HISNET-FF: Hierarchical identification of species using a network with fused cranial and dental features

Zhong Cao[1,#], Qiu-Le Tang[2,#], Wei-Qi Zeng[1], Kun-Hui Wang[1,3], Quentin Martinez[4], Ze-Ling Zeng[5], Si-Ning Xie[5], Qiu-Qin Lu[5], Shi-Yun Liu[5], Xiao-Yun Zheng[5], Wen-Hua Yu[5], Jun-Jie Hu[5], Zhong-Zheng Chen[6], Shao-Ying Liu[7], Song Li[8], Fei-Yun Tu[9], Zi-Wen Hong[1], Ming Bai[10,11,12*], Kai He[5,*]

1. School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, Guangdong 510006, China
2. School of Computer Science and Cyber Engineering, Guangzhou University, Guangzhou, Guangdong 510006, China
3. School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, Beijing 101400, China
4. Staatliches Museum für Naturkunde Stuttgart, Stuttgart 70191, Germany
5. South China Biodiversity Research Center, School of Life Sciences, Guangzhou University, Guangzhou, 510006, China
6. Collaborative Innovation Center of Recovery and Reconstruction of Degraded Ecosystem in Wanjiang Basin Co-founded by Anhui Province and Ministry of Education, School of Ecology and Environment, Anhui Normal University, Wuhu, Anhui 241002, China
7. Sichuan Academy of Forestry, Chengdu, Sichuan 610081, China
8. State Key Laboratory of Genetic Evolution & Animal Models, Kunming Institute of Zoology, Chinese Academy of Sciences, Kunming, Yunnan 650023, China
9. Ministry of Education Key Laboratory for Ecology of Tropical Islands, Key Laboratory of Tropical Animal and Plant Ecology of Hainan Province, College of Life Sciences, Hainan Normal University, Haikou, Hainan 571158, China
10. State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
11. Academy of Plateau Science and Sustainability, Qinghai Normal University, Xining, Qinghai 810016, China
12. Hebei Key Laboratory of Animal Diversity, Langfang Normal University, Langfang, Hebei 065000, China

Corresponding Author:

Ming Bai, Email: baim@ioz.ac.cn

Kai He, Email: hekai@gzhu.edu.cn

#Authors contributed equally to this work

Abstract

Accurate species identification from mammalian craniodental features is essential but traditionally slow and requires specialized expertise. We address this by developing HISNET-FF, a deep learning framework featuring a dual-branch architecture to fuse global features from the cranium and local features from the teeth and auditory bullae. The network employs a hierarchical pipeline, first classifying to genus and then to species. Tested on a comprehensive image dataset of the Family Talpidae (18 genera, 51 species), HISNET-FF achieved exceptional accuracy at both the genus (99.6±0.4%) and species (96.5±1.3%) levels. This species-level accuracy significantly outperforms single-modality approaches, including both flat (up to 91.2±2.3% accuracy) and hierarchical (up to 93.9±2.1% accuracy) strategies. To enable a fully automated workflow, we also developed a YOLO-based tool that annotates diagnostic features with high performance, achieving 97.8% recall, 97.9% precision, and 81.5% mean average precision (mAP@[.50:.95]). This automation resulted in a minor drop in final identification accuracy of 1.9%. HISNET-FF thus provides a robust and highly accurate framework that can accelerate morphology-based research, with strong potential for broader application.

Keywords: Craniodental morphology; Deep learning; Feature fusion; Hierarchical classification; Species identification

INTRODUCTION

Taxonomy underpins numerous biological disciplines, including biodiversity, conservation, and genetics. However, this fundamental field faces significant challenges in the 21st century (Britz et al., 2020). A decreasing number of specialized taxonomists (Wägele et al., 2011), and an ever-growing demand for accurate species identification have created a pressing need for innovative approaches to species identification. While DNA barcoding has revolutionized species identification and classification (Moritz & Cicero, 2004), there remains a pressing requirement for efficient technologies to enhance the speed and accuracy of morphology-based species identification (Orr et al., 2021).

The classification as well as identification of mammals, particularly among speciose small mammal groups such as rodents, bats, and eulipotyphlans, have predominantly relied upon cranial and dental characteristics. This approach has historical roots dating back to pioneers such as Thomas Oldfield (Hinton, 1929), who recognized their diagnostic potential. Unlike external morphological traits, which can be highly variable within a taxon or exhibit similarities across distinct taxa, the detailed and subtle features on skull and teeth offer more consistent and reliable markers for species identification (Dayan et al., 2002).

Traditional approaches of species identification/delimitation rely on either discrete diagnostic characters or quantitative analyses, the latter include both traditional morphometrics, which is based on linear measurements, and geometric morphometrics, which captures complex variations in overall shape (Mutanen & Pretorius, 2007). These methodologies are invaluable for identifying extant species and are especially critical for fossil taxa, where teeth and skulls are often the only available materials**.** Despite their efficacy, morphology-based identification requires a high level of specialized expertise and is time-consuming, limiting the pace and scale of modern taxonomic research (Zamani et al., 2022).

In recent years, deep learning (DL) has emerged as a powerful tool for classification tasks across scientific disciplines (Lecun et al., 2015), including biological taxonomy (Badirli et al., 2023; Valan et al., 2019). The foundation for this revolution, particularly for image-based analysis, was laid by Convolutional Neural Networks (CNNs). This potential was realized in 2012 when AlexNet dramatically outperformed traditional methods such as support vector machines (SVMs) and random forests (RFs) in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). This demonstrated the profound ability of CNNs to learn diagnostic patterns directly from raw pixels, making them a natural fit for morphology-based species identification. This suggests they have broad potential for application not only in taxonomic research but also benefit research of ecology and evolution that relying on morphology-based species identification (Fortelius et al., 2002; Lyons et al., 2016). Success has been achieved in identifying plants and invertebrates, based on external morphology (Lee et al., 2015; Zhao et al., 2023). Surprisingly, its application to vertebrates (Gill et al., 2024), especially mammals, remains scarce in the literature. The few existing studies are often limited in scope, typically focusing only on discriminating between a small number of closely related genera or species (Miele et al., 2020; Pinho et al., 2022).

Our recent work introduced HIS-NET, a CNN-based method for species identification that achieved high accuracy rates of 95% (genus) and 90% (species) within the Family Talpidae (He et al., 2025). Analysis using Gradient-weighted Class Activation Mapping (Grad-CAM) confirmed that HIS-NET correctly focused on taxonomically informative regions of the skulls. However, the Grad-CAM heatmaps also revealed that HIS-NET utilized only a limited portion of the skull, and did not fully leverage the fine-grained diagnostic information present on the teeth. This observation suggested an opportunity for further enhancement.

Feature fusion is a technique that combines data from multiple sources or characteristic of different perspective to exploit their complementary information (Caci et al., 2013; Dai et al., 2021). In image classification, this often involves fusing "global" features that capture the overall context of an image, with "local" features that capture fine-grained details, thereby enhancing model accuracy (Peng et al., 2021). This dual-scale approach has delivered notable gains, such as improving tree species recognition by up to percentage 10 points through bark-and-leaf feature fusion (Bertrand et al., 2018).

In this study, we developed a new hierarchical species identification network using feature fusion (HISNET-FF). This network uses a dual-stream architecture to separately process and then combine embeddings from both the cranium, teeth and auditory bullae. Given the well-established principle that dental and auditory bullae morphology encodes subtle, species-specific traits, we hypothesized that our new dual-stream model would yield enhanced accuracy in species identification compared with a stand-alone model. We applied this network to the Family Talpidae, aiming to demonstrate its effectiveness in improving species recognition accuracy**.** Furthermore, we also developed and tested a method using a YOLO-based object detection model to streamline the time-consuming process of annotating key diagnostic features including teeth and auditory bullae.

MATERIALS AND METHODS

Specimen accession and photography

We focused our study on the mammal Family Talpidae, that includes 19 genera and 68 recognized species worldwide (Burgin et al., 2025). The photography collection utilized was consistent with He et al. (2025). We photographed specimens housed in natural history museums in China, Japan, Germany, Vietnam, and the USA. We followed a consistent protocol for our equipment and setup. All photographs were taken using a DSLR camera on a camera stand. Specimens were positioned on a level platform 10-30 cm below the lens, and we used a bubble level on both the camera and platform to ensure a consistent, perpendicular perspective. To minimize shadows on diagnostic features, particularly the teeth, two lights were positioned on both sides of the cranium, angled downwards at approximately 60 degrees. To further compensate for the variations in light sources, we manually adjusted the camera's F-stop (f/13–f/23) and ISO (200–1250) settings based on our experience, with the goal of achieving images that were as uniform as possible in quality and exposure. After removing specimens that we could not identify confidently, we collected 747 photographs of the ventral view of the cranium, of which 674 were intact and 73 were partially damaged. The data encompassed 18 genera and 51 species (including putative species; **Supplementary Appendix I**). The number of photos per species ranged from 3 to 48 (**Supplementary Table S1**).

Data labeling and manipulation and augmentation

Each image was annotated with the genus, species, and museum voucher information. For compatibility with the deep learning models, each cranial image was cropped and subsequently padded with black pixels on its shorter sides to create a square input (**Figure S1**). The full dataset was then split into training (594 images) and testing (153 images) sets with an 8:2 ratio (dataset CA0).

The shapes of teeth and the auditory bullae (AB) are crucial for talpid species identification and diagnosis. Therefore, to isolate these features, we created a dedicated dental image set (TA0) using a multi-step process. First, using LabelImg v.1.8.6, we annotated each image by placing a distinct rectangular bounding box around every individual tooth and auditory bulla. Each tooth was precisely identified and categorized as incisor (I1−I3), canine (C1), premolar (P1−P4), or molar (M1−M3) following the dental formulae provided by Wilson & Mittermeier (2018). Subsequently, we cropped each bounding box and stitched these regions together onto a new canvas, using their original coordinates to preserve their precise spatial relationships (**Supplementary Figure S1**). It should be noted that while the initial bounding boxes for adjacent teeth could have marginal overlaps, the cropped regions themselves were placed as distinct, non-overlapping images in the final composite. This final composite image was then padded to a square and added to the TA0 dataset, which was subsequently split into training and testing sets at an 8:2 ratio (**Supplementary Table S2**).

To reduce overfitting, we employed a series of image augmentation strategies (Maharana et al., 2022). Initially, we expanded the dataset CA0 fivefold by utilizing techniques such as 90-degree rotation, the addition of Gaussian noise, random information dropout, and random cropping with coverage of 200 pixels, resulting in 2 970 images (dataset CA5). Subsequently, we further expanded the dataset tenfold by applying five additional methods, including rotations of 45-degrees, random adjustments to image properties (brightness and hue), and horizontal translation (right shift by 100 pixels), local patch rotation (4 equal patches), yielding 5 940 images (dataset CA10). Finally, we introduced ten augmentation techniques: 180-and 270-degree rotation, two instances of random information dropout, the addition of complex noise (salt-and-pepper and Poisson noise), random adjustments to image properties (saturation and contrast), vertical translation (down shift by 100 pixels), and a broader range of random cropping and covering (covering 300 pixels). This expanded the dataset to 20 times its original size, resulting in a total of 11 880 images (dataset CA20). The same strategies were also applied to the teeth data set to obtain a tenfold augmented data set (dataset TA10) (**Supplementary Table S2**). All specimens were augmented uniformly.

Natural history collections are inherently imbalanced, a characteristic reflected in our dataset. Specifically, the number of photographs per species was highly uneven, ranging from 3 to 48, with 15 species represented by five or fewer specimens. To determine if this imbalance negatively affected the identification of rare species (e.g., those with <5 specimens), we created a specially balanced training set, CA-EQ. This set was generated using a differential augmentation strategy where underrepresented species received higher augmentation rates (from 6-fold to 50-fold). This process normalized the number of training images for each species to a consistent range of 100 to 250 (**Supplementary Table S3**).

Feature fusion model architecture design

Cranial photographs capture "global" morphological differences between species, while teeth often contain more subtle, "local," yet crucial diagnostic variations. To effectively integrate these dual-scale morphological features, we propose a core network employs a dual-branch network architecture. Our implementation of this architecture consists of three key modules (**Figure 1**). The process begins with the feature extraction module, which utilizes two independent and parallel branches: one branch is dedicated to analyzing all features present in the entire cranial image, while the other focuses exclusively on the fine-grained features from the teeth and AB. Following this independent extraction, the high-level feature outputs from both branches are combined in the feature fusion module. This module performs intermediate fusion, achieved via a straightforward concatenation process: the feature map from each branch is first condensed via pooling and flattened into a feature vector, and these two vectors are then joined end-to-end to create a single, unified feature vector. This unified vector, which represents the combined information from both cranial and dental sources, is passed to the classification module for identification.

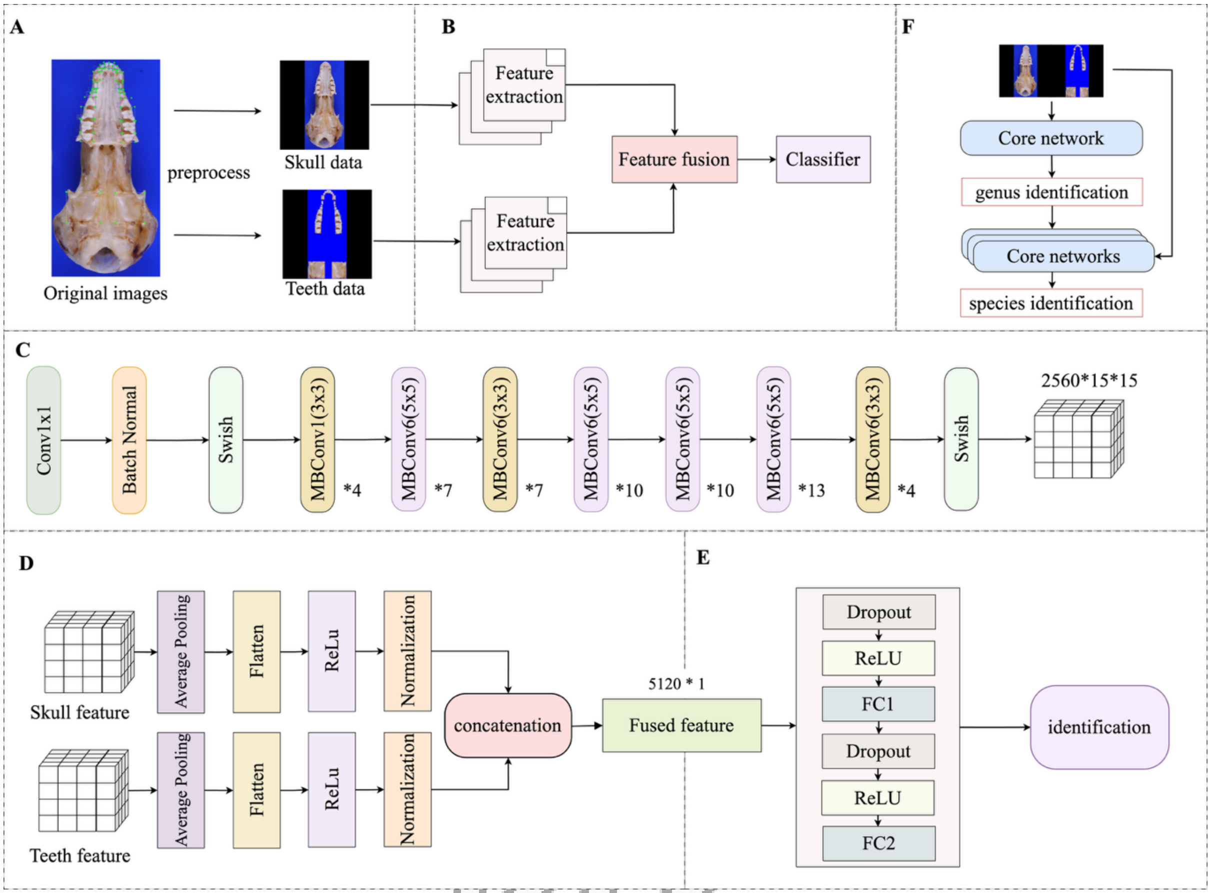

Figure 1. The architecture and workflow of the HISNET-FF model. The framework processes a single ventral cranial image through a multi-stage pipeline. A: Data Preprocessing - the original image is processed to generate two distinct inputs: one for global cranial features and a composite image for local dental and auditory bullae (AB) features. B: Core Network - two inputs are fed into a dual-branch network where features are extracted in parallel, fused, and then classified. C: Feature Extraction - each branch uses an EfficientNetB7-based architecture to produce a 2560×15×15 feature tensor. D: Feature Fusion - two feature tensors are processed (pooled, flattened, activated) and then concatenated into a single 5120-dimensional feature vector. E: Classification - this final vector is fed into a Multi-Layer Perceptron (MLP) classifier to yield the final identification. F: Hierarchical identification - a core network is first applied to identify genera, and for polytypic genera, a second specialized network is used to identify species.

Feature extraction module

Our first step was to choose the most effective CNN that could best extract key morphological features from the images. We conducted an evaluation using the ten-fold augmented cranial (CA10) dataset. We began by comparing various state-of-the-art networks, including EfficientNetB0 (Tan & Le, 2019), ResNet50 (Targ et al., 2016), ShuffleNetV2 (Zhang et al., 2018), MobileNetV2 (Howard, 2017), MnasNet (Tan et al., 2019), GoogleNet (Al-Qizwini et al., 2017), DenseNet121 (Iandola et al., 2014) as well as a Vision Transformer - Large with 32×32 patch size (ViT-L/32) (Zhai et al., 2022). We standardized the image resolution to 224×224 pixels for both training and testing datasets to enhance computational efficiency. To ensure a fair comparison of their transfer learning capability, all candidate models were initialized with their standard weights pre-trained on the ImageNet dataset (Krizhevsky et al., 2012) and fine-tuned all layers on our training data. The performance of each model was evaluated using the Top-1 accuracy metric (**Supplementary Text S1**). In this initial test, EfficientNetB0 performed best, achieving an accuracy of 88.9%, surpassing other networks by 1.3% to 10.5% (**Supplementary Table S4; Supplementary Figure S2**).

Next, we explored the entire EfficientNet family (from B0 to B7). Larger models in this series have more layers and can perceive finer details, though they require higher resolution images (Tan & Le, 2019). After evaluating each model with its recommended image resolution, we found that the most complex model, EfficientNetB7 (EB7), achieved the highest accuracy (**Supplementary S4; Supplementary Figure S2**). Therefore, we selected EB7 as our base feature extractor.

To identify the optimal augmentation level, we evaluated the EB7 model's accuracy on datasets with no (CA0), 5-fold (CA5), 10-fold (CA10), and 20-fold (CA20) augmentation. Both the CA10 and CA20 datasets achieved a peak accuracy of 91.5% (**Supplementary Figure S3**), while a specially balanced dataset (CA-EQ) did not offer an advantage (90.8%). Considering the lack of improvement beyond 10-fold augmentation, we selected the more computationally efficient tenfold strategy (CA10) for all subsequent analyses.

The core architecture of EB7 consists of eight stages including an initial convolutional block, known as the "Stem", followed by seven main stages that are built from repeating Mobile Inverted Bottleneck Convolution (MBConv) blocks (**Figure 1C**). The initial block captures basic visual patterns such as edges and textures. As data flows through the subsequent, more complex blocks, these simple elements are progressively combined to recognize abstract and taxonomically meaningful structures, such as the specific shape of a tooth. The entire sequence culminates in a final Swish activation function that refines the output from the last MBConv block. We extract the feature tensor immediately after this terminal activation. The resulting 2560×15×15 tensor represents the most distilled, high-level summary of the visual features the model has learned.

The Module for Feature Fusion

To test our hypothesis that combining cranial and dental information would improve accuracy, we designed a module to fuse the two tensors generated by the EB7 model (**Figure 1D**). First, we condense each 2560×15×15 data array down to 2560×1×1 (average pooling). The step is to shrink the data's spatial dimensions while preserving the most important "global" information. Next, transforms this tensor into a simple, one-dimensional list, or vector, with a size of 2560×1 (flatten). This is a necessary formatting step to prepare the data for the subsequent module. Each number in the feature list was passed through an activation function called the Rectified Linear Unit (ReLU) (Agarap, 2018). The ReLU function is a simple rule: if a feature value is positive, it passes through unchanged; if it is negative, it is set to zero. This can be summarized using the function $f(x) = \max(0, x)$. This allows the network to learn non-linear relationships inherent in biologic shapes. Afterward, we applied normalization to each feature list through MinMaxScaler to adjust feature values to the range [0, 1]. Finally, the two normalized lists are concatenated into a vector, with a size of 5120×1, which was then used to train a new, lightweight classification module.

The Module for classification

To optimize the classification module, we evaluated eight classifiers using the fused 5120×1 vectors from the CA10 and TA10 datasets: logistic regression, decision tree, random forest, multinomial and Gaussian Naïve Bayes classifier, support vector machine (SVM), single-layer perceptron (SLP), and multi-layer perceptron (MLP). While SLP, used in EfficientNet, comprises a Dropout layer and a fully connected layer (Tan & Le, 2019), MLP extends this structure with two additional ReLU activations, another Dropout layer, and a second fully connected layer (Kruse et al., 2022; Fig. 1F). Because this training step only involves the new classifier, the underlying EB7 feature extractors were frozen, and their weights were not updated. Our results showed that the MLP achieved the accuracy of 93.5%, higher the other classifiers (ranged from 92.8% to 69.9%; **Supplementary Figure S4**) and thus was selected as our classifier.

A hierarchical structure for identification

To account for the taxonomic structure of Talpidae, in which seven of the 18 studied genera are polytypic (comprising multiple species), we implemented a hierarchical classification pipeline. This approach uses a sequence of distinct, specialized classifiers to first identify the genus and then the species. The process begins with a primary genus-level classifier, trained to distinguish among all 18 genera. An input image is first assigned to its most likely genus. If that genus is monotypic, the identification process is complete. If the image is assigned to one of the seven polytypic genera (e.g., *Euroscaptor* or *Talpa*), it is then forwarded to a dedicated species-level classifier. To this end, we trained seven separate species-level classifiers, one for each polytypic genus. Each of these classifiers was trained exclusively on the image subset of the species within that specific genus (**Figure 1F**). This sequential design means that a misclassification at the genus level will prevent a correct final identification. However, as our genus classifier achieved 99.6% accuracy (up to one is mis-identified, see Result), this risk of error propagation is minimal.

Model Training, Implementation and Validation

To leverage robust, pre-existing feature representations, we initialized all models that was pre-trained on the ImageNet dataset (Krizhevsky et al., 2012). Following this initialization, we fine-tuned each model on our datasets. This fine-tuning process was performed independently to the baseline flat species classifiers, the genus-level classifiers for our hierarchical model, and seven distinct species-level classifiers, per polytypic genus.

During the model fine-tuning phase, we used the Stochastic Gradient Descent (SGD) optimizer, an algorithm that iteratively adjusts the model's parameters to minimize prediction error. The model's training process involved 50 epochs, meaning the network reviewed the entire training dataset 50 times to progressively learn the features. During this process, images were processed in groups of 16 (a batch size of 16). After each batch, the model's internal parameters were adjusted to improve its accuracy. The magnitude of these adjustments was controlled by a dynamic learning rate, which started at 0.01 and was reduced using a StepLR scheduler. This scheduler multiplied the learning rate by a factor of 0.8 every 10 epochs. To ensure the model learned generalizable patterns rather than simply memorizing the training data, we applied a regularization technique called dropout, which randomly ignored 50% of the network's neurons during each training step (**Supplementary Table S4**). Similarly, the classifiers in the module for classification such as SLP and MLP were trained using the same parameters except that we set the dropout rate to 0.7. All experiments were conducted using a GeForce RTX 3090 GPU (VRAM 24GB), within a Conda v.23.3 environment running Python v.3.9.

To validate our inherently hierarchical, feature fusion network, we benchmarked it against two single-modality baseline models trained on either the cranial (CA10) or dental (TA10) datasets exclusively. To isolate the contribution of the hierarchical method itself, these baseline models were evaluated using both a direct flat and a two-step hierarchical classification strategy. To evaluate the consistency of network performance and reduce the impact of data partitioning on the results, we employed a five-fold cross-validation approach. We specifically selected k=5 due to the highly imbalanced nature of our dataset, which contains numerous rare species (15 with ≤5 specimens). This choice provides larger test folds (20% of the data) compared to a higher k-value, ensuring that rare species are represented in each split, thus yielding more stable and reliable performance estimates (Wong & Yeh, 2020). In this protocol, the dataset was partitioned into five subsets of approximately equal size, with each subset used once for testing while the other four were used for training.

Given the varying number of images per species and the presence of partially damaged specimens, we investigated how these factors might affect identification accuracy. To assess this, we conducted correlation analyses using three machine learning regression models, namely, Decision Tree (Kotsiantis, 2013), Gradient Boosting (Bentéjac et al., 2021), and Random Forest (Paul et al., 2018).

Object detection-based automatic annotation for teeth

To streamline the time-consuming process of annotation, we developed a automated pipeline using the state-of-the-art YOLOv5 family of object detection models (Jocher et al., 2022). The objective was to train a model that could accurately replicate our manual annotation process.

We began by creating a ground-truth dataset (DS0), where each of the 594 cranial images in our training set was manually annotated in LabelImg with distinct bounding boxes for every individual tooth and auditory bulla (AB). To ensure the model's robustness, we then expanded this dataset tenfold to create DS10, using a variety of data augmentation methods including mirroring (up-down, left-right, and mixed), random interpolation resizes (600×600, 960×960, and 1600×1600 pixels), 90-degree rotations (clockwise and counterclockwise), and the addition of Gaussian noise (Wan et al., 2023). It is of note that YOLO training and testing does not require cropping the teeth and AB to generate a new image.

The evaluation was based on standard object detection metrics, including precision, recall, and mean average precision (mAP) (**Supplementary Text 1**). This involved two sequential steps. First, to select the best architecture, we compared several pre-trained YOLOv5 variants at a standardized 640×640 pixel resolution. The largest model, YOLOv5x, was chosen as it achieved the best overall performance (**Supplementary Table S5**). Second, recognizing that object detection accuracy is highly sensitive to input resolution, we benchmarked the selected YOLOv5x model across a range of resolutions (from 640×640 to 1600×1600). Our findings revealed a clear performance trade-off; we selected 1280×1280 as the optimal resolution as it achieved the highest score on the rigorous mAP@[.50:.95] metric (81.5%) with only a negligible sacrifice in peak recall (**Supplementary Table S5; Figure S5**).

Finally, the trained YOLOv5x model (at 1280×1280 resolution) was used to predict bounding box coordinates on our test set of cranial images. These automatically generated coordinates were then used to create a new dental image set, DT0, by following the exact same "crop and stitch" procedure used for the manually annotated TA0 dataset. Finally, to measure the downstream impact of this automation, we assessed all models that rely on dental imagery. We compared the species identification accuracy of the teeth-only and feature fusion networks when using the automatically generated DT0 dataset against their respective baseline performances with the manually annotated TA0 dataset.

RESULTS

We evaluated identification accuracy using cranial (CA10) and dental (TA10) datasets separately with EB7, as well as a combined approach utilizing our feature fusion network that integrates both cranial and dental data. The flat models trained individually on the cranial and dental datasets achieved accuracies of 90.5±1.6% and 91.2±2.3%, respectively. In comparison, the flat feature fusion-based model achieved an average accuracy of 92.5±1.8%, though the improvement is marginal (paired t-test: p>0.06) (**Table 1**).

Table 1. Identification accuracy achieved using EB7 trained individually on the cranium and teeth, as well as feature fusion networks utilizing both cranium and teeth. We employed both flat and hierarchical strategies for identification. Bold indicates the accuracy obtained using HISNET-FF.

		EB7-cranium accuracy (%)	EB7-teeth accuracy (%)	feature fusion accuracy (%)
Flat ident.	All species	90.5±1.6	91.2±2.3	92.9±1.8
Hierarchical identification	All genera	98.8±0.9	98.3±1.2	**99.6±0.4**
	All species	93.9±2.1	93.2±2.4	**96.5±1.3**
	Euroscaptor (9 spp.)	79.7±2.9	75.1±8.5	**88.3±6.1**
	Mogera (9 spp.)	89.0±3.6	89.0±6.0	**93.2±3.7**
	Parascaptor (3 spp.)	92.7±7.4	94.70±4.4	**96.4±8.1**
	Scapanus (4 spp.)	98.0±4.5	98.0±4.5	**98.0±4.5**
	Scaptonyx (3 spp.)	79.7±6.3	86.4±2.2	**91.1±4.1**
	Talpa (7 spp.)	100.0±0.0	96.3±3.2	**100.0±0.0**
	Uropsilus (5 spp.)	93.3±7.0	88.3±13.9	**95.0±7.2**

Under the hierarchical strategy, all models achieved exceptional genus-level accuracy, with HISNET-FF performing best (99.6±0.4%; **Supplementary Table S6**). At the species level, all three hierarchical models outperformed their respective flat identification counterparts, improving the accuracy by 2.0—4.0% (paired t-test: EB7-cranium, p=0.021; EB7-teeth, p=0.008; HISNET-FF, p=0.0035). The HISNET-FF model maintained consistently high performance (mean=96.5±1.3%; **Figure 2)** significantly outperformed both the hierarchical cranium (p=0.027) and the teeth (p=0.029) model. This superiority was consistent across all seven polytypic genera, where specific error patterns are visualized in confusion matrices (**Figure 3**).

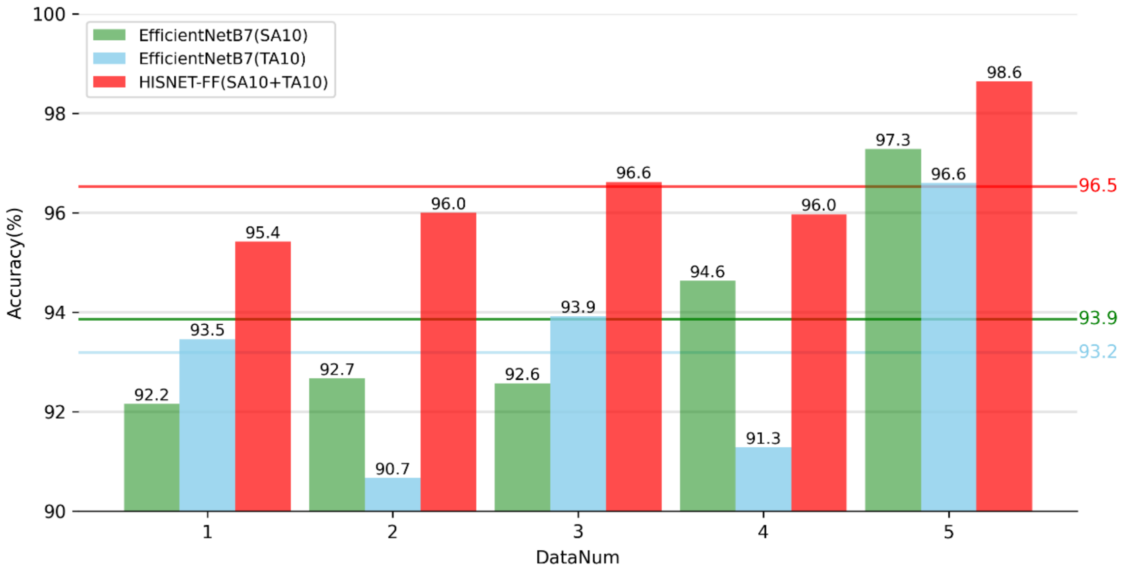

Figure 2. Comparison of species-level identification accuracy for the feature-fusion (HISNET-FF) and single-modality (EB7-cranium, EB7-teeth) hierarchical models across five cross-validation folds. Each bar represents the accuracy achieved in one of the five folds for the respective model. The horizontal lines denote the mean accuracy for each model across all folds.

To understand the mechanism behind HISNET-FF's superior accuracy, we analyzed the misclassified specimens from our five-fold cross-validation (**Supplementary Table S7**). Overall, 26 out of 747 specimens were misidentified, with all errors occurring within polytypic genera, primarily talpine *Euroscaptor* and *Mogera*, which might be due to subtle interspecific morphological differentiation that is characteristic of these diverse genera. A comparative analysis of errors between HISNET-FF and the hierarchical single-modality models trained on cranium and teeth revealed benefits of feature fusion. Where at least one of the two single-modality models made a correct identification, HISNET-FF also produced the correct classification in almost every case, with only two exceptions. On the other hand, we identified five specific cases where both the hierarchical EB7-cranium and teeth models failed, HISNET-FF made the correct identification.

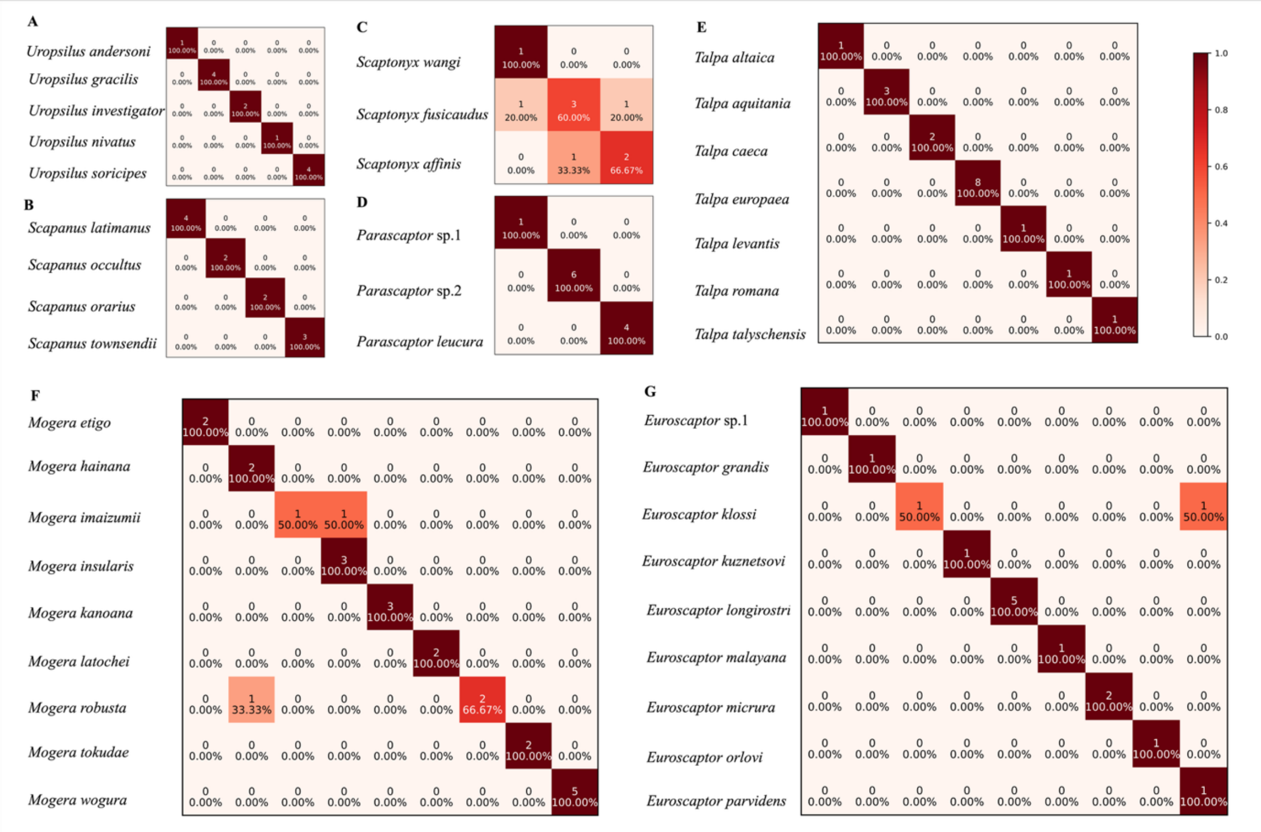

Figure 3. Confusion matrix heatmaps detailing the species-level identification performance of HISNET-FF within the seven polytypic genera. Each panel shows the classification results for a specific genus from a single test set, illustrating typical error patterns. The color intensity corresponds to the accuracy of the species identification.

To assess whether our model was biased against rare species, we tested for a correlation between sample size and identification accuracy. We found no evidence of a significant negative relationship (Spearman's ρ=–0.188, p=0.187), indicating that the model is robust to the inherent data imbalance. For instance, of the 14 species with six or fewer specimens, all but one (*Uropsilus nivatus*, 80% accuracy) were classified perfectly (**Supplementary Table S8**). We therefore conclude that sample size is not a primary limiting factor for achieving high identification accuracy with our framework.

A significant portion of small mammal skulls in museum collections are partially damaged due to the use of snap traps during collection, with about 10% affected in our case. HISNET-FF achieved 91.6% accuracy for damaged cranium compared to 97.0% for intact ones (**Supplementary Table S9; Supplementary Figure S6**), indicating a modest but noticeable impact of damage on identification accuracy. Despite this reduction, HISNET-FF's performance significantly surpasses EB7 individually trained on craniums and teeth, which yields 80.8% and 78.1% accuracy, respectively.

Building on these individual analyses, we investigated the combined effect of sample size and specimen damage on accuracy using three machine learning-based correlation analyses. The results revealed a synergistic effect: while each factor alone had a low correlation with accuracy ($R^2<0.3$; **Supplementary Table S10**), their combined influence was more substantial ($R^2>0.4$). This interaction was strongest for the cranial-based and HISNET-FF models ($R^2>0.5$).

Automatic annotation of teeth streamlining species identification

We implemented YOLOv5 to automate the annotation of teeth and auditory bullae (AB). After comparing several variants, the YOLOv5x model at a 1280x1280 resolution was selected for its superior performance (**Supplementary Figure S7**). This optimized model proved highly effective, achieving a precision of 97.9%, a recall of 97.8%, and a rigorous mAP@[.50:.95] of 81.5% (**Supplementary Table S5**). The lowest recall rates were for the AB (95.8%), the second upper premolar (P2, 96.8%), and the second upper incisor (I2, 96.9%), which is expected given that AB are often damaged and I2/P2 are the smallest teeth (**Supplementary Table S11**).

To evaluate the downstream effect of our automatic annotation strategy, we compared the final species identification accuracy of our network configurations when using manually annotated dental images (TA0) versus those generated by our automated YOLOv5x pipeline (DT0). The results show that for HISNET-FF, using the automated annotations yielded a species identification accuracy of 93.5%, representing a marginal decrease from the 95.4% achieved with manual annotations. This performance drop can be attributed to three additional misclassifications unique to the automated dataset; all seven errors from the manual set were replicated in the automated results. automated This trend of a slight performance reduction was consistent across all tested approaches (**Figure 4).**

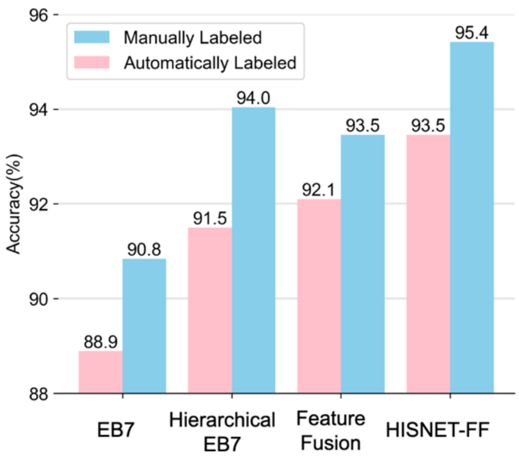

Figure 4. Comparison of species identification accuracy using manually versus automatically annotated dental images. The grouped bar chart displays the final accuracy for four different model configurations. For each configuration, performance is compared when using manually labeled dental images against those generated by our automated YOLOv5x pipeline.

To assess the potential broad application of our automatic annotation method, we tested it on diverse insectivorous mammals not trained previously. These included a recently described talpid mole (*Alpiscaptulus medogensis*), five erinaceids (*Erinaceidae*), five shrews (*Soricidae*), a tree shrew (*Scandentia*) as well as three *Afrotherian* species. YOLOv5x model effectively recognized 92.0% of the teeth and AB (**Supplementary Figure S8**), though a high percentage of teeth were not assigned to the proper labels (recall=0.46, precision=0.41, mAP@0.5=0.39, mAP@[.50:.95]=0.33).

DISCUSSION

Our results establish that an integrated framework combining feature fusion and a hierarchical strategy is essential for overcoming the inherent limitations of standard single-stream CNNs in complex morphological identification. Such models are limited in their ability to resolve global cranial architecture while capturing fine-grained dental features, often failing to register critical diagnostic details like cusps and cingula (Lin et al., 2021; Singha et al., 2024). Our dual-stream fusion architecture directly addresses this by creating a more comprehensive, dual-scale representation.

The efficacy of this approach is underscored by several analyses. First, Grad-CAM heatmaps reveal that a cranium-only model is effectively blind to the dental region, confirming the need for a dedicated feature stream (**Supplementary Figure S9**). Second, our error analysis shows that the fusion model successfully leverages complementary information: it not only reinforces correct classifications when one of the single-modality models succeeds but can also discern novel diagnostic patterns to correct instances where both baselines fail (**Supplementary Table S7**). Furthermore, HISNET-FF shows strong performance on rare species (13 of 14 species with≤6 images were identified perfectly), suggesting our approach is not compromised by class imbalance which might benefit from pretraining on ImageNet. However, we consider this result preliminary and believe more fine-grained experiments are needed to definitively validate the model's performance on rare taxa. In conclusion, HISNET-FF achieves its superior performance by effectively integrating non-overlapping diagnostic information combined with a hierarchical strategy to produce a more accurate classification.

We recently developed HIS-NET, another EfficientNet-based hierarchical classifier that utilized up to four different cranial and mandibular views per specimen (dorsal, ventral, lateral cranium, and lateral mandible; He et al., 2025). While both models share a hierarchical strategy, HISNET-FF introduces a novel architecture: it replaces the multi-view input with a single-image workflow, achieving superior accuracy through the feature fusion. The high performance of HISNET-FF on talpids suggests its potential for broader application, particularly for other speciose mammals (e.g., rodents, bats and primates) where diagnostic characters present in craniodental morphology.

The success of this feature-fusion framework opens up several avenues for future research. An immediate next step is to synthesize the strengths of our prior multi-view HIS-NET and the current HISNET-FF into a more comprehensive multi-view feature fusion architecture using a multi-task approach (Liu et al., 2019). Such a model would integrate features from dorsal, ventral, and lateral views simultaneously, creating a holistic morphological representation. Furthermore, beyond just teeth and auditory bullae, other diagnostic regions could be annotated, and dental features themselves could be refined into their constituent components, such as the trigonid and talonid, or even individual cusps. Ultimately, creating a holistic taxonomic identification system would require moving beyond the fusion of image-based features. This would likely involve leveraging more advanced architectures, such as Large Language Models (LLMs), to integrate the morphological outputs from our network with disparate data types such a DNA sequences, distributional and ecological information (Pyron, 2023).

While powerful, the current framework is fundamentally an identification tool for known species. It is warranted to move beyond the current supervised learning paradigm to develop models capable of Novel Category Discovery (NCD)(Vaze et al., 2022), which not only recognize known species but also to flag novel or anomalous specimens that fall outside the known morphological space, thereby transforming this tool from a simple identifier into an engine for taxonomic discovery (Badirli et al., 2023).

Our automated annotation tool, powered by YOLOv5x, proves to be a highly effective and practical component of our workflow. The use of automated annotations results in a final species identification accuracy of 93.5%, which is an acceptable trade-off compared to the 95.4% achieved with labor-intensive manual labeling. The implications of this minor accuracy drop should be considered in context. Our error analysis reveals that the vast majority of misidentifications are between closely related, congeneric species (genus level accuracy 99.6%). While this requires caution for formal taxonomic work, this level of precision is often sufficient for many large-scale ecological and macro-evolutionary studies where genus-level patterns are of primary interest. Furthermore, as discussed previously, this residual error rate could likely be reduced even further by incorporating non-morphological data such as DNA sequences and geographic distribution into a more integrative framework.

Our tests of the automated annotation tool on untrained insectivorous mammals underscored both its potential and its current limitations. The model proved highly effective at the general task of localizing craniodental features, detecting 92% of teeth and auditory bullae. However, its performance on classifying these features to specific tooth positions was poor (precision: 41%; recall: 46%), demonstrating that the model, trained exclusively on talpids, lacks broad taxonomic generalizability. This limitation is not a conceptual flaw, but a data-driven one. We are confident that by incorporating a diverse array of annotated specimens from other key mammalian orders (e.g., rodents, bats, shrews), the YOLOv5 framework can learn the varied dental formulas and morphologies necessary for high-accuracy, cross-family annotation.

The potential scientific impact of our tools is notable. Large-scale research in evolutionary biology and ecology often relies on numerous specimens, yet is frequently bottlenecked by the laborious process of manual identification (Pineda-Munoz et al., 2021; Saarinen & Lister, 2023). Having demonstrated high efficacy on a challenging taxonomic group, our automated species identification approach will accelerate research and enable new, large-scale, data-driven investigations.

DATA AVAILABILITY

The data and code of HISNET-FF is freely available through https://github.com/Onlyroad2n/HISNET-FF

SUPPLEMENTARY DATA

Supplementary data to this article can be found online.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHORS' CONTRIBUTIONS

ZC, MB and KH designed and supervised the project. KH, WHY, ZZC, SYL, SL and FYT collected specimens and prepare skull for photographs. KH and QM photographed the specimens. ZC, QLT, WQZ, KHW, and ZWH performed deep learning analyses. ZLZ, QQL, SYL, XYZ and SNX conducted data annotation for deep learning. JJH conduct statistical analyses.

ACKNOWLEDGEMENTS

We thank the five anonymous reviewers for their professional review, constructive suggestions, and helpful corrections that have greatly strengthened this paper. We thank museum curators and staff at the Guangdong Institute of Zoology (JH Zheng), National Museum of Natural Science at Taichung (YJ Chen), American Museum of Natural History (NB Simmons, E Westwig), Smithsonian National Museum of Natural History (KM Helgen, D Lunde), Field Museum (L Heaney), Museum of Comparative Zoology (HE Hoekstra and J Chupasko), National Museum of Nature and Science of Japan (S-I Kawada), Hokkaido University Natural History Museum (T Tsubota), Institute of Ecology and Biological Resources of Vietnam (TS Nguyen), the State Museum of Natural History Stuttgart (E Amson, C Leidenroth, S Merker) and the University of Montpellier (PH Fabre) for allowing us to examine specimens and take photos of specimens under their curation. We thank J Decher, C Montermann for kindly providing us photos of *Galemys*, and T Martin, C Steinweg for kindly providing us photos of *Desmana*.

Foundation items:

This work was supported by the National Natural Science Foundation of China (32170452), GuangDong Basic and Applied Basic Research Foundation (2022B1515020033), Key Program of the National Natural Science Foundation of China Regional Innovation and Development Joint Fund (U23A20161), the Open Project of Ministry of Education Key Laboratory for Ecology of Tropical Islands, Hainan Normal University (HNSF-OP-2024-3), the Guangzhou Higher Education Teaching Quality and Teaching Reform Engineering Special Talent

Training Plan Project (2022ZXRCPR007), and the Bundesministerium für Bildung und Forschung (BMBF Project KI-Morph 05D2022), National Key R&D Program of China (2022YFC2601200) .

References

Agarap AF. 2018. Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375.

Al-Qizwini M, Barjasteh I, Al-Qassab H, et al. 2017. Deep learning algorithm for autonomous driving using googlenet. Pages 89-96 in Proc. 2017 IEEE intelligent vehicles symposium (IV). IEEE.

Badirli S, Picard CJ, Mohler G, et al. 2023. Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. Methods Ecol Evol., **14**(6): 1515-1530.

Bentéjac C, Csörgő A, Martínez-Muñoz G. 2021. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, **54**(3): 1937-1967.

Bertrand S, Ben Ameur R, Cerutti G, et al. 2018. Bark and leaf fusion systems to improve automatic tree species recognition. Ecological Informatics, **46**: 57-73.

Britz R, Hundsdoerfer A, Fritz U. 2020. Funding, training, permits—the three big challenges of taxonomy. Megataxa, **1**: 49-52.

Burgin C, Zijlstra J, Becker M, et al. 2025. How many mammal species are there now? Updates and trends in taxonomic, nomenclatural, and geographic knowledge.

Caci G, Biscaccianti AB, Cistrone L, et al. 2013. Spotting the right spot: computer-aided individual identification of the threatened cerambycid beetle Rosalia alpina. J. Insect Conserv., **17**: 787-795.

Dai Y, Gieseke F, Oehmcke S, et al. 2021. Attentional feature fusion. Pages 3560-3569 in Proc. Proceedings of the IEEE/CVF winter conference on applications of computer vision.

Dayan T, Wool D, Simberloff D. 2002. Variation and covariation of skulls and teeth: modern carnivores and the interpretation of fossil mammals. Paleobiology, **28**(4): 508-526.

Fortelius M, Eronen J, Jernvall J, et al. 2002. Fossil mammals resolve regional patterns of Eurasian climate change over 20 million years. Evol. Ecol. Res., **4**(7): 1005-1016.

Gill KS, Gupta R, Malhotra S, et al. 2024. Classification of Reptiles and Amphibians Using Transfer Learning and Deep Convolutional Neural Networks. Pages 1-5 in Proc. 2024 IEEE 9th International Conference for Convergence in Technology (I2CT).

He K, Li A, Martinez Q, et al. 2025. Sky islands of Southwest China. II: Unraveling hidden species diversity of talpid moles using phylogenomics and skull-based deep learning. bioRxiv: 2025.2003.2006.641773.

Hinton MaC. 1929. MR. M. R. Oldfield Thomas, F.R.S. Nature, **124**(3116): 101-102.

Howard AG. 2017. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861.

Iandola F, Moskewicz M, Karayev S, et al. 2014. Densenet: Implementing efficient convnet descriptor pyramids. arXiv preprint arXiv:1404.1869.

Jocher G, Chaurasia A, Stoken A, et al. 2022. ultralytics/yolov5: v6. 2-yolov5 classification models, apple m1, reproducibility, clearml and deci. ai integrations. Zenodo.

Kotsiantis SB. 2013. Decision trees: a recent overview. Artificial Intelligence Review, **39**(4): 261-283.

Krizhevsky A, Sutskever I, Hinton GE. 2012. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, **25**.

Kruse R, Mostaghim S, Borgelt C, et al. 2022. Multi-layer perceptrons. In Computational intelligence: a methodological introduction: Springer, 53-124.

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

Lee SH, Chan CS, Wilkin P, et al. 2015. Deep-plant: Plant identification with convolutional neural networks. Pages 452-456 in Proc. 2015 IEEE international conference on image processing (ICIP). IEEE.

Lin B, Su H, Li D, et al. 2021. PlaneNet: an efficient local feature extraction network. PeerJ Computer Science, **7**: e783.

Liu M, Zhang J, Adeli E, et al. 2019. Joint Classification and Regression via Deep Multi-Task Multi-Channel Learning for Alzheimer's Disease Diagnosis. IEEE Trans. Biomed. Eng., **66**(5): 1195-1206.

Lyons SK, Amatangelo KL, Behrenstneyer AK, et al. 2016. Holocene shifts in the assembly of plant and animal communities implicate human impacts. Nature, **529**(7584): 80-U183.

Maharana K, Mondal S, Nemade B. 2022. A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, **3**(1): 91-99.

Miele V, Dussert G, Cucchi T, et al. 2020. Deep learning for species identification of modern and fossil rodent molars. bioRxiv: 2020.2008.2020.259176.

Moritz C & Cicero C. 2004. DNA barcoding: Promise and pitfalls. PLoS Biol., **2**(10): 1529-1531.

Mutanen M & Pretorius E. 2007. Subjective visual evaluation vs. traditional and geometric morphometrics in species delimitation: a comparison of moth genitalia. Syst. Entomol., **32**(2): 371-386.

Orr MC, Ferrari RR, Hughes AC, et al. 2021. Taxonomy must engage with new technologies and evolve to face future challenges. Nature Ecology & Evolution, **5**(1): 3-4.

Paul A, Mukherjee DP, Das P, et al. 2018. Improved Random Forest for Classification. IEEE Transactions on Image Processing, **27**(8): 4012-4024.

Peng Z, Huang W, Gu S, et al. 2021. Conformer: Local features coupling global representations for visual recognition. Pages 367-376 in Proc. Proceedings of the IEEE/CVF international conference on computer vision.

Pineda-Munoz S, Wang Y, Lyons SK, et al. 2021. Mammal species occupy different climates following the expansion of human impacts. Proceedings of the National Academy of Sciences, **118**(2): e1922859118.

Pinho C, Kaliontzopoulou A, Ferreira CA, et al. 2022. Identification of morphologically cryptic species with computer vision models: wall lizards (Squamata: Lacertidae: *Podarcis*) as a case study. Zool. J. Linn. Soc., **198**(1): 184-201.

Pyron RA. 2023. Unsupervised machine learning for species delimitation, integrative taxonomy, and biodiversity conservation. Mol. Phylogenet. Evol., **189**: 107939.

Saarinen J & Lister AM. 2023. Fluctuating climate and dietary innovation drove ratcheted evolution of proboscidean dental traits. Nature Ecology & Evolution, **7**(9): 1490-1502.

Singha T, Pham D-S, Krishna A. 2024. Effi-Seg: Rethinking EfficientNet Architecture for Real-Time Semantic Segmentation. Pages 55-68 in Proc. Neural Information Processing. Springer Nature Singapore, Singapore.

Tan M, Chen B, Pang R, et al. 2019. Mnasnet: Platform-aware neural architecture search for mobile. Pages 2820-2828 in Proc. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Tan M & Le Q. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. Pages 6105-6114 in Proc. International conference on machine learning. PMLR.

Targ S, Almeida D, Lyman K. 2016. Resnet in resnet: Generalizing residual architectures. arXiv preprint arXiv:1603.08029.

Valan M, Makonyi K, Maki A, et al. 2019. Automated Taxonomic Identification of Insects with Expert-Level Accuracy Using Effective Feature Transfer from Convolutional Networks. Syst. Biol., **68**(6): 876-895.

Vaze S, Han K, Vedaldi A, et al. 2022. Generalized category discovery. Pages 7492-7501 in Proc. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Wägele H, Klussmann-Kolb A, Kuhlmann M, et al. 2011. The taxonomist - an endangered race. A practical proposal for its survival. Frontiers in Zoology, **8**(1): 25.

Wan D, Lu R, Xu T, et al. 2023. Random interpolation resize: A free image data augmentation method for object detection in industry. Expert Systems with Applications, **228**: 120355.

Wilson DE & Mittermeier RA. 2018. Handbook of the mammals of the world: Insectivores, Sloths and Colugos. Editon. Barcelona Lynx Edicions.

Wong T-T & Yeh P-Y. 2020. Reliable Accuracy Estimates from k-Fold Cross Validation. IEEE Transactions on Knowledge and Data Engineering, **32**(8): 1586-1594.

Zamani A, Dal Pos D, Fric ZF, et al. 2022. The future of zoological taxonomy is integrative, not minimalist. Syst. Biodivers., **20**(1): 1-14.

Zhai X, Kolesnikov A, Houlsby N, et al. 2022. Scaling vision transformers. Pages 12104-12113 in Proc. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.

Zhang X, Zhou X, Lin M, et al. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. Pages 6848-6856 in Proc. Proceedings of the IEEE conference on computer vision and pattern recognition.

Zhao Z, Lu Y, Tong Y, et al. 2023. PENet: A phenotype encoding network for automatic extraction and representation of morphological discriminative features. Methods Ecol Evol., **14**(12): 3035-3046.

Supplementary text: The evaluation metrics used in this study

To validate the performance of the method proposed in this paper, we adopted four commonly used evaluation metrics: Top-1 accuracy, precision, recall, and mean average precision (mAP@0.5) and mAP@[.5:.95].

1. Top-1 accuracy

This metric measures the standard classification accuracy of the species identification models. An image is considered correctly classified if the single class with the highest predicted probability (the "top 1" prediction) matches the true label of the specimen. The final accuracy is the ratio of correctly classified images to the total number of images in the test set.

$$\text{Top1 accuracy} = \frac{\text{TP+TN}}{\text{Total number of samples}} \tag{1}$$

In this formula, TP refers to true positives and TN refers to true negatives.

2. **Precision**

Precision evaluates the accuracy of the object detection model's predictions. It answers the question that of all the bounding boxes the model predicted, what fraction were correct. A high precision score indicates a low rate of false positives (i.e., the model rarely detects objects that are not actually there).

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{2}$$

In this formula, TP refers to true positives and FP refers to false positives.

3. **Recall**

Recall evaluates the completeness of the object detection model's predictions. It answers the question that of all the actual objects that exist, what fraction did the model successfully find? A high recall score indicates a low rate of false negatives (i.e., the model rarely misses existing objects).

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{3}$$

In this formula, TP refers to true positives and FN refers to false negative.

4. **Mean average precision (mAP)**

mAP is the primary metric for evaluating the overall performance of an object detection model, as it provides a single value that summarizes both precision and recall across all object classes. It is calculated by averaging the average precision (AP) over all classes.

mAP@0.5: This metric calculates the mAP using a fixed Intersection over Union (IoU) threshold of 0.50. An IoU of 0.50 means a predicted bounding box is only considered a True Positive if its overlap with the ground-truth box is 50% or greater. This metric provides a good measure of the model's general detection capability.

$$mAP_{50} = \frac{1}{C}\sum_{i=1}^{C} AP_i(\mathrm{IoU} = 0.5) \tag{4}$$

mAP@[.5:.95]: This is a more rigorous and comprehensive metric. It calculates the mAP at ten different IoU thresholds (from 0.50 to 0.95, in steps of 0.05) and then averages these values. It rewards models that produce more precise and tightly-fitting bounding boxes, providing a more thorough assessment of localization accuracy.

$$mAP_{50-95} = \frac{1}{10}\sum_{i=1}^{10} AP_i(\mathrm{IoU} = 0.5 + 0.05\mathrm{x}(i-1)) \tag{5}$$

Supplementary figures

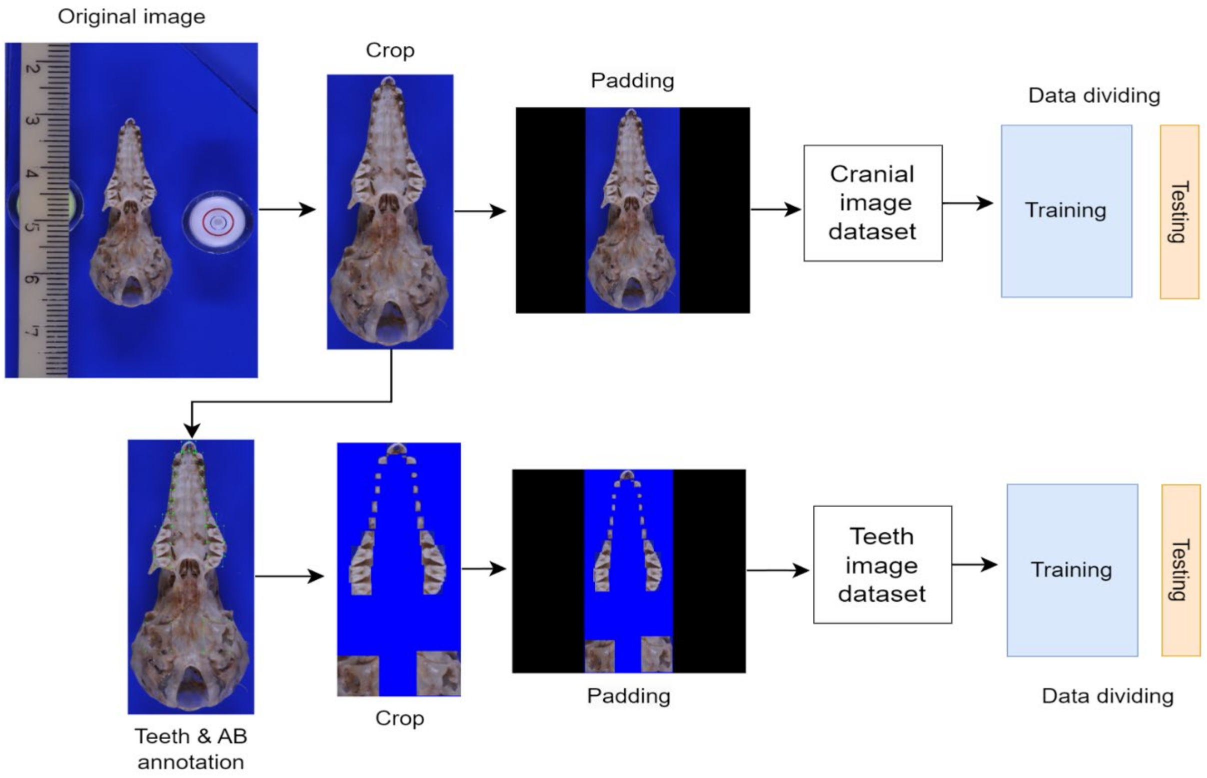

Figure S1. Workflow for the preparation of the cranial and dental image datasets. The diagram illustrates the preprocessing pipeline used to generate two parallel datasets from the original specimen photographs. The original image is cropped to isolate a single cranium, removing the background. The resulting rectangular image is then padded with black pixels to create a square aspect ratio. This final image becomes part of the "Cranial image dataset," which is subsequently divided into training and testing sets. The Teeth and Auditory bullae (AB) Dataset starting with the same cropped cranial image, all teeth and auditory bullae are first manually annotated. These annotated regions are then isolated by masking the remainder of the cranium. This masked image is also padded to create a square aspect ratio, forming the "Teeth image dataset." This dataset is then similarly divided into training and testing sets.

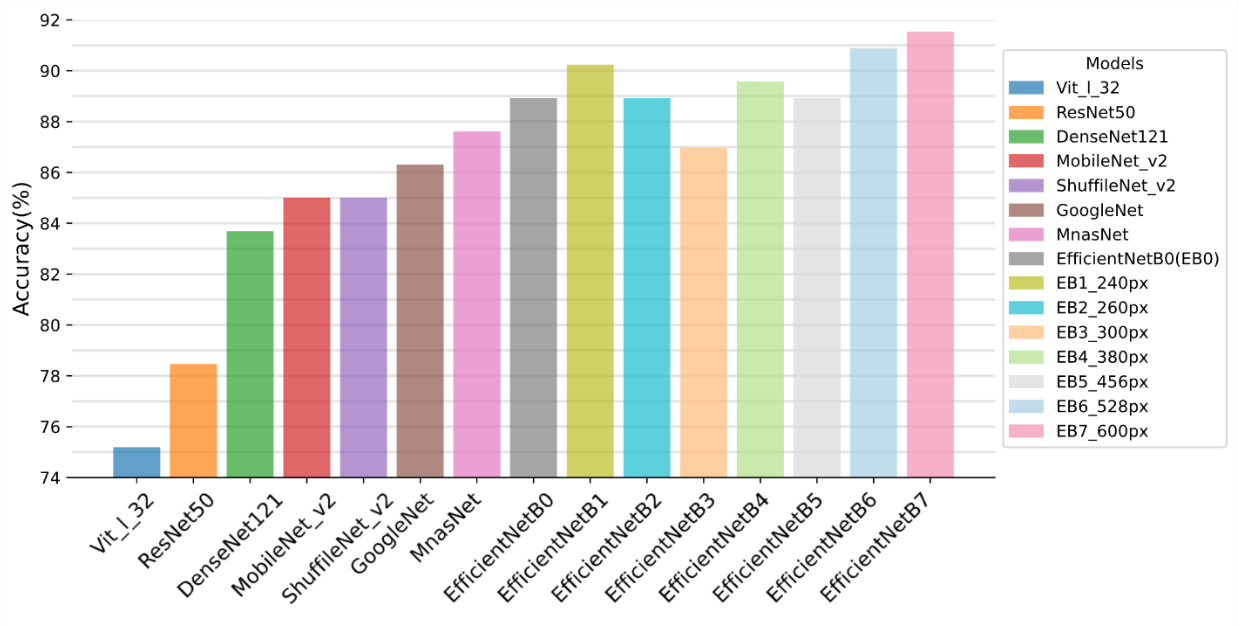

Figure S2. Performance comparison of different Convolutional Neural Network (CNN) architectures for feature extraction. The bar chart displays the Top-1 species identification accuracy for various state-of-the-art models. All models were fine-tuned on the ten-fold augmented cranial dataset (CA10). For EfficientNetB0, ResNet50, ShuffleNetV2, MobileNetV2, MnasNet, GoogleNet, DenseNet121 and Vit_l_32, we used images with resolution of 224×224 pixels. For EfficientNet B0-B7 series, each model was evaluated using its officially recommended input image resolution, ranging from 224×224 pixels to 600×600 pixels. The results show that the EfficientNet family, particularly EfficientNetB7, provided the highest accuracy.

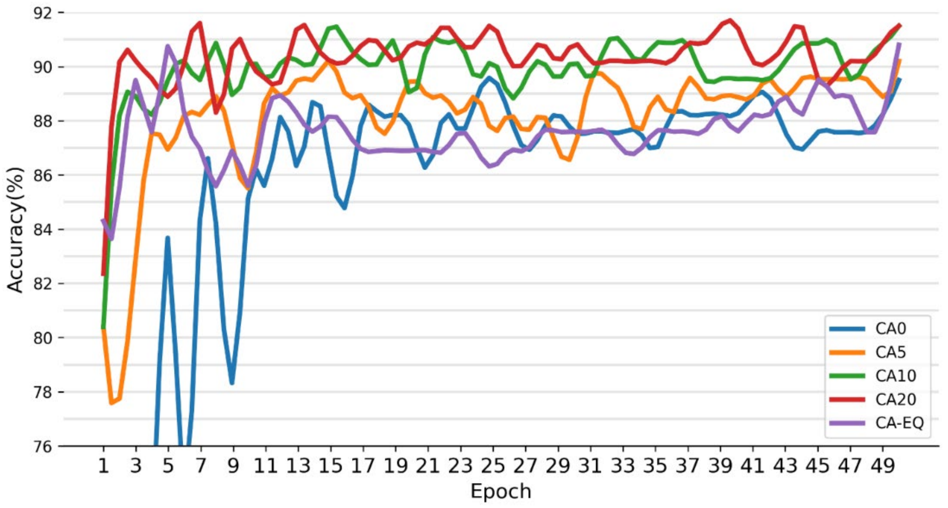

Figure S3. Effect of data augmentation on the training performance of the EfficientNet-B7 model. Each curve represents the species identification accuracy on the test set over 50 training epochs, using different training data strategies: no augmentation (CA0), 5-fold augmentation (CA5), 10-fold augmentation (CA10), 20-fold augmentation (CA20), and a class-balanced augmentation (CA-EQ). The results indicate that the 10-fold and 20-fold augmentation strategies achieved the highest and most stable final accuracy (91.5%).

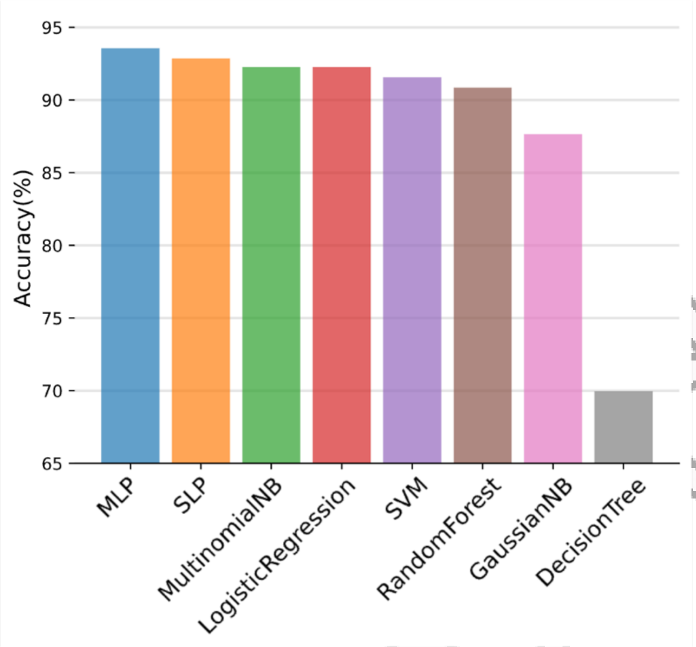

Figure S4. Performance comparison of different classifiers for the feature fusion module. The bar chart shows the final species identification accuracy achieved by eight different classification algorithms when applied to the fused feature vector from the HISNET-FF core network. The Multi-Layer Perceptron (MLP) achieved the highest accuracy (93.5%), outperforming all other tested classifiers.

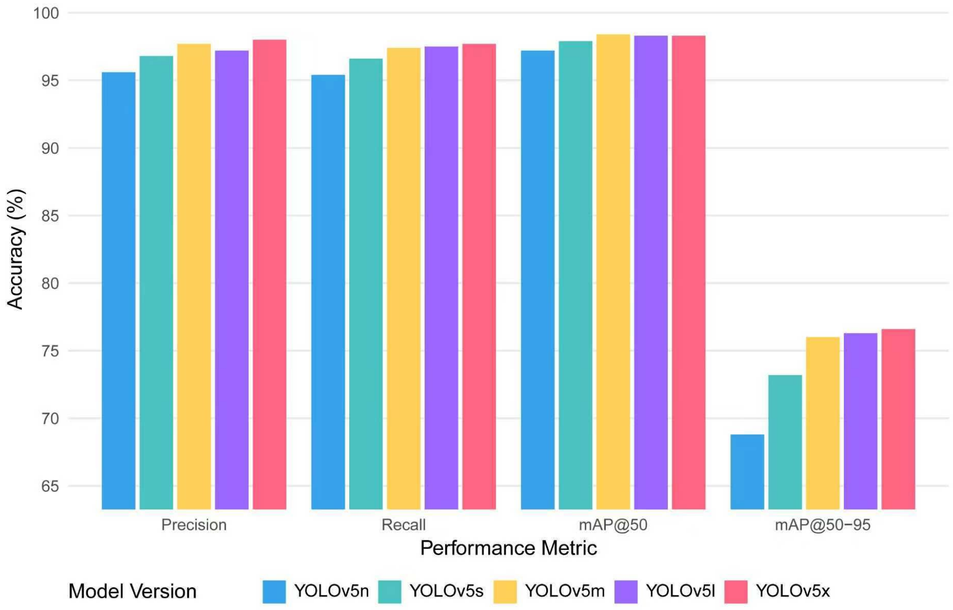

Figure S5. Performance comparison of different YOLOv5 model architectures at a fixed 640x640 resolution. The grouped bar chart displays the performance of five YOLOv5 variants (n, s, m, l, and x) across four standard object detection metrics: Precision, Recall, mean Average Precision at an IoU threshold of 0.50 (mAP@0.5), and mean Average Precision averaged over IoU thresholds from 0.50 to 0.95 (mAP@[.5:.95]).

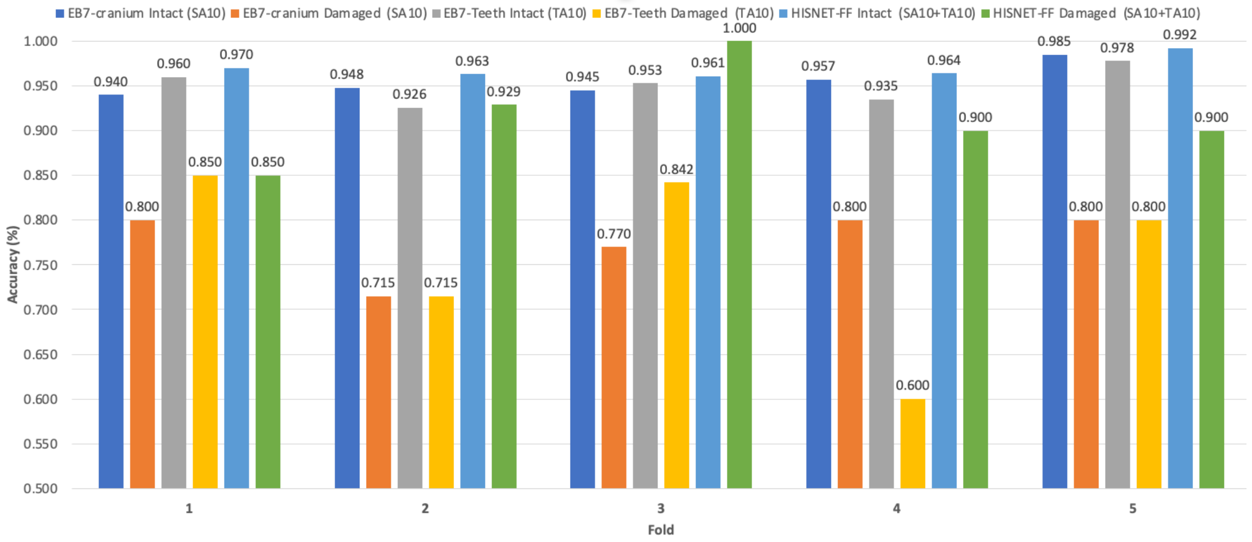

Figure S6. Species identification accuracy for intact versus partially damaged crania across the five cross-validation folds. The chart compares the performance of three hierarchical models: EB7-cranium, EB7-teeth, and the HISNET-FF fusion model. For each fold, the accuracy on intact specimens is shown alongside the accuracy on damaged specimens. The results consistently show that while damage reduces accuracy, the HISNET-FF model maintains the highest performance on both intact and damaged specimens.

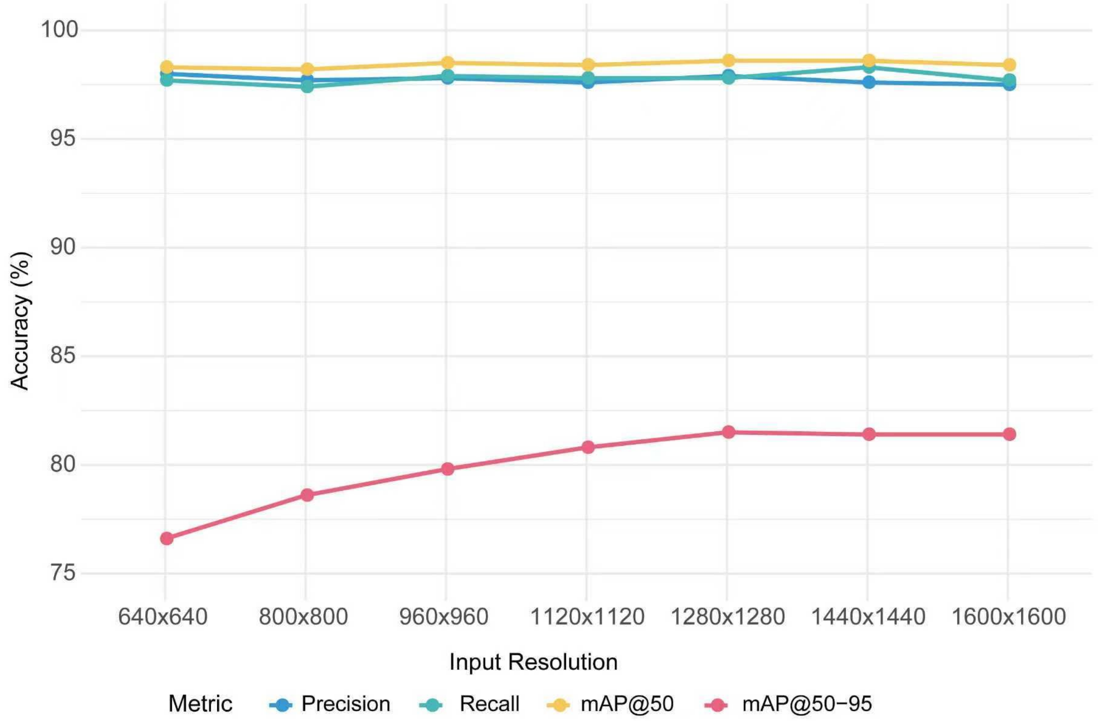

Figure S7. Effect of varying input resolutions on the performance of the YOLOv5x model. This analysis identifies 1280×1280 as the optimal resolution, providing the best balance of performance across all metrics with only a negligible trade-off in peak recall.

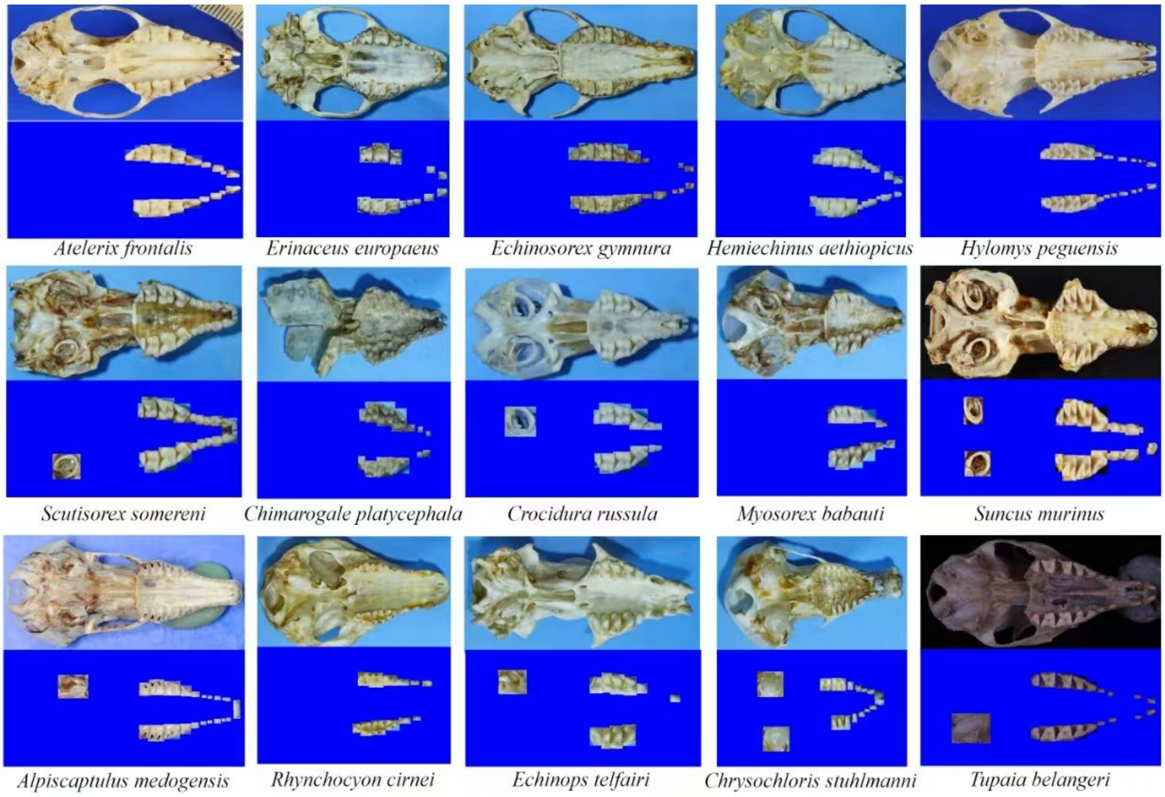

Figure S8. Qualitative results of the Talpidae-trained YOLOv5x model applied to untrained insectivorous mammal taxa. We included five erinaceids (Erinaceidae, first row), five shrews (Soricidae, second row), a talpid mole (*Alpiscaptulus medogensis*), a tree shrew (*Tupaia belangeri*), and three Afrotherian species (*Rhynchocyon cirnei*, *Echinops telfairi* and *Chrysochloris stuhlmanni*). The model demonstrated a detection rate of 92% for teeth and auditory bullae, calculated as the ratio of detected structures to the total number present. The other Performance metrics are recall=0.46, precision=0.41, mAP@0.5=0.39, mAP@[.5:.95]=0.33. These results suggest the YOLOv5x can detect teeth and auditory bullae in other groups of insectivorous mammals. The figure shows the model's ability to locate teeth and auditory bullae (bottom row of each pair) on cranial images (top row) from species it was not trained on. The model successfully localizes the features, but as noted in the text, its ability to assign correct specific labels is limited in taxa with different dental formulas.

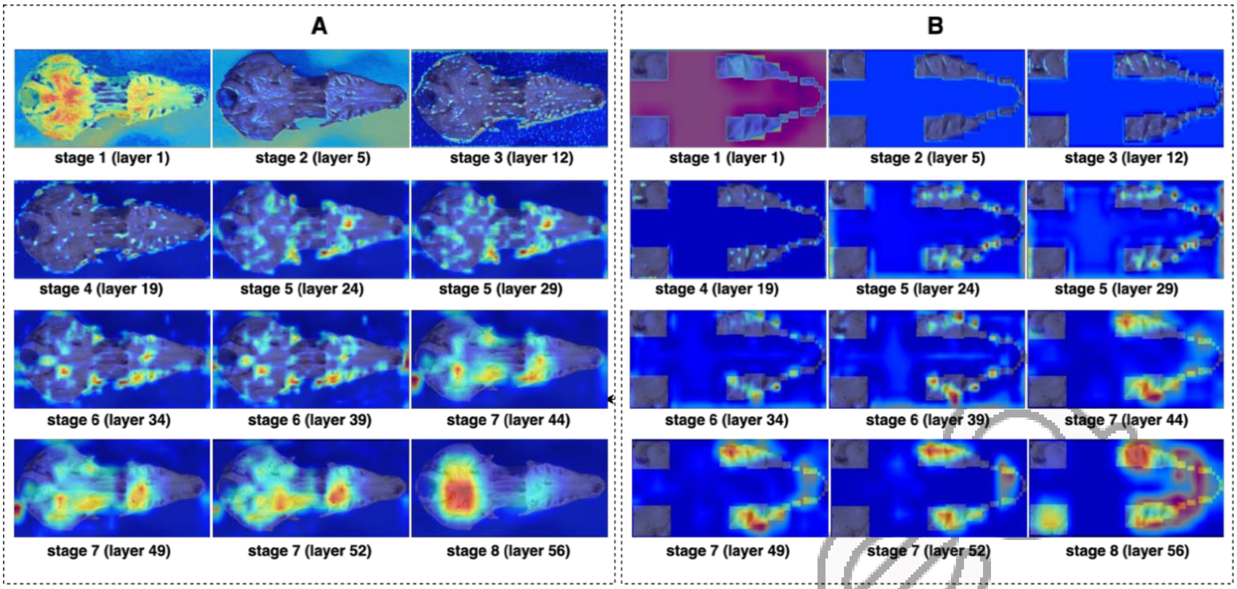

Figure S9. Gradient-weighted Class Activation Mapping (Grad-CAM) heatmaps illustrating the feature focus of single-modality models. The color scale indicates regions of high (red) to low (blue) importance for the model's classification decision. A: Heatmaps from an EB7 model trained only on cranium data, showing that the model learns features across the cranium but largely overlooks the dental region; B: Heatmaps from an EB7 model trained only on dental data, showing that the model focuses exclusively on the teeth and auditory bullae.