Machine Learning for Paleontology Education: Two Outside the Classroom Case Studies

Carlos Peredo
Western Governors University

Rebecca Strauch
Miami University

Zara Mendivil
California State University San Marcos

Michael Ye
Ivy Mind Consulting

Machine learning is rapidly revolutionizing the biological sciences and the geosciences. It has the potential to advance paleontology education by providing students with hands-on experience in data analysis and problem-solving. This study explores the application of machine learning techniques in two case studies: one using random forests to infer morphological measurements from fossils and another using computer vision to classify mammal skulls. The students involved in these projects demonstrated the value of hands-on, project-based learning to understand how modern computer science research techniques can be applied to questions about the fossil record. Both students successfully navigated technical challenges, acquired new skills, and gained confidence in their ability to solve problems outside of the scope of traditional classroom learning. While the models developed in these case studies are modest in scope, they serve as valuable foundational projects for students to explore paleontological research questions. These experiences provide a solid understanding of machine learning concepts and techniques, which can be applied to more complex research questions in the future. By integrating machine learning projects into educational curricula, we can enhance the preparedness of students for careers in paleontology and data science, while also contributing to the broader research enterprise.

Introduction

The fossil record serves as a window into Earth's past, lending important historical context for understanding modern Earth systems. Although it is inherently incomplete and biased, the fossil record can serve as a reliable model for reconstructing past ecosystems and forecasting future patterns (Abdelhady et al., 2024; Alroy, 2001, 2003). In recent years, the emergence of digital repositories with vast datasets has spurred a methodological revolution in paleontology (Abdelhady et al., 2024; Yu et al., 2024). Many of these techniques transcend paleontology and are truly interdisciplinary, with origins as widely distributed as physics (Puttonen et al., 2018), medicine (Padian et al., 2004), and statistics (Langley, 2011).

Beyond paleontology, machine learning is revolutionizing research efforts in the biological sciences and geosciences. Machine learning is a technique where a computer is taught to learn like humans, typically through a combination of statistics and artificial intelligence (AI). The core concept behind machine learning is that the machine is constantly learning and improving from new input data. Typically, researchers produce a model using training data, and then apply the model to forecast patterns on test data (Langley, 2011). Unlike traditional statistics, machine learning algorithms typically do not identify association between variables (Mohri, 2018). Machine learning techniques have been applied broadly in the biological sciences and geosciences in recent years (Dramsch, 2020; Ghosh & Dasgupta, 2022; Karpatne et al., 2018; Lary et al., 2016; Vanaja & Yella, 2022). Here, we will focus on two methods of machine learning: random forests and computer vision. Specifically, we will explore two case studies using machine learning as an educational tool for teaching paleontology in outside-the-classroom settings.

Computer vision is a type of artificial intelligence that aims to train computers to identify, interpret, and analyze visual data, typically in the form of photographs or videos (Szeliski, 2022). Traditionally, paleontology has relied on manual methods for identifying, classifying, and interpreting fossils. Although this process is foundational for training paleontologists, it is susceptible to human error and inherently subjective. Computer vision models have been successfully trained to recognize, categorize, and analyze fossils with efficacy comparable to that of human experts (De Baets, 2021; Ferreira-Chacua & Koeshidayatullah, 2023; Liu et al., 2023). Unlike human methods, computer vision models are inherently reproducible and scalable, making them more accessible beyond just the subject matter experts.

Random forests, sometimes called random decision forests, are popular machine learning algorithms that make predictions and classifications using a series of decision trees (Parmar et al., 2019). Random forests excel at classifying subsets of large datasets to create regression models for predictive analytics (Parmar et al., 2019; Pranckevičius & Marcinkevičius, 2017; Svetnik et al., 2003). Because they excel at classification and prediction, random forests can be readily applied to fossils (Abdelhady et al., 2024; Yu et al., 2024). Given sufficiently large datasets, random forests can be used to develop models that can predict missing values for fragmentary datasets, such as the fossil record.

Here, we examine the machine learning as a tool for teaching paleontology in non-traditional, outside-the-classroom settings. We present two case studies of individualized learning through machine learning projects. The first was a summer Research Experience for Undergraduates (REU) at Miami University. The student (ZIM) created several random forest models to infer morphological measurements from fragmentary fossils. The second of these was an advanced high school research experience conducted through Ivy Mind Consulting. The student (MY) created a computer vision model to identify and classify mammal skulls into taxonomic bins. Both students were supervised by the lead author (CMP). Here, we discuss each case study and draw conclusions about the efficacy of machine learning as an educational tool for outside-the-classroom research experiences in higher education.

CASE STUDY 1

Objectives

The fossil record is inherently incomplete. Machine learning models excel at classification and prediction, even when datasets are sparse or fragmentary (Abdelhady et al., 2024; Yu et al., 2024). Traditional regression models often require large datasets, but random forest decision trees are increasingly able to uncover patterns and make predictions even when datasets are incomplete (Hapfelmeier & Ulm, 2014; Loh et al., 2019). Although the research questions vary across these studies, the end goal is the same: to uncover patterns from partial datasets that can be used to infer potential information about missing data.

Geometric morphometric analyses typically rely on several morphological measurements to compare across a study sample (Zelditch et al., 2012). However, this methodology is limited by the fact that most fossils are incomplete. Fossils are often broken or otherwise damaged by taphonomy and diagenesis (Behrensmeyer, 1978; Behrensmeyer et al., 2000; Behrensmeyer & Miller, 2012). The goal of the project in this case study was to train a random forest model using decision trees to accurately infer missing data from fossil and modern specimens. A secondary goal of the project was to develop a web application that could be operated from a mobile phone. This would make it convenient for a researcher in a museum collection to enter known measurements from a specimen and quickly predict values for missing measurements.

The project in this case study was structured to maximize the student's learning experience. This project required the student to engage in both front-end and back-end development. As a result, the student oversaw the project through the complete lifecycle, from conception to implementation.

Methods

The dataset for this case study, which consists of photographs of fossil and modern whale mandibles, was sourced by the second author (RJS). This dataset was initially sourced for a separate research project, but repurposed for this case study. The REU student digitally measured nine morphological measurements using ImageJ (Table 1). The final morphometric dataset included 63 complete specimens with all nine measurements, and 139 fragmentary specimens with only some values. This dataset spanned 124 extant and extinct whale species.

Table 1: List of measurements used in Case Study 1, with precise anatomical definitions and the view from which they were measured.

Value	Definition	View
SL	Straight anteroposterior length from rostral tip to articular process	Lateral
SY	Anteroposterior length of the mandibular symphysis	Dorsal
M1	Posterior margin of the articular process to tip of the coronoid process	Lateral
M2	Posterior margin of the articular process to posteriormost tooth	Lateral
M3	Posterior margin of the articular process to posterior end of the symphysis	Dorsal
M4	Broadest transverse width across the articular process	Dorsal

M5	Anterior tip of the rostrum to center of the 3rd tooth (if present)	Dorsal
M6	Transverse width of the mandible at the level of the 3rd tooth	Dorsal
M7	Transverse width of the mandible at the posterior end of the symphysis	Dorsal

The student initially coded the random forest model using decision trees in a local environment using Python. This approach allowed for fast iteration and provided an opportunity to learn troubleshooting in a local setting. For preliminary data exploration and visualization, the student utilized the Seaborn library. They then executed the random forest model using the SciKit-Learn library. SciKit-Learn is an open source data analysis library that is beginner friendly and commonly used for teaching entry level machine learning concepts.

Once the model was refined, the student moved to the online platform Kaggle. Kaggle is an online community platform for data science competitions which offers notebooks for coding in Python in a virtual environment. In the Kaggle environment, the student fine-tuned the model for increasingly precise binary splits, optimizing for lower mean absolute error values. Code for the final model is available as supplementary files in the Appendix.

Results

To evaluate the accuracy of the model, the student retained measurements from 30 complete whale mandibles that were not included in the training dataset. This dataset was equally distributed among taxonomic groups. The random forest model was optimized for the lowest possible mean absolute error (a measure of error between the predicted values and the actual values in the test dataset for each variable). The model was able to accurately predict linear measurements with a mean margin of error of approximately 5% of the estimated value.

Additionally, the student added code to assign an importance score to each of the nine variables. This score indicates the contribution of each variable to successfully predicting the other eight. The results indicate that M7, M3, and SL collectively comprise nearly 90% of the model's predictive power. This suggests that the other six variables are of limited utility as proxies for missing data and may not be worth collecting in future studies.

Once the model was finalized, the student built a web application for users to interact with it. The application was built and hosted on the web platform HuggingFace: an open source platform for hosting and testing generative AI models. The web application built by the student is mobile friendly for convenient use in museum collections. The user inputs given morphological measurements taken from a physical specimen and then receives as an output a prediction of the missing values, as well as a margin of error. The web application was built primarily in Python, with minimal coding in JavaScript to design the user interface.

CASE STUDY 2

Objectives

Some of the earliest attempts to use machine learning for paleontological research involve the identification and analysis of images (Liu et al., 2023; Puttonen et al., 2018). The specific objectives vary from analyzing morphology to taxonomic classification, and the images in question range from photographs or videos to SEM or CT scans. However, the end goal is the same: to train a computer to recognize paleontological images with near-human accuracy so that the process can be automated (Szeliski, 2022; Yu et al., 2024).

Although taxonomic identification is crucial for research purposes, it is also critical for curation and management of voucher specimens in research repositories. The goal of the project in this case study was to train a computer vision model that could reliably categorize photographs of mammal skulls into their respective mammalian Order. A secondary goal of the project was to develop a web application that could be operated from a mobile phone, making identification and curation of skulls quick and convenient without sophisticated software.

As in the first case study, this project was designed to maximize the student’s learning experience. The project required the student to engage in both front-end and back-end development. This structure ensured that the student oversaw the project through the complete lifecycle, from conception to implementation.

Methods

The dataset for this case study was sourced by the lead author and PI (CMP). It is comprised of photographs of mammal skulls from the Smithsonian National Museum of Natural History (USNM), University of

Michigan Museum of Zoology (UMMZ), and the Hefner Museum of Natural History (MU). The full dataset included 738 artiodactyls, 1134 terrestrial carnivorans, 240 bats, 166 lagomorphs, 2006 pinnipeds, and 671 rodents. The photographs include a range of different views (dorsal, ventral, lateral, and oblique) and further differ in the presence or absence of the mandible. These differences were preserved to mimic real conditions under which a user may attempt to use the model.

The student initially created a prototype model in a local environment with Python using only a subset of the data set. Working in a local environment allowed for faster iteration and provided valuable troubleshooting experience for the student. The student built this prototype using the fast.ai library with ResNet18 as the base model. The ResNet models are deep learning image recognition models of varying complexity; ResNet18 is a low-resource model that is ideal for fast iteration of preliminary code.

Once the prototype model was operational and ready to scale up to the full dataset, the student moved to the online platform Kaggle. Among its many features are notebooks for coding in Python in a virtual environment, which use machines hosted by Kaggle. These machines had faster GPUs that could more efficiently process both the complete dataset and a more complex ResNet base model. Code for the final model, which uses ResNet152 as its base, is available as supplementary files in the Appendix.

Results

To evaluate the accuracy of the model, we retained 60 photographs of mammal skulls that were not included in the training dataset. This test dataset was equally distributed across the six orders included in the training dataset. The final model trained on the ResNet152 base correctly assigned all 60 photographs to the appropriate order, and never dipped below 90% certainty. These scores are comparable to human experts, and more than sufficient for both research and curatorial purposes.

Once the model was fully trained and tested, the student built a web application with a minimal user interface for interacting with the model. This web application was built and hosted on the platform HuggingFace. This web application is mobile phone compatible, allowing for easy access to a camera. The user inputs a photograph of a skull and receives a prediction of the mammalian order to which the skull belongs.

The web application was built primarily in Python, with minimal code in JavaScript to design the user interface.

Discussion

In this study, we examine the application of machine learning techniques as teaching tools for teaching paleontology in outside-the-classroom settings. We present two distinct case studies—one focusing on random forests for inferring morphological measurements, and the other on computer vision for classifying mammal skulls. Our case studies demonstrate the value of hands-on, project-based educational experiences for students interested in paleontology. Here, we examine the educational outcomes derived from these projects, highlighting how hands-on engagement with machine learning can foster a deeper understanding of the subject matter and develop key coding and research skills in junior researchers. By comparing the two case studies, we draw broader conclusions about the role of machine learning in science education.

The first case study was executed by a summer REU student (ZIM) at Miami University. The student acquired technical skills in Python and gained a strong understanding of how random forests operate, particularly in the context of handling incomplete datasets. The student faced several important challenges, such as troubleshooting errors during model development and refining the model for accuracy. Overcoming these challenges was a critical part of developing the student's problem solving abilities for future research endeavors. By designing a prototype web application, the student was able to successfully see a project through from conception to implementation. In addition to its value as a research experience, this work serves as a critical portfolio piece for students interested in data science, software development, or project management careers.

The second case study was conducted by a high school student (MY) through an advanced research program at Ivy Mind Consulting. During this project, the student developed technical proficiency in Python and gained a deep understanding of computer vision techniques, specifically using the fast.ai library. The student encountered and overcame challenges related to training and scaling deep learning models, which significantly contributed to their ability to troubleshoot and optimize complex algorithms. By building a prototype model and eventually scaling it to handle a full dataset, the student demonstrated strong project management skills and the ability to iteratively improve their work. The final product, a web application for mammal skull classification, not only served as a capstone to their learning experience but also provided a tangible asset for their portfolio. This research experience is invaluable for a researcher at such a junior stage, as their career trajectory is still undefined. The capstone product resulting from this project will be instrumental in securing a future in data science, machine learning, or software engineering.

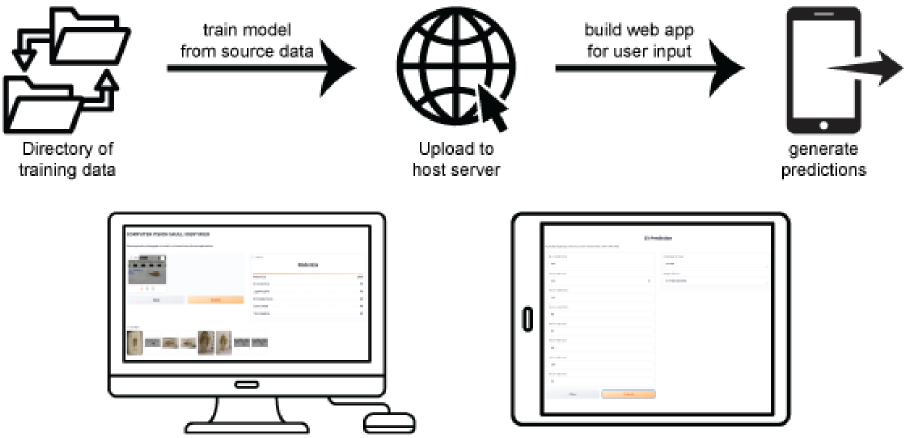

Figure 1: *Generalized procedure for both machine learning case studies. Each project began with a directory of training data, which was used to train the model, and was then uploaded to a host serve (www.kaggle.com). Once the model was trained, the student built a web application with a user interface and hosted the web application on an online platform (www.huggingface.com). The end result were simple web applications that users can provide an input to in order to query the model for predictions*.

Parallels across both case studies highlight the value of key educational objectives (Figure 1). Both students developed foundational skills in Python and gained hands on experience with advanced machine learning techniques. Both students initially built prototype models in local environments where they could iterate quickly. Similarly, both students faced challenges in scaling up the model to full datasets and subsequently optimizing the model for minimal error. By designing both case studies with similar approaches, the students were able to benefit from discussing their projects with one another and helping each other problem solve. Both students successfully managed their project from conception to implementation, resulting in tangible products for them to use as resume and portfolio pieces.

Both students self-reported high levels of satisfaction with their learning experience. Despite having limited prior exposure to Python and AI projects, they were able to successfully overcome technical challenges. Both students emphasized the importance of the mentorship role for answering technical questions, providing foundational reading, and encouragement. Both students reported elevated levels of confidence in their capability to handle complex, technical, projects and their capacity to problem solve through challenges.

These case studies underscore the broader implications of integrating machine learning into paleontology education. By engaging in hands-on, project-based learning, students not only develop technical skills but also gain confidence in applying these skills to real-world problems. The iterative nature of the projects, from prototype development to full-scale implementation, mirrors the workflow of professional research, providing students with a realistic understanding of the scientific process. Moreover, the collaborative element—where the students discussed challenges and solutions with each other—highlights the importance of peer learning in scientific education. These experiences suggest that integrating machine learning projects into educational curricula can significantly enhance the preparedness of students for careers in paleontology and more broadly in data science, software development, and experimental design.

Our results underscore the effectiveness of machine learning as a teaching tool in paleontology, particularly in non-traditional, outside the classroom settings. Both students successfully navigated the complexities of their research projects, from initial concept to product launch. Both students successfully built prototype models that could serve as the foundation for more sophisticated, research-applicable models in the future. In doing so, both students gained valuable professional development for their early career stage. As the field of paleontology continues to evolve, incorporating technological approaches into educational practices will be crucial for preparing the next generation of researchers. Future initiatives should continue to explore and expand the role of machine learning in both formal and informal educational settings.

Finally, we would like to acknowledge the capacity by these early researchers to contribute broadly to the research enterprise. The models developed in these case studies are simple in scope and ambition, and do not directly address research questions. Nevertheless, they can serve as foundational projects to inspire and guide future research endeavors. Engaging early career scientists in these types of projects will build a foundation of project ideas and solutions that can inspire research-quality projects down the line. In this way, even early career scientists can contribute to the broader research enterprise.

Author Contributions

CMP and RJS conceived and executed the project. They designed both case studies, prepared, wrote, and revised the manuscript. ZIM conducted the case study 1: building the random forest model to infer missing data. MY performed case study 2: building the computer vision model to categorize skulls. CMP supervised, mentored, and assisted both students with their projects.

Acknowledgements

The REU program at Miami University was funded by the National Science Foundation (NSF Award #0754991). We would like to thank Michelle Boone and Dave Berg, who served as program coordinators for the REU experience. We would also like to thank Betty Wang, of IvyMind Consulting, who served as the program coordinator for the advanced high school research program. We are grateful for Jeremy Howard and the team at fast.ai for their efforts building robust, open source educational content for teaching machine learning. Finally, we would like to thank the Student Success Analytics team at Western Governors University, for their encouragement and support of this research.

References

Abdelhady, A. A., Seuss, B., Jain, S., Abdel-Raheem, K. H., Elsheikh, A., Ahmed, M. S., Elewa, A. M., & Hussain, A. M. (2024). New and emerging technologies in paleontology and paleobiology: A horizon scanning review. *Journal of African Earth Sciences*, 105155.

Alroy, J. (2001). A multispecies overkill simulation of the end-Pleistocene megafaunal mass extinction. *Science*, *292*(5523), 1893-1896.

Alroy, J. (2003). Global databases will yield reliable measures of global biodiversity. *Paleobiology*, *29*, 26-29.

Behrensmeyer, A. K. (1978). Taphonomic and ecologic information from bone weathering. *Paleobiology*, *4*(2), 150-162.

Behrensmeyer, A. K., Kidwell, S. M., & Gastaldo, R. A. (2000). Taphonomy and paleobiology. *Paleobiology*, *26*(S4), 103-147.

Behrensmeyer, A. K., & Miller, J. H. (2012). Building links between ecology and paleontology using taphonomic studies of recent vertebrate communities. *Paleontology in ecology and conservation*, 69-91.

De Baets, K. (2021). Performance of machine-learning approaches in identifying ammonoid species based on conch properties. *Peer Community in Paleontology*, *1*, 100010.

Dramsch, J. S. (2020). 70 years of machine learning in geoscience in review. *Advances in geophysics*, *61*, 1-55.

Ferreira-Chacua, I., & Koeshidayatullah, A. I. (2023). ForamViT-GAN: exploring new paradigms in deep learning for micropaleontological image analysis. *IEEE Access*, *11*, 67298-67307.

Ghosh, S., & Dasgupta, R. (2022). *Machine Learning in Biological Sciences*. Springer.

Hapfelmeier, A., & Ulm, K. (2014). Variable selection by Random Forests using data with missing values. *Computational Statistics & Data Analysis*, *80*, 129-139.

Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H. A., & Kumar, V. (2018). Machine learning for the geosciences: Challenges and opportunities. *IEEE Transactions on Knowledge and Data Engineering*, *31*(8), 1544-1554.

Langley, P. (2011). The changing science of machine learning. *Machine learning*, *82*(3), 275-279.

Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, *7*(1), 3-10.

Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y., & Song, H. (2023). Automatic taxonomic identification based on the Fossil Image Dataset (> 415,000 images) and deep convolutional neural networks. *Paleobiology*, *49*(1), 1-22.

Loh, W.-Y., Eltinge, J., Cho, M. J., & Li, Y. (2019). Classification and regression trees and forests for incomplete data from sample surveys. *Statistica Sinica*, *29*(1), 431-453.

Mohri, M. (2018). Foundations of machine learning. MIT press.

Padian, K., Horner, J. R., & De Ricqlès, A. (2004). Growth in small dinosaurs and pterosaurs: the evolution of archosaurian growth strategies. *Journal of Vertebrate Paleontology*, *24*(3), 555-571.

Parmar, A., Katariya, R., & Patel, V. (2019). A review on random forest: An ensemble classifier. (Ed.),^(Eds.). International conference on intelligent data communication technologies and internet of things (ICICI) 2018.

Pranckevičius, T., & Marcinkevičius, V. (2017). Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic Journal of Modern Computing*, *5*(2), 221.

Puttonen, A., Harzhauser, M., Puttonen, E., Mandic, O., Szekely, B., Molnar, G., & Pfeifer, N. (2018). Automatic determination of 3D orientations of fossilized oyster shells from a densely packed Miocene shell bed. *Int J Earth Sci*, *107*(6), 2125-2142. https://doi.org/10.1007/s00531-018-1591-0

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random forest: a classification and regression tool for compound classification and QSAR modeling. *Journal of chemical information and computer sciences*, *43*(6), 1947-1958.

Szeliski, R. (2022). *Computer vision: algorithms and applications*. Springer Nature.

Vanaja, A., & Yella, V. R. (2022). Evolution of machine learning in biosciences: A bibliometric network analysis. *Journal of Applied Biology & Biotechnology. Published online July*, *20*, 45-51.

Yu, C., Qin, F., Watanabe, A., Yao, W., Li, Y., Qin, Z., Liu, Y., Wang, H., Jiangzuo, Q., Hsiang, A. Y., Ma, C., Rayfield, E., Benton, M. J., & Xu, X. (2024). Artificial intelligence in paleontology. *Earth-Science Reviews*, *252*. https://doi.org/10.1016/j.earscirev.2024.104765

Zelditch, M., Swiderski, D., & Sheets, H. D. (2012). *Geometric morphometrics for biologists: a primer*. academic press.