Using Deep Learning to Identify Fossils of the Atlantic Coastal Plain

Abstract

Fossils of Eastern North America are notable for their abundance despite their relatively fragmentary nature compared to similar-aged fossils found elsewhere in the world. Those of the Atlantic Coastal Plain are frequently collected by avocational palaeontologists, yet few public resources exist to assist both professionals and amateurs in the independent identification of these often-sparse remains. While several publicly available tools using deep learning image classification techniques have recently emerged, very few have been developed specifically for fossilized material or publicly released outside research circles. This project will assess whether deep learning-based image classification can reliably identify fossilized material from fossil-bearing strata across the Cretaceous–Paleogene boundary of the Mid-Atlantic, and whether such methods are suitable for heterogeneous, community-contributed datasets. Publicly available Python libraries will be used to construct an image classifier trained on fossil image data of varying quality and quantity. An accompanying image modification toolkit will be developed in parallel to automatically augment the size of the training dataset while introducing realistic noise. Variations in the classifier's accuracy when identifying known fossils will be used to evaluate the impact of data quality, dataset size, and augmentation techniques on model training. From this, a pipeline optimized for fossil identification under a growing, dynamic training set will be derived. Project results are intended to support the development of a publicly accessible image classifier that can incorporate community-contributed data and serve as a practical tool for fossil collectors across the region.

Background: Fragmentary Fossils of the Eastern United States

Paleontology and natural history have been an omnipresent science in the Eastern United States since the country's founding. Thomas Jefferson famously studied fossilized bones from Virginia, believing that mammoth and ground sloth remains he described belonged to animals still roaming the continent's interior (Jefferson, 1799, p. 252). Just sixteen years after Richard Owen coined the term *Dinosauria* in 1842, Joseph Leidy described *Hadrosaurus foulkii* from fossils found in Haddonfield, New Jersey. The famous "bone wars" of the late 19th century, fought between Edward Drinker Cope and Othniel Charles Marsh, were first waged among bones recovered from southern New Jersey marl pits (Gallagher, 1997, p. 35-38). Cope and Marsh were ultimately drawn westward as the Late Mesozoic record of Eastern North America, preserved mainly in the marine deposits of the Atlantic Coastal Plain, held fewer charismatic dinosaurs than the richer riverine deposits out west (Schwimmer, 1997). Nearshore marine environments are prime areas for fossilization but, with the ever-present threat of storms and wave action, are particularly hostile to the preservation of articulated specimens more commonly found in terrestrial environments (Boessenecker et al., 2014). The resulting fragmentary and isolated nature of East Coast fossils, combined with widespread suburbanization and a decline in large-scale earthworks across the region, left Eastern North America comparatively understudied over the ensuing century.

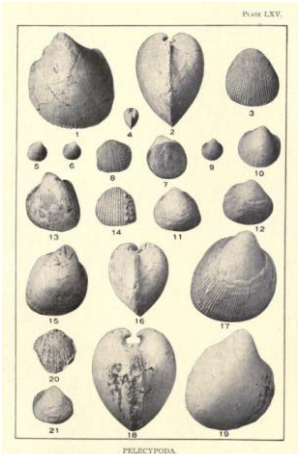

Figure 1. The Atlantic Coastal Plain still preserves a rich diversity of marine fossils. Shown here are Late Cretaceous bivalves (*Pelecypoda*) (Weller, 1907, p. 1012-1013).

This 'hiatus' overlapped with rapid technical innovation in other branches of biology. In particular, the rise of computing and bioinformatics in the late 20th and early 21st centuries have significantly expanded the scientific toolset available to researchers. Machine learning, a form of algorithmic, predictive modelling and a rapidly evolving branch of computer science, has demonstrated notable success in fields such as genomics, species identification, and morphometric classification (Christin et al., 2019;

Libbrecht & Noble, 2015; Pillay et al., 2024). Encouraged by these advances, a handful of studies have begun applying machine learning methods to paleontology. For example, a broad study by Liu et al. (2022) achieved approximately 90% clade-level accuracy in classifying fossil images using a dataset of over 415,000 specimens. More constrained analyses have also been proven, with Hendrickx et al. (2023) successfully distinguishing unique theropod tooth morphologies from high-resolution images of just 36 isolated Moroccan dinosaur teeth. Building on these efforts, this project will follow a similar trajectory by targeting distinct and recurring genera across the fossil-bearing rock units of the Northeast. If automated image classification is effective on an evolving, heterogeneous image dataset, it may yet prove that the more fragmentary fossil assemblages of the Atlantic Coastal Plain are still ideal grounds for more focused regional study.

Significance: Serving the Revived Interest in Eastern Paleontology

Interest in the fossils of Eastern North America is once again growing. The rise of social media has enabled like-minded amateurs and professionals to cultivate online communities and share images of specimens they've collected *in situ* despite limited institutional attention to the region. Many of these individuals collect across historically significant fossil sites in the Mid-Atlantic, ranging from Cretaceous brooks in New Jersey to beach-side cliff exposures in Maryland and fossiliferous marshes of the Carolinas. Academic interest across the Atlantic Coastal Plain has grown in recent years as well. In Maryland, Dinosaur Park, an open dig site excavating the Arundel Formation and managed by the Prince George's County Department of Recreation, accepts volunteers and continues to reveal significant dinosaur material from the East, including a recently described sub-adult *Acrocanthosaurus*, a T. rex-sized theropod (Carrano, 2024). A similar arrangement exists at the newly opened Jean and Ric Edelman Fossil Park and Museum in Mantua, New Jersey where visitors can register for public digs into the upper Hornerstown Formation. Situated atop the former Inversand Company marl quarry, this site is well known for its diverse invertebrate and vertebrate assemblage documenting species transitions during the Cretaceous-Paleogene mass extinction (Boles et al., 2024). Yet despite the rise of guided digs and community engagement, fossil identification remains largely dependent on individuals navigating a patchwork of online resources or consulting experts directly. A publicly available image classifier, capable of training on a dynamic set of images contributed by both professionals and amateurs, could help close this gap. Such a tool would offer immediate support for fossil identification, improving both fieldwork and post-collection analysis. It could not only provide timely suggestions for fossil identification but also help flag potentially rare or scientifically significant specimens that might otherwise go unrecognized due to limited experience or incomplete preservation.

Project Design: Using an Expanding Image Set to Train a Deep Learning Image Model

The intent of this project is to evaluate how classification accuracy of fossil image data responds to variations in dataset volume, artificial augmentation and noise, class imbalance, and the number of taxonomic classes represented. Fossils will be drawn from geological formations spanning roughly 70 million years, from the Campanian epoch of the Late Cretaceous to the Miocene epoch of the Neogene, an interval coinciding with much of the primary marine-shelf depositional window in the Mid-Atlantic Coastal Plain (Ator, 2005, pp. 40-41). Building a robust image dataset is critical to classifier performance and will remain a continuous focus throughout the project. Class imbalances in the dataset may be due to shelly invertebrates, such as bivalves and brachiopods, and vertebrate teeth, which occupy a comparatively large portion of the dataset. This is primarily driven by chondrichthyans like sharks, owing to

Figure 2. Fossil images to be artificially manipulated and cropped for model training. Clockwise from top right: *Exogyra costata*, *Belemnitella americana*, *Scapanorhynchus texanus*, and *Squalicorax pristodontis*. All images taken by the author.

their lifelong tooth replacement and abundance in shallow marine habitats (Boles et al., 2024; Höltke et al., 2024).

Image acquisition will draw from multiple sources, with additional taxonomic classes added continuously. Initial testing will rely on public and private image sets, including Wikimedia Commons and digitized specimens from the collections of the Yale Peabody Museum and the Edelman Fossil Park and Museum. The author's personal collection of confidently identified taxa will contribute substantially, alongside curated submissions from amateur and professional collectors where identification confidence is high. Attribution will be required for all images and will be hard-coded into the model architecture.

Three primary tools will be scripted in Python. The first will be an image augmentation script designed to expand and diversify the dataset by applying transformations and introducing artificial noise. The second is an image classification script based on deep learning ResNet-18 architecture, capable of training on both original and augmented fossil images. Deep learning is a type of machine learning algorithm capable of automatically extracting important features and structures across multiple processing layers, often using labelled data to perform tasks such as image classification (LeCun et al., 2015). ResNet-18, a deep learning convolutional neural network, has demonstrated strong performance in image classification of fossils in carbonate rock, offering high accuracy with relatively low computational cost; this makes ResNet-18 well-suited for projects with limited training data (Tao et al., 2024). The third tool will apply the trained model to a curated test set of known specimens, enabling quantitative evaluation of performance across average accuracy, class count, and dataset size. Tools two and three will be implemented using the PyTorch torchvision library and all generated code and training set architecture will be shared internally via GitHub.

Dataset Size	Classes	Augmentation Applied	Class Balance	Avg. Accuracy
250 images	8	No	Balanced	68%
500 images	8	Yes	Balanced	72%
1000 images	14	Yes	Balanced	81%
1000 images	14	Yes	Imbalanced	79%
1500 images	19	Yes	Imbalanced	83%

Figure 3. Example dataset illustrating the expected effects of dataset size, class count, class balance, and augmentation on classification accuracy. Classification accuracy generally increases with dataset size, with variation influenced by augmentation and class distribution. This set is illustrative and not drawn from real data.

Expected Outcomes: A Working Model Prepared for a Citizen Science Application

Several outcomes are anticipated from an image classifier trained on a robust dataset. First, the classifier is expected to accurately distinguish between major clades (e.g., a tooth from the shark *Cretalamna* will not be mistaken for the shell of an extinct oyster like *Exogyra*). Second, classification accuracy is expected to improve with training set size; that is, accuracy should increase as the number of base images in each class grows, including when an augmentation pipeline is applied. Third, the classifier should be sufficiently robust to approximate the identity of taxa not explicitly present in the training set (e.g., a tooth from *Hexanchus* should still be aligned with other shark genera even if it is absent from the training set). Fourth, and most critically, classification accuracy can remain stable or improve with the continuous introduction of additional taxonomic classes, provided the model is retrained appropriately and class imbalance is managed.

If these hypotheses hold, the result will be a strong foundation for a functional bioinformatics toolkit. The model may even be adapted for broader use beyond the initial scope of this project. iNaturalist, a widely used citizen science platform, incorporates a similar machine learning-driven image classifier to assist users in identifying species from submitted observations. According to iNaturalist (2022), they maintain a continuously updated computer vision model that is periodically retrained on millions of community-contributed observations to automate species suggestions. A comparable tool could be developed from the scripts produced here, with the eventual goal of building a public-facing application to support fossil identification in an accessible, user-friendly format.

Bibliography

Ator, S. W. (2005). *A Surficial Hydrogeologic Framework for the Mid-Atlantic Coastal Plain* (No. 1680). U.S. Department of the Interior, U.S. Geological Survey.

Boessenecker, R. W., Perry, F. A., & Schmitt, J. G. (2014). Comparative taphonomy, taphofacies, and bonebeds of the Mio-Pliocene Purisima Formation, Central California: Strong physical control on marine vertebrate preservation in shallow marine settings. *PLOS ONE, 9*(3), e91419.

Boles, Z., Ullmann, P. V., Putnam, I., Ford, M., & Deckhut, J. T. (2024). New vertebrate microfossils expand the diversity of the chondrichthyan and actinopterygian fauna of the Maastrichtian-Danian Hornerstown Formation in New Jersey. *Acta Palaeontologica Polonica.*

Carrano, M. T. (2024). First definitive record of *Acrocanthosaurus* (Theropoda: Carcharodontosauridae) in the Lower Cretaceous of Eastern North America. *Cretaceous Research, 157*, 105814.

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution, 10*(10), 1632-1644.

Gallagher, W. B. (1997). *When Dinosaurs Roamed New Jersey*. Rutgers University Press.

Hendrickx, C., Trapman, T. H., Wills, S., Holwerda, F. M., Stein, K. H. W., Rauhut, O. W. M., Melzer, R. R., van Woensel, J., & Reumer, J. W. F. (2024). A combined approach to identify isolated theropod teeth from the Cenomanian Kem Kem Group of Morocco: Cladistic, discriminant, and machine learning analyses. *Journal of Vertebrate Paleontology, 43*(4), 2024.

Höltke, O., Maxwell, E. E., & Rasser, M. W. (2024). A review of the paleobiology of some Neogene sharks and the fossil records of extant shark species. *Diversity, 16*(3), 147.

iNaturalist. (2022, April 12). The latest computer vision model updates. *iNaturalist Blog.* https://www.inaturalist.org/blog/63931-the-latest-computer-vision-model-updates

Jefferson, T. (1799). A memoir on the discovery of certain bones of a quadruped of the clawed kind in the western parts of Virginia. *Transactions of the American Philosophical Society, 4*, 246-260.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444.

Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. *Nature Reviews Genetics, 16*(6), 321-332.

Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y., & Song, H. (2022). Automatic taxonomic identification based on the Fossil Image Dataset (> 415,000 images) and deep convolutional neural networks. *Paleobiology, 49*(1), 1-22.

Pillay, A. B., Pathmanathan, D., Dabo-Niang, S., Abu, A., & Omar, H. (2024). Functional data geometric morphometrics with machine learning for craniodental shape classification in shrews. *Scientific Reports, 14*, 15579.

Schwimmer, D. R. (1997). Late Cretaceous dinosaurs in eastern USA—A taphonomic and biogeographic model of occurrences. In D. L. Wolberg, E. Stump, & G. D. Rosenberg (Eds.), *Dinofest*

International: Proceedings of a Symposium Sponsored by Arizona State University (pp. 203-211). Academy of Natural Sciences.

Tao, Y., Bao, Z., Ma, F., Gao, D., He, Y., & Wang, F. (2024). Image recognition of carbonate fossils and abiotic particles based on deep convolutional neural network mode.

Weller, S. (1907). *A Report on the Cretaceous Paleontology of New Jersey* (Vol. 4). MacCrellish & Quigley, State Printers.