Bollettino della Società Paleontologica Italiana, 63 (3), 2024, 215-227. Modena

An explainable Convolutional Neural Network approach to fossil shark tooth identification

Andrea Barucci, Giulia Ciacci, Pietro Liò, Tiago Azevedo, Andrea Di Cencio, Marco Merella, Giovanni Bianucci, Giulia Bosio*, Simone Casati & Alberto Collareta

A. Barucci, Istituto di Fisica Applicata "Nello Carrara", CNR-IFAC, Via Madonna del Piano 10, I-50019 Sesto Fiorentino (FI), Italy; a.barucci@ifac.cnr.it
G. Ciacci, Istituto di Fisica Applicata "Nello Carrara", CNR-IFAC, Via Madonna del Piano 10, I-50019 Sesto Fiorentino (FI), Italy; giuliaciacci8@gmail.com
P. Liò, Department of Computer Science and Technology, University of Cambridge, William Gates Building 15 JJ Thomson Avenue, CB3 0FD Cambridge, UK; pl219@cam.ac.uk
T. Azevedo, Department of Computer Science and Technology, University of Cambridge, William Gates Building 15 JJ Thomson Avenue, CB3 0FD Cambridge, UK; tiago.azevedo@cst.cam.ac.uk
A. Di Cencio, Gruppo Avis Mineralogia e Paleontologia Scandicci, Piazza Vittorio Veneto 1, Badia a Settimo, I-50018 Scandicci (FI), Italy; Istituto Comprensivo "Vasco Pratolini", Via G. Marconi 11, I-50018 Scandicci (FI), Italy; Studio Tecnico Geologia e Paleontologia, Via Pietro Annigoni 16A, I-50025 Montespertoli (FI); andrea.dicencio@gmail.com
M. Merella, Dipartimento di Scienze della Terra, Università di Pisa, Via S. Maria 53, I-56126 Pisa, Italy; marco.merella@phd.unipi.it
G. Bianucci, Dipartimento di Scienze della Terra, Università di Pisa, Via S. Maria 53, I-56126 Pisa, Italy; Museo di Storia Naturale, Università di Pisa, Via Roma 79, I-56011 Calci (PI), Italy; giovanni.bianucci@unipi.it
G. Bosio, Dipartimento di Scienze della Terra, Università di Pisa, Via S. Maria 53, I-56126 Pisa, Italy; giulia.bosio.giulia@gmail.com **corresponding author*
S. Casati, Gruppo Avis Mineralogia e Paleontologia Scandicci, Piazza Vittorio Veneto 1, Badia a Settimo, I-50018 Scandicci (FI), Italy; sim.casati@gmail.com
A. Collareta, Dipartimento di Scienze della Terra, Università di Pisa, Via S. Maria 53, I-56126 Pisa, Italy; Museo di Storia Naturale, Università di Pisa, Via Roma 79, I-56011 Calci (PI), Italy; alberto.collareta@unipi.it

KEY WORDS - Deep Learning, Pattern recognition, Elasmobranchii, explainability, palaeoichthyology, taxonomic determination.

ABSTRACT - This study explores the capability of Convolutional Neural Networks (CNNs), a particular class of Deep Learning algorithms specifically crafted for computer vision tasks, to classify images of isolated fossil shark teeth gathered from online datasets as well as from the authors' experience on Peruvian Miocene and Italian Pliocene fossil assemblages. The shark tooth images that are included in the final, composite dataset (which consists of more than one thousand images) are representative of both extinct and extant genera, namely, Carcharhinus*,* Carcharias*,* Carcharocles*,* Chlamydoselachus*,* Cosmopolitodus*,* Galeocerdo*,* Hemipristis*,* Notorynchus*,* Prionace *and* Squatina*. We compared the classification performances of two CNNs, namely: SharkNet-X, a specifically tailored neural network that was developed and trained from scratch; and VGG16, which was trained using the transfer learning paradigm. Furthermore, in order to understand and explain the behaviour of the two CNNs, while providing a palaeontologist's perspective on the results, we firstly elaborated a visualisation of the features extracted from the images using the last dense layer of each CNN, which was achieved through the application of the t-distributed Stochastic Neighbor Embedding (t-SNE) clustering technique. Then, we introduced the explainability method SHAP (SHapley Additive exPlanations), which is a game theoretic approach to explain the output of any Machine Learning model. The results show that VGG16 outperforms SharkNet-X in most scenarios, especially when trained with data augmentation techniques, achieving high accuracy (93%-97%) in tooth classification. In addition, the SHAP heatmaps revealed that the CNNs relied heavily on tooth margins and inner regions for identification, offering insights into the automated classification process. Overall, this study demonstrates that Deep Learning techniques can effectively assist in identifying isolated fossil shark teeth, paving the way for developing automated tools for fossil recognition and classification.*

INTRODUCTION

Applications of Artificial Intelligence (AI) are growing in popularity across a broad spectrum of scientific fields (Wang et al., 2023), from fundamental physics (Carleo et al., 2019; Ciacci et al., 2024) and biology (Hassoun et al., 2021; Piazza et al., 2021; Richards et al., 2022) to healthcare (Barucci et al., 2021a; D'Andrea et al., 2023) and cultural heritage (Barucci et al., 2021b; Bickler, 2021; Cucci et al., 2021; Guidi et al., 2023). In the broad field of AI, Machine Learning (Domingos, 2012) and in particular Deep Learning (LeCun et al., 2015) algorithms are de facto driving the AI revolution by enabling researchers to harness the power of data, automate processes, improve decision-making and create innovative solutions that have a profound impact on society, economy and technology.

Palaeontology is starting to experience the same kind of trend with AI, as highlighted by some recent works and projects (Beaufort & Dollfus, 2004; Itaki et al., 2020; Liu & Song, 2020; Tetard et al., 2020; Zhang et al., 2020; Hou et al., 2021; Burton, 2022; Antonenko & Abramowitz, 2023; MacFadden et al., 2023; Tetard et al., 2023).

For instance, Mimura et al. (2024) applied the Deep Learning model YOLO-v7 to the detection of microscopic fish teeth and denticles, while Marchant et al. (2020) used a Convolutional Neural Network (CNN) to classify foraminifera. Other interesting applications in micropalaeontology were reported by Carlsson et al. (2023) and Marret (2023), whilst Perez et al. (2023) developed an array of Machine Learning models capable of classifying images of extinct shark teeth, aiming to examine the benefits and drawbacks of Roboflow and Google's Teachable Machine (two free online

platforms) as well as to identify techniques for enhancing palaeontological datasets.

In Liu et al. (2023), three CNNs were trained for taxonomic identification purposes with about 415 thousand photos from 50 different fossil clades, while in Ho et al. (2023) a CNN was introduced to hierarchically classify carbonate skeletal grains, thus enabling Linnaean taxonomic classification from a single image. A dataset of Computer Tomography pictures of protoceratopsian dinosaurs from the Gobi Desert (Mongolia) was developed by Yu et al. (2022) to evaluate the fossil segmentation capabilities of the Deep Learning autoencoder architecture U-net. A Deep Learning application for the identification of uncommon Cambrian microfossils was provided by Wang B. et al. (2022), whilst the identification of fossil brachiopods was the main emphasis of Wang H. et al. (2022). Work presented by Xu et al. (2020) employed Machine Learning to identify palaeontological photomicrographs, while Liu & Song (2020) used CNNs to identify fossil and abiotic grains. Recently, Yu et al. (2024) reviewed over 70 palaeontological AI investigations conducted since the 1980s, encompassing tasks like prediction, image segmentation, and the categorization of micro- and macrofossils.

Fostering the integration of AI techniques for data analysis in palaeontology is crucial due to the complexity of the field. The area of study at the intersection of these two disciplines may yield valuable insights from several perspectives. On the one hand, palaeontology can, for the first time, leverage AI-based powerful tools to examine data; on the other hand, AI can use palaeontological data to build, test and refine its methods. Thus, palaeontology has all the potential to become one of the new frontiers in Artificial Intelligence applications today.

Here, we applied a Deep Learning approach to the problem of the genus-level classification of images of isolated fossil shark teeth based on CNNs, a class of algorithms specifically tailored to deal with computer vision tasks. Thanks to the development of a specific dataset of shark tooth images, we were able to compare the performances of two CNNs. We developed and trained from scratch a CNN, named SharkNet-X, tailored on the complexity of our problem, keeping into account the simplicity of the network architecture and ease of training. At the same time, leveraging on the transfer learning paradigm (Zhuang et al., 2020), we trained the famous VGG16 architecture (Simonyan & Zisserman, 2014).

Our dataset allocates more than one thousand images, representing both extinct and extant shark genera. The dataset used to train the two CNNs was built by merging publicly available fossilised shark teeth (sub)datasets along with images of Peruvian Miocene and Italian Pliocene fossil materials gathered by the authors.

Thanks to this dataset, a comparison between the two CNNs was performed in terms of classification performance. It is worth noting that data augmentation (Shorten & Khoshgoftaar, 2019; Mumuni & Mumuni, 2022) was applied, in particular by varying the rotation angle of the images, thus aiming to develop a CNN model able to work in real settings, with samples oriented in all directions. Furthermore, performing feature extraction from the last dense layer of SharkNet-X and VGG16, allowed us to verify the presence of clusters among our data thanks to the t-distributed Stochastic Neighbor Embedding (t-SNE) clustering technique (Van der Maaten & Hinton, 2008; Kobak & Berens, 2019). This approach allowed us to make considerations about the similarity between genera and species.

Moreover, we introduced the explainability of the CNNs by using the SHapley Additive exPlanations (SHAP) method (Lundberg & Lee, 2017; Lundberg, 2018; Azevedo, 2022). This powerful tool allows data scientists to gain insights into the factors driving predictions in Machine Learning models. Such an explainable approach to the results of the developed CNNs enabled us to make a step towards bridging the gap between a pure Machine Learning approach and the palaeontological approach of human taxonomists.

Our main goal here is to demonstrate the Deep Learning paradigm's ability to assist the identification of ancient shark teeth, setting the groundwork for the creation of information tools for the automatic recognition and categorization of objects in the field of palaeontology.

MATERIAL AND METHODS

Fossil shark teeth and their use in systematic palaeontology

The marine vertebrates that comprise the Class Chondrichthyes (also known as chondrichthyans, or cartilaginous fishes) belong in two subclasses: the poorly diverse Holocephali (currently consisting of a few more than 50 species of chimaeras) and the much more speciose Elasmobranchii, which in turn include the extant superorders Selachimorpha (more than 500 extant species of sharks in eight orders) and Batomorphii (almost 700 extant species of rays in four orders; Serena et al., 2020, and references therein) (Fig. 1). The extant chondrichthyan diversity is still only partially known, with new species being described on a yearly basis (White et al., 2023). The cartilaginous fishes have roamed the Earth's global ocean for more than 400 million years (Andreev et al., 2022), leaving a vast fossil record in their wake. Most of this record is represented by dental and dermal remains: this is due to the fact that teeth and scales are heavily mineralised in chondrichthyans, unlike the endoskeleton, which is largely cartilaginous. In particular, the elasmobranch teeth are coated with a stiff, strong, tough layer of enameloid (a calcium phosphate material; Wilmers et al., 2021) that makes them good candidates for being preserved as fossils. Further enhancing the fossil record of the elasmobranchs is their peculiar pattern of dental replacement, whereby their teeth are shed rapidly and replaced continuously throughout life (Kemp, 1999; Gillis & Donoghue, 2007; Tucker & Fraser, 2014; Berkovitz & Shellis, 2017), so much so that some extant sharks can loss thousands or even tens of thousands of teeth during their lifetime (Brisswalter, 2009). Luckily for palaeontologists, the shark dentitions often exhibit diagnostic morphological characters, hence the utility of isolated fossil shark teeth for taxonomic identifications to the genus or even species level (e.g., Sáez & Pequeño, 2010; Cappetta, 2012; Pollerspöck & Straube, 2018). Ultimately, the painstaking palaeontological study of the fossil shark tooth record allows for reconstructing a comprehensive and detailed picture of the evolutionary history and

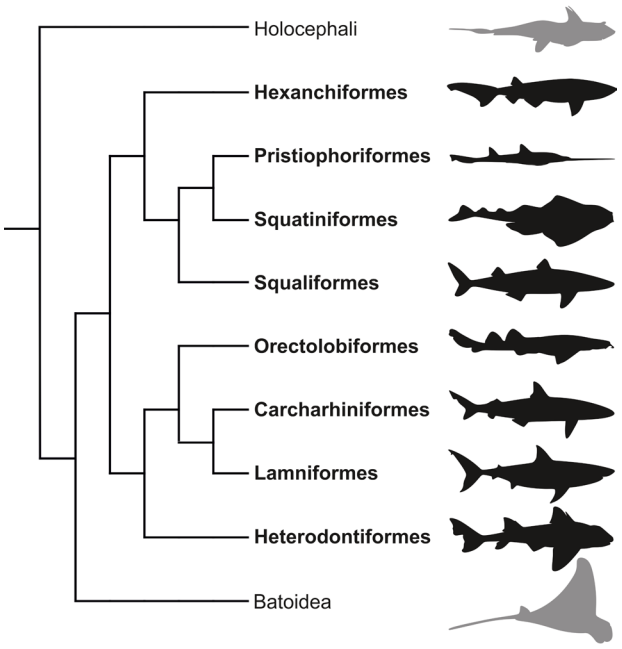

Fig. 1 - Hypothetical phylogenetic tree of the Class Chondrichthyes, showing the relationship between the eight shark superorders (names highlighted in bold and black silhouettes). Tree topology after Díaz-Jaimes et al. (2016).

palaeobiodiversity of one of the most charismatic groups of marine vertebrates.

Dataset of shark tooth images

The dataset developed in this work to train and test the two CNNs (namely, SharkNet-X and VGG16) was built by combining eight separate (sub)datasets. Six such (sub)datasets are publicly accessible and contain labelled pictures of shark teeth coming mostly from palaeontological collections of the Calvert Marine Museum and Florida Museum of Natural History, while the remaining two were created by the authors based on the Italian Pliocene collection of the G.A.M.P.S. (Gruppo Avis Mineralogia e Paleontologia Scandicci) permanent exhibition (Badia a Settimo, Scandicci, Italy; e.g., Cigala Fulgosi et al., 2009; Collareta et al., 2018, 2020) as well as on the authors' first-hand experience on Peruvian Miocene fossil assemblages from the East Pisco Basin stored at the Museo de Historia Natural de la Universidad Nacional Mayor de San Marcos (Lima, Peru; e.g., Landini et al., 2017, 2019; Collareta et al., 2021; Bosio et al., 2022). Information about each (sub)dataset is reported in Tab. 1, while some examples of the tooth images included therein are shown in Fig. 2.

Firstly, all the images comprising the eight (sub) datasets were merged into a single composite dataset containing about 1800 images. Then, a deep manual check was performed, during which duplicate images and photos of taxa represented by very few specimens were removed. Additionally, photos displaying hand-held or weirdly oriented teeth as well as severely damaged specimens were also eliminated. The resulting dataset numbers about 1400 pictures of teeth in either labial or lingual view. All these images underwent a pre-processing step, including the removal of the background to exclude associated elements such as museum labels and scale bars, and the creation of a black uniform background. Examples of the resulting pictures are shown in Fig. 2, with other examples being reported in the Supplementary Online Material (Fig. S1).

The shark taxa allocated in the final dataset include both extinct and extant genera, namely: *Carcharhinus*, *Carcharias*, *Carcharocles*, *Chlamydoselachus*, *Cosmopolitodus*, *Galeocerdo*, *Hemipristis*, *Notorynchus*, *Prionace* and *Squatina.* The number of images available for each genus is reported in Tab. 2. This image collection helps to increase the generalizability of our model by taking into consideration variations in terms of size, orientation, light condition, and intrageneric variability (with some relevant exceptions such as the as-yet monotypic *Prionace*, most of the studied genera are represented by more than a single species in our dataset).

It is also important to keep in mind that some of the considered genera, including *Hemipristis* and *Notorynchus*, are characterised by a high degree of heterodonty, which means that major morphological differences exist between the upper/lower (i.e., dignathic heterodonty) and/or anterior/posterolateral teeth (i.e., monognathic heterodonty), thus further complicating their automated identification.

Convolutional Neural Network architectures for shark tooth classification

In order to understand how Convolutional Neural Networks can perform on our classification task, we compared two of them. A CNN called "SharkNet-X" was built and trained from scratch, while a second CNN (VGG16) was selected from among the most renowned and top-performing architectures.

It is worth noting that both the CNNs were trained, validated and tested on the same dataset, optimising the hyperparameters on the validation set, using the "class_weight" option of TensorFlow for weighting the loss function (during training only) on the base of ratio

Name	Number of images
Shark Tooth Model Dataset	700
Shark teeth Dataset	280
AI and Natural History Shark Tooth Model Dataset	115
Shark Tooth Data Computer Vision Project	46
MHS Data. Megalodon or not megalodon dataset	41
Shark sorting Dataset	40
G.A.M.P.S. (Casati-Zanaga collection of Tuscan Pliocene shark teeth)	428
Departamento de Paleontología de Vertebrados MUSM (collection of Peruvian Miocene shark teeth)	140

Tab. 1 - Overview of the (sub)datasets of shark tooth images for classification used in this paper.

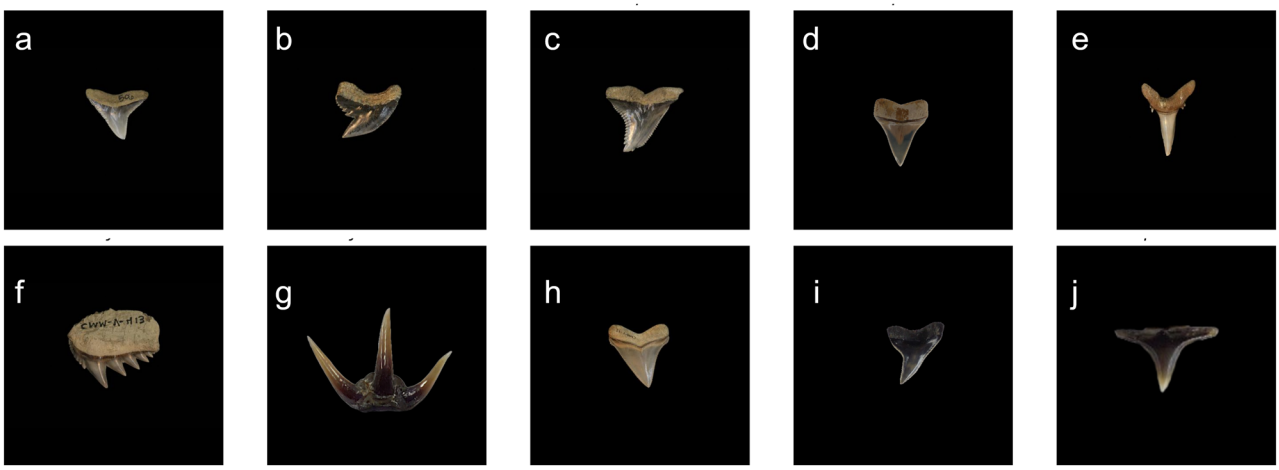

Fig. 2 - Sample images from our composite dataset used to train the CNNs, depicting one specimen for each studied genus. a) *Carcharhinus*. b) *Galeocerdo*. c) *Hemipristis*. d) *Cosmopolitodus*. e) *Carcharias*. f) *Notorynchus*. g) *Chlamydoselachus*. h) *Carcharocles*. i) *Prionace*. j) *Squatina*. Note that the teeth are shown in either lingual or labial view as well as without a scale bar, and set against a black background. This approach was employed to eliminate potential biases and standardise the images for consistent analysis.

between classes in our dataset (aiming to alleviate the class imbalance issue).

Model performances were measured by using Accuracy, Balanced Accuracy, Precision, Recall and F1 score (a detailed description of these metrics is reported in the Supplementary Online Material).

SharkNet-X architecture - SharkNet-X is a Convolutional Neural Network that was implemented in Python, using Keras (Keras, 2015; Chollet, 2021) and Tensorflow. It consists of convolutional (2D), max pooling, flatten, drop-out and dense layers (Abadi et al., 2016). A sufficiently small size of 224 × 224 pixels was empirically selected for the input images, which demonstrated to lessen the computational load without compromising the classification results while enabling easy detection of the shark tooth characteristics. Details on the network's architecture, layer by layer, showing the layer type, the output shape and the number of parameters is reported in Tab. S1 of the Supplementary Online Material.

SharkNet-X architecture was empirically derived by exploring different hyperparameter configurations. Specifically, we explored random selection of batch sizes (i.e., 16, 32, 64, 128), number of epochs (from 10 to 100) and initial learning rates (0.01, 0.001, 0.0005, 0.0001, 0.00005, 0.00001). The model hyperparameters were selected given the best performance on the validation set. Best performances were achieved with SharkNet-X trained using the Adam optimiser (Kingma & Ba, 2014) with an initial learning rate of 0.0001, batch size of 32, 45 epochs, and sparse categorical cross-entropy as loss function. In addition, we implemented early stopping to prevent overfitting and ensure optimal model performance.

VGG16 architecture and transfer learning - VGG16 is a CNN architecture that was proposed by the Visual Geometry Group at the University of Oxford, and it is characterised by its simplicity and depth. It consists of 16 layers, including 13 convolutional layers, two fully connected layers, and a softmax layer. The architecture uses small, 3 × 3 convolution filters throughout the network, and applies them in a consistent manner to increase the depth while keeping the computational requirements manageable. VGG16 has been widely adopted thanks to its excellent performance on image recognition tasks, thus achieving high accuracy in competitions like the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Its pre-trained models are frequently used for transfer learning in various computer vision applications. We utilised a pre-trained VGG16 model from the TensorFlow's Keras API, initialised with ImageNet weights, as the base feature extractor for our classification task with ten classes. The base model layers were set to non-trainable to preserve the learned features. We constructed a sequential model by adding the base model, followed by a flattening layer, a dense layer with 32 ReLU-activated units, and a final dense layer with softmax activation to produce class probabilities. The network was trained with the following hyperparameters: as optimizer Adam with an initial learning rate of 0.0001, batch size of 64, number of

Genus	Number of images (train/test)
Carcharhinus	266/25
Carcharias	232/21
Carcharocles	100/8
Chlamydoselachus	58/10
Cosmopolitodus	71/8
Galeocerdo	127/10
Hemipristis	208/21
Notorynchus	108/10
Prionace	45/8
Squatina	23/8
Total number of images	1367 (1238/129)

Tab. 2 - Composition of the final, composite dataset of shark tooth images used to train the two CNN architectures.

epochs equal to 150 and sparse categorical cross-entropy as loss function. Additionally, we implemented early stopping to prevent overfitting and ensure optimal model performance. Early stopping was used during training to prevent overfitting.

Dataset for training - In order to properly train the two network architectures and assess the resulting model generalisation capabilities, the performances of the two CNNs were compared by splitting the data into three subsets: training, validation and test sets. This enabled the training and evaluation of the final model on distinct sets of data (Bishop, 2006; Bishop & Bishop, 2023). Some details about these definitions are reported in the Supplementary Online Material. In the case of our dataset, we were in presence of class imbalance; thus, in order to alleviate this possible issue, we applied two strategies: firstly, the training loss of the CNNs was weighted using the parameter "class_weight"; secondly, we used data augmentation techniques (described below). The train/test division was then performed as in many classification problems, maintaining an approximate 9/1 ratio between the training and test sets (see Tab. 2). Nevertheless, we decided for the classes with fewer examples to keep a minimum of eight images in the test set. Consequently, this decision led to deviations from the 9/1 ratio for minority classes. The taxonomic distribution of the shark tooth images allocated in the training and test sets is reported in the Supplementary Online Material (Fig. S3). We compared the two CNNs in terms of performance metrics and then selected the model that we considered to be the most suitable for the classification task of our study; this model of choice is referred to as "best model".

Moreover, for this best model we performed another measure of the generalisation performance using the 5-fold cross-validation scheme (Ng, 1997; Hastie et al., 2009; Raschka & Mirjalili, 2019). It is important to note that the images must be pre-processed to be scaled and normalised according to the necessary network input shape in order to feed the CNN model.

Data augmentation - Data augmentation is a technique used in Machine Learning to increase the diversity and size of a training dataset without actually collecting new data. Generally speaking, this is achieved by applying various transformations such as rotations, translations, flips, zooms, and colour adjustments to existing data. By introducing variability and preventing overfitting, data augmentation can help the model to better generalise to new, unseen data. We implemented data augmentation using Keras' "ImageDataGenerator" function, applying the following transformations: random rotations as well as horizontal flipping. It is important to note that, to ensure the CNNs generalize well to real-world images, we used two distinct rotation ranges: a minor perturbation of ±20° and a full rotation of ±180°. This approach enables the network to classify teeth regardless of the photograph's orientation. Consequently, the two CNN architectures were trained under three conditions: no data augmentation, augmentation with rotations up to ±20° and horizontal flipping, and augmentation with rotations up to ±180° and horizontal flipping.

Examples of images for the training/validation/test set in different rotation positions are provided in the Supplementary Online Material (Fig. S2).

Feature clustering - The employment of clustering methods applied to features extracted by CNNs has garnered increasing attention due to their potential in revealing underlying structures within data (Guérin et al., 2021). Features extracted by a CNN through its convolutional layers capture hierarchical representations of input images, while the last dense layer provides a high-dimensional feature vector that can be used for clustering similar images together, thus enabling the unsupervised analysis and organisation of image datasets.

Here, we employed an unsupervised clustering method called t-SNE to the last dense layer of SharkNet-X and VGG16. This idea was explored in the field of palaeontology also by Liu et al. (2023) and Niu et al. (2024). t-SNE is a popular dimensionality reduction technique used in Machine Learning and data visualisation. It aims to map high-dimensional data points into a lower-dimensional space, typically 2D or 3D, while preserving the local structure of the data as much as possible. t-SNE works by modelling the similarity between data points in the original high-dimensional space and then representing these similarities in the lower-dimensional embedding. This helps to reveal clusters and patterns in the data that might not be easily discernible in the original high-dimensional space.

Exaggeration and perplexity are two crucial t-SNE hyperparameters that affect the final visualisation, influencing how the algorithm balances between capturing local and global structures in the data. While perplexity controls the effective number of neighbours considered for each data point, exaggeration enhances the separation between clusters in the lower-dimensional embedding space, making them more visually distinct. A low perplexity value causes t-SNE to focus more on local structure, while a high perplexity value considers a larger neighbourhood for each data point. Typically, perplexity values in the range of 5 to 50 are used, while it is a common practice to use exaggeration with a factor of 4 or 12. It is often necessary to experiment with different values of perplexity and exaggeration to empirically find the optimal combination that reveals the structure of the data.

CNN Explainability - Nowadays, the field of explainable artificial intelligence (XAI; Samek et al., 2017) comprises an important research field. In fact, explainability of the results of a Neural Network is one of the most remarkable and complex aspects of Deep Learning. This step is necessary in order to move these algorithms into applications in science, giving to the field experts not only a performance score, but also an explanation of why certain results are obtained. Explainability can also provide a feedback mechanism, wherein experts can help to improve the Deep Learning model looking at the explained results, thus contributing to modify the algorithm architecture and/or the training procedure if necessary. Today, much of the research in XAI continues to rely on CNNs for their comparatively more interpretable representations and established methods for explainability (Zeiler & Fergus, 2014). CNNs have been extensively studied and have well-established techniques such as feature visualisation, occlusion analysis, and saliency mapping, which facilitate understanding of model behaviour and decision-making processes. For example, Grad-CAM, LIME and SHAP are well-known

such methods (Ribeiro et al., 2016; Selvaraju et al., 2020; Van Zyl et al., 2024), which have been extensively studied and confronted (Panati et al., 2022), also in the field of palaeontology (e.g., Hou et al., 2023).

For the purposes of the present work, we implemented SHAP - SHapley Additive exPlanations, an algorithm based on Shapley values, a concept from cooperative game theory. Shapley values have been introduced in 1953 by Lloyd Shapley (Shapley, 1953) and have been used in the past to compute explanations on model predictions (Lipovetsky & Cocklin, 2001; Štrumbelj & Kononenko, 2014). SHAP is able to explain the output of any Machine Learning model, quantifying the contribution of each feature to the prediction made by the model, thus providing insights on how the model arrived at a particular prediction. The core idea behind SHAP is rooted in the concept of Shapley values, which assigns a value to each player in a cooperative game based on their contribution to the overall outcome. SHAP computes Shapley values for each feature, indicating how much each feature contributes to the difference between the actual and average predictions. SHAP values then provide interpretable explanations for individual predictions, highlighting which features pushed the prediction towards a certain outcome, and which in turn pulled it away. These explanations can help users to understand and trust complex Machine Learning models, identify influential features, detect biases, and debug model behaviour.

In the context of palaeontology, SHAP can help fill the gap between a purely data-driven approach to classification and the necessity of a palaeontological explanation of the results obtained by the model. In the case of images, which is relevant to our task of automating the classification of shark tooth photographs, SHAP provides insights on why a particular image was classified in a certain way by attributing the contribution of each pixel to the model decision. SHAP analyses the importance of individual pixels or regions of the image in influencing the model prediction. It computes SHAP values for each pixel, indicating how much a pixel contributes to the difference between the model's prediction for the image and its average prediction. Positive SHAP values indicate that a pixel contributes to a higher prediction score for a particular class, while negative values indicate the opposite. By visualising these SHAP values, users can gain insights into which parts of the image are most influential in the model decision-making process. This can help in understanding why the model classified the image in a certain way, identifying important features or patterns in the image, and assessing the model strengths and weaknesses.

We applied SHAP to the classification results of SharkNet-X and VGG16, thus obtaining for each CNN a heatmap of the SHAP values for image inputs of representative examples of each genus. These heatmaps provide a visual representation of the importance or contribution of each pixel to the model prediction for a particular image. A colour gradient is used to indicate the magnitude of the contribution, with warmer colours (e.g., red) representing higher or positive contributions, and cooler colours (e.g., blue) representing lower or negative contributions. The heatmap often reveals localised patterns or regions of interest within the image that are particularly influential in the model decision-making process. These regions may correspond to specific features, objects or structures that are relevant to the classification task. Overall, the heatmaps generated by SHAP provide valuable visual insights into the inner workings of image classification models, helping users to interpret and understand model predictions in a transparent and interpretable manner.

The warranted consistency of SHAP values helps explaining why it gained so much popularity in recent years. Indeed, these and other advantages have been highlighted in recent literature reviews (Arrieta et al., 2020; Heuillet et al., 2021; Vilone & Longo, 2021; Bodria et al., 2023). SHAP is presented by Linardatos et al. (2020, p. 12) as "the most complete method, providing explanations for any model and any type of data, doing so at both a global and local scope", and "[together with LIME (Ribeiro et al., 2016)], by far, the most comprehensive and dominant across the literature methods for visualising feature interactions and feature importance". Here, we used DeepExplainer, a variant of SHAP designed to work with Deep Learning models such as those implemented in frameworks like TensorFlow and Keras. To calculate Shapley values for model explainability, we selected 70 random samples from the training set to form the background dataset, ensuring diversity and representativeness by selecting seven samples from each class. We then used the SHAP DeepExplainer with this background dataset to explain the model's predictions. Specifically, we applied the explainer to test images, thus generating Shapley values that quantify the contribution of each image pixel to the model's output.

CNN training and computing environments

All the analysis described above were executed on a workstation with the following characteristics: CPU Intel Core i9-7940X CPU, RAM 128 GB, GPU NVIDIA Quadro P6000 24 GB, OS Ubuntu 22.04.4 LTS. An epoch of training of VGG16 with data augmentation required about 50 seconds, which for 120-150 epochs means hours of training. Exploring the space of hyperparameters of the CNN, such as learning rate or changing the data augmentation settings, can exponentially increase the time required for training. In turn, a single run of SHAP can take up to five minutes (depending on the number of samples used) to provide a heatmap as output.

RESULTS

Classification results

Table 3 compares the two CNNs under different data augmentation conditions (rotations of 0°, ±20° and ±180°) for multiple classification metrics.

VGG16 demonstrates robustness across various rotations, maintaining strong performance in all metrics. In particular, VGG16 shows consistently higher accuracy and balanced accuracy across all rotations compared to SharkNet-X, maintaining an accuracy between 93-97% and a balanced accuracy between 91-95% regardless of the rotation, which indicates robustness to image transformations. Additionally, VGG16 consistently

	SharkNet-X			VGG16		
Rotation [degrees]	**0**	**20**	**180**	**0**	**20**	**180**
Accuracy (%)	80	89	77	97	95	93
Bal. Accuracy (%)	76	85	75	95	94	91
Precision (%)	82	91	80	97	96	95
Recall (%)	80	89	77	97	95	93
F1 (%)	78	88	76	97	95	93

Tab. 3 - Classification performance metrics for the two CNNs, SharkNet-X and VGG16. Results are reported for the three cases considered, namely, no data augmentation, and data augmentation with horizontal flipping and rotation angles ±20° and ±180°.

demonstrates higher precision (95-97%), meaning it produces fewer false positives across all rotations, while also maintaining a high recall rate (93-97%), which indicates its effectiveness in identifying true positives regardless of image orientation. This results in a higher overall F1 score (93-97%) that reflects an excellent balance between precision and recall.

SharkNet-X performs reasonably well for minor rotations but shows noticeable drops in performance when classifying images rotated by 180°. This suggests it may be more sensitive to significant transformations in the input data, possibly due to a less robust internal representation of rotational variance. In real-world applications, where handling rotated images may prove necessary, VGG16 appears to be a more suitable option due to its consistent performance across various transformations. However, SharkNet-X may still be an efficient model under normal or minimally rotated conditions of the input data.

In light of the aforementioned, we selected VGG16 with data augmentation using rotation angles up to ±180° as our "best model". Despite not being the highest-performing model in terms of classification metrics, we consider it the best compromise between robust performance and generalizability to real-world scenarios. The learning curves (loss function and accuracy) for the training and validation sets and a performance matrix (evaluated on the test set) for the best model are reported in Fig. 3 and Fig. 4, respectively, while the corresponding figures for the other models are provided in the Supplementary Online Material (Figs S4, S5 and S6). In addition, the best model was cross-validated obtaining an average balanced accuracy of 89 ± 3% (std). Additional cross-validation results are reported in Tab. S2 of the Supplementary Online Material.

t-SNE and SHAP results

We used the last dense layer of the best model CNN to extract 128 features for each image in the dataset. The resulting feature data matrix (1367 images × 128 features) was given as input to the t-SNE algorithm implemented with SciKit-Learn (Pedregosa et al., 2011). Results are shown in Fig. 5 for the combination of hyperparameters given by a perplexity of 50 and an exaggeration of 50. t-SNE results for the other CNN models (including SharkNet-X) are provided in the Supplementary Online Material (Fig. S7).

Figure 6 displays the SHAP heatmaps for the best model CNN, as evaluated for a few representative examples.

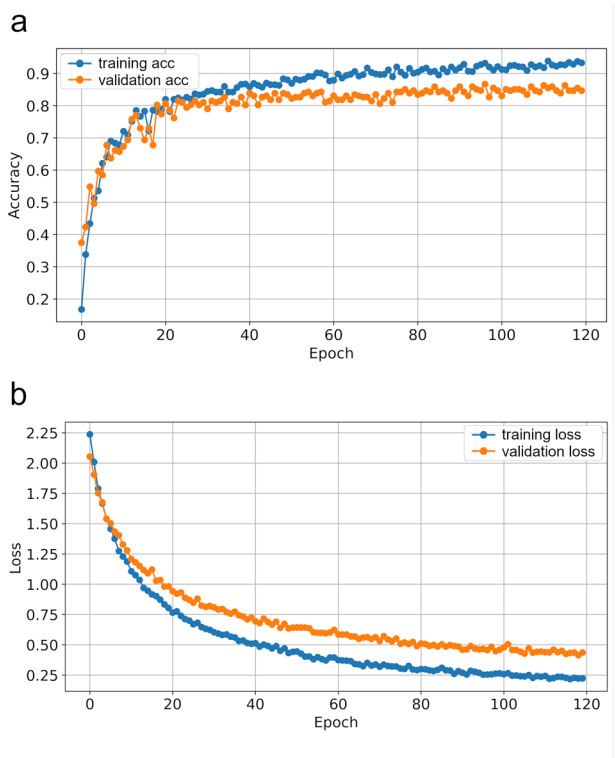

Fig. 3 - Train and Validation learning curves for the VGG16 with data augmentation (rotation range ± 180°): a) Accuracy; b) Loss.

DISCUSSION

The performance results summarised in Tab. 3 show good performances for all the models. The SharkNet-X architecture appears to work better when trained using data augmentation with small rotation angles (up to ±20°), while VGG16 shows about the same performances when trained without data augmentation and when using data augmentation with small rotation angles (up to ±20°). Using instead data augmentation with complete rotation angles (up to ±180°) results in slightly decreasing the VGG16 model performance. However, in general the performances of VGG16 architecture are always better than those of SharkNet-X. Aiming to obtain a model with good performances on real scenarios, where shark teeth can appear in different orientations, we selected as best model the VGG16 architecture trained with data augmentation with angles up to ±180°.

Results for the test set using the best model are summarised by the Confusion Matrix shown in Fig. 4, which shows that the most frequent mispredictions regarded: 1) teeth of *Carcharocles* being misinterpreted as belonging to *Carcharhinus* and *Prionace*; and 2) teeth of *Galeocerdo* being misinterpreted as belonging to *Prionace*. The former issue may be due to the fact that some upper teeth of *Prionace* and broad-toothed *Carcharhinus* species (i.e., the "bull group" sensu Cappetta, 1987) are superficially reminiscent of those

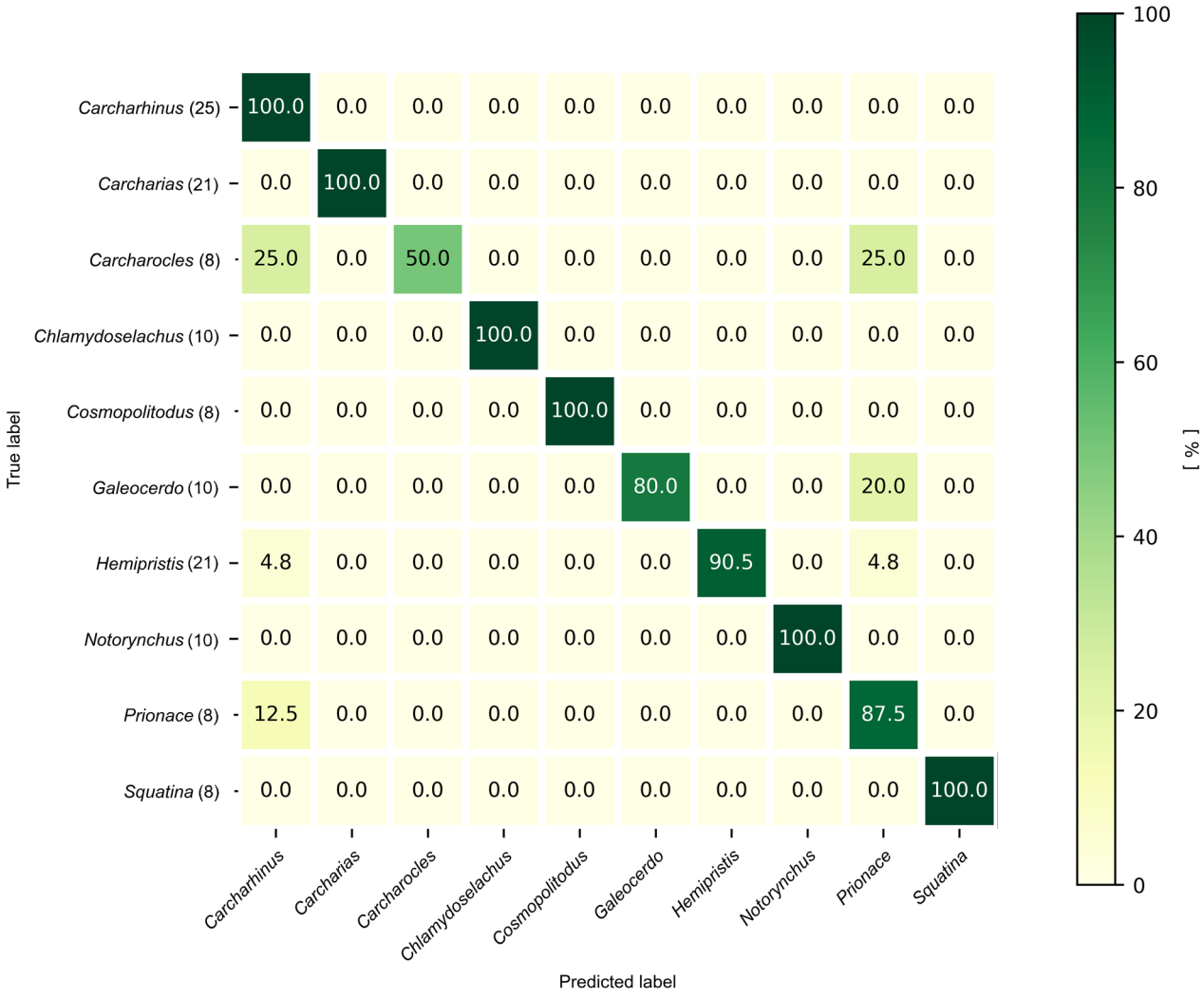

Fig. 4 - Heatmap visualization of the classification performance of the VGG16 best model with data augmentation (rotation range ± 180°), showing the percentage of correct and incorrect predictions for each class made on the test set. The rows represent the true labels (genera), and the columns represent the predicted labels (genera), with each cell indicating the percentage of prediction. Along the rows, next to each genus name, the number in parentheses represents the total number of images belonging to that specific class. Diagonal cells show the model's accuracy for each class, while off-diagonal values reveal the proportion of misclassifications, helping to identify specific genera that are frequently confused with others. The color gradient from light yellow to dark green corresponds to increasing percentage values.

of *Carcharocles*, especially when size and profile views are not taken into account. Such superficial similarities concern the overall outline, presence of serrations along the cutting edges, absence of cusplets, and relatively short root lobes, although many major differences do obviously exist. That two teeth of *Galeocerdo* out of ten were misinterpreted as belonging to *Prionace* is more difficult to explain based on dental similarities alone.

The high-dimensional representation learned by the last dense layer of the CNN is confirmed by the 2D t-SNE visualisation shown in Fig. 5. The extracted features exhibit clustering patterns corresponding to the shark genera, indicating the effectiveness of the CNN in capturing meaningful distinctions within the dataset. The best individualised clusters appear to be those concerning *Chlamydoselachus* and *Notorynchus*. This is not surprising as the tricuspid teeth of *Chlamydoselachus* and the comb-like teeth of *Notorynchus* are among the most distinctive of the whole dataset. Conversely, the cluster that describes the genus *Carcharhinus* partly overlaps with a few others. Once again, this comes as no surprise considering that *Carcharhinus* spp. are characterised by a cutting-clutching dentition wherein the upper and lower teeth are often different from each other in general morphology (e.g., Bourdon, 1991), and that different species of *Carcharhinus* (including both narrow- and broad-toothed forms) are present in our dataset, which concurs to create a wide intrageneric variability in terms of tooth shape.

SHAP results are shown in Fig. 6 for some representative examples. From a data analysis point of view, it is important to outline some characteristics of the obtained SHAP heatmaps. In the figure, SHAP values are given for each example and some possible outcomes. Generally speaking, the distribution of warm- and cool-coloured pixels indicates that the characters

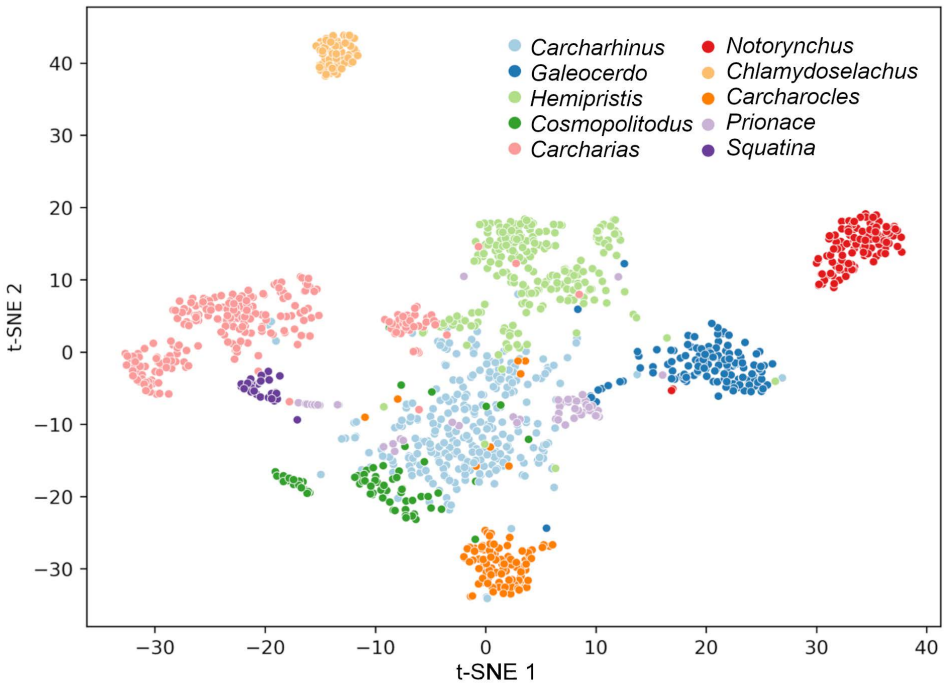

Fig. 5 - t-SNE 2D visualization of the features extracted from the last dense layer of the VGG16 best model with data augmentation (rotation range ± 180°), using the entire dataset.

that the CNN keeps as most diagnostic cluster along the tooth margins, and secondarily inside the margins themselves, suggesting that it is the tooth outline (and especially the crown outline) that leads the process of automated identification. In some cases, however, we found such pixels outside the main region of the image. This outcome is to be expected, as this method focuses exclusively on highlighting what the CNN is "looking at" in the input image in order to discriminate different genera of shark teeth rather than explaining why a specific prediction was made.

Although the performances achieved by the best model are good, the clustering of t-SNE meaningful and the results of the SHAP explainability heatmaps promising, it is important to point out the limitations of our work in order to indicate future improvements and advancements. Firstly, one of the limitations consists in the small number of specimens included in some of the considered genera (e.g., *Chamydoselachus*, *Prionace* and *Squatina*) and class imbalance. Addressing this challenge might result in improving the model's performances. It is also worth noting that a more systematic analysis will be necessary in future works to assess the consistency of the t-SNE results due to the stochasticity and hyperparameter dependency of the method.

Another limitation concerns the interpretability of the results provided by SHAP. While the SHAP heatmaps confirm that the CNN is "looking at" the right places (i.e., the tooth borders and inner regions), the method does not explain how this information is used by the network in order to perform the prediction. This limitation is inherent to the SHAP post-hoc approach to Explainability.

By acknowledging these limitations, we pave the way for future research endeavours aimed at addressing these challenges.

CONCLUSIONS

We explored the efficacy of two Convolutional Neural Networks, SharkNet-X and VGG16, for the genus-level classification of fossil shark teeth. We compared the two CNNs and chose a best model, namely, the VGG16 trained with data augmentation. This model achieved good performances. Additionally, the t-SNE clustering analysis conducted on the last dense layer of the CNN reaffirmed the quality of the learned representation, clearly demonstrating distinct clustering patterns corresponding to different genera. Hopefully, this model will serve as an initial step toward the creation of research software, such as the already available ParticleTrieur (Marchant et al., 2020), dedicated to specimen classification and the analysis of tooth morphology.

In addition, we introduced the SHAP method for CNN explainability, which allowed us to obtain a visual heatmap to interpret the output of the CNN. Such maps were investigated by palaeontologists in order to understand what kind of patterns/features the neural network was looking at during the identification process.

Although certain limitations persist, we are confident that future research will significantly advance our comprehension and utilization of Deep Learning techniques in the field of shark tooth classification and beyond.

DATA AND CODE AVAILABILITY

The complete dataset generated and analysed in the current study is available from the first and corresponding authors. Additionally, the fine-tuned VGG16 Model (trained with rotation up to 180 degree) for our application

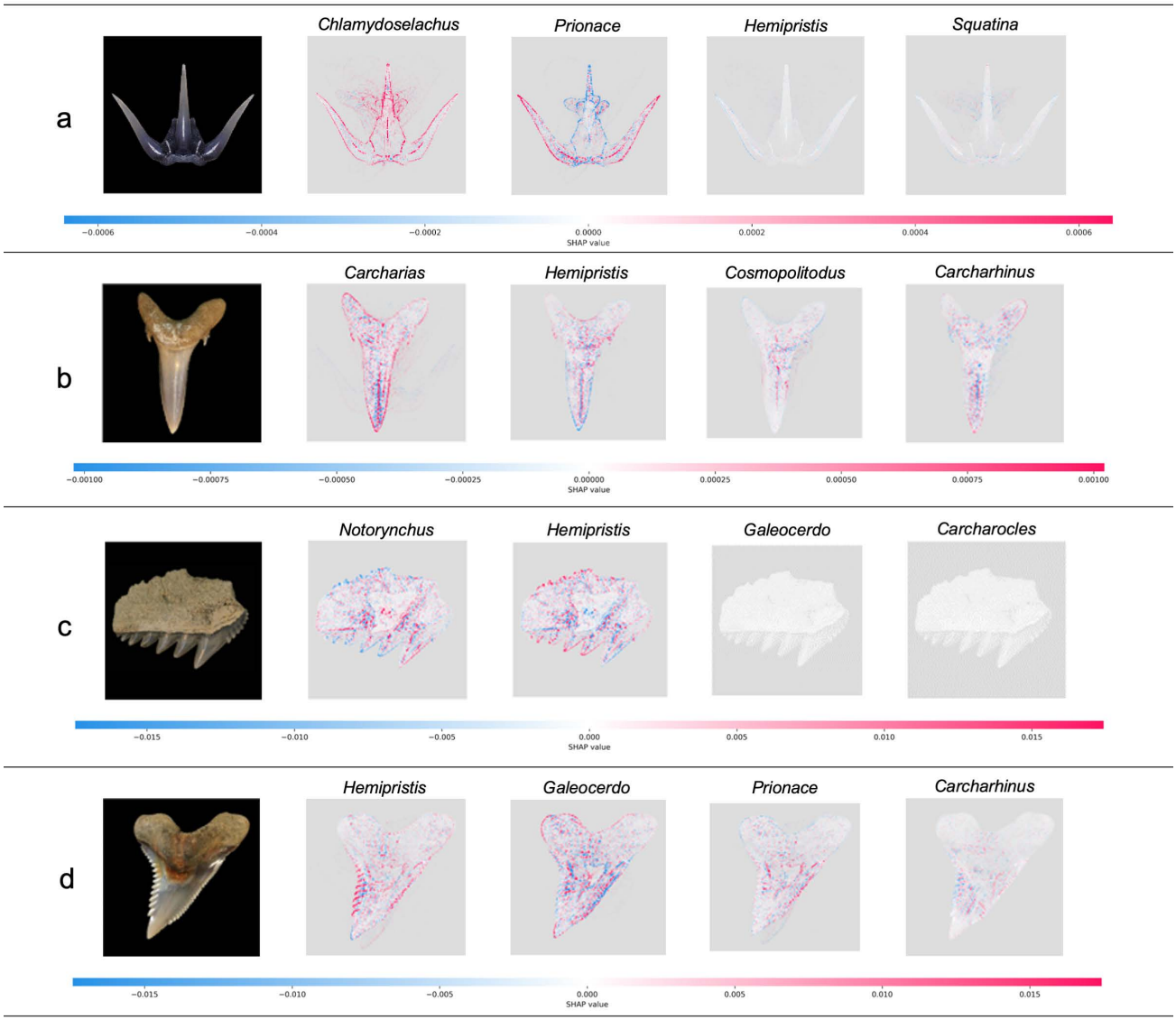

Fig. 6 - Close-ups of the SHAP heatmaps, obtained using the VGG16 best model with data augmentation (rotation range ± 180°), for four sample images in our dataset, showing the original image (left column) besides four relevant heatmaps. The genera depicted in the four original images are the following: a) *Chlamydoselachus*; b) *Carcharias*; c) *Notorynchus*; d) *Hemipristis*.

and part of the dataset are available on github at https://github.com/GAIA-IFAC-CNR/SharkNet-X.

SUPPLEMENTARY ONLINE MATERIAL

Supplementary data of this work are available on the BSPI website at: https://www.paleoitalia.it/bollettino-spi/bspi-vol-633/

ACKNOWLEDGEMENTS

We gratefully acknowledge J. Agresti, C. Cucci, G. Magni, D. Mugnai, F. Rossi, S. Siano and P. Di Maggio from CNR-IFAC, S. Fiore (CNR), C. Canfailla, D. Baracchi, L. Seidenari and F. Argenti from DINFO-UniFI. We thank all the Signal Processing and Communications Laboratory (LESC) from DINFO-UniFI, providing support for the computational resources for the CNNs training and test. Special thanks go to A.M. Addabbo, S. Taruffi and to all the students from the "Istituto di Istruzione Superiore Tecnica e Liceale Russell Newton" (Scandicci, Italy), who supported the G.A.M.P.S. dataset creation. Furthermore, we are grateful to the many researchers who made publicly available the six tooth image datasets used in this work. The BSPI Associate Editor Silvia Danise and two anonymous reviewers are graciously acknowledged for their constructive comments.

M.M., G.Bi., G.Bo. and A.C. acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.1, Call for tender No. 104 published on 02/02/2022 by the Italian Ministry of University and Research (MUR), funded by the European Union - NextGenerationEU - Project Title: BIOVERTICES (BIOdiversity of VERTebrates In the CEnozoic Sea) - CUP I53D23002070 006 - Grant Assignment Decree No. 965 adopted on 30/06/2023 by the Italian Ministry of University and Research (MUR).

REFERENCES

Abadi M., Agarwal A., Barham P., Brevdo E., Chen Z., Citro C., Corrado G.S., Davis A., Dean J., Devin M., Ghemawat S.,

Goodfellow I., Harp A., Irving G., Isard M., Jia Y., Jozefowicz R., Kaiser L., Kudlur M., Levenberg J., Mane D., Monga R., Moore S., Murray D., Olah C., Schuster M., Shlens J., Steiner B., Sutskever I., Talwar K., Tucker P., Vanhoucke V., Vasudevan V., Viegas F., Vinyals O., Warden P., Wattenberg M., Wicke M., Yu Y. & Zheng X. (eds) (2016). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. http://tensorflow.org (accessed 30.04.2024).

AI and Natural History Shark Tooth Model Dataset (ed.) (2023). https://universe.roboflow.com/ai-and-natural-history-d1nwu/shark-tooth-model-x8teb (accessed 30.04.2024).

Andreev P.S., Sansom I.J., Li Q., Zhao W., Wang J., Wang C.-C., Peng L., Jia L., Qiao T. & Zhu M. (2022). The oldest gnathostome teeth. *Nature*, 609: 964-968.

Antonenko P. & Abramowitz B. (2023). In-service teachers' (mis) conceptions of artificial intelligence in K-12 science education. *Journal of Research on Technology in Education*, 55: 64-78.

Arrieta A.B., Díaz-Rodríguez N., Del Ser J., Bennetot A., Tabik S., Barbado A., Garcia S., Gil-Lopez S., Molina D., Benjamins R., Chatila R. & Herrera F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58: 82-115.

Azevedo T. (2022). Data-driven Representations in Brain Science: Modelling Approaches in Gene Expression and Neuroimaging Domains. 166 pp. Ph.D. Thesis, Cambridge. https://doi.org/10.17863/CAM.86924

Barucci A., D'Andrea C., Farnesi E., Banchelli M., Amicucci C., de Angelis M., Hwang B. & Matteini P. (2021a). Label-free SERS detection of proteins based on machine learning classification of chemo-structural determinants. *Analyst*, 146: 674-682.

Barucci A., Cucci C., Franci M., Loschiavo M. & Argenti F. (2021b). A deep learning approach to ancient egyptian hieroglyphs classification. *Ieee Access*, 9: 123438-123447.

Beaufort L. & Dollfus D. (2004). Automatic recognition of coccoliths by dynamical neural networks. *Marine Micropaleontology*, 51: 57-73.

Berkovitz B. & Shellis P. (2017). The teeth of non-mammalian vertebrates. 342 pp. Academic Press, Cambridge.

Bickler S.H. (2021). Machine learning arrives in archaeology. *Advances in Archaeological Practice*, 9: 186-191.

Bishop C.M. (2006). Pattern recognition and machine learning. Information Science and Statistics. 738 pp. Springer, New York.

Bishop C.M. & Bishop H. (2023). Deep Learning: Foundations and Concepts. 649 pp. Springer, Cham.

Bodria F., Giannotti F., Guidotti R., Naretto F., Pedreschi D. & Rinzivillo S. (2023). Benchmarking and survey of explanation methods for black box models. *Data Mining and Knowledge Discovery*, 37: 1719-1778.

Bosio G., Bianucci G., Collareta A., Landini W., Urbina M. & Di Celma C. (2022). Ultrastructure, composition, and $^{87}Sr/^{86}Sr$ dating of shark teeth from lower Miocene sediments of southwestern Peru. *Journal of South American Earth Sciences*, 118: 103909.

Bourdon J. (ed.) (1999). The Life and Times of Long Dead Sharks. www.elasmo.com (accessed 27.08.2022).

Brisswalter G. (2009). Inventaire des Élasmobranches (requins, raies, chimères) des dépôts molassiques du Sud-Luberon (Miocène supérieur), Cabrières d'Aigues (Vaucluse) France. *Courriers scientifiques du Parc naturel Régional du Lubéron*, Hors Série: 1-100.

Burton R. (ed.) (2022). UF Earns Grant to Teach Middle Schoolers About Shark Teeth Using AI. https://www.floridamuseum.ufl.edu/earth-systems/blog/uf-earns-grant-to-teach-middle-schoolers-about-shark-teeth-using-ai/ (accessed 11.04.2024).

Cappetta H. (1987). Chondrichthyes II. Mesozoic and Cenozoic Elasmobranchii. *In* Schultze H.-P. (ed.), Handbook of Paleoichthyology, Volume 3B. 193 pp. Gustav Fischer Verlag, Stuttgart, New York.

Cappetta H. (2012). Chondrichthyes. Mesozoic and Cenozoic Elasmobranchii: Teeth. *In* Schultze H.-P. (ed.), Handbook of Paleoichthyology, Volume 3E. 512 pp. Pfeil, Munich.

Carleo G., Cirac I., Cranmer K., Daudet L., Schuld M., Tishby N., Vogt-Maranto L. & Zdeborová L. (2019). Machine learning and the physical sciences. *Reviews of Modern Physics*, 91: 045002.

Carlsson V., Danelian T., Tetard M., Meunier M., Boulet P., Devienne P. & Ventalon S. (2023). Convolutional neural network application on a new middle Eocene radiolarian dataset. *Marine Micropaleontology*, 183: 102268.

Chollet F. (2021). Deep learning with Python. 504 pp. Simon and Schuster, New York.

Ciacci G., Barucci A., Di Ruzza S. & Alessi E.M. (2024). Asteroids co-orbital motion classification based on Machine Learning. *Monthly Notices of the Royal Astronomical Society*, 527: 6439-6454.

Cigala Fulgosi F., Casati S., Orlandini A. & Persico D. (2009). A small fossil fish fauna, rich in *Chlamydoselachus* teeth, from the Late Pliocene of Tuscany (Siena, central Italy). *Cainozoic Research*, 6: 3-23.

Collareta A., Casati S. & Di Cencio A. (2018). The porbeagle shark, *Lamna nasus* (Elasmobranchii: Lamniformes), from the late Pliocene of the central Mediterranean Basin. *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen*, 287: 307-316.

Collareta A., Merella M., Mollen F.H., Casati S. & Di Cencio A. (2020). The extinct catshark *Pachyscyllium distans* (Probst, 1879) (Elasmobranchii: Carcharhiniformes) in the Pliocene of the Mediterranean Sea. *Neues Jahrbuch für Geologie und Paläontologie, Abhandlungen*, 295: 129-139.

Collareta A., Di Celma C., Bosio G., Pierantoni P.P., Malinverno E., Lambert O., Marx F.G., Landini W., Urbina M. & Bianucci G. (2021). Distribution and paleoenvironmental framework of Middle Miocene marine vertebrates along the western side of the lower Ica Valley (East Pisco Basin, Peru). *Journal of Maps*, 17: 7-17.

Cucci C., Barucci A., Stefani L., Picollo M., Jiménez-Garnica R. & Fuster-Lopez L. (2021). Reflectance hyperspectral data processing on a set of Picasso paintings: Which algorithm provides what? A comparative analysis of multivariate, statistical and artificial intelligence methods. *In* Liang H. & Groves R. (eds), Optics for Arts, Architecture, and Archaeology VIII, *SPIE Proceedings*, 11784: 1-10.

D'Andrea C., Cazzaniga F.A., Bistafa E., Barucci A., de Angelis M., Banchelli M., Farnesi E., Polykretis P., Marzi C., Indaco A., Tiraboschi P., Giaccone G., Matteini P. & Modaet F. (2023). Impact of seed amplification assay and surface-enhanced Raman spectroscopy combined approach on the clinical diagnosis of Alzheimer's disease. *Translational Neurodegeneration*, 12: 35.

Díaz-Jaimes P., Bayona-Vásquez N.J., Adams D.H. & Uribe-Alcocer M. (2016). Complete mitochondrial DNA genome of bonnethead shark, *Sphyrna tiburo*, and phylogenetic relationships among main superorders of modern elasmobranchs. *Meta Gene*, 7: 48-55.

Domingos P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55: 78-87.

Gillis J.A. & Donoghue P.C. (2007). The homology and phylogeny of chondrichthyan tooth enameloid. *Journal of Morphology*, 268: 33-49.

Guérin J., Thiery S., Nyiri E., Gibaru O. & Boots B. (2021). Combining pretrained CNN feature extractors to enhance clustering of complex natural images. *Neurocomputing*, 423: 551-571.

Guidi T., Python L., Forasassi M., Cucci C., Franci M., Argenti F. & Barucci A. (2023). Egyptian Hieroglyphs Segmentation with Convolutional Neural Networks. *Algorithms*, 16: 79.

Hassoun S., Jefferson F., Shi X., Stucky B., Wang J. & Rosa E Jr. (2021). Artificial intelligence for biology. *Integrative and Comparative Biology*, 61: 2267-2275.

Hastie T., Tibshirani R. & Friedman J.H. (2009). The elements of statistical learning: data mining, inference, and prediction. Volume 2. 758 pp. Springer, New York.

Heuillet A., Couthouis F. & Díaz-Rodríguez N. (2021). Explainability in deep reinforcement learning. *Knowledge-Based Systems*, 214: 106685.

Ho M., Idgunji S., Payne J.L. & Koeshidayatullah A. (2023). Hierarchical multi-label taxonomic classification of carbonate skeletal grains with deep learning. *Sedimentary Geology*, 443: 106298.

Hou C., Lin X., Huang H., Xu S., Fan J., Shi Y. & Lv H. (2023). Fossil image identification using deep learning ensembles of data augmented multiviews. *Methods in Ecology and Evolution*, 14: 3020-3034.

Hou Y., Canul-Ku M., Cui X., Hasimoto-Beltran R. & Zhu M. (2021). Semantic segmentation of vertebrate microfossils from computed tomography data using a deep learning approach. *Journal of Micropalaeontology*, 40: 163-173.

Itaki T., Taira Y., Kuwamori N., Saito H., Ikehara M. & Hoshino T. (2020). Innovative microfossil (radiolarian) analysis using a system for automated image collection and AI-based classification of species. *Scientific Reports*, 10: 21136.

Kemp N.E. (1999). Integumentary system and teeth. *In* Hamlett W. (ed.), Sharks, Skates and Rays. *The Biology of Elasmobranch Fishes*. John Hopkins University Press, Baltimore: 43-68.

Kingma D.P. & Ba J. (2014). Adam: A method for stochastic optimization. *arXiv preprint,* arXiv: 1412.6980.

Kobak D. & Berens P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10: 5416.

Landini W., Altamirano-Sierra A., Collareta A., Di Celma C., Urbina M. & Bianucci G. (2017). The late Miocene elasmobranch assemblage from Cerro Colorado (Pisco Formation, Peru). *Journal of South American Earth Sciences*, 73: 168-190.

Landini W., Collareta A., Di Celma C., Malinverno E., Urbina M. & Bianucci G. (2019). The early Miocene elasmobranch assemblage from Zamaca (Chilcatay Formation, Peru). *Journal of South American Earth Sciences*, 91: 352-371.

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

Linardatos P., Papastefanopoulos V. & Kotsiantis S. (2020). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23: 18.

Lipovetsky S. & Conklin M. (2001). Analysis of regression in game theory approach. *Applied stochastic models in business and industry*, 17: 319-330.

Liu X. & Song H. (2020). Automatic identification of fossils and abiotic grains during carbonate microfacies analysis using Deep Convolutional Neural Networks. *Sedimentary Geology*, 410: 105790.

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: 1-22.

Lundberg S. (ed.) (2018). SHAP software. https://shap.readthedocs.io/en/latest/ (accessed 30.04.2024).

Lundberg S.M. & Lee S.I. (2017). A unified approach to interpreting model predictions. *In* Guyon I., Von Luxburg U., Bengio S., Wallach H., Fergus R., Vishwanathan S. & Garnett R. (eds), Advances in Neural Information Processing Systems 30 (NIPS 2017). 31st Conference on Neural Information Processing Systems, California: 1-10.

MacFadden J.W.V.P.B. & Antonenko P. (2023). Integrating AI Machine Learning into the Teaching of Paleontology Using Fossil Shark Teeth in Middle Schools. https://stelar.edc.org/projects/23884/profile/integrating-ai-machine-learning-teaching-paleontology-using-fossil-shark (accessed 30.04.2024).

Marchant R., Tetard M., Pratiwi A., Adebayo M. & de Garidel-Thoron T. (2020). Automated analysis of foraminifera fossil records by image classification using a convolutional neural network. *Journal of Micropalaeontology*, 39: 183-202.

Marret F. (2023). The impact of artificial intelligence systems in micropalaeontology. *Evolving Earth*, 1: 100022.

MHS Data. Megalodon or not megalodon dataset (ed.) (2022). *Megalodon* or not *Megalodon* dataset. https://universe.roboflow.com/mhs-data-e4pkl/megalodon-or-not-megalodon (accessed 30.04.2024).

Mimura K., Nakamura K., Yasukawa K., Sibert E.C., Ohta J., Kitazawa T. & Kato Y. (2024). Applicability of object detection to microfossil research: Implications from deep learning models to detect microfossil fish teeth and denticles using YOLO-v7. *Earth and Space Science*, 11: e2023EA003122.

Mumuni A. & Mumuni F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, 16: 100258.

Ng A.Y. (1997). Preventing "overfitting" of cross-validation data. *International Conference on Machine Learning*, 97: 245-253.

Niu Z.B., Jia S.Y. & Xu H.H. (2024). Automated graptolite identification at high taxonomic resolution using residual networks. *Iscience*, 27: 108549.

Panati C., Wagner S. & Brüggenwirth S. (2022). Feature relevance evaluation using grad-CAM, LIME and SHAP for deep learning SAR data classification. *In* 2022 23rd International Radar Symposium (IRS), *Proceedings of the IEEE*: 457-462.

Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M. & Duchesnay E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12: 2825-2830.

Perez V., Groff S., Hintermeister M. & MacFadden B.J. (2023). Exploration of Free Web-based Machine Learning Platforms for Paleontology Applications. *Society of Vertebrate Paleontology*, Program and Abstracts, 2023: 53.

Piazza G., Valsecchi C. & Sottocornola G. (2021). Deep learning applied to SEM images for supporting marine coralline algae classification. *Diversity*, 13: 640.

Pollerspöck J. & Straube N. (2018). An identification key to elasmobranch genera based on dental morphological characters. Part A: Squalomorph sharks (Superorder Squalomorphii). *Bulletin of Fish Biology*, 18: 77-105.

Raschka S. & Mirjalili V. (2019). Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. 770 pp. Packt Publishing, Mumbai.

Ribeiro M.T., Singh S. & Guestrin C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *In* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining: 1135-1144.

Richards B., Tsao D. & Zador A. (2022). The application of artificial intelligence to biology and neuroscience. *Cell*, 185: 2640-2643.

Roboflow 100 (ed.) (2023). Shark teeth Dataset. https://universe.roboflow.com/roboflow-100/shark-teeth-5atku (accessed 30.04.2024).

Sáez S. & Pequeño G. (2010). Taxonomic dental keys for the Chilean taxa of the Superorder Squalomorphi (Chondricthyes: Elasmobranchii). *Latin American Journal of Aquatic Research*, 38: 474-484.

Samek W., Wiegand T. & Müller K.R. (2017). Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. *arXiv preprint*, arXiv: 1708.08296.

Selvaraju R.R., Cogswell M., Das A., Vedantam R., Parikh D. & Batra D. (2020). Grad-CAM: visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128: 336-359.

Serena F., Abella A.J., Bargnesi F., Barone M., Colloca F., Ferretti F., Fiorentino F., Jenrette J. & Moro S. (2020). Species diversity, taxonomy and distribution of Chondrichthyes in the Mediterranean and Black Sea. *European Zoological Journal*, 87: 497-536.

Shapley L.S. (1953). A value for n-person games. *In* Shapley L. (ed.), Classics in Game Theory. Princeton University Press, Princeton: 307-317.

Shark sorting Dataset (ed.) (2022); https://universe.roboflow.com/shark-sorting/shark-sorting (accessed 30.04.2024).

Shark teeth Dataset (ed.) (2024); https://universe.roboflow.com/roboflow-100/shark-teeth-5atku (accessed 30.04.2024).

Shark Tooth Data Computer Vision Project (ed.) (2024); https://universe.roboflow.com/mhs-data-e4pkl/shark-tooth-data; 2023 (accessed 30.04.2024).

Shark Tooth Model Dataset (ed.) (2023); https://universe.roboflow.com/sharks/shark-tooth-model-9bbox Roboflow Universe (accessed 30.04.2024).

Shorten C. & Khoshgoftaar T.M. (2019). A survey on image data augmentation for deep learning. *Journal of big data*, 6: 1-48.

Simonyan K. & Zisserman A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint*, arXiv: 1409.1556.

Štrumbelj E. & Kononenko I. (2014). Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41: 647-665.

Tetard M., Marchant R., Cortese G., Gally Y., de Garidel-Thoron T. & Beaufort L. (2020). A new automated radiolarian image acquisition, stacking, processing, segmentation, and identification workflow. *Climate of the Past*, 16: 2415-2429.

Tetard M., Carlsson V., Meunier M. & Danelian T. (2023). Merging databases for CNN image recognition, increasing bias or improving results? *Marine Micropaleontology*, 185: 102296.

Tucker A.S. & Fraser G.J. (2014). Evolution and developmental diversity of tooth regeneration. *Seminars in Cell & Developmental Biology*, 25: 71-80.

Van der Maaten L. & Hinton G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9: 2579-2605.

Van Zyl C., Ye X. & Naidoo R. (2024). Harnessing eXplainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. *Applied Energy*, 353: 122079.

Vilone G. & Longo L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76: 89-106.

Wang B., Sun R., Yang X., Niu B., Zhang T., Zhao Y., Zhang Y., Zhang Y. & Han J. (2022). Recognition of rare microfossils using transfer learning and deep residual networks. *Biology*, 12: 16.

Wang H., Li C., Zhang Z., Kershaw S., Holmer L.E., Zhang Y., Wei K. & Liu P. (2022). Fossil brachiopod identification using a new deep convolutional neural network. *Gondwana Research*, 105: 290-298.

Wang H., Fu T., Du Y., Gao W., Huang K., Liu Z., Chandak P., Liu S., Van Katwyk P., Deac A., Anandkumar A., Bergen K., Gomes C.P., Ho S., Kohli P., Lasenby J., Leskovec J., Liu T.-Y., Manrai A., Marks D., Ramsundar B., Song L., Sun J., Tang J., Veličković P., Welling M., Zhang L., Coley C.W., Bengio Y. & Zitnik, M. (2023). Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972): 47-60.

White W.T., Mollen F.H., O'Neill H.L., Yang L. & Naylor G.J. (2023). Species in disguise: A new species of Hornshark from Northern Australia (Heterodontiformes: Heterodontidae). *Diversity*, 15: 849.

Wilmers J., Waldron M. & Bargmann S. (2021). Hierarchical microstructure of tooth enameloid in two lamniform shark species, *Carcharias taurus* and *Isurus oxyrinchus*. *Nanomaterials*, 11: 969.

Xu Y., Dai Z., Wang J., Li Y. & Wang H. (2020). Automatic recognition of palaeobios images under microscope based on machine learning. *IEEE Access*, 8: 172972-172981.

Yu C., Qin F., Li Y., Qin Z. & Norell M. (2022). CT segmentation of dinosaur fossils by deep learning. *Frontiers in Earth Science*, 9: 805271.

Yu C., Qin F., Watanabe A., Yao W., Li Y., Qin Z., Liu Y., Wang H., Jiangzuo Q., Hsiang A.Y. & Ma C. (2024). Artificial intelligence in paleontology. *Earth-Science Reviews*, 252: 104765.

Zeiler M.D. & Fergus R. (2014). Visualizing and understanding convolutional networks. *In* Computer Vision–ECCV 2014, *Proceedings of 13th European Conference*, Springer International Publishing, Zurich: 818-833.

Zhang T., Wang B., Li D., Niu B., Sun J., Sun Y., Yang X., Luo J. & Han J. (2020). Artificial intelligence identification of multiple microfossils from the Cambrian Kuanchuanpu Formation in southern Shaanxi, China. *Acta Geologica Sinica*, 94: 189-197.

Zhuang F., Qi Z., Duan, K., Xi D., Zhu Y., Zhu H., Xiong H. & He Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109: 43-76.

Manuscript submitted 2 May 2024
Revised manuscript accepted 2 October 2024
Published online 16 November 2024
Editor Silvia Danise