Classification of Theropod Fossil Images using Fine Grained Image Analysis

Christopher B McCutcheon[1] and Jennifer N Andriot[2]

[1]Department of Computer Science, East Carolina University

[2]REU, East Carolina University

Abstract—Fossil classification is an important but time-consuming process that requires experts in a variety of specializations. Without accurate fossil identification, the information we delineate about the history of our world will become harder to discern. This research focuses on the classification of fossil images belonging to the clade *Theropod*. The goal of this research is to determine if deep learning algorithms can be used to accurately classify images of *Theropod* fossils into their respective geneses. We used convolutional neural networks and Fine-Grained Image Analysis to create machine learning models that automatically classify fossil images.

Index Terms—heropod, Clade, Genus, Convolutional Neural Network, Fine Grained Image Analysisheropod, Clade, Genus, Convolutional Neural Network, Fine Grained Image AnalysisT

I. Introduction

The study of fossils gives us insight into the history of our planet. From evolution to environmental change, the information preserved in fossils is critical to our understanding of the history of life. The accurate identification of fossils is paramount in extracting information about our past. Unfortunately, fossil identification is an arduous, time-consuming task that requires expert paleontologists who specialize in a variety of taxon. Additionally, despite the work that goes into fossil identification, it is not always accurate which leads to more lost time and effort. It is reasonable to believe that the automation of fossil identification will save time and allow paleontologists to focus their efforts elsewhere.

The answer for implementing automatic fossil classification lies in Machine Learning [11]. Efforts to apply machine learning to image classification are vast and all-encompassing [14]. Methods vary circumstantially but image classification is performed with shallow learning methods like Support Vector Machines (SVM), and deep learning methods such as Convolutional Neural Networks (CNN). What machine learning methods are used depends on the datasets and hardware available. Shallow learning uses far less hardware resources but can only process a relatively small amount of data. Alternatively, deep learning methods can process relatively large datasets, but require more powerful hardware to run. The application of Deep Learning to automate tasks across many disciplines has accelerated over the last decade [8]. One such field is paleontology which has seen image classification brought to all manner of fossils from microorganisms like Globotruncana [13] to land vertebrates such as Sauropods [11].

Currently, the effort to classify fossil images is focused on the species level classification of micro fossils with a limited dataset or the broad classification of a multitude of fossil types into large groups (clades) [11][13][2]. The effort to classify land vertebrates on a species level is not a major focus for many research groups. Perhaps this is due to the relative ease of identification of large land vertebrates as opposed to microorganisms which require special imaging equipment to examine.

However, the instant classification of land vertebrates is still a useful tool to attempt to develop as there are many land vertebrates whose minor differences make them difficult to identify, especially for non-experts. The goal of creating a model that could classify almost any known fossil instantly on the species level should be the ultimate objective as such a system would save time for any paleontologist as well as provide insight for those unfamiliar with paleontology. Consider an application, perhaps for mobile devices, that allows anyone to take a picture of a fossil/skeleton and return the classification of that fossil/skeleton image as well as information regarding the species. Such technology is currently beyond our reach as we have yet to develop a model capable of such largescale image classification. Great strides have been taken towards such as model, like the work done by [11] where a collection of approximately 415, 000 fossil images were collected and classified into 50 different clades. However, such models are hindered by the datasets available. To achieve successful deep learning models large, accurately labeled datasets are required. For this project, we aim to develop and test multiple deep learning models to classify images of fossils belonging to the clade Theropod into approximately 100 geneses within the clade.

II. Related Work

The information relevant to this project can be categorized as either domain information or related studies. Domain information refers to all background information related to the project material, including the taxonomic tree, the clade Theropod, and machine learning. Related studies refer to the results of other papers whose work tested the effectiveness of machine learning models in different applications.

A. Domain Information

To provide some important background information regarding the domain of this project, the taxonomic tree should be mentioned. Broadly understood, taxonomy is the classification of life, though it is most often focused on describing species, their genetic variability, and their relationships to one another [4]. Within a taxonomic tree, there are many subcategories describing the relationships of organisms to each other. One of those classifications is a clade, which can be simply defined as a group of organisms who share a common ancestor. The clade that will be this project's central focus is theropod or beast-foot. Members of the clade Theropod are bipedal carnivorous saurischians which include examples like Tyrannosaurus, Deinonychus, Allosaurus and so forth. In addition to the biological information necessary for this project, we must also discuss the computational side of machine learning.

Machine learning can be defined as an evolving branch of computational algorithms that are designed to emulate human intelligence by learning from the surrounding environment [3]. Machine learning can also be broken down into separate categories such as shallow and deep learning or supervised and unsupervised learning. These terms refer to methods surrounding any machine learning model. For example, a supervised machine learning model refers to a model that requires labeled data, that being data that the users of the model have assigned a value to. Think of an image of a dog being assigned the value 'Dog.' Unsupervised learning refers to a model that does not require labeled data but instead categorizes groups of data based on their similarities to each other. Think of a

collection of images of dogs being grouped into a category based on their similarities. Shallow and Deep learning refers to the number of 'layers' in a

model architecture that data must go through for the model to complete data processing. Shallow learning models typically encompass the household names in machine learning like support vector machines or linear discriminant analysis. Deep learning models include several types of neural networks, such as the convolutional neural network used to classify image inputs. The vast expanse of information regarding machine learning is far too broad to provide a comprehensive overview of in this paper, know that this project will focus on deep learning convolutional neural networks.

Convolutional Neural Networks can be understood as an artificial neural network specialized in the field of image processing through the use of convolutional, pooling, and fully connected layers [1]. CNNs (Convolutional Neural Networks) take 3-dimensional data derived from images and filter it through multiple layers of input processing to determine a final categorical weight for that image. They are used in a wide variety of image classification tasks from the medical field to biological studies [5][9]. Some of the most common model architectures for CNNs have been trained on the ImageNet database, a collection of random images that include dogs, cars, and balls. CNNs are accurate when classifying discrete images like a ball and car, but they struggle in more specific applications such as identifying the differences between a golden retriever and a border collie. Since our project will require the classification of visually similar images, we must find better performance in image classification by looking into a subset of computer vision, Fine Grained Image Analysis [15] [6]. FGIA (Fine Grained Image Analysis) is the process of breaking an image down into sections and identifying the whole image based on those sections. Instead of a CNN looking at an image of bird and classifying it as such, the model separates the image into pieces focused on key parts of the bird such as the beak or wings, allowing it to identify a sparrow or cardinal. Our project will require our model to identify a tyrannosaurus fossil from an allosaurus fossil which will require precise image analysis to accomplish successfully. FGIA implementations are relatively new in the field of machine learning. One example is Bilinear Convolutional Neural Networks (BCNN) which combines two CNNs to process separate parts of an image and combine their results into one output. Other examples include architectures like CBAM Resnet v50 and Squeeze and Excitation Resnet v50 [2]. Though the use of FGIA is uncommon, it suits the needs of our project and has been used in similar projects within our domain.

B. Related Studies

This section will focus on the implications of other's work in regards to this project and its domain. Our first study used fine grained convolutional neural networks to determine if they could effectively classify images of conodonts, jawless micro fish [2]. In the study, five finegrained models were tested including Bilinear VGG16, Bilinear Resnet18, Bilinear Resnet50,
CBAM (Convolutional Block Attention Module) Resnet50, and Squeeze and Excitation Resnet50. The model accuracy results respectively are 0.618, 0.676, 0.642, 0.658, and 0.622. Accuracies for the models tested are poor across the board, however, these results in the context of this experiment bode well for future studies of similar projects. Consider that the dataset used in the conodont study consisted of 613 images of mostly fractal conodont fossils. When using complex neural networks for image classification, it is ideal to have as large a dataset as reasonably possible. Tens of thousands or even millions of images are ideal for training CNNs to be accurate [11]. Additionally, if the fossils are in poor condition, as they were in the conodont study, it becomes more difficult for the models to classify them as key features of a particular species may be destroyed or unrecognizable. Taking

dataset quality into consideration, the results of all models staying with the 0.60 range are impressive and may be a sign that studies with larger, more complete datasets will have greater success.

The study performed by Liu et al. In 2023 has been the most influential on our project as it is the source of our project direction and dataset. This study's objective was to test the effectiveness of convolutional neural networks in classifying a fossil image database into 50 different clades. The dataset was a collection of over 415,000 images generated by web scrapers that was then organized into 50 different clades to be classified. The study tested the effectiveness of three DCNN's (Deep Convolutional Neural Networks) including inception v4, inception Resnet v2, and PNASNet-5-large. The final metrics used include a top-one score, the ability of a CNN to correctly classify an image into one category, and top-three score, the ability of a CNN to correctly guess an image's category from one of three predictions. The top one and three scores for the inception v4 model are 0.89 and 0.96. The top one and three scores for the inception Resnet v2 model are 0.90 and 0.97. The top one and three scores for the PNASNet-5-large model are 0.88 and 0.96. All three models performed remarkably well overall with inception Resnet v2 having the highest performance. Precision, Recall, and F1 scores of the individual clades were taken which shed some light on the performance of the models. For instance, in the inception Resnet v2 architecture, the clades conodonts, radiolarians, shark teeth, and trilobites all scored an average F1 score of more than 0.90. However, the clades of sponges and bryozoans scored lower than 0.70. This could be due to the number of images available in each clade for the models to learn from, or it could also be that sponges and bryozoans are harder to classify due to their amorphous nature. The clade Theropod achieved an F1 score of 0.9158 boding well for the model's ability to determine distinct features about theropods such as their bipedalism, typically small arms, and skull shape. The next step Liu and their colleagues determined was classifying at the genus/species level, which is considerably more difficult, but given the advent of models capable of FGIA and the inception Resnet v2 architecture's ability to classify theropod fossil images, does appear to be feasible. The goal of this project then is to build upon the work of Liu et al by creating a model capable of classifying the theropod fossil images collected in this study into over 100 separate geneses.

CNNs in the domain of paleontology are frequently used to classify fossil images of microorganisms [16]. One such study tested the effectiveness of machine learning algorithms in classifying images of the genus of microorganism Globotruncana into three separate species [13].
Four models were tested including Support Vector Machine (SVM), a K Nearest Neighbor (KNN), Linear Discriminant Analysis (LDA), and one CNN. Note that the CNN used in this study was made and trained from scratch. The shallow model SVM, KNN, and LDA outperformed the CNN, but the researchers attributed this to the small size of the dataset used (180 images) and recommended that future studies attempt to aggregate a larger dataset for testing deep learning methods like CNN [13].

Another study focused on the classification of microorganisms that was one done in 2019 focused on the classification of planktonic foraminifera to create a software application that would act as a 'brain' for robotic device designed to extract certain foraminifera from a sample [12]. Using transfer learning for two model architectures, VGG16 and Resnet 50, they compared the results of automated machine classification to that of experts and novices in the field of foraminifera. Using precision, recall, and F1 score as metrics, it was determined that the model performance

averaged an F1 score of approximately 0.86 consistently. The consistent performance from the model contrasted with the performances of the human novices and experts whose F1 scores ranged from anywhere between 0.40 to 0.80. The human results can be explained as a combination of both the subjects not being experienced with the particular form of foraminifera used in this study as well as the subjects being overly cautious and choosing to not classify images that they were unsure about, a choice the machine model did not have. Regardless of human performance, the ability of the CNNs to classify only 1437 images of foraminifera with an F1 score 0.86 bodes well for their application in other niche projects with difficult images to classify.

A study directly relevant to the clade theropod was written in 2021 where shallow machine learning models were tested as classifiers for theropod teeth. In fossil recovery and extraction, one of the many challenges paleontologist faces is that the fossil is incomplete or fragmentary, therefore, paleontologist must learn to gleam information with what little samples they can. The nature of paleontology also makes the aggregation of large datasets difficult because the fossil examples from species-to-species can be drastically different. This study wanted to see if the fossilized teeth of theropods could be used to identify specific theropods and if machine learning algorithms could learn to use teeth images to classify the fossils [16]. Six models were tested in this study including linear discriminant analysis, linear regression, mixture discriminant analysis, Na¨ıve Bayes, random forest, and C5.0. All the aforementioned models are shallow, meaning that they only have 3 or less layers of data processing. Additionally, some of these models are unsupervised, meaning that the dataset used to train them is unlabeled and the models grouped the images based on similarities. All models were trained on two datasets from other studies, one with 800 images and the other with 3000 images, and results for each dataset were kept separate. On the dataset with approximately 3000 images, two permutations of training were used, one in which 17 classes were sorted and the other in which 4 classes were sorted. On the dataset with approximately 800 images, two permutations of training were used, one in which 32 classes were sorted and the other in which 17 classes were sorted. If more images were used and less classes were created, then all the models performed well with decision trees being the top performer. Depending on the permutation of any given round of testing, model accuracies were anywhere between 0.70 and 0.98. The conclusions of this study were that, for shallow models, decision trees should be considered ideal as they handle missing information the best when it comes to non-ideal datasets. In terms of our research, this study demonstrates that even shallow learning models can be used to classify images with minor differences.

In summary, the goals for this project are to build upon the research of Liu et al 2023 by creating neural networks that can classify fossil images on the genus/species level. To accomplish this task, we plan to incorporate the models, like Bilinear VGG16, used in Fine Grained Image Analysis (FGIA) to classify approximately 20,000 fossil images into 103 classes. Additionally, we intend to test both traditional shallow and deep learning models on our dataset to compare with results with our FGIA architecture.

III. Data and Methods

Within this section, we will detail the acquisition and augmentation of our data as well as provide a brief overview of the models we evaluated and the metrics we used to evaluate them.

A. Data Management

1) Fossil Images: Fossil images do not pertain to strictly images of fossils. The primary focus of this study was to classify images of fossil reconstructions based on partial remains, including primarily skeletal reconstructions of *Theropods*. The reasons we ideally want to compare skeletal reconstructions include following:

1. There is greater abundance, in general, ofskeletal reconstruction images as opposed to real fossil images.
2. Skeletal reconstructions offer a complete view

Fig. 1. Image Examples from dataset

of the organism, creating more distinct features for the models to extract.

3. Typically, the skeletal reconstructions offermore photographic views of the same subject from different angles, as opposed to fossils which are typically pressed into the rock they are found, leading most images to be taken from the same side and slightly different angles.

However, this does not mean that we excluded fossil images from our dataset. For some genera, such as Scipionyx, we were unable to find a skeletal reconstruction, however the fossil preservation of the genus was complete enough for our purposes to consider testing on the models. Additionally, some paleoart seemed suitable for classification, such as Sinosauropteryx, who's feather pigmentation was preserved and allowed for accurate, consistent paleoart to be created. In the case of Sinosauropteryx, we were curious to see if the models would be able to classify the paleoart, given that most of the paleoart had consistent features to extract.

2) Data Collection: Our data, was aggregated using the FID (Fossil Image Dataset) created by Liu et al. [11] and images scraped from the Google Chrome search engine. From the FID, we pulled the *Theropod* folder originally containing 113 separate class folders with approximately 20,000 images, mostly denoting fossil genera. After parsing through the dataset, we realized that there were some folders that we believed were harmful for our model training. Those being 10 folders including either box characters or unknown labels as their class names. These unidentified folders contained images of fossils that belonged in other classes and we could not discern why the folders were present so we removed them. We removed these folders as we were unsure of their purpose and did not know how to categorize the data within them. At this point in our dataset there were a total of 103 classes and approximately 18,000 images. An inventory was taken of the dataset to determine how many images were in each class and, after the count was concluded, every class that had 50 or less images were removed. We removed these classes because we believed that the small amount of images would create a negative bias towards them during model training and that the image scraper would not acquire enough images to offset this bias. This process left us with 82 folders at

approximately 19,000 images. Additionally, there was a folder containing all images sourced from the Bing search engine that totaled approximately 8,000 images of any *Theropod*. The Bing folder in particular created a lot of bias in our model's performance as it contained a vast majority of the total images that belonged in other classes. Therefore, we determined that we had to remove the folder and supplement our dataset with additional images to makeup for the roughly 8,000 lost. We implemented a Google Chrome image scraper to collect images for our remaining genera using keywords such as 'Dilophosaurus skeleton', 'Dilophosaurus fossil', and 'Diliphosaurus skull'. Our web scraper obtained approximately 15,000 fossil images with an average of 200 images per class of 80 classes.

3) Data Cleaning: When we scraped the web, upon observing the images we had acquired, we realized that the images obtained would have to be manually cleaned as they contained many inaccurate categorizations. It should be noted that this team does not have any expert or novice paleontologist and therefore could not accurately classify every image that belonged to every class. It would be difficult for us to discern the differences between tarbosaurus and tyrannosaurus at glance through an image given that many images don't have a good reference for size. That being said, there were very obvious outliers that we could identify. The Google Chrome image scraper would grab any image with the aforementioned search keys. Those images could contain anything from *Sauropods* to *Anklyosaurids* which did not have any relation to the subject matter our research focused on and could be safely identified and removed. Additionally, for every genera, a brief background search of the class was done to determine what holotypes existed of each so that we could remove any obvious, unlabeled diagrams which did not match the available fossil records of that genus. After all classes had been manually cleaned they were ported over into the actual dataset where we removed any duplicate images. Once the duplicates were removed, the dataset was in its final iteration, version 5 at this point, and all model results were generated using this version of the dataset.

B. Machine Learning Models for Classification

All of models tested in this study are CNNs imported via transfer learning from the ImageNet database. Those models include Xception, VGG16, Inception Resnet v3, and Bilinear VGG-16. These models were chosen for both their ease of implementation and success in other research. VGG-16 and Inception Resnet were both frequently used in other studies with promising results [9] [11] [12] [13]. Bilinear VGG-16 was pulled from Duan's study on the classification of conodont species in which multiple FGIA models were tested, however the VGG-16 model was the only architecture we were able to find and implement for our ongoing research [2].

1) CNNs (Convolutional Neural Networks): Convolutional Neural Networks are used in computer vision to examine visual data for a variety of applications including, classification, segmentation, and detection. CNNs consist of three primary components for visual recognition including convolutional layers, pooling layers, and Fullyconnected layers. The central component of a CNN are the convolutional layers that use filters, or kernels, to parse a matrix of values which represents the image being examined. The convolutional layer uses the kernel to perform calculations on a selection of pixels so that the CNN

can extract feature information about the image and begin recognizing the patterns in the image pixels, which the model uses to classify images. Pooling layers are used to downsample the dimensions of the feature maps produced by the convolutional layers. The Fully-connected layers are found at the top of the CNN and are used to connect the neurons of the previous layer with the output neurons of the final layer so that the CNN can make predictions based on the features learned during training [1]. CNNs attempt to replicate the human process of visual recognition by recognizing patterns using matrices as opposed to visualization. The focus of CNNs being pattern recognition makes them effective tools for image classification, hence their widespread use in many image classification problems [5] [8] [9] [11] [13].

2) BCNN (Bilinear Convolutional Neural Network): Bilinear Convolutional Neural Networks are a specialized form of CNNs that aim to capture fine-grained feature information about images to identify subtle difference between similar subjects. BCNNs utilize a special form of pooling known as bilinear pooling that aims to combine two separate feature maps produced by the CNN so that interactions between features can be captured. By calculating the outer dot product of the feature vectors produced by the CNN, a new feature representation is produced in a fixed matrix that details the interactions of the features in the original feature maps [6] [7] [15] [10] [17]. With more detailed information about what features are present and how they interact, a more accurate spatial understanding of different subjects can be built which allows BCNNs to make categorical predictions with greater precision.

C. Metrics

Due to technical difficulties involving the sklearn library used in python, the only metric we could use to consistently compare the models performance was the categorical accuracy on the validation set each round of training. The validation categorical accuracy was pulled from the categorical cross-entropy argument used during model training. All of the validation scores represent the accuracy of the final epoch after model layers were unfrozen. These scores reflect the highest accuracy achieved for each model under different hyperparameter tuning permutations.

IV. Findings and Discussion

The results of our validation scores post testing were as such, Xception with 0.347, VGG16 with 0.2402, Inception Resnet v3 with 0.1948, and Bilinear VGG-16 with 0.4457. The results are visualized in Figure 2. Every model that was evaluated overfit the dataset we used. Overfitting is a term used in machine learning to indicate that a model performs well in training, but when tested on information new to the model in a practical setting, the performance is worse. This is likely due to the nature of the dataset. When training CNNs, and other similar models, it is ideal to have as much data as feasibly possible so that the models can fine tune the factors that influence their predictions. The ImageNet dataset, for instance, has approximately 1.2 million images for these neural networks to train on. For our results, and due to a number of factors including time constraints and data availability, we were using roughly 16,000 images. We believe that with more data, our models would not have been prone to overfitting.

V. Conclusions

Despite the consistent overfitting across all of the models tested, the Bilinear VGG-16 model

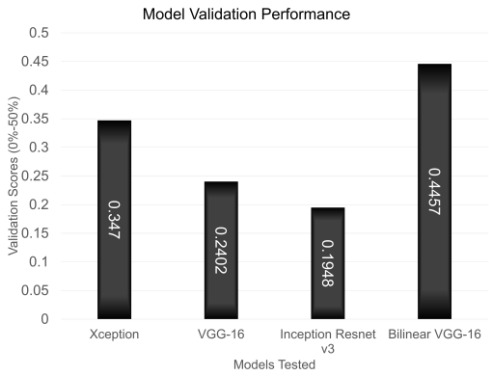

Fig. 2. Chart of Model Validation Scores

consistently achieved a higher validation score than any of the standard CNN models tested with it. These results, while not conclusive enough to determine that FGIA consistently performs better than typical CNN's, do demonstrate that FGIA has the potential to eclipse CNNs in the classification of *Theropod* images and potentially other datasets containing images of similar subjects. For future work, we recommend the following:

1.That the researches obtain a more comprehensive dataset with accurate class organization.

2.That the researches implement more models of both CNN and FGIA architecture.

3.That the researchers include more metrics for comparison such as F1 score, precision, and recall.

We hope to continue this work and would encourage others to look into the applications of FineGrained Image Analysis as well as the classification of fossil images.

References

[1] Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. 2017 International Conference on Engineering and Technology (ICET) p. 1–6 (2017). DOI https://doi.org/10.1109/icengtechnol.2017.8308186. URL https://ieeexplore.ieee.org/document /8308186/

[2] Duan, X.: Automatic identification of conodont species using fine-grained convolutional neural networks. Frontiers in Earth Science 10 (2023). DOI 10.3389/feart.2022.1046327. URL https://www.frontiersin.org/articles/10.3389/feart.2022.1046327

[3] El Naqa, I., Murphy, M.J.: What Is Machine Learning?, pp. 3–11. Springer International Publishing, Cham (2015). DOI 10.1007/978-3-319-18305-3 1. URL https://doi.org/10.1007/978-3-319-18305-3_1

[4] Enghoff, H.: What is taxonomy? – an overview with myriapodological examples. SOIL ORGANISMS 81(3), 441 (2009). URL https://www.soil-organisms.org /index.php/SO/article/view/39

[5] Gupta, G., Kshirsagar, M., Zhong, M., Gholami, S., Ferres, J.L.: Comparing recurrent convolutional neural networks for large scale bird species classification. Scientific Reports 11(1) (2021). DOI https://doi.org/10.1038/s41598-021-96446-w

[6] Hossain, S., Umer, S., Rout, R.K., Tanveer, M.: Fine-grained image analysis for facial expression recognition using deep convolutional neural networks with bilinear pooling. Applied Soft Computing 134, 109,997 (2023). DOI https://doi.org/10.1016/j.asoc.2023.109997. URL https://www.sciencedirect.com/science/article/pii/S1568494 623000157

[7] Huang, M., Wang, Z., Zhu, S.: A fine-grained image classification method combining yolov7 and bilinear multi-level feature fusion. In: Proceedings of the 2022 11th International Conference on Networks, Communication and Computing, ICNCC '22, p. 1–5. Association for Computing Machinery, New York, NY, USA (2023). DOI 10.1145/3579895.3579950. URL https://doi.org/10.1145/3579895.3579 950

[8] LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015). DOI https://doi.org/10.1038/nature14539

[9] Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D., Chen, M.: Medical image classification with convolutional neural network. In: 2014 13th International Conference on Control Automation Robotics Vision (ICARCV), pp. 844–848 (2014). DOI 10.1109/ICARCV.2014.7064414

[10] Lin, T.Y., RoyChowdhury, A., Maji, S.: Bilinear convolutional neural networks for fine-grained visual recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 40(6), 1309–1322 (2018). DOI 10.1109/TPAMI.2017.27234 00

[11] Liu, X., Jiang, S., Wu, R., Shu, W., Hou, J., Sun, Y., Sun, J., Chu, D., Wu, Y., Song, H.: Automatic taxonomic identification based on the fossil image dataset (¿415,000 images) and deep convolutional neural networks. Paleobiology p. 1–22 (2022). DOI https://doi.org/10.1017/pab.2022.14

[12] Mitra, R., Marchitto, T.M., Ge, Q., Zhong, B., Kanakiya, B., Cook, M.S., Fehrenbacher, J.S., Ortiz, J.D., Tripati, A., Lobaton, E.: Automated species-level identification of planktic foraminifera using convolutional neural networks, with comparison to human performance. Marine Micropaleontology 147, 16–24 (2019). DOI https://doi.org/10.1016/j.marmicro.2 019.01.005. URL https://www.sciencedirect.com/science/article/abs/pii/S0377839818301 105

[13] Ozer, I., Ozer, C.K., Karaca, A.C., Gorur, K., Kocak, I., Cetin, O.: 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 (2022). DOI https://doi.org/10.1007/s11042 -022-13810-2

[14] Wang, P., Fan, E., Wang, P.: Comparative analysis of image classification algorithms based on traditional machine learning and deep learning. Pattern Recognition Letters 141 (2020). DOI https://doi.org/10.1016/j.patrec.2020.07.042

[15] Wei, X., Song, Y., Aodha, O.M., Wu, J., Peng, Y., Tang, J., Yang, J., Belongie, S.J.: Fine-grained image analysis with deep

learning: A survey. CoRR abs/2111.06119 (2021). URL https://arxiv.org/abs/2111.06119

[16] Wills, S., Underwood, C.J., Barrett, P.M.: Learning to see the wood for the trees: machine learning, decision trees, and the classification of isolated theropod teeth. Palaeontology (2020). DOI https://doi.org/10.1111/pala.12512

[17] Wu, Q., Miao, S., Chai, Z., Guo, G.: Fine-grained image classification with global information and adaptive compensation loss. IEEE Signal Processing Letters 29, 36–40 (2022). DOI 10.1109/LSP.2021.3123453