achieved a test accuracy of **95.64%**, with a top-3 accuracy of **99.6%** and an area under the ROC curve (AUC) of **0.998** across all species. To facilitate practical deployment, we developed an interactive advanced dashboard that supports real-time slice classification and 3D slice matching using advanced similarity metrics, including SSIM, NCC, and the Dice coefficient. This work establishes new benchmarks for AI-assisted micropaleontological identification and provides a fully reproducible framework for foraminifera classification research, bridging the gap between deep learning and applied geosciences.

Keywords foraminifera classification, deep learning, transfer learning, 2D slices, 3D micro-CT data

1 Introduction

Paleontology is undergoing a transformative data revolution, catalyzed by the widespread adoption of high-resolution, non-destructive imaging technologies such as micro-Computed Tomography (micro-CT) [Knutsen and Kononov, 2024, Edie et al., 2023]. These imaging methods have become indispensable for exploring the internal morphology of fossil specimens, offering unprecedented access to delicate anatomical structures that would otherwise be compromised by traditional mechanical preparation techniques [Heřmanová et al., 2020, Cunningham et al., 2014]. However, this

technological leap has introduced a significant bottleneck in post-processing: the manual segmentation of fossils from surrounding rock matrices. As imaging resolutions increase, the volumetric datasets grow exponentially, making segmentation—especially for low-contrast specimens like calcareous fossils in carbonate-rich matrices—the most labor-intensive and error-prone step in the workflow [Edie et al., 2023].

To address this challenge, the field has increasingly turned to deep learning, a subset of machine learning that excels in automated image analysis. Early applications demonstrated that convolutional neural networks (CNNs), particularly U-Net architectures, could segment fossils with accuracies rivaling manual annotation, but in a fraction of the time [Knutsen and Konovalov, 2024, Edie et al., 2023, Itaki et al., 2020a]. These breakthroughs have laid the foundation for a broader integration of artificial intelligence (AI) into paleontological research, extending beyond segmentation to include fossil classification and reconstruction, as summarized in Table 1.

One of the most promising directions in this domain is the classification of microfossils using transfer learning. Several studies have shown that pretrained CNN architectures—such as VGG16, ResNet50, MobileNetV2, InceptionV3, and Xception—can be fine-tuned to classify microfossils from low-resolution optical images with remarkable accuracy. For instance, classification of the genus *Globotruncanita* achieved up to **96.97%** accuracy and an AUC of **0.978** [Ozer et al., 2024], while broader genus-level classification reached **99.78%** accuracy [Ozer et al., 2023a]. These approaches are particularly valuable for laboratories with limited computational resources, offering high performance with minimal infrastructure.

Beyond transfer learning, researchers have explored custom architectures built from scratch, combining CNNs with recurrent neural networks (LSTM, BiLSTM) to classify species within the genus *Globotruncana*, achieving up to **97.35%** accuracy and an AUC of **0.968** [Ozer et al., 2023b]. Classical machine learning models such as SVM, LDA, and kNN have also demonstrated strong performance when paired with tailored CNNs, validating the utility of lightweight models in resource-constrained settings [Gorur et al., 2023].

In parallel, semantic segmentation and 3D reconstruction of microfossils have advanced through hybrid architectures combining U-Net and ResNet, enabling precise segmentation in CT and micro-CT images with IoU scores exceeding **94%** and segmentation accuracies up to **98%** [Hou et al., 2021, Carvalho et al., 2020]. These methods preserve spatial fidelity and support detailed reconstructions essential for paleoenvironmental and geological analyses.

Finally, synthetic data generation has emerged as a powerful tool to overcome the scarcity of labeled paleontological datasets. The ForamiViT-GAN framework, which integrates Vision Transformers (SwinIR) with GANs (StyleGAN2), produces high-fidelity synthetic images of foraminifera [Ferreira-Chacua and Koeshidayatullah, 2023]. Few-shot segmentation using only five examples further demonstrates the model’s efficiency. Complementary work using fractal pretraining (FDSL) with Vision Transformers has enabled multiclass classification of **32 radiolarian species** with **86.3%** accuracy, showing that mathematically generated data can rival real images in training efficacy [Mimura et al., 2024].

Together, these innovations mark a paradigm shift in paleontological research, where deep learning not only accelerates traditional workflows but also opens new avenues for fossil discovery, classification, and interpretation.

Foraminifera (forams) are unicellular aquatic organisms that produce calcareous tests (shells), which fossilize abundantly in carbonate rocks, making them essential for reconstructing Earth’s evolutionary history and past climate [Faulkner et al., 2025]. The introduction of 30 μm -thick thin sections by Henry Clifton Sorby in 1852 enabled the first systematic study of foraminifera from carbonate rocks—a technique still central to biostratigraphy today [Chidsey et al., 2015, Ueno, 2022]. While this work focuses on foram micro-CT data, the proposed workflow is broadly applicable to other microfossils, providing a framework to evaluate and optimize deep learning pipelines for automated classification.

2 Methods

2.1 Dataset Creation and Curation

We constructed a scientifically rigorous dataset from 97 micro-CT scanned fossil specimens representing 27 foraminifera species. Twelve species with at least four 3D models each were selected to ensure robust training (we use the term 3D model to refer to individual 3D micro-CT image stacks). We provide a general overview of this AI-based workflow for the reader in Figure 1. The selected species are summarized in Figure 2 and Table 2. To prevent data leakage [Yagis et al., 2021], we implemented a strict specimen-level splitting protocol, assigning all slices from a single 3D model exclusively to one of the training, validation, or test sets. This approach avoids the train-test contamination common in slice-level splits, which can overestimate accuracy by up to 55%.

Slice extraction employed Otsu’s [Otsu, 1979] thresholding to remove low-content slices, followed by advanced segmentation to isolate fossil silhouettes. The final dataset comprises 109,617 high-quality 224×224 RGB images:

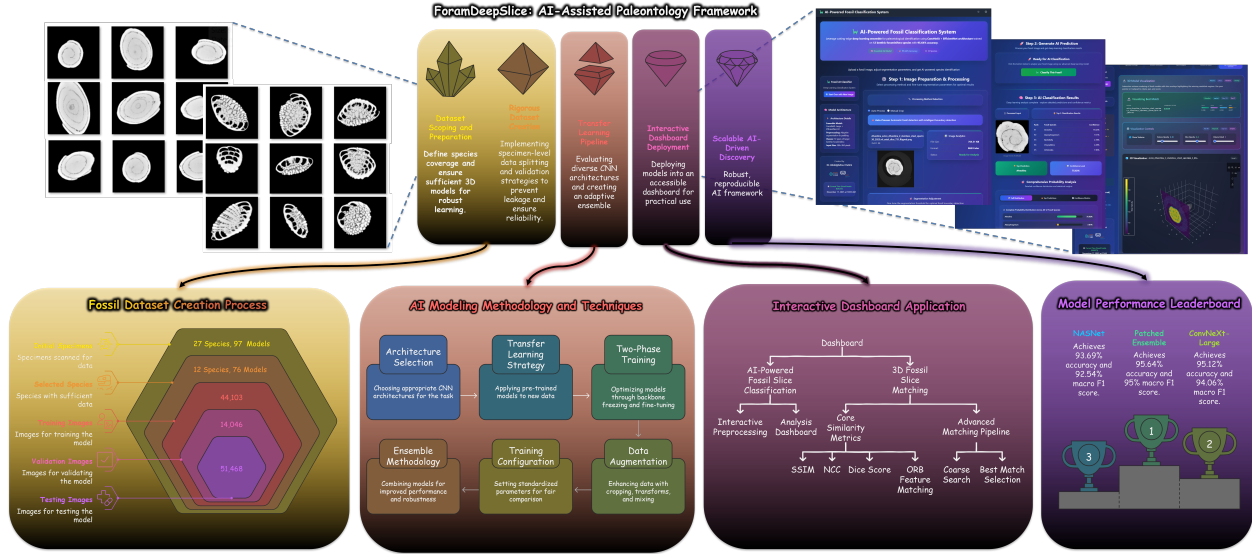


Figure 1: **Deep learning pipeline for automated foraminifera classification.** We curated 97 micro-CT scanned specimens across 27 species, selecting 12 for robust training. Using specimen-level data splitting, 109,617 2D slices were processed through seven CNN architectures. Our PatchEnsemble model achieved **95.64%** accuracy and **AUC = 0.998**. An interactive dashboard enables real-time slice classification and 3D matching, setting new benchmarks for AI-assisted micropaleontology.

44,103 for training, 14,046 for validation, and 51,468 for testing (Table 2). Dataset balance was optimized using brute-force coefficient of variation minimization (Table 3).

Disclaimer: While every effort was made to ensure accurate classification of the data samples used in this study, there remains a possibility of misidentification or mislabeling of species or classes due to limitations in the source data, annotation tools, or expert consensus. These potential discrepancies should be considered when interpreting the results, and future work may involve revalidation or refinement of sample classifications.

2.2 Model Architecture and Training Strategy

We evaluated seven state-of-the-art 2D CNN architectures: ConvNeXt [Liu et al., 2022] (Base and Large), Efficient-NetV2[Tan and Le, 2021] (Small and Large), NASNet [Zoph et al., 2018], MobileNet [Howard et al., 2019], and ResNet101V2[He et al., 2016]. Transfer learning was applied using ImageNet-pretrained weights. Training followed a two-phase strategy:

- **Phase 1: Backbone Freezing** — 10 epochs with frozen pretrained weights; only classifier head trained. Learning rate: $1e-4$ with cosine scheduling.
- **Phase 2: Fine-Tuning** — 20 epochs with full network unfrozen; learning rate: $5e-5$ with warm restarts. Mixed precision (fp16) training was used for efficiency. Early stopping was applied with patience=5 based on validation loss.

All models used standardized configurations: input resolution (224×224 or 384×384), batch size 32, AdamW optimizer [Loshchilov and Hutter, 2017], and categorical crossentropy loss with label smoothing (0.1) [Szegedy et al., 2016]. Training was conducted on NVIDIA GPUs with fixed seeds and deterministic operations for reproducibility.

2.3 Data Augmentation and Preprocessing

To enhance generalization, we applied intelligent bounding-box cropping and a suite of augmentations:

- **Geometric:** rotations ($\pm 45^\circ$), scaling (0.8–1.2 $\times$), horizontal flips.
- **Color:** brightness/contrast adjustments, jittering.
- **Regularization:** CutMix[Yun et al., 2019] and Mixup[Zhang et al., 2017] strategies.

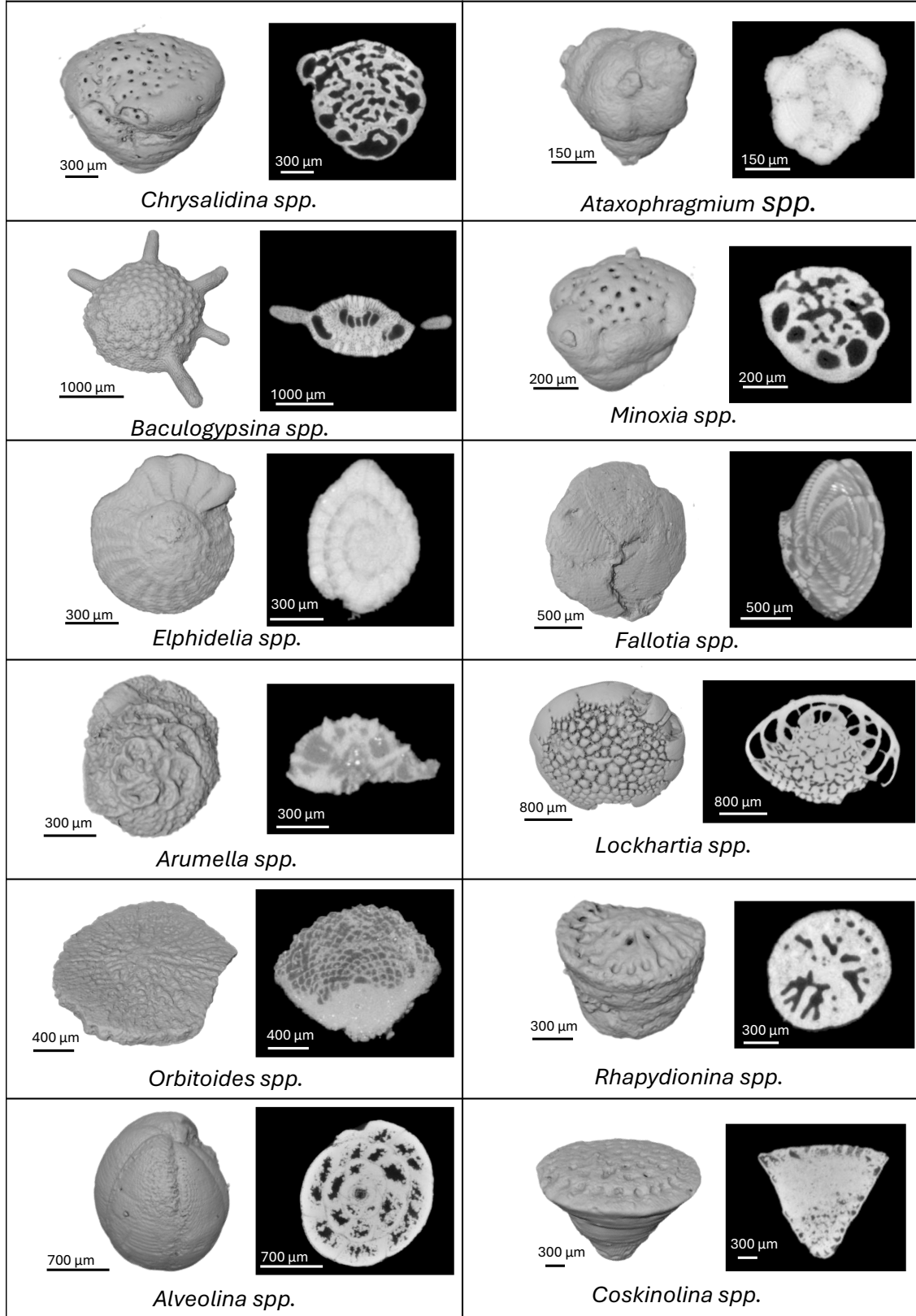


Figure 2: **3D volume rendering and sample 2D slice from micro-CT data for each species used in the study.** We used Avizo software for visualization and conversion of micro-CT data to NIfTI files used for deep learning workflows.

2.4 Ensemble Modeling

Our final model is a weighted PatchEnsemble combining ConvNeXt-Large and EfficientNetV2-Small. PatchEnsemble refers to an ensemble strategy that combines predictions from multiple models using confidence-based switching logic [Zhou, 2025]. This adaptive approach leverages complementary strengths across architectures and improves robustness. The ensemble achieves 95.64% test accuracy and 99.6% top-3 accuracy.

2.5 Interactive Dashboard Deployment

We deployed the trained models via a dashboard with two core applications:

- **Real-Time Slice Classification:** interactive interface with ranked predictions, confidence metrics, and preprocessing tools.
- **3D Slice Matching:** similarity analysis using Structural Similarity Index (SSIM) [Zhou Wang et al., 2004], Normalized Cross-Correlation (NCC) [Lewis, 1994], Dice coefficient [Dice, 1945], and Oriented FAST and Rotated BRIEF (ORB) feature matching [Rublee et al., 2011].

The dashboard bridges research and field application, enabling real-time fossil identification and correspondence analysis.

2.6 Reproducibility and Environment

A Docker environment is provided for reproducibility, with minimum system requirements: Python 3.8+, 16GB RAM, 8GB VRAM GPU, and 100GB storage.

The following Docker command launches the full environment required for dataset generation, model training, and dashboard deployment:

```
docker run --gpus all --ipc=host --ulimit memlock=-1 --ulimit stack=67108864 \
  --rm -p 10000:8888 -p 8501:8501 -v ${PWD}:/workspace/mycode \
  abdelghafour1/ngc_tf_rapids_25_01_vscodetorch:2025-v3 \
  jupyter lab --ip=0.0.0.0 --allow-root
```

3 Results

3.1 State-of-the-art Models Performance Overview

We systematically evaluated seven state-of-the-art 2D convolutional neural network (CNN) architectures on a balanced dataset comprising 109,617 micro-CT slices. To benchmark model performance, we computed all standard classification metrics, including accuracy, F1-score, and area under the receiver operating characteristic curve (AUC) (Figure 3).

As a baseline, we selected ResNet101V2, against which all other models were compared. Classification accuracy across models ranged from 86% to 95%, with ConvNeXt-Large achieving the highest accuracy of 95%. The F1-score, used as a proxy for classification quality, revealed that ConvNeXt-Large and NASNet exhibited the most favorable balance between precision and recall, indicating robust performance across species.

All models demonstrated excellent calibration, with AUC values consistently exceeding **0.98**, underscoring their reliability in distinguishing between classes.

3.1.1 Species-Level Performance

To assess model performance at a finer granularity, we computed three key classification metrics precision, recall, and F1-score at the species level, treating each species as an independent class.

Most models achieved F1-scores exceeding 90%, with **Fallotia** reaching a peak performance of **99.66%**. However, certain species posed greater classification challenges. Notably, *Baculogypsina* and *Orbitoides* exhibited lower F1-scores of **75.22%** and **87.17%**, respectively (Fig. 4).

A closer inspection of precision and recall revealed distinct error patterns. *Baculogypsina* consistently showed the lowest precision across models, indicating a high rate of false positives. In contrast, *Orbitoides* suffered from low recall, suggesting a substantial number of false negatives. These deficiencies are reflected in their reduced F1-scores, which signal a decline in overall classification quality for these species.

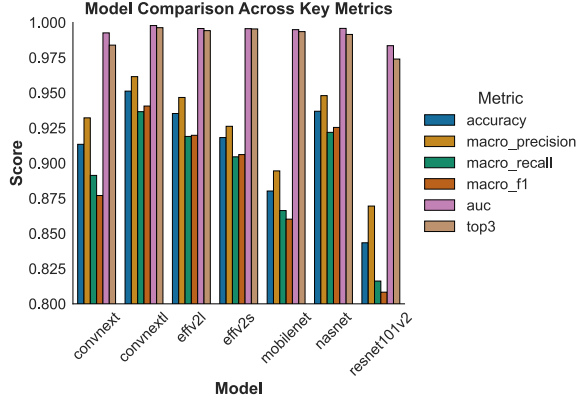


Figure 3: **Performance comparison of seven state-of-the-art 2D CNN architectures.** Each model was evaluated on a balanced dataset of 109,617 micro-CT slices using standard classification metrics. Accuracy, F1-score, and AUC are reported for each architecture. ConvNeXt-Large achieved the highest accuracy, while both ConvNeXt-Large and NASNet demonstrated superior F1-scores. All models exhibited excellent calibration, with AUC values exceeding 0.98.

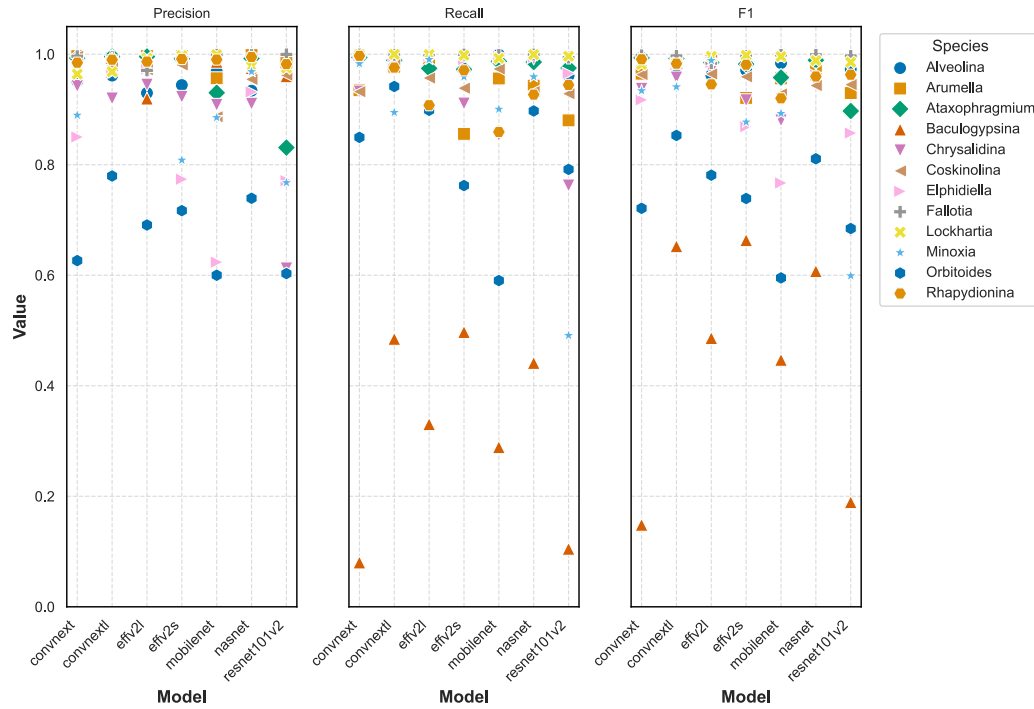


Figure 4: **Species-level classification metrics across models.** F1-score, precision, and recall were computed for each species, revealing strong overall performance but notable weaknesses for *Baculogypsina* and *Orbitoides*. These species exhibited distinct error patterns—high false positives for *Baculogypsina* and high false negatives for *Orbitoides*—resulting in reduced F1-scores.

3.2 PatchEnsemble Model Performance

To further enhance classification performance, particularly in challenging cases, we implemented a *PatchEnsemble* strategy. Usually, the PatchEnsemble uses different models for different patches of one image, and their outputs are aggregated to form a final prediction. Unlike traditional ensemble methods, which typically rely on majority voting or averaging across models trained on the full image. PatchEnsemble introduces diversity by leveraging specific context, thereby improving robustness and sensitivity for the classification.

In our study, we used the F1-score as a guiding metric to identify the most difficult cases and selectively apply the PatchEnsemble strategy to support the best-performing fine-tuned model. This targeted approach led to a measurable improvement in classification performance. In contrast, a conventional ensemble method based on majority voting did not yield any significant gains, highlighting the advantage of spatially-aware ensemble techniques in this domain.

We compare the performance metrics of the traditional ensemble combination of the Top-2(ConvNeXt-Large + NASNet), Top-3 (Top-2 + EfficientNetV2-Large), and Top-4 (Top-3 + EfficientNetV2-Small) models according to the results of our systematic analysis of the state-of-the-art models and the PatchEnsemble model of ConvNeXt-Large and EfficientNetV2-Small.

The final PatchEnsemble model, combining ConvNeXt-Large and EfficientNetV2-Small, was named *ForamDeepSlice*, achieved **95.64%** test accuracy and **99.6%** top-3 accuracy (Table 8)

3.2.1 Species-specific classification challenges

Species-level analysis revealed notable disparities in classification performance, particularly for *Baculogypsina* and *Orbitoides*. While most species achieved F1-scores above 90%, these two taxa consistently underperformed across ensemble configurations. *Baculogypsina* exhibited exceptionally high precision (up to 0.999), yet suffered from markedly low recall (as low as 0.227), indicating that the model was highly conservative in predicting this class and frequently missed true instances. Conversely, *Orbitoides* maintained strong recall values (above 0.90), but its precision was comparatively low (e.g., 0.666 in Top-3), suggesting a tendency to misclassify other species as *Orbitoides*, leading to false positives.

These imbalances resulted in reduced F1-scores, highlighting the limitations of conventional ensemble methods in handling species with ambiguous or overlapping morphological features (Figure 5; Figure 6). The PatchEnsemble strategy mitigated some of these issues by leveraging localized image regions, leading to improved recall for *Baculogypsina* and better precision for *Orbitoides*. This targeted enhancement underscores the importance of spatially-aware ensemble techniques in fine-grained biological classification tasks.

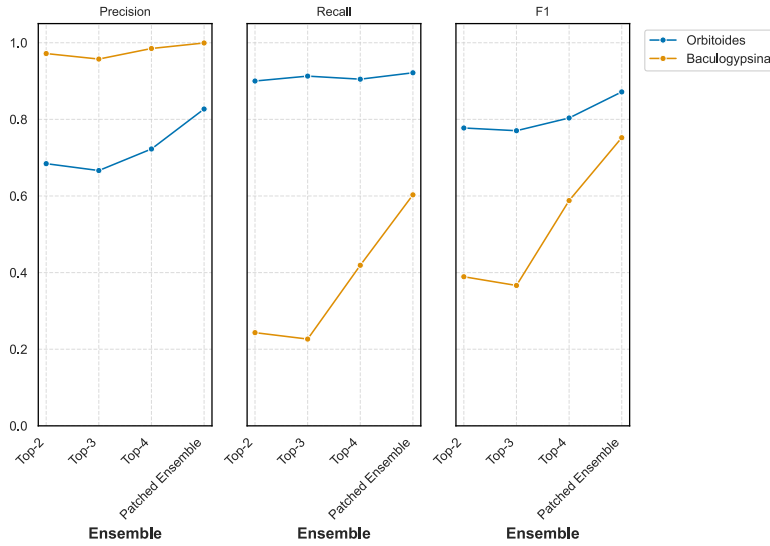
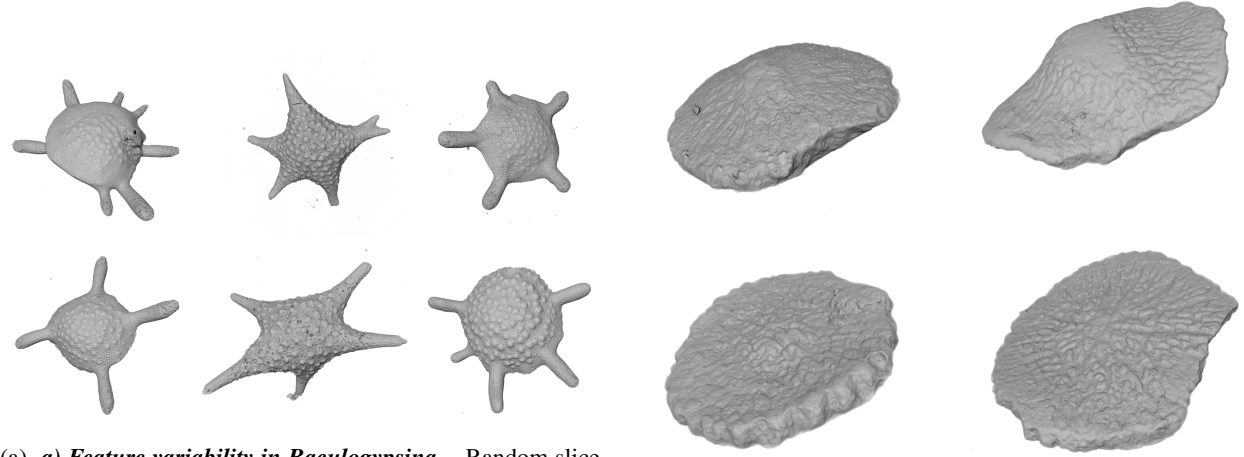


Figure 5: **PatchEnsemble improves classification of difficult species.** Compared to conventional ensemble methods, the PatchEnsemble strategy significantly boosts precision and recall for *Baculogypsina* and *Orbitoides*, which are otherwise prone to misclassification.



(a) ***a) Feature variability in Baculogypsina.*** Random slice sampling captures highly variable features due to the species' asymmetry, such as differing numbers of radial extensions, complicating deep learning classification.

(b) ***b) Extraction-induced damage in Orbitoides.*** Physical deformations—fractures, missing fragments, and cracks—negatively impact classification accuracy. PatchEnsemble mitigates these challenges by leveraging localized predictions.

Figure 6: **Challenges in micro-CT slice-based classification of foraminifera.** (a) Intrinsic morphological variability in *Baculogypsina* introduces inconsistencies across slices. (b) Specimen damage in *Orbitoides* further complicates classification, highlighting the need for robust strategies such as PatchEnsemble.

3.3 User Interface Development for Domain Scientists

To facilitate the practical deployment of fossil classification tools, we developed an interactive dashboard tailored to the needs of paleontologists and domain scientists. The interface does not require coding background and is designed to be intuitive and accessible, lowering the barrier to entry for non-technical users while providing advanced functionality for expert analysis. The supplementary video provides a complete view of the interactive dashboard as well as an example workflow for foraminifera species classification and 3D slice matching. The supplementary video can be found here: <https://youtu.be/betkJ3gsmNQ>.

The application is structured into three main layers (shown in Figure 7):

1. Landing Page Upon launching the tool, users are greeted with a landing page that introduces the system's capabilities and a quick start guide. The page describes the datasets and the overall classification performance. It also provides the entry points for classification and slice-matching pages.

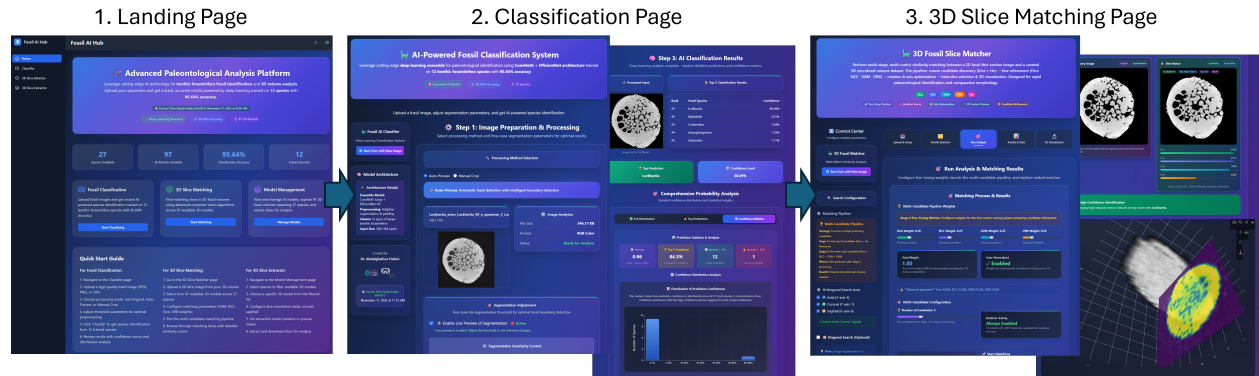


Figure 7: **Our interactive dashboard provides a user-friendly access to the classifier and 3D matching algorithm.** A complete view of the dashboard in action is provided in our supplementary video (<https://youtu.be/betkJ3gsmNQ>).

2. Classification Page This page features a drag-and-drop interface for uploading micro-CT images and provides basic preprocessing options, including contrast enhancement, image cropping, and image sharpening, to enhance input quality before analysis. Once an image is uploaded, an automatic cropping, denoising, and segmentation is performed to isolate the foram. A preview of the result is shown on the page. The user can adjust this result by manually cropping and adjusting the segmentation sensitivity control. Once the user is satisfied with the input image, the classifier can be executed - this uses our ForamDeepSlice model to perform species-level classification. The result interface displays ranked predictions with associated confidence scores, probability distributions for each class, and additional metadata. The top-5 predictions are presented to support expert validation and further exploration, allowing users to interpret results in the context of their domain knowledge.

3. 3D Slice Matching Page Once the correct species is identified, users may choose to perform 3D slice matching to find the approximate slice orientation of the input image slice within a 3D model. This module treats the input image as a slice from a volumetric micro-CT scan and compares it against a curated database of 3D fossil models. The matching process employs a two-stage pipeline:

- **Coarse Matching:** Based on Dice coefficient and Hu moments.
- **Fine Matching:** Incorporates rotation-invariant analysis using Structural Similarity Index (SSIM) [Zhou Wang et al., 2004], Normalized Cross-Correlation (NCC) [Lewis, 1994], and Oriented FAST and Rotated BRIEF (ORB) feature matching [Rublee et al., 2011].

Users can optionally restrict the search to a subset of models based on prior knowledge or hypotheses, enabling targeted exploration. The best matching slice can also be visualized in the context of the 3D model using a simplified 3D visualization.

3.4 Example Results

In addition to the performance metrics reported earlier for the deep learning workflows, we also provide some example use-case results. The first example is a micro-CT slice from a *Lockhartia* test specimen illustrated in Figures 8 and 9. The user provides as input a raw micro-CT image of the specimen, which is then automatically pre-processed via the dashboard. This includes cropping, denoising, and segmentation of the foram specimen. The cleaned-up and enhanced image is then fed into the AI model for classification. As shown in Figure 8, the model correctly predicts the correct species (*Lockhartia*) with a confidence of 80.49%. The user then proceeds to find the input image’s slice orientation within the test model, which is indeed found by the algorithm, along with identifying the most similar slice in other models.

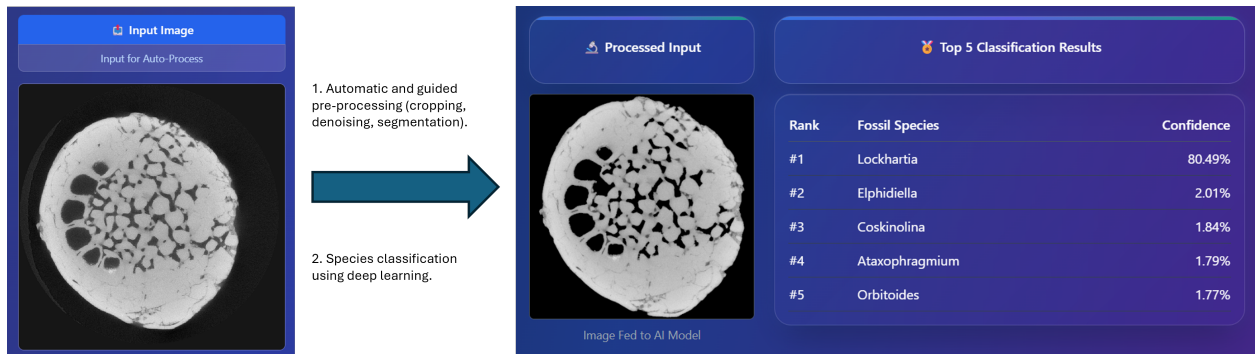


Figure 8: *Our AI model correctly classifies the input micro-CT image as belonging to a Lockhartia test specimen.*

While our AI model was trained on micro-CT image data, we show that it can still provide some interesting results even for other image types such as optical microscopy images, as shown in Figure 10. This is most likely due to the ability of the deep learning models to learn and capture the important structures of the foraminifers both as a whole and in parts due to their multi-scale nature. However, the introduction of artifacts such as cracks and breaks in fossils, as shown in the example, can be an interesting area to investigate and fully address in future iterations of this work.

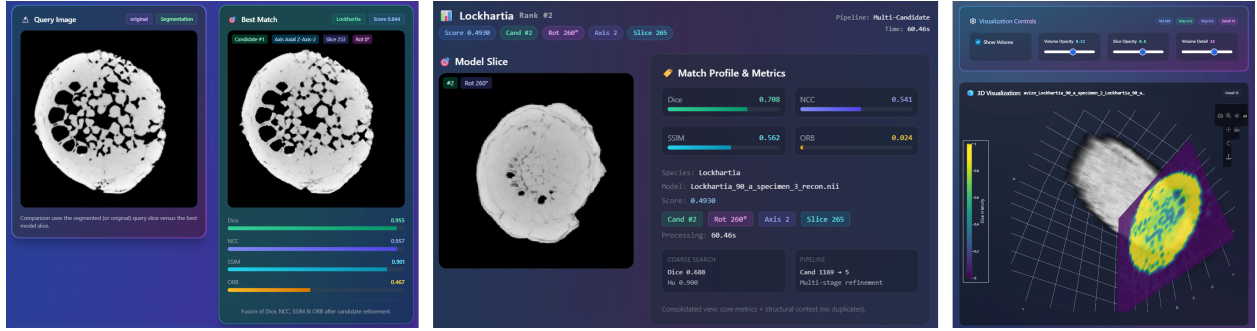


Figure 9: *Our 3D matcher finds the exact slice and orientation of the input image within the sample specimen (left), but also finds the most similar slice from other samples of the same species (middle). The 3D view displays the slice in the context of the 3D sample visualization (right).*

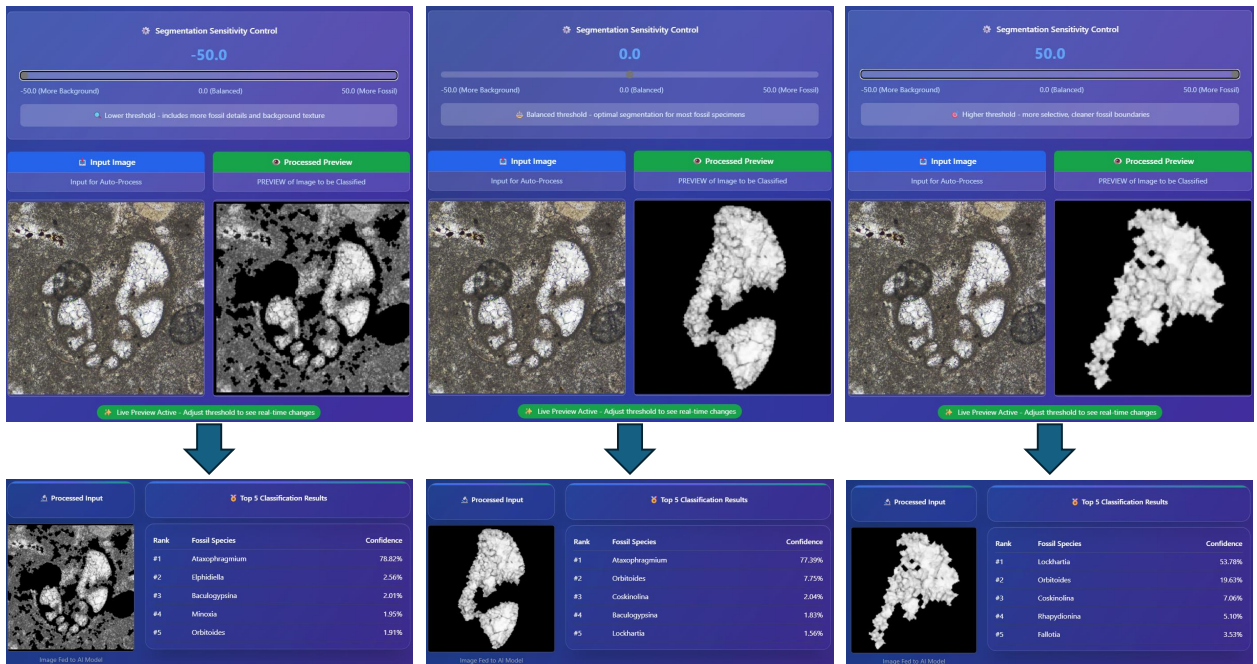


Figure 10: *Our classifier can sometimes yield good results even for non-micro-CT images, such as optical microscopy images in this example, as long as some of the unique distinguishing foram features are visible.* This particular example also showcases how the segmentation sensitivity control affects the details retained in the input image and the final classification results. In this case, extreme background removal (rightmost) removes most of the foram features, leading to an incorrect classification, while minimal (leftmost) and balanced (middle) settings yield the correct classification of *Ataxophragmium*.

4 Discussion

The benefits of computed tomography (CT) scanning in paleontology are undeniable, offering unprecedented access to the internal and external morphology of fossil specimens. In particular, micro-CT imaging enables the detailed study of complex structures in foraminifera, allowing experts to classify specimens with greater accuracy and based on a richer set of morphological features than traditional 2D imaging techniques.

In this study, we addressed the challenge of classifying 12 unique species of foraminifera using micro-CT data. Given the similarity of this task to medical image classification, we adopted a transfer learning approach, leveraging pretrained convolutional neural networks (CNNs) to benchmark performance on this domain-specific dataset. All tested models

outperformed the baseline (ResNet101V2), with ConvNeXt-Large emerging as the top-performing architecture in terms of overall accuracy and F1-score.

However, species-level analysis revealed that the models struggled with two particularly challenging taxa: *Orbitoides* and *Baculogypsina*. We attribute this diminished performance to class imbalance and limited representation of these species in the training data. To address this, we implemented a targeted strategy using a PatchEnsemble approach, which combines spatially localized predictions from multiple models to enhance classification robustness. Specifically, we patched the best-performing model (ConvNeXt-Large) with models that showed superior performance on the difficult species.

In contrast, a traditional ensemble approach—combining the Top-2 to Top-4 models via majority voting—did not yield improved results. We hypothesize that this is due to the propagation of false positives and false negatives across models, which diluted the ensemble’s ability to correct errors. The PatchEnsemble strategy, by contrast, enabled us to selectively integrate complementary strengths, resulting in improved overall performance and notably better classification of the most challenging species.

While this AI-based workflow represents an important step toward automated micropaleontological classification, it does not replace careful taxonomic analysis. Geological context remains indispensable, particularly in cases of homeomorphy—where unrelated foraminifera from different ages evolve similar morphologies. For example, confusion can arise between homeomorphic taxa such as *Lepidorbitoides* spp. from the Cretaceous and *Lepidocyclina* spp. from the Cenozoic. Although homeomorphic taxa were not included in this study, it is critical to acknowledge that the algorithm is trained only on the listed species and may misclassify specimens that resemble others outside the training set. Thus, the system should be viewed as a decision-support tool, guiding experts toward the most likely matches rather than providing definitive identifications.

5 Acknowledgements

The authors gratefully acknowledge the support of the King Abdullah University of Science and Technology (KAUST) Core Labs, particularly the Supercomputing Core Lab for providing computational resources on the Ibex cluster. We thank Domingo Lattanzi-Sanchez of KAUST Geo-Energy and Mineral Resources Platform for micro-CT scanning support. Their contributions were essential to the success of this research.

6 Author contributions statement

A.H. conceived the experiment(s) and conducted the experiment(s), R.S. and A.A. produced and obtained the raw data, A.H. and D.B.B. analyzed the results and managed the project. All authors reviewed the manuscript.

References

- L.E. Carvalho, G. Fauth, S. Baecker Fauth, G. Krah, A.C. Moreira, C.P. Fernandes, and A. Von Wangenheim. Automated Microfossil Identification and Segmentation using a Deep Learning Approach. *Marine Micropaleontology*, 158:101890, June 2020. ISSN 03778398. doi:10.1016/j.marmicro.2020.101890. URL <https://linkinghub.elsevier.com/retrieve/pii/S0377839819300830>.
- Thomas C. Chidsey, Michael D. Vanden Berg, and David E. Eby. Petrography and characterization of microbial carbonates and associated facies from modern Great Salt Lake and Uinta Basin’s Eocene Green River Formation in Utah, USA. *Geological Society, London, Special Publications*, 418(1):261–286, January 2015. ISSN 0305-8719, 2041-4927. doi:10.1144/SP418.6. URL <https://www.lyellcollection.org/doi/10.1144/SP418.6>.
- John A. Cunningham, Imran A. Rahman, Stephan Lautenschlager, Emily J. Rayfield, and Philip C.J. Donoghue. A virtual world of paleontology. *Trends in Ecology & Evolution*, 29(6):347–357, June 2014. ISSN 01695347. doi:10.1016/j.tree.2014.04.004. URL <https://linkinghub.elsevier.com/retrieve/pii/S0169534714000871>.
- Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, July 1945. ISSN 0012-9658, 1939-9170. doi:10.2307/1932409. URL <https://esajournals.onlinelibrary.wiley.com/doi/10.2307/1932409>.
- Stewart M. Edie, Katie S. Collins, and David Jablonski. High-throughput micro-CT scanning and deep learning segmentation workflow for analyses of shelly invertebrates and their fossils: Examples from marine Bivalvia. *Frontiers in Ecology and Evolution*, 11:1127756, March 2023. ISSN 2296-701X. doi:10.3389/fevo.2023.1127756. URL <https://www.frontiersin.org/articles/10.3389/fevo.2023.1127756/full>.

- Katherine Faulkner, Christopher Lowery, Rowan Clare Martindale, Carl Simpson, and Andrew Jeffrey Fraass. Record of Foraminifera test composition throughout the Phanerozoic. *Proceedings of the Royal Society B: Biological Sciences*, 292(2044):20250221, April 2025. ISSN 1471-2954. doi:10.1098/rspb.2025.0221. URL <https://royalsocietypublishing.org/doi/10.1098/rspb.2025.0221>.
- Ivan Ferreira-Chacua and Ardiansyah Koeshidayatullah. ForamiViT-GAN: Exploring New Paradigms in Deep Learning for Micropaleontological Image Analysis, 2023. URL <https://arxiv.org/abs/2304.04291>. Version Number: 1.
- Kutlucan Gorur, Caner Kaya Ozer, Ilyas Ozer, Ali Can Karaca, Onursal Cetin, and Ismail Kocak. Species-Level Microfossil Prediction for Globotruncana genus Using Machine Learning Models. *Arabian Journal for Science and Engineering*, 48(2):1315–1332, February 2023. ISSN 2193-567X, 2191-4281. doi:10.1007/s13369-022-06822-5. URL <https://link.springer.com/10.1007/s13369-022-06822-5>.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi:10.1109/CVPR.2016.90. URL <http://ieeexplore.ieee.org/document/7780459/>.
- Zuzana Heřmanová, Jana Bruthansová, Katarína Holcová, Radek Mikuláš, Martina Kočová Veselská, Tomáš Kočí, Jan Dudák, and Martin Vohník. Benefits and limits of x-ray micro-computed tomography for visualization of colonization and bioerosion of shelled organisms. *Palaeontologia Electronica*, 2020. ISSN 19353952, 10948074. doi:10.26879/1048. URL <https://palaeo-electronica.org/content/2020/3032-micro-ct-bioerosion>.
- Chengbin Hou, Xinyu Lin, Hanhui Huang, Sheng Xu, Junxuan Fan, Yukun Shi, and Hairong Lv. Fossil image identification using deep learning ensembles of data augmented multiviews. *Methods in Ecology and Evolution*, 14(12):3020–3034, December 2023. ISSN 2041-210X, 2041-210X. doi:10.1111/2041-210X.14229. URL <https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14229>.
- Yemao Hou, Xindong Cui, Mario Canul-Ku, Shichao Jin, Rogelio Hasimoto-Beltran, Qinghua Guo, and Min Zhu. ADMorph: A 3D Digital Microfossil Morphology Dataset for Deep Learning. *IEEE Access*, 8:148744–148756, 2020. ISSN 2169-3536. doi:10.1109/ACCESS.2020.3016267. URL <https://ieeexplore.ieee.org/document/9165546/>.
- Yemao Hou, Mario Canul-Ku, Xindong Cui, Rogelio Hasimoto-Beltran, and Min Zhu. Semantic segmentation of vertebrate microfossils from computed tomography data using a deep learning approach. *Journal of Micropalaeontology*, 40(2):163–173, October 2021. ISSN 2041-4978. doi:10.5194/jm-40-163-2021. URL <https://jm.copernicus.org/articles/40/163/2021/>.
- Andrew Howard, Mark Sandler, Bo Chen, Weijun Wang, Liang-Chieh Chen, Mingxing Tan, Grace Chu, Vijay Vasudevan, Yukun Zhu, Ruoming Pang, Hartwig Adam, and Quoc Le. Searching for MobileNetV3. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1314–1324, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-7281-4803-8. doi:10.1109/ICCV.2019.00140. URL <https://ieeexplore.ieee.org/document/9008835/>.
- Takuya Itaki, Yosuke Taira, Naoki Kuwamori, Toshinori Maebayashi, Satoshi Takeshima, and Kenji Toya. Automated collection of single species of microfossils using a deep learning–micromanipulator system. *Progress in Earth and Planetary Science*, 7(1):19, December 2020a. ISSN 2197-4284. doi:10.1186/s40645-020-00332-4. URL <https://progearthplanetsci.springeropen.com/articles/10.1186/s40645-020-00332-4>.
- Takuya Itaki, Yosuke Taira, Naoki Kuwamori, Hitoshi Saito, Minoru Ikehara, and Tatsuhiko Hoshino. Innovative microfossil (radiolarian) analysis using a system for automated image collection and AI-based classification of species. *Scientific Reports*, 10(1):21136, December 2020b. ISSN 2045-2322. doi:10.1038/s41598-020-77812-6. URL <https://www.nature.com/articles/s41598-020-77812-6>.
- Espen M. Knutsen and Dmitry A. Konovalov. Accelerating segmentation of fossil CT scans through Deep Learning. *Scientific Reports*, 14(1):20943, September 2024. ISSN 2045-2322. doi:10.1038/s41598-024-71245-1. URL <https://www.nature.com/articles/s41598-024-71245-1>.
- JP Lewis. Fast template matching. In D Laurendeau and C Cedras, editors, *VISION INTERFACE '95, PROCEEDINGS, PROCEEDINGS - CANADIAN IMAGE PROCESSING AND PATTERN RECOGNITION SOCIETY*, pages 120–123, 430 KING ST W, STE 205, TORONTO ON M5V 1L5, CANADA, 1994. Canadian Image Proc & Pattern Recognit Soc, CANADIAN INFORMATION PROCESSING SOC. Vision Interface 95 Conference (VI 95), QUEBEC CITY, CANADA, MAY 16-19, 1995.
- Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976,

- New Orleans, LA, USA, June 2022. IEEE. ISBN 978-1-6654-6946-3. doi:10.1109/CVPR52688.2022.01167. URL <https://ieeexplore.ieee.org/document/9879745/>.
- Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2017. URL <https://arxiv.org/abs/1711.05101>. Version Number: 3.
- K. Mimura, K. Nakamura, K. Yasukawa, E. C. Sibert, J. Ohta, T. Kitazawa, and Y. Kato. Applicability of Object Detection to Microfossil Research: Implications From Deep Learning Models to Detect Microfossil Fish Teeth and Denticles Using YOLO-v7. *Earth and Space Science*, 11(1):e2023EA003122, January 2024. ISSN 2333-5084, 2333-5084. doi:10.1029/2023EA003122. URL <https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2023EA003122>.
- Nobuyuki Otsu. A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1):62–66, January 1979. ISSN 0018-9472, 2168-2909. doi:10.1109/TSMC.1979.4310076. URL <http://ieeexplore.ieee.org/document/4310076/>.
- Ilyas Ozer, Ismail Kocak, Onursal Cetin, Ali Can Karaca, Caner Kaya Ozer, and Kutlucan Gorur. Towards investigation of transfer learning framework for Globotruncanita genus and Globotruncana genus microfossils in Genus-Level and Species-Level prediction. *Engineering Science and Technology, an International Journal*, 48:101589, December 2023a. ISSN 22150986. doi:10.1016/j.jestech.2023.101589. URL <https://linkinghub.elsevier.com/retrieve/pii/S2215098623002677>.
- Ilyas Ozer, Caner Kaya Ozer, Ali Can Karaca, Kutlucan Gorur, Ismail Kocak, and Onursal Cetin. Species-level microfossil identification for globotruncana genus using hybrid deep learning algorithms from the scratch via a low-cost light microscope imaging. *Multimedia Tools and Applications*, 82(9):13689–13718, April 2023b. ISSN 1380-7501, 1573-7721. doi:10.1007/s11042-022-13810-2. URL <https://link.springer.com/10.1007/s11042-022-13810-2>.
- Ilyas Ozer, Ali Can Karaca, Caner Kaya Ozer, Kutlucan Gorur, Ismail Kocak, and Onursal Cetin. The exploration of the transfer learning technique for Globotruncanita genus against the limited low-cost light microscope images. *Signal, Image and Video Processing*, 18(8-9):6363–6377, September 2024. ISSN 1863-1703, 1863-1711. doi:10.1007/s11760-024-03322-x. URL <https://link.springer.com/10.1007/s11760-024-03322-x>.
- Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*, pages 2564–2571, Barcelona, Spain, November 2011. IEEE. ISBN 978-1-4577-1102-2 978-1-4577-1101-5 978-1-4577-1100-8. doi:10.1109/ICCV.2011.6126544. URL <http://ieeexplore.ieee.org/document/6126544/>.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi:10.1109/CVPR.2016.308. URL <http://ieeexplore.ieee.org/document/7780677/>.
- Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller Models and Faster Training, 2021. URL <https://arxiv.org/abs/2104.00298>. Publisher: arXiv Version Number: 3.
- Katsumi Ueno. Carboniferous fusuline Foraminifera: taxonomy, regional biostratigraphy, and palaeobiogeographic faunal development. *Geological Society, London, Special Publications*, 512(1):327–496, January 2022. ISSN 0305-8719, 2041-4927. doi:10.1144/SP512-2021-107. URL <https://www.lyellcollection.org/doi/10.1144/SP512-2021-107>.
- Ekin Yagis, Selamawet Workalemahu Atnafu, Alba García Seco De Herrera, Chiara Marzi, Riccardo Scheda, Marco Giannelli, Carlo Tessa, Luca Citi, and Stefano Diciotti. Effect of data leakage in brain MRI classification using 2D convolutional neural networks. *Scientific Reports*, 11(1):22544, November 2021. ISSN 2045-2322. doi:10.1038/s41598-021-01681-w. URL <https://www.nature.com/articles/s41598-021-01681-w>.
- Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6022–6031, Seoul, Korea (South), October 2019. IEEE. ISBN 978-1-7281-4803-8. doi:10.1109/ICCV.2019.00612. URL <https://ieeexplore.ieee.org/document/9008296/>.
- Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond Empirical Risk Minimization, 2017. URL <https://arxiv.org/abs/1710.09412>. Version Number: 2.
- Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC press, 2025.
- Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004. ISSN 1057-7149, 1941-0042. doi:10.1109/TIP.2003.819861. URL <https://ieeexplore.ieee.org/document/1284395/>.

Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. Learning Transferable Architectures for Scalable Image Recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8697–8710, Salt Lake City, UT, June 2018. IEEE. ISBN 978-1-5386-6420-9. doi:10.1109/CVPR.2018.00907. URL <https://ieeexplore.ieee.org/document/8579005/>.

Table 1: Summary of State-of-the-Art Deep Learning Applications in Fossil Analysis. Table adjusted to full page width.

Reference	Fossil Type	Imaging Modality / Data Type	Task / Model	Key Contribution
Knutsen et al., 2024 [Knutsen and Konovalov, 2024]	Vertebrate limb bone	Synchrotron micro-CT	Segmentation / U-Net	Dice similarity of 0.96 with <2% training data
Edie et al., 2023 [Edie et al., 2023]	Calcareous shelly invertebrates	Micro-CT	Segmentation / Deep Learning	Workflow for low-contrast fossil segmentation
Ferreira et al., 2023 [Ferreira-Chacua and Koeshidayatullah, 2023]	Foraminifera microfossils	2D Images	Synthetic Data / ViT + StyleGAN	High-quality synthetic datasets for classification and segmentation
Mimura et al., 2024 [Mimura et al., 2024]	Radiolarians (32 classes)	Automatic Segmentation	Vision Transformers / ViT + FDSL	Accuracy of 86.3%; synthetic data matches real image performance
Hou et al., 2020 [Hou et al., 2020]	Vertebrate microfossils	3D Models	Classification / VGG16 + LSTM-RNN + SVM	Automated classification of vertebrate microfossils
Hou et al., 2021 [Hou et al., 2021]	Fish microfossils	CT Images	Semantic Segmentation / U-Net + ResNet34	Precise segmentation of fish microfossils
Hou et al., 2023 [Hou et al., 2023]	Fusulinid microfossils	2D Images	Classification / Multi-view Ensemble	High agreement with experts; robust with limited data
Ozer et al., 2023 [Ozer et al., 2023b]	Globotruncana (3 species)	Optical Microscopy	Hybrid Model / CNN + LSTM / BiLSTM	Accuracy of 97.35%; AUC of 0.968
Ozer et al., 2023 [Ozer et al., 2023a]	Globotruncana and Globotruncanita	2D Optical Images	Transfer Learning / ResNet50, Xception, InceptionV3, VGG16, MobileNet	Genus-level classification with 99.78% accuracy; species-level 81.19%
Ozer et al., 2024 [Ozer et al., 2024]	Globotruncanita (3 species)	Optical Microscopy	Transfer Learning / VGG16, ResNet50, InceptionV3, MobileNetV2, Xception	Accuracy of 96.97%; AUC of 0.978
Carvalho et al., 2020 [Carvalho et al., 2020]	Carbonate rock microfossils	MicroCT Images	Semantic Segmentation / U-Net + ResNet (18, 34, 50, 101)	Automatic identification of microfossils in depositional context
Itaki et al., 2020 [Itaki et al., 2020a,b]	Radiolarians (siliceous)	Transmitted Light (CCD, 2M and 5M pixels)	Automated Identification / CNN (RAPID ML, Cdv%v2 and Cdv%v6R)	Classification of 5 classes; manual count correlation $r = 0.972$
Gorur et al., 2023 [Gorur et al., 2023]	Foraminifera (Globotruncana)	Optical Microscopy (160×160 px)	Supervised Classification / SVM, LDA, kNN, Custom CNN	Low-cost classification of 3 species: G. arca, G. linneiana, G. ventricosa

Table 2: Summary of species used in the study, including the number of 3D models and slice counts for training, validation, and testing. Note that we use the term 3D model to refer to a 3D micro-CT image stack.

Species	3D Models	Training	Validation	Test	Total Slices
Chrysalidina	16	3,782	1,079	5,111	9,972
Ataxophragmium	7	3,943	993	4,898	9,834
Baculogypsina	7	2,243	997	2,980	6,220
Minoxia	6	4,156	1,034	3,557	8,747
Elphidiella	6	4,048	1,012	3,385	8,445
Fallotia	6	4,102	987	5,009	10,098
Arumella	5	3,597	1,143	2,163	6,903
Lockhartia	5	3,899	1,234	5,359	10,492
Orbitoides	5	3,705	1,074	5,015	9,794
Rhapydionina	5	3,743	1,186	3,802	8,731
Alveolina	4	3,535	1,724	5,129	10,388
Coskinolina	4	3,350	1,583	5,060	9,993
TOTAL	76	44,103	14,046	51,468	109,617

Table 3: Dataset balance metrics across training, validation, and test splits. The coefficient of variation (CV) indicates the relative variability in class distribution.

Split	Total Images	CV (%)	Balance Quality
Training	44,103	13.91	Excellent
Validation	14,046	20.58	Good
Test	51,468	24.72	Acceptable

Table 4: Performance metrics for *Baculogypsina* and *Orbitoides* across ensemble methods.

Ensemble Method	F1-score	Precision	Recall
<i>Baculogypsina</i>			
Patched Ensemble	0.752	0.999	0.603
Top-2	0.389	0.972	0.243
Top-3	0.366	0.957	0.227
Top-4	0.588	0.985	0.419
<i>Orbitoides</i>			
Patched Ensemble	0.872	0.827	0.922
Top-2	0.777	0.684	0.900
Top-3	0.770	0.666	0.913
Top-4	0.804	0.723	0.905

Table 5: F1-scores for all species across ensemble methods.

Species	Patched	Top-2	Top-3	Top-4
Alveolina	0.973	0.970	0.974	0.968
Arumella	0.987	0.968	0.983	0.987
Ataxophragmium	0.992	0.991	0.995	0.994
Baculogypsina	0.752	0.389	0.366	0.588
Chrysalidina	0.958	0.952	0.974	0.974
Coskinolina	0.982	0.950	0.963	0.965
Elphidiella	0.976	0.951	0.969	0.973
Fallotia	0.997	0.998	0.993	0.996
Lockhartia	0.984	0.984	0.995	0.998
Minoxia	0.941	0.972	0.989	0.985
Orbitoides	0.872	0.778	0.770	0.804
Rhapydionina	0.983	0.977	0.972	0.977
Average	0.950	0.907	0.912	0.934

Table 6: Precision values for all species across ensemble methods.

Species	Patched	Top-2	Top-3	Top-4
Alveolina	0.950	0.946	0.952	0.938
Arumella	0.993	0.997	0.999	0.999
Ataxophragmium	0.996	0.992	0.996	0.996
Baculogypsina	0.999	0.972	0.957	0.985
Chrysalidina	0.920	0.915	0.951	0.950
Coskinolina	0.979	0.975	0.976	0.982
Elphidiella	0.956	0.910	0.943	0.950
Fallotia	0.994	0.997	0.987	0.994
Lockhartia	0.968	0.970	0.990	0.996
Minoxia	0.992	0.973	0.987	0.988
Orbitoides	0.827	0.684	0.666	0.723
Rhapydionina	0.990	0.995	0.996	0.991
Average	0.964	0.944	0.950	0.958

Table 7: Recall values for all species across ensemble methods.

Species	Patched	Top-2	Top-3	Top-4
Alveolina	0.997	0.994	0.998	0.999
Arumella	0.981	0.941	0.968	0.976
Ataxophragmium	0.989	0.989	0.994	0.992
Baculogypsina	0.603	0.243	0.226	0.419
Chrysalidina	0.999	0.993	0.997	0.998
Coskinolina	0.986	0.927	0.950	0.949
Elphidiella	0.996	0.995	0.996	0.998
Fallotia	0.999	1.000	0.999	0.999
Lockhartia	1.000	0.999	1.000	1.000
Minoxia	0.895	0.970	0.990	0.982
Orbitoides	0.922	0.900	0.913	0.905
Rhapydionina	0.976	0.960	0.949	0.963
Average	0.945	0.909	0.915	0.932

Table 8: Comprehensive performance comparison across all evaluated models.

Model	Accuracy	Precision	Recall	F1-Score	Top-3 Acc	AUC
Ensemble (Final)	0.956	0.964	0.945	0.950	0.996	0.998
ConvNeXt-Large	0.951	0.962	0.937	0.941	0.996	0.998
NASNet	0.937	0.948	0.922	0.925	0.992	0.996
EfficientNetV2-L	0.935	0.947	0.919	0.920	0.994	0.996
EfficientNetV2-S	0.918	0.926	0.904	0.906	0.995	0.996
ConvNeXt	0.913	0.932	0.891	0.877	0.984	0.993
MobileNet	0.880	0.894	0.866	0.860	0.994	0.995
ResNet101V2	0.843	0.870	0.816	0.808	0.974	0.984