*Quest Journals*
*Journal of Software Engineering and Simulation*
*Volume 11 ~ Issue 3 (2025) pp: 01-06*
*ISSN(Online) :2321-3795 ISSN (Print):2321-3809*
www.questjournals.org

**Research Paper**

# Machine Learning in Paleontology: Automating Fossil Identification and Analysis

Michael Ye
*Webb Schools, 1175 W Baseline Rd, Claremont, CA 91711*

***Abstract***
*Machine learning (ML) is revolutionizing paleontology by addressing longstanding challenges in fossil identification and analysis. This study explores the application of ML techniques, particularly computer vision, in automating the classification of mammalian skulls. Using a meticulously curated dataset sourced from renowned natural history institutions, an ML model was trained to identify six mammalian orders with a high degree of accuracy. The tool, hosted on the Huggingface platform, simplifies skull identification through a user-friendly interface, enabling rapid and reliable classification. While constrained by computational resources and limited training scope, the model demonstrates immense potential for improving the efficiency and objectivity of paleontological research. Future developments aim to expand the dataset to encompass all 29 mammalian orders, integrate confidence scoring, and enhance computational capabilities. These advancements promise to make ML an indispensable asset in the study of evolutionary relationships and biodiversity. The findings underline the transformative role of ML in paleontology, fostering new insights and expediting research in a field traditionally reliant on manual methods.*

***Key words***
*Machine Learning; Paleontology; Fossil Identification; Computer Vision; Evolutionary Biology Biodiversity Analysis*

***R**eceived 27 Feb., 2024; **R**evised 04 Mar., 2025; **A**ccepted 06 Mar., 2025* *© The author(s) 2025.
**P**ublished with open access at **www.questjournas.org***

## I. Introduction

Machine Learning (ML) has emerged as a transformative technology in paleontology, offering researchers powerful tools to analyze large, complex datasets that reveal the intricate and ancient history of life on Earth. Traditionally, paleontology has relied heavily on manual methods for identifying, classifying, and interpreting fossils—a process that, while foundational, is not only time-consuming but also prone to human error and subjectivity. The introduction of ML into paleontological research is revolutionizing the field, enabling scientists to process and interpret vast amounts of data with unprecedented speed, precision, and accuracy. This paradigm shift is not just a technological upgrade; it is a fundamental change in how paleontologists uncover and understand patterns and relationships that were previously hidden within the data.

One of the primary and most impactful applications of ML in paleontology is the classification of fossils. By training ML models on extensive datasets of fossil images, researchers can develop sophisticated algorithms capable of automatically identifying and categorizing fossils based on subtle features that may be challenging for even the most experienced human experts to discern. For instance, ML has been utilized to classify ape species by analyzing their dental structures, which are often complex and highly variable. The ability of ML to detect these nuanced differences and make accurate classifications underscores its potential to significantly enhance the precision, reliability, and efficiency of paleontological research. This capability is particularly valuable in a field where the accurate identification of fossils is critical to constructing reliable evolutionary timelines and understanding biodiversity across geological epochs.

While ML is a subset of Artificial Intelligence (AI), it is distinct in its approach and methodology. Specifically, ML involves training AI models using large datasets, enabling the model to "learn" from the data in a manner that mirrors human learning processes. This training involves feeding the model with vast amounts of data, allowing it to identify patterns, correlations, and relationships within the information. Once the model has been adequately trained, it can make predictions or classifications on new, previously unseen data based on the patterns it has learned. This process is particularly advantageous in paleontology, where ML can be applied to a variety of tasks, including species identification, pattern recognition, and complex data analysis. These applications are not only accelerating research but also opening new avenues for discoveries that were previously inaccessible due to the limitations of traditional methods.

A specific and highly effective application of ML in paleontology is computer vision, a branch of ML that uses images as the primary data input for training AI models. Computer vision enables the automatic identification and analysis of images by recognizing and interpreting patterns within them. This technology has been employed in numerous paleontological studies to analyze fossil images, detect morphological features, and classify species with remarkable accuracy. For example, in a recent project, a computer vision model was developed to classify mammal skulls based on their morphological characteristics. This model, implemented as a web application with a user-friendly interface, allows users to upload images of mammal skulls, which the AI model then categorizes into one of six mammalian orders: Artiodactyla, Carnivora, Chiroptera, Lagomorpha, Pinnipedia, and Rodentia. The code for this model was written in Python, utilizing libraries such as pathlib and fastai.vision to build and deploy the AI model. The primary goal of this project was to create a tool that could accurately and efficiently identify the order to which a mammalian skull belongs, thereby streamlining the process of cataloging and studying these fossils. This application is not just a proof of concept but a practical tool that has the potential to significantly impact how paleontological research is conducted.

The integration of ML into paleontology is not merely a recent innovation; it has already been applied in various studies with remarkable success. For example, a study conducted by Abdelhady et al. demonstrated that machine learning models could achieve an accuracy rate of 85% in identifying taxa from paleontological images. This level of accuracy is comparable to that of human experts, highlighting the potential of ML to complement, and in some cases, surpass traditional paleontological methods. Another notable study by De Baets in 2021 utilized ML to identify patterns in ammonoid species based on their conch shapes. By analyzing these morphological patterns, the ML model was able to classify different ammonoid species with a high degree of accuracy, providing insights that may have been overlooked due to human biases or limitations. These examples illustrate how ML can not only match human performance but also uncover new patterns and relationships that were previously inaccessible, offering a more comprehensive understanding of the fossil record.

However, the effectiveness of ML models in paleontology, as in other scientific disciplines, is heavily dependent on the quality and quantity of the training data. High-quality, well-curated datasets are essential for training accurate and reliable models. If the training data contains errors, inconsistencies, or biases, the model's predictions are likely to be flawed. Additionally, the size of the training dataset is crucial; larger datasets generally allow for more robust and generalizable models. This aspect is particularly important in paleontology, where fossil data can be sparse, fragmented, or incomplete, making it challenging to gather comprehensive datasets for training ML models. Nevertheless, as more fossil data is digitized and made available for research, the potential for ML in paleontology continues to grow, offering new possibilities for the analysis and interpretation of the fossil record.

Beyond ammonites and other invertebrates, ML has also been successfully applied to the study of mammalian fossils. In a 2024 study by Dominguez-Rodrigo et al., researchers used computer vision to analyze tooth marks left by African carnivores on bones. The algorithm developed in this study was capable of distinguishing between tooth marks made by different carnivore species with high accuracy, thereby reducing the subjectivity that often accompanies such analyses. This application of ML not only enhances the accuracy of fossil analysis but also provides a more objective and systematic approach to studying predator-prey interactions in the fossil record. Such applications demonstrate the versatility of ML in paleontology, capable of addressing a wide range of research questions and improving the reliability of paleontological interpretations.

ML has also proven effective in the study of smaller, less conspicuous fossils. In a 2020 study by Xu et al., researchers developed a computer vision algorithm to rapidly and accurately identify microscopic palaeobios. The study utilized convolutional neural networks (CNNs), a type of ML model particularly well-suited for image analysis, to classify images of palaeobios. The model was trained on a large dataset of images and then validated on a separate set of images to assess its accuracy. The results demonstrated that the ML model could identify and classify palaeobios images with high precision, significantly reducing the time required for analysis compared to traditional manual methods. This application of ML highlights its versatility and effectiveness in studying a wide range of fossil types, from large vertebrates to tiny microorganisms, thereby broadening the scope of paleontological research.

Moreover, ML models are capable of handling exceptionally large datasets, which is particularly advantageous in paleontology. A 2023 study by Liu et al. exemplified this capability by using CNNs to gdevelop an automatic taxonomic identifier for fossils. The study utilized a dataset of over 415,000 images, encompassing a wide variety of fossil taxa. The sheer size and diversity of the dataset allowed the ML model to become a comprehensive and all-purpose identifier with high accuracy. The model's performance in this study underscored its effectiveness by outperforming traditional manual classification methods, significantly accelerating the pace of research and enabling paleontologists to process large volumes of data quickly and efficiently. This advancement in ML technology has the potential to revolutionize paleontological research by automating the identification and classification of fossils on a scale that was previously unimaginable, thereby transforming the way paleontologists work.

The creation of a machine-learning model capable of identifying mammal skulls has profound implications for museums, research institutions, and educational settings. Museums, in particular, could benefit greatly from such a tool, as it would enable curators to quickly and accurately catalog skulls in their collections. Instead of relying on the slow and labor-intensive process of manual identification, curators could use the ML model to automatically identify the order of each skull, allowing for more efficient cataloging and management of collections. This would not only save time but also ensure a higher level of consistency and accuracy in the classification process. Additionally, such a tool could be used in educational settings to teach students about mammalian diversity and evolution, providing a hands-on learning experience that combines cutting-edge technology with traditional paleontological methods. By integrating ML into the educational curriculum, students would gain valuable insights into both the history of life on Earth and the technological advancements that are shaping the future of paleontological research.

As ML technology continues to advance, its applications in paleontology are likely to expand even further. Future developments could include the integration of ML models with other emerging technologies, such as 3D imaging, augmented reality (AR), and virtual reality (VR), to create immersive and interactive experiences for studying fossils. For example, researchers could use ML models to analyze 3D scans of fossils, providing detailed insights into their morphology and allowing for more accurate reconstructions of extinct species. AR and VR could be used to create virtual fossil labs where students and researchers can interact with digital fossils in a simulated environment, enhancing the learning experience and making paleontology more accessible to a broader audience.

Additionally, as more fossil data is digitized and made available for research, the potential for ML to uncover new insights into the history of life on Earth will only increase. In this context, ML represents not just a tool for paleontologists, but a new paradigm for understanding the complex and dynamic processes that have shaped life on our planet. The future of paleontology is likely to be one where ML plays a central role, driving new discoveries, transforming our understanding of the past, and paving the way for the next generation of paleontologists. With the continued integration of ML into paleontological research, we are entering a new era of discovery, where the secrets of the fossil record are revealed with greater clarity

## II. Methods

The development of the machine-learning model for identifying mammalian skulls was conducted on a 2021 iMac running Sonoma 14.5, providing a stable computational environment for this complex task. The model was initially coded on the Kaggle platform, which was chosen for its extensive tools in data management and model training. Subsequently, the model was integrated into the Huggingface Machine Learning Platform, which facilitated further refinement and deployment of the model.

The training dataset was meticulously curated using Kaggle’s dataset feature, ensuring that the data was both comprehensive and organized for optimal machine-learning performance. The images utilized in the training process were sourced from the Smithsonian Museum of Natural History and the Hefner Museum of Natural History, institutions renowned for their extensive and diverse collections. These images were initially stored in a Google Drive folder to maintain accessibility and security before being transferred to the local computing environment for further processing.

Prior to the primary coding effort, a preliminary study of computer vision principles was undertaken. This was facilitated by a video tutorial focused on basic computer vision techniques, where a simple model was constructed to differentiate between images of birds and forests. Despite limited prior experience with computer vision, the replication of this tutorial was successful, serving as an essential preparatory step for the more complex task of skull identification. This initial success provided a foundational understanding, but the skull identification model required a significantly more advanced approach and deeper technical expertise.

A major challenge encountered during the project was the management and upload of the training dataset to Kaggle. Given the necessity for large datasets in effective AI training, this task was critical. However, the process was complicated by the substantial size of the Google Drive folder containing the images, necessitating file compression to meet platform constraints. This time-intensive process highlighted the logistical difficulties inherent in managing large datasets.

After successful compression and upload of the files, an additional challenge emerged in establishing a reliable data path within the Kaggle environment. For a period of two weeks, progress was impeded by a persistent TypeError indicating that a NoneType object was not iterable. This error presented significant difficulties, as its underlying cause took time to apparent, even with expert consultation. The complexity of the error message and the lack of clear diagnostic information necessitated an iterative troubleshooting approach.

Ultimately, it was determined that the dataset's folder structure was incompatible with the model's requirements, preventing the code from correctly accessing the training images. This misalignment between the dataset organization and the model's expectations was identified as the root cause of the TypeError. Upon reorganization of the dataset to conform to the model's specifications, the error was resolved, allowing the project to proceed.

This experience underscored the critical importance of dataset management and the challenges associated with large-scale machine-learning projects. The resolution of these issues not only facilitated the successful development of the model but also provided valuable insights into the complexities of machine learning. The process highlighted the necessity of rigorous data organization and the potential pitfalls that can arise in the absence of careful planning. Through meticulous troubleshooting and iterative refinement, the project was ultimately successful, contributing to the broader understanding of machine-learning applications in biological classification.

## III. Results

The application, hosted on the Huggingface platform, represents a sophisticated integration of advanced machine learning techniques into a user-friendly tool designed specifically for identifying mammalian skulls. This tool, aptly named "skullidentifier," is crafted to be both accessible and powerful, providing users with a seamless experience that simplifies the complex process of skull identification. The application is particularly valuable for researchers, educators, and enthusiasts who require a reliable and efficient means of identifying and classifying mammalian skulls based on their morphological features.

At the core of the application is a straightforward input box where users can easily upload a picture of a skull they wish to identify. This simplicity in design ensures that users of varying technical expertise can engage with the tool without needing specialized knowledge or extensive training. Once an image is uploaded, the model rapidly processes it using state-of-the-art image recognition algorithms that have been trained on a diverse and comprehensive dataset of mammalian skull images. These algorithms are capable of detecting and analyzing intricate patterns within the image, comparing these features against the extensive training data to generate an identification.

The output generated by the model is expressed as a percentage, which reflects the degree of similarity between the uploaded image and the various mammalian orders that the model has been trained to recognize. This percentage serves as an indication of the model's confidence in its identification. Importantly, the percentages provided by the model do not always total 100%. This design choice allows the model to reflect the reality of biological diversity, where certain skull features may be shared across multiple mammalian orders. As a result, the model may suggest that the characteristics of a skull share similarities with two or more orders, acknowledging the complex and sometimes overlapping morphological traits that exist across different mammalian families.

This nuanced approach to identification is particularly useful in cases where skull features are not distinctive enough to clearly belong to a single order, thereby providing users with a more informed and comprehensive understanding of the potential classifications. For instance, a skull might exhibit features that are common to both Carnivora and Pinnipedia, and the model would reflect this by providing a percentage match for both orders. This capability is crucial for researchers who are exploring evolutionary relationships or studying species with convergent traits, as it allows for a more sophisticated analysis of the data.

Beneath the input box, the application includes a selection of sample images that users can utilize to test and explore the capabilities of the AI model. These sample images serve multiple important functions within the application. First and foremost, they offer a practical demonstration of the model's abilities, allowing users to see how the tool processes and identifies skulls from different mammalian orders. By experimenting with these examples, users can gain a clearer understanding of how the model interprets various morphological features and how it arrives at its percentage-based outputs.

In addition to demonstrating the model's functionality, these sample images also provide an educational component that is particularly valuable for students, educators, and researchers. By observing how the model handles skulls with ambiguous or overlapping features, users can deepen their understanding of the diversity and complexity of mammalian skull structures. This educational aspect is further enhanced by the ability to see how the model responds to images with mixed or partial matches, providing insights into the challenges and intricacies of skull identification in the context of evolutionary biology.

The inclusion of sample images also enhances user engagement by offering an interactive and hands-on experience. Instead of passively reading about the model's capabilities, users can actively participate in the identification process, experimenting with different images and observing the results. This interactive approach not only makes the application more engaging but also fosters a deeper connection between users and the tool, encouraging them to explore and learn more about the fascinating world of mammalian skulls.

Moreover, the application’s design could be expanded in the future to include additional features that would further enhance its utility and educational value. For example, the inclusion of detailed explanations or annotations for each identification result could provide users with context about why the model assigned certain percentages to specific orders. These explanations could highlight particular morphological traits that influenced the model's decision, offering users a deeper understanding of the identification process and the biological significance of these features.

Another potential enhancement could be the integration of a feedback system, allowing users to contribute their observations or corrections to improve the model over time. Such a feature would not only refine the model’s accuracy but also create a collaborative environment where users can share knowledge and insights. Additionally, features like side-by-side skull comparisons or the ability to track changes in identification confidence as the model is updated with new data could further enrich the user experience, making the application a more comprehensive and versatile tool for research and education.

## IV. Discussions

Although our model is not yet trained to recognize all mammalian orders, it has demonstrated a high degree of accuracy in identifying skulls from the six orders it has been trained on. This success is noteworthy, especially given the complexity and subtlety involved in distinguishing between different mammalian skulls. However, as promising as these results are, there remain several areas where the model could be significantly enhanced, particularly if it is to be made available for widespread use in academic, research, and practical settings.

The most pressing issue with the current model is its limited scope. At present, it is capable of identifying only six mammalian orders out of the 29 recognized in the mammal class. This limitation is primarily due to the time constraints and the computational limitations of the computer used to store and process the training files. During the development phase, these constraints necessitated focusing the model's training on a smaller subset of mammalian orders, preventing the inclusion of the full spectrum of diversity within the mammal class.

This restricted training set poses a significant challenge when the model encounters a skull from an order it hasn't been trained on. In such cases, the model is likely to produce inaccurate results, as it may incorrectly classify the skull into one of the six orders it recognizes. This potential for misclassification is a serious concern, as it could lead to significant errors in research and analysis, thereby undermining the reliability and credibility of the tool. For instance, in a paleontological study where accurate classification of fossils is crucial, such errors could skew the results and lead to incorrect conclusions about evolutionary relationships or species diversity.

However, despite these limitations, the model still offers considerable advantages. One of its most valuable benefits is the ability to significantly reduce the time required for research involving mammal skulls. While a human expert might take hours or even days to accurately identify a skull, the model can perform this task in a fraction of the time. This speed is particularly beneficial in research settings where time is a critical factor, such as in the cataloging of museum collections or the analysis of large datasets in evolutionary biology. By automating the identification process, the model allows researchers to focus on higher-level analysis and interpretation, potentially accelerating the pace of scientific discovery.

To make the model more robust and suitable for broader application, future development efforts should focus on two primary areas. First, expanding the training dataset to cover all 29 mammalian orders is essential. Including the full range of mammalian diversity in the training data would ensure that the model can accurately identify skulls from any order within the mammal class. This expansion would not only increase the model's accuracy but also its utility across a wider range of research scenarios. For example, in a comprehensive study of mammalian evolution, the ability to accurately identify all 29 orders would be invaluable.

Second, there is a critical need to address the computational limitations that currently restrict the model's capabilities. The initial development was constrained by the computational power available, which limited the size and complexity of the training dataset. To overcome these barriers, future work could involve investing in more powerful hardware capable of handling larger datasets and more complex models. Alternatively, optimizing the existing model to run more efficiently on current hardware could also provide a solution, allowing for the inclusion of more data without the need for significant additional resources. Advanced techniques such as model pruning, quantization, or the use of cloud-based computing resources could be explored to enhance the model's performance and scalability.

In addition to these core improvements, it would be highly beneficial to integrate a confidence scoring system into the model. This system would provide users with a measure of the model's certainty in its predictions, allowing researchers to assess the reliability of each identification. For example, if the model classifies a skull with a high confidence score, the user can be more assured of the accuracy of the result. Conversely, a low confidence score could indicate that the skull is from an order not well represented in the training data, prompting further investigation or a manual review. Such a feature would not only improve the model's transparency but also its usability, enabling researchers to make more informed decisions based on the model's output.

These enhancements would collectively improve the accuracy, reliability, and versatility of the model, making it a more dependable tool for researchers. By expanding the training dataset, addressing computational constraints, and incorporating confidence scores, the model would be better equipped to handle the diverse and complex nature of mammalian skulls, reducing the potential for errors and enhancing the efficiency of studies in this field. Ultimately, these improvements could help establish the model as a valuable asset in paleontological research, museum curation, and educational applications, contributing to a deeper understanding of mammalian evolution and diversity.

In conclusion, while the current model has shown great promise, there are clear avenues for further development that could significantly enhance its capabilities. By addressing the limitations of its training scope and computational constraints, and by integrating features such as confidence scoring, the model could evolve into a powerful tool for biological classification. As machine learning continues to advance, the potential applications of such models in paleontology and related fields are vast, offering exciting opportunities for future research and discovery.

## V. Conclusions

Ultimately, the project successfully developed an app that can identify the order that a mammal skull belongs to, a significant step forward in the application of machine learning to biological classification. Although the AI was not trained using data from every single order in Mammalia—there are 29 orders in total—it is currently capable of identifying skulls from 6 distinct mammalian orders. The functionality of the app is straightforward: it requires users to upload an image of a skull, and it then processes this image to output a probability percentage, indicating which order it believes the skull belongs to based on its trained parameters.

However, if the skull is not from one of the 6 mammalian orders the AI has been trained to recognize, the app will still attempt to make an identification. This guess is based on the patterns and similarities it detects when compared to the training images. Due to this limitation, there's a higher chance of incorrect identification in such cases, as the model's predictions are constrained by its training data.

Looking ahead, there are plans to enhance the capabilities of this app significantly. One of the primary goals is to expand the database so that the model can recognize and accurately identify skulls from a broader range of mammalian orders. By covering more orders, the app will become more versatile and useful for a wider array of educational, research, and conservation tasks. Additionally, future updates will focus on improving the model's output by reporting a confidence score for each identification, along with a margin of error. These features will provide users with a clearer understanding of the reliability of each identification, making the app a more powerful tool for studying and cataloging mammalian diversity.

## References

[1]. Monson, Tesla A., et al. "Using Machine Learning to Classify Extant Apes and Interpret the Dental Morphology of the Chimpanzee-Human Last Common Ancestor." PaleoBios, vol. 35, no. 0, 24 Aug. 2018, escholarship.org/uc/item/84d1304f.

[2]. Abdelhady, Ahmed A., et al. "New and Emerging Technologies in Paleontology and Paleobiology: A Horizon Scanning Review." Journal of African Earth Sciences, vol. 210, 1 Feb. 2024, p. 105155, www.sciencedirect.com/science/article/pii/S1464343X2300328X, https://doi.org/10.1016/j.jafrearsci.2023.105155.

[3]. Domínguez-Rodrigo, M., Pizarro-Monzo, M., Cifuentes-Alcobendas, G. et al. Computer vision enables taxon-specific identification of African carnivore tooth marks on bone. Sci Rep 14, 6881 (2024). https://doi.org/10.1038/s41598-024-57015-z

[4]. De Baets, K. (2021) Performance of machine-learning approaches in identifying ammonoid species based on conch properties. Peer Community in Paleontology, 100010. **10.24072/pci.paleo.100010**

[5]. Y. Xu, Z. Dai, J. Wang, Y. Li and H. Wang, "Automatic Recognition of Palaeobios Images Under Microscope Based on Machine Learning," in IEEE Access, vol. 8, pp. 172972-172981, 2020, doi: 10.1109/ACCESS.2020.3024819.

[6]. Liu X, Jiang S, Wu R, et al. Automatic taxonomic identification based on the Fossil Image Dataset (>415,000 images) and deep convolutional neural networks. Paleobiology. 2023;49(1):1-22. doi:10.1017/pab.2022.14g