Cao et al. *Zool. Res.* 2026, 47(2): 404−413
https://doi.org/10.24272/j.issn.2095-8137.2025.156
https://cstr.cn/32248.14.zr.2025.156

ZR | Zoological Research

Article Open Access

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*, *Guangdong* 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*

ABSTRACT

Accurate taxonomic identification based on mammalian craniodental features remains critical for evolutionary, ecological, and paleontological research, yet conventional approaches are time-intensive and demand expert input. To overcome these limitations, a deep learning framework, HISNET-FF, was developed with a dual-stream architecture that integrates global cranial morphology with local diagnostic signals from teeth and auditory bullae. The model operates within a hierarchical classification pipeline, processing from genus-level discrimination to species-level resolution. Evaluation on an extensive image dataset encompassing 51 species across 18 genera of Talpidae achieved exceptional accuracy at both the genus (99.6%±0.4%) and species (96.5%±1.3%) levels. This species-level performance substantially exceeded that of single- stream models employing either flat (91.2%±2.3%) or hierarchical (93.9%±2.1%) strategies. To support end-to-end automation, a YOLO-based annotation module was implemented to localize key morphological traits with 97.8% recall, 97.9% precision, and 81.5% mean average precision (mAP@[.50:.95]). Incorporating this module incurred only a marginal reduction of 1.9% in identification accuracy. Thus, HISNET-FF offers a robust and accurate framework that accelerates morphology-based species identification and enables automated taxonomic classification, with strong potential for broader implementation across diverse biological research domains.

Received: 11 September 2025; Accepted: 23 October 2025; Online: 24 October 2025

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), Open Project of Ministry of Education Key Laboratory for Ecology of Tropical Islands, Hainan Normal University (HNSF-OP-2024-3), Guangzhou Higher Education Teaching Quality and Teaching Reform Engineering Special Talent Training Plan Project (2022ZXRCPR007), Bundesministerium für Bildung und Forschung (BMBF Project KI-Morph 05D2022), and National Key R&D Program of China (2022YFC2601200)

[#]Authors contributed equally to this work

*Corresponding authors, E-mail: baim@ioz.ac.cn; hekai@gzhu.edu.cn

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

INTRODUCTION

Taxonomic classification serves as a foundational pillar across biological disciplines, from biodiversity monitoring and conservation management to evolutionary biology and genomics. However, the field faces significant challenges in the 21st century (Britz et al., 2020), driven by a global decline in trained taxonomists (Wägele et al., 2011) and an escalating need for accurate and scalable species identification. Although DNA barcoding has transformed molecular systematics and species delimitation (Moritz & Cicero, 2004), there remains a critical demand for high-throughput, morphology-based technologies capable of delivering precise taxonomic resolution, particularly when genetic material is unavailable or degraded (Orr et al., 2021).

In mammalian systematics, particularly within speciose small mammal groups such as rodents, bats, and eulipotyphlans, craniodental morphology has long served as the primary diagnostic foundation. Pioneering work by Thomas Oldfield (Hinton, 1929) formally established the utility of skull and dental traits in species discrimination. Notably, unlike external traits, which often show high intraspecific variation or convergence among distinct taxa, the fine-scale anatomical features of the cranium and dentition provide consistent and reliable characters for resolving species boundaries (Dayan et al., 2002).

Species identification frameworks have traditionally relied on either qualitative assessment of discrete morphological traits or quantitative approaches such as traditional morphometrics, which are based on linear measurements, and geometric morphometrics, which capture complex shape variations using landmark-based methods (Mutanen & Pretorius, 2007). These techniques remain indispensable, particularly for paleontological contexts where cranial and dental remains frequently represent the only preserved material. However, morphology-based identification requires a high level of expertise and labor-intensive protocols, limiting the pace and scale of modern taxonomic research (Zamani et al., 2022).

In recent years, deep learning (DL) has emerged as a transformative tool for classification tasks across scientific domains (Lecun et al., 2015), including biological taxonomy (Badirli et al., 2023; Valan et al., 2019). This advance has been driven largely by Convolutional Neural Networks (CNNs), which established a new paradigm for image-based analysis. A pivotal milestone occurred in 2012, when AlexNet surpassed traditional classifiers such as support vector machines (SVMs) and random forests (RFs) in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Krizhevsky et al., 2012). This achievement demonstrated the profound ability of CNNs to extract diagnostic representations directly from raw pixel data, establishing these models as a powerful framework for morphology-based species identification. Such capacity holds promise not only for taxonomic research but also for ecological and evolutionary studies that rely on accurate morphological classification (Fortelius et al., 2002; Lyons et al., 2016). Although substantial progress has been achieved for plants and invertebrates using external morphology (Lee et al., 2015; Zhao et al., 2023), applications to vertebrates, especially mammals, remains scarce (Gill et al., 2024). Existing studies are often limited in scope, focusing on discrimination among a small number of closely related genera or species (Miele et al., 2020; Pinho et al., 2023).

Our recent work introduced HIS-NET, a CNN-based framework that achieved genus- and species-level accuracies of 95% and 90%, respectively, within Talpidae (He et al., 2025). Gradient-weighted Class Activation Mapping (Grad-CAM) confirmed that the model correctly focused on taxonomically informative cranial regions. However, heatmaps also indicated reliance on restricted skull areas, with limited exploitation of fine-grained diagnostic information encoded in dentition, highlighting an opportunity for methodological enhancement.

Feature fusion integrates complementary representations derived from multiple sources or perspectives to improve classification performance (Caci et al., 2013; Dai et al., 2021). In image classification, this strategy commonly combines “global” descriptors that capture overall structure with “local” descriptors that capture fine anatomical detail, thereby strengthening predictive accuracy (Peng et al., 2021). This dual-scale integration has produced substantial gains in diverse applications, including tree species recognition, where fusion of bark and leaf features improved accuracy by approximately 10 percentage points (Bertrand et al., 2018).

In this study, a new hierarchical species identification network incorporating feature fusion, termed HISNET-FF, was established. The network employs a dual-stream architecture to independently encode cranial morphology and localized traits from teeth and auditory bullae, followed by joint embedding integration. Given that dental and auditory bullae morphology encode subtle, species-specific traits, this design was expected to enhance taxonomic resolution relative to single-stream models. The framework was evaluated using Talpidae to demonstrate improvements in species recognition accuracy. In parallel, a YOLO-based object detection pipeline was implemented to automate annotation of key diagnostic features, including teeth and auditory bullae, thereby streamlining the annotation of diagnostic structures and facilitating more efficient morphology-based species identification.

MATERIALS AND METHODS

Specimen accession and photography

This study focused on the mammalian family Talpidae, comprising 19 genera and 68 recognized species worldwide (Burgin et al., 2025). The image dataset was consistent with the protocol described in He et al. (2025). Specimens were photographed in natural history museums located in China, Japan, Germany, Vietnam, and the United States. A standardized photography protocol was applied across all sites. Images were captured using a DSLR camera mounted on a fixed stand. Each specimen was placed on a level platform positioned 10–30 cm below the lens. A bubble level was used on both the camera and the platform to ensure a consistent perpendicular perspective. To minimize shadowing of diagnostic features—especially dentition—two adjustable lights were positioned on both sides of the cranium and angled downwards at approximately 60°. To account for lighting variability across institutions, the F-stop (f/13–f/23) and ISO (200–1250) settings of the camera were manually adjusted to

produce consistent image quality and exposure. Specimens with uncertain identification were excluded. A total of 747 ventral-view cranial photographs were obtained, including 674 intact and 73 partially damaged specimens, representing 18 genera and 51 species (including putative taxa; Supplementary Appendix I). The number of photographs per species ranged from three to 48 (Supplementary Table S1).

Data annotation, processing, and augmentation

Each cranial image was annotated with genus, species, and museum voucher information. For compatibility with the deep learning architecture, all images were cropped and symmetrically padded with black pixels to generate standardized square inputs (Supplementary Figure S1). The full dataset was then partitioned into training (594 images) and testing (153 images) subsets at an 8:2 ratio, yielding the baseline dataset CA0.

Given the diagnostic importance of tooth and auditory bulla morphology in talpid taxonomy, a specialized dental image dataset (TA0) was constructed through a multi-step process. Using LabelImg v1.8.6, rectangular bounding boxes were manually placed around each individual tooth and auditory bulla. Teeth were identified and categorized based on standard dental formulae (Wilson & Mittermeier, 2018) as incisors (I1−I3), canine (C1), premolars (P1−P4), and molars (M1−M3). Cropped bounding box regions were then reassembled onto a new canvas, preserving original spatial relationships through coordinate mapping (Supplementary Figure S1). While adjacent bounding boxes occasionally overlapped during annotation, the final composite images contained distinct, non-overlapping placements. Resulting composites were square-padded and added to the TA0 dataset, which was subsequently split into training and testing sets at an 8:2 ratio (Supplementary Table S2).

To mitigate overfitting and enhance model generalization, progressive image augmentation strategies were implemented (Maharana et al., 2022). The dataset CA0 was initially expanded five-fold (CA5) using rotation (90°), Gaussian noise addition, random information dropout, and random cropping (200 pixel coverage), resulting in 2 970 images. A further expansion to 10-fold (CA10) was achieved by introducing 45° rotations, brightness and hue adjustments, horizontal translation (100 pixels), and local patch rotations across four quadrants, yielding 5 940 images. Final augmentation generated a 20-fold dataset (CA20) by applying 180° and 270° rotations, dual random dropout, complex noise types (salt-and-pepper, Poisson), saturation and contrast shifts, vertical translation (100 pixels), and enhanced random cropping (300-pixel coverage), producing 11 880 images. Identical augmentation protocols were applied to the dental dataset, resulting in a 10-fold expanded version (TA10) (Supplementary Table S2). All specimens underwent augmentation under uniform conditions.

Reflecting the inherent imbalance of natural history collections, the dataset exhibited considerable disparity in species representation, with image counts per species ranging from three to 48 and 15 species represented by five or fewer specimens. To assess the impact of this imbalance on model performance for underrepresented species (less than five specimens), a balanced training dataset (CA-EQ) was constructed. This was achieved through differential augmentation, where rare species received augmented images expansions ranging from six-fold to 50-fold, standardizing representation across species to between 100 and 250 training images (Supplementary Table S3).

Feature fusion model architecture

Cranial images encode broad structural variation across taxa, while the teeth and auditory bullae harbor localized traits that often determine species boundaries. To effectively integrate these complementary dimensions of morphological features, a dual-stream neural architecture was constructed to perform hierarchical feature fusion. The architecture comprises three sequential modules (Figure 1). The feature extraction module utilizes two independent and parallel convolutional streams. One stream processes the entire cranial image to extract global morphological descriptors, while the second stream id dedicated to capturing fine-grained features from the teeth and auditory bullae. Each branch independently encodes its input into a high-dimensional representation. These two outputs are integrated in the feature fusion module using intermediate-level concatenation. Specifically, each feature map is pooled and flattened into a vector, and the two vectors are then concatenated end-to-end to form a unified feature representation that combines global cranial context with localized dental and auditory information. The resulting composite vector is passed to the classification module, which performs species identification based on the fused morphological embedding.

Feature extraction module

To determine the most effective convolutional backbone for morphological feature extraction, a comparative evaluation was conducted using the 10-fold augmented cranial dataset (CA10). Several state-of-the-art architectures were assessed, 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), and Vision Transformer-Large with 32×32 patch size (ViT-L/32) (Zhai et al., 2022). Image resolution was standardized to 224×224 pixels for both training and testing datasets to ensure consistent computational efficiency. All networks were initialized with ImageNet-pretrained weights (Krizhevsky et al., 2012) and fully fine-tuned on the training dataset. Performance was evaluated using Top-1 accuracy (Supplementary Text). EfficientNetB0 achieved the highest performance in this benchmark, with an accuracy of 88.9%, surpassing other networks by 1.3% to 10.5% (Supplementary Table S4 and Figure S2).

Subsequent evaluation focused on the entire EfficientNet series (B0–B7), which scales depth, width, and resolution to capture increasingly fine-grained image detail (Tan & Le, 2019). After evaluating each model with its recommended image resolution, the most complex model, EfficientNetB7 (EB7), achieved the highest accuracy and was thus selected as the feature extractor for all subsequent modules (Supplementary Table S4 and Figure S2).

To optimize augmentation intensity, EB7 was trained on four datasets representing different augmentation levels: CA0 (no augmentation), CA5 (five-fold augmentation), CA10 (10-fold augmentation), and CA20 (20-fold augmentation). Peak accuracy (91.5%) was achieved with both CA10 and CA20 (Supplementary Figure S3), while the balanced dataset (CA-EQ) provided no further gain (90.8%). Due to comparable performance and lower computational burden, the 10-fold-augmented CA10 dataset was selected for all downstream

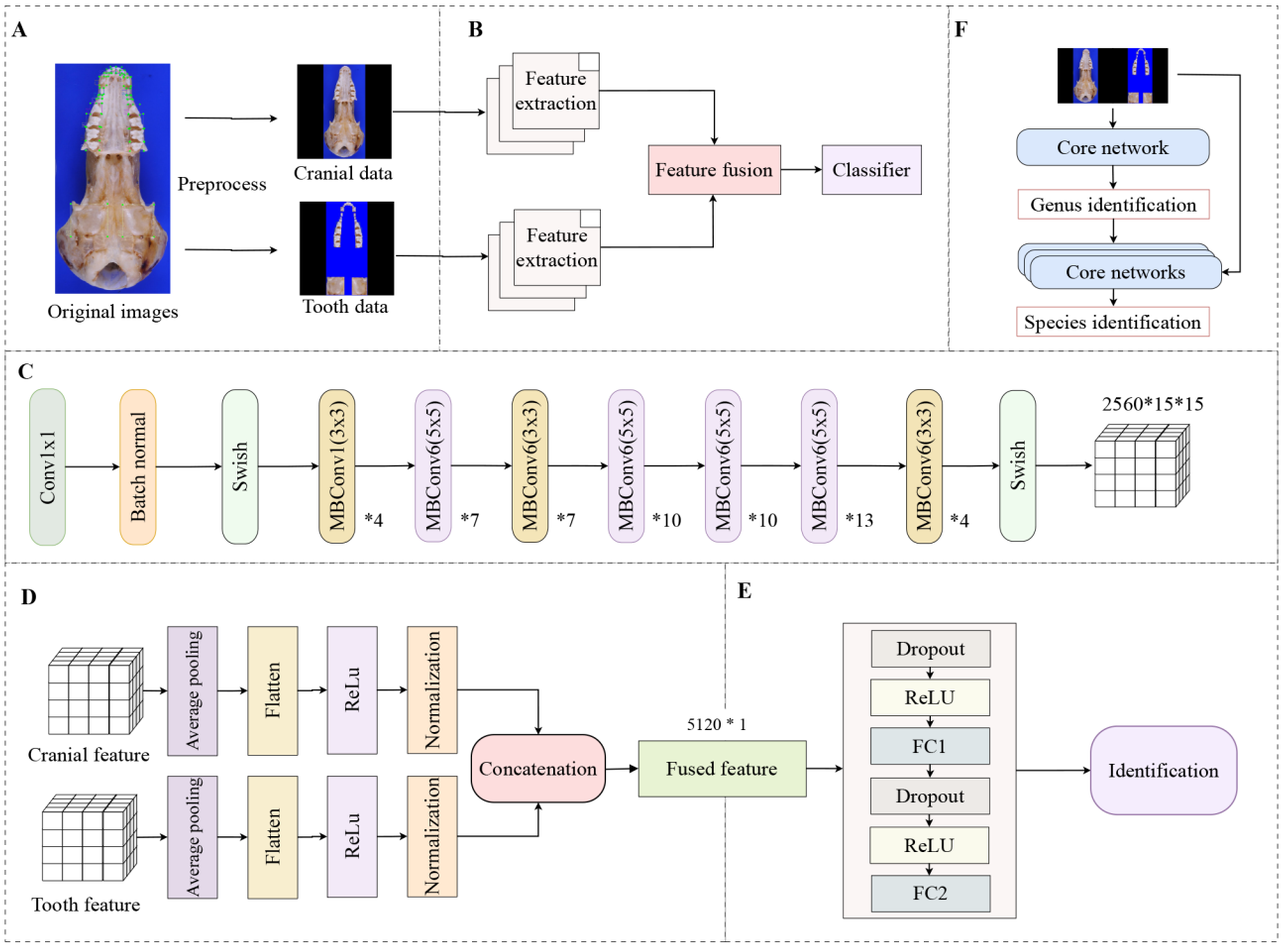

Figure 1 Architecture and workflow of the HISNET-FF model

The framework operates on a single ventral cranial image through a sequential, multi-stage pipeline. A: Data preprocessing converts the original image into two distinct inputs: a full cranial image capturing global morphology and a composite image encoding localized traits from teeth and auditory bullae. B: Core network processing routes these inputs into a dual-stream architecture, where features are extracted in parallel, integrated through fusion, and forwarded for classification. C: Feature extraction in each stream is performed using an EfficientNetB7-based architecture, producing a 2 560×15×15 feature tensor per stream. D: Feature fusion applies pooling, flattening, and activation to both tensors, followed by concatenation into a unified 5 120-dimensional feature vector. E: Classification is carried out by a multi-layer perceptron (MLP), which produces the final taxonomic assignment. F: Hierarchical identification proceeds in two stages: an initial network resolves genus-level identity, followed by a genus-specific network for species-level classification within polytypic genera.

tasks.

The core architecture of EB7 consists of eight sequential stages. The initial “Stem” convolutional block captures fundamental image patterns such as edges and gradients. This is followed by seven compound stages composed of repeated Mobile Inverted Bottleneck Convolution (MBConv) blocks (Figure 1C). As data flow through the subsequent, more complex blocks, low-level features are progressively integrated into high-level semantic representations, such as specific tooth morphologies. These layers culminate in a Swish activation function applied to the output of the final MBConv block. The feature tensor is immediately extracted following this terminal activation. The resulting 2560×15×15 tensor encodes a condensed, high-dimensional summary of taxonomically informative visual features for subsequent fusion and classification.

Feature fusion module

To test the hypothesis that integrating global cranial morphology with localized dental and auditory features would improve accuracy, a dedicated fusion module was constructed to combine the paired feature tensors generated by EB7 (Figure 1D). Each 2 560×15×15 tensor was first reduced to 2 560×1×1 via global average pooling, collapsing spatial dimensions while preserving the most important global information. These pooled outputs were then reshaped into one-dimensional vectors of length 2 560 (flattened) to enable downstream processing. Each feature value was passed through a Rectified Linear Unit (ReLU) activation function (Agarap, 2018), defined as:

$$f(x) = max(0, x) \tag{1}$$

which retains positive values and sets negative values to zero. This nonlinearity enables the network to capture complex and biologically relevant morphological relationships. Following activation, feature values were normalized using MinMaxScaler to adjust them to the range [0, 1], ensuring numerical stability and harmonization across feature channels. Finally, the two normalized vectors were concatenated to form a unified 5120-dimensional feature representation that jointly encodes both global cranial and local diagnostic information. This composite vector was then used to train a lightweight classification module for taxonomic prediction.

Classification module

To identify the most effective classifier for final taxonomic prediction, eight candidate algorithms were evaluated using the fused 5 120×1 feature vectors derived from the CA10 and TA10 datasets. These included logistic regression, decision tree, random forest, multinomial and Gaussian Naïve Bayes classifiers, support vector machine (SVM), single-layer perceptron (SLP), and multi-layer perceptron (MLP). The SLP architecture, as implemented in EfficientNet, comprises a Dropout layer and a fully connected layer (Tan & Le, 2019). MLP extends this structure by incorporating two additional ReLU activations, an additional Dropout layer, and a second fully connected layer (Kruse et al., 2022; Figure 1F). During this evaluation, the EB7 feature extraction layers remained frozen to isolate classifier performance, and their weights were not updated during training. Among all tested classifiers, MLP achieved the highest accuracy (93.5%), outperforming all other classifiers, which ranged from 92.8% to 69.9% (Supplementary Figure S4). Based on this result, the MLP was selected as the final classification module.

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), a hierarchical classification pipeline was implemented. This sequential design first resolves genus identity then refines species-level classification within polytypic groups. The process begins with a primary genus-level classifier trained to discriminate among all 18 genera. Each input image is initially assigned to the most probable genus. For monotypic genera, this step concludes the identification. In cases were the assigned genus contains multiple species (e.g., *Euroscaptor*, *Talpa*), the image is subsequently routed to a genus-specific classifier trained exclusively on that group. Seven specialized species-level classifiers were constructed, each restricted to a single polytypic genus (Figure 1F). This hierarchical architecture introduces the risk of error propagation, as an incorrect genus-level prediction precludes accurate species assignment. However, the genus-level model achieved 99.6% accuracy (see Results), indicating that such misclassification events are rare and unlikely to impact overall performance.

Model training, implementation, and validation

To leverage robust, pre-existing feature representations, all models were initialized with pre-trained weights from the ImageNet dataset (Krizhevsky et al., 2012). Fine-tuning was subsequently conducted on the project-specific datasets. This procedure was applied independently to the baseline flat species-level classifiers, the genus-level classifiers with the hierarchical framework, and seven genus-specific species-level classifiers corresponding to each polytypic lineage.

During training, model parameters were optimized using Stochastic Gradient Descent (SGD), an iterative algorithm designed to minimize prediction error through gradient-based updates. Training proceeded 50 epochs, with the entire training dataset presented to the model in each epoch. Images were processed in mini-batches of 16. After each batch, network parameters were adjusted to improve accuracy. A dynamic learning rate was employed, beginning at 0.01 and decayed every 10 epochs by a factor of 0.8 using a StepLR scheduler. To mitigate overfitting and promote generalization, dropout regularization was applied during training, which randomly ignored 50% of the network neurons during each training step (Supplementary Table S4). For classification modules such as SLP and MLP, the dropout rate was set to 0.7. All model training and inference were conducted on a GeForce RTX 3090 GPU (VRAM 24GB) within a Conda v.23.3 environment running Python v.3.9.

To evaluate the performance of the hierarchical feature fusion network, two single-modality baseline models were implemented: one trained exclusively on cranial images (CA10) and the other on dental composites (TA10). These baseline models were evaluated under both flat and two-step hierarchical classification strategies to isolate the contribution of hierarchical structure itself. To evaluate the consistency of network performance and reduce the impact of data partitioning, a five-fold cross-validation approach was employed. The selection of k=5 was made to accommodate dataset imbalance, particularly the presence of rare species with fewer than five specimens. This design ensured that each test fold contained sufficient representation of underrepresented taxa, thereby stabilizing accuracy estimates (Wong & Yeh, 2020). In each iteration, the dataset was partitioned into five equal subsets, with one subset used for testing and four for training.

To further assess performance sensitivity, analyses were conducted to examine how variation in sample size and specimen integrity influenced identification accuracy. Correlation analyses were performed using three machine learning regression models: Decision Tree (Kotsiantis, 2013), Gradient Boosting (Bentéjac et al., 2021), and Random Forest (Paul et al., 2018).

Object detection-based automatic annotation of teeth and auditory bullae

To automate the labor-intensive annotation of dental and auditory structures, an object detection pipeline was constructed using the YOLOv5 architecture (Jocher et al., 2022; Zhang et al., 2025) with the aim to replicate manual annotation processes with high fidelity and scalability.

A ground-truth dataset (DS0) was first established by manually annotating all 594 cranial training images using LabelImg. Each image was labeled with distinct bounding boxes for every individual tooth and auditory bulla. To enhance model robustness and generalization, this dataset was expanded 10-fold through data augmentation to create DS10. Augmentation methods included mirroring (up-down, left-right, and mixed), random interpolation resizing (600×600, 960×960, and 1 600×1 600 pixels), 90° rotations (clockwise and counterclockwise), and Gaussian noise injection (Wan et al., 2023). Notably, the YOLO framework does not require image cropping during training or testing.

Model performance was evaluated using standard object detection metrics, including precision, recall, and mean average precision (mAP) (Supplementary Text). Evaluation proceeded in two stages. First, several pre-trained YOLOv5 variants were benchmarked at a standardized 640×640 resolution to identify optimal architecture. YOLOv5x, the largest variant, demonstrated the best overall performance and was selected for further tuning (Supplementary Table S5). Second, given that object detection accuracy is highly sensitive to input resolution, the selected YOLOv5x model was benchmarked across a range of resolutions ranging from 640×640 to 1 600×1 600. The 1 280×1 280 resolution achieved the highest score on the comprehensive mAP@[.50:.95] metric (81.5%) with minimal loss in recall and was thus

adopted as the operational setting (Supplementary Table S5 and Figure S5).

Finally, the trained YOLOv5x model (1 280×1 280 resolution) was then applied to the cranial test set to generate bounding box predictions. These predicted coordinates were then used to construct a new automatically annotated dental dataset (DT0), following the same crop-and-stitch procedure used for the manually annotated TA0 dataset. To evaluate downstream impact of this automated workflow, species identification performance was compared between models using DT0 and their original counterparts trained on TA0. This included both teeth-only and dual-stream fusion networks, allowing direct assessment of the impact of automated annotation on final classification accuracy.

RESULTS

Identification accuracy was assessed independently for cranial (CA10) and dental (TA10) datasets using the EB7 backbone, and in combination with the dual-stream feature fusion network. Flat classifiers trained on cranial and dental inputs achieved mean accuracies of 90.5%±1.6% and 91.2%±2.3%, respectively. The flat feature fusion model yielded a slightly higher accuracy of 92.5%±1.8%, although the improvement did not reach statistical significance (paired *t*-test: $P>0.06$) (Table 1).

Under hierarchical classification, all models demonstrated exceptional genus-level accuracy, with the HISNET-FF model achieving the highest accuracy (99.6%±0.4%; Supplementary Table S6). At the species level, all three hierarchical models outperformed their flat counterparts, with accuracy gains ranging from 2.0% to 4.0% (paired *t*-test: EB7-cranium, $P=0.021$; EB7-teeth, $P=0.008$; HISNET-FF, $P=0.0035$). HISNET-FF maintained consistently high performance (mean=96.5%±1.3%; Figure 2), significantly outperforming both the hierarchical cranium-based model ($P=0.027$) and teeth-based model ($P=0.029$). This superior performance was consistent across all seven polytypic genera, with misclassification patterns visualized in confusion matrices (Figure 3).

To explore the mechanism underlying the superior accuracy of HISNET-FF, misclassification events were analyzed based on five-fold cross-validation (Supplementary Table S7). A total of 26 misidentified specimens were recorded out of 747 specimens, all from within polytypic genera, predominantly the talpine groups *Euroscaptor* and *Mogera*, suggesting that subtle interspecific morphological overlap may have contributed to classification errors. Comparative analysis indicated that HISNET-FF consistently corrected errors made by single-modality models: in all but two cases where one of the single-modality models yielded a correct identification, HISNET-FF did so as well. Notably, five specimens misclassified by both EB7-cranium and EB7-teeth models were correctly identified by HISNET-FF, underscoring the added value of feature fusion.

To evaluate potential bias against underrepresented taxa, identification accuracy was correlated with sample size using Spearman's rank correlation. No significant negative association was detected (Spearman's $\rho=-0.188$, $P=0.187$), indicating good model generalization despite inherent data imbalance. Among the 14 species represented by six or fewer specimens, all but one (*Uropsilus nivatus*, 80% accuracy) were classified without error (Supplementary Table S8), supporting the robustness of the framework to limited sample availability.

Approximately 10% of cranial specimens from museum collections exhibited partial damage, typically due to the use of snap traps during collection. HISNET-FF achieved 91.6% accuracy on damaged specimens, compared to 97.0% on intact ones (Supplementary Table S9 and Figure S6), indicating a modest but noticeable accuracy reduction. Despite this, HISNET-FF substantially outperformed the corresponding EB7-cranium and EB7-teeth models, which yielded 80.8% and 78.1% accuracy, respectively, on the same damaged subset.

To assess the joint impact of sample size and specimen condition on accuracy, correlation analyses were performed using three machine learning regressors: decision tree, gradient boosting, and random forest. While each factor alone exhibited weak correlation with classification accuracy ($R^2<0.3$; Supplementary Table S10), their combined influence yielded higher explanatory power ($R^2>0.4$), with the strongest interaction observed for the HISNET-FF and cranial-based models ($R^2>0.5$).

Automatic annotation of teeth enhances workflow efficiency in species identification

To streamline the annotation of dental and auditory features, YOLOv5 was implemented for automated detection of teeth and auditory bullae. Among the evaluated architectures, the YOLOv5x variant operating at 1 280×1 280 resolution exhibited the highest performance (Supplementary Figure S7) and was selected for downstream application. This optimized model achieved a precision of 97.9%, recall of 97.8%, and

Table 1 Species identification accuracy obtained using EB7 trained separately on cranial and dental datasets, as well as dual-stream feature fusion model integrating both modalities

		EB7-cranium accuracy (%)	EB7-teeth accuracy (%)	Feature fusion accuracy (%)
Flat identification	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 (nine spp.)	79.7±2.9	75.1±8.5	**88.3±6.1**
	Mogera (nine spp.)	89.0±3.6	89.0±6.0	**93.2±3.7**
	Parascaptor (three spp.)	92.7±7.4	94.70±4.4	**96.4±8.1**
	Scapanus (four spp.)	98.0±4.5	98.0±4.5	**98.0±4.5**
	Scaptonyx (three spp.)	79.7±6.3	86.4±2.2	**91.1±4.1**
	Talpa (seven spp.)	100.0±0.0	96.3±3.2	**100.0±0.0**
	Uropsilus (five spp.)	93.3±7.0	88.3±13.9	**95.0±7.2**

Both flat and hierarchical strategies were used for identification; bold indicates accuracy obtained using HISNET-FF

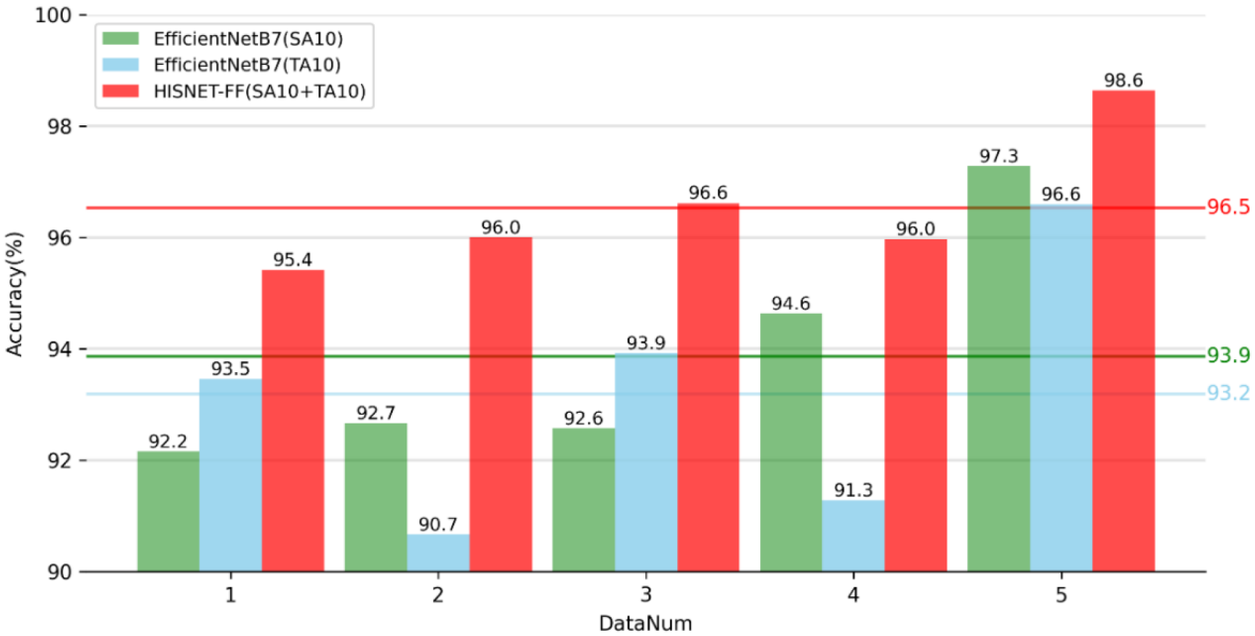

Figure 2 Comparison of species-level identification accuracy among hierarchical models, including feature-fusion network (HISNET-FF) and single-modality models (EB7-cranium and EB7-teeth) based on five-fold cross-validation

Bars represent accuracy values for individual folds, and horizontal lines represent mean accuracy across all folds for each model.

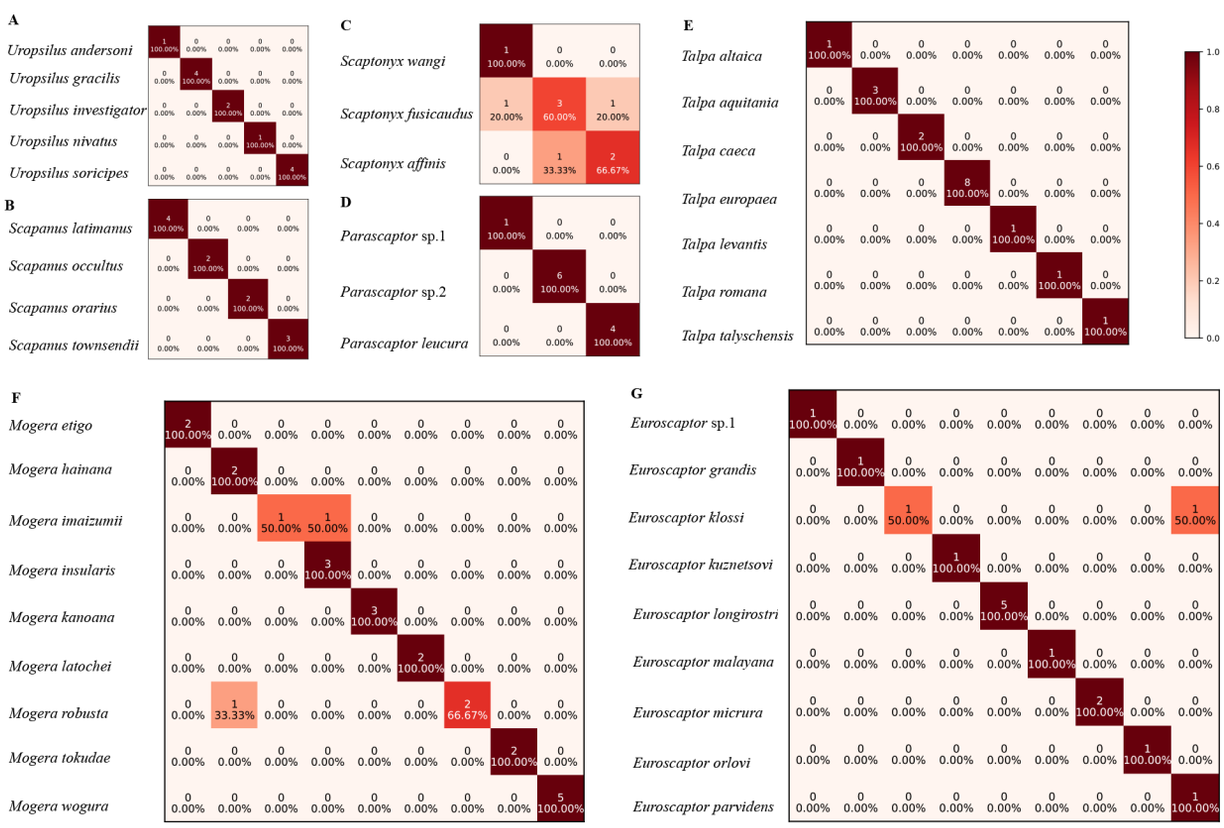

Figure 3 Confusion matrix heatmaps illustrating species-level identification performance of HISNET-FF across seven polytypic genera

Each panel corresponds to a representative test set for one genus, highlighting characteristic misclassification patterns. Color intensity reflects per-species identification accuracy.

mAP@[.50:.95] of 81.5% (Supplementary Table S5). The lowest recall rates were observed for the auditory bullae (95.8%), second upper premolar (P2, 96.8%), and second upper incisor (I2, 96.9%)—a trend likely reflecting the fragility of auditory bullae and the small size of I2 and P2 (Supplementary Table S11).

To evaluate the downstream impact of this automated annotation strategy, species identification accuracy was

compared between network configurations trained on manually annotated dental images (TA0) and those generated via the automated YOLOv5x pipeline (DT0). For HISNET-FF, automated annotation yielded a final accuracy of 93.5%, representing a marginal decrease from the 95.4% achieved with manual annotation. This difference was attributable to three additional misclassifications unique to the automated dataset, while all seven errors from the manual set were replicated in the automated results. This modest reduction in performance was consistent across all tested network configurations (Figure 4).

To assess the broader applicability of the YOLOv5x automatic annotation framework, the model was tested on a phylogenetically diverse set of insectivorous mammals not included in training. This external test set included the recently described talpid mole (*Alpiscaptulus medogensis*), five hedgehogs (*Erinaceidae*), five shrews (*Soricidae*), one tree shrew (*Scandentia*), and three afrotherian taxa. The model successfully detected 92.0% of teeth and auditory bullae across this dataset (Supplementary Figure S8), although labeling accuracy declined markedly (recall=0.46, precision=0.41, mAP@0.5=0.39, mAP@[.50:.95]=0.33).

DISCUSSION

The results of this study demonstrated that integrating a hierarchical classification pipeline with a dual-stream feature fusion architecture substantially enhanced species-level classification from craniodental morphology, overcoming inherent limitations of standard single-stream convolutional networks in complex morphological identification. Such models often fail to resolve global cranial architecture while capturing fine-grained dental features, frequently missing critical diagnostic elements such as cusps, crests, and cingula (Lin et al., 2021; Singha et al., 2024). The dual-stream architecture developed here overcomes these limitations by creating a multi-scale morphological representation that combines global and local features within a unified predictive framework.

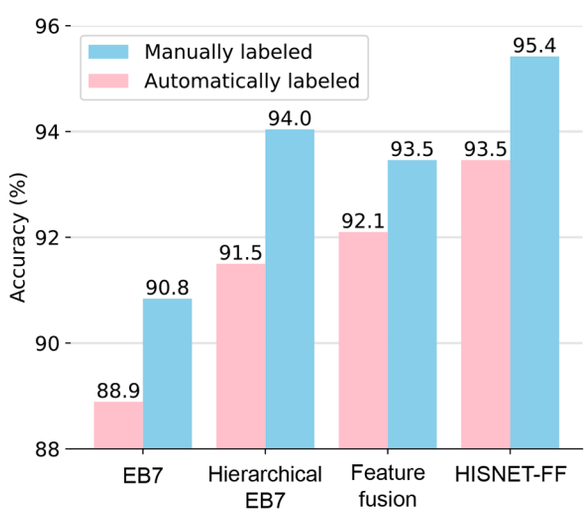

Figure 4 Comparison of species identification accuracy using manually versus automatically annotated dental images

Grouped bar chart displays final accuracy across four model configurations. For each configuration, accuracy was compared between dental inputs generated via manual annotation and those generated by the automated YOLOv5x pipeline.

Multiple lines of evidence support the efficacy of this approach. Grad-CAM heatmaps confirmed that cranium-only models were effectively blind to the dental region, underscoring the need for a dedicated feature channel to capture taxonomically informative structures (Supplementary Figure S9). Error analysis further revealed that the fusion model successfully integrated complementary information across modalities, not only reinforcing correct predictions when one input modality was sufficient but also resolving ambiguous cases where both single-modality models failed (Supplementary Table S7). Furthermore, HISNET-FF maintained high classification accuracy even for underrepresented taxa, correctly identifying 13 of the 14 species represented by six or fewer images. This suggests that the architecture is not compromised by class imbalance, a feature potentially enhanced by transfer learning from large-scale image datasets such as ImageNet. However, further targeted analyses are required to validate generalization performance on rare taxa with high morphological similarity.

By integrating spatially distinct diagnostic features within a hierarchical prediction structure, HISNET-FF achieved superior performance across complex taxonomic boundaries. This distinguishes it from previously developed models, including the EfficientNet-based HIS-NET (He et al., 2025), which relied on multiple cranial and mandibular views per specimen. Unlike HIS-NET, the current model employs a single-image workflow and attains superior accuracy through deep feature fusion rather than increased view sampling. Given its high performance on Talpidae, HISNET-FF holds promise for broader application to other speciose mammalian clades, such as Rodentia, Chiroptera, and Primates, where detailed craniodental morphology serves as a primary axis of taxonomic differentiation.

The success of the feature-fusion framework establishes a foundation for several promising directions in future research. A logical next step involves integrating the complementary advantages of the previously developed multi-view HIS-NET with the current HISNET-FF model into a more comprehensive multi-view feature fusion architecture using a multi-task learning strategy (Liu et al., 2019). Such a model would concurrently process dorsal, ventral, and lateral cranial views, enabling construction of a more holistic and spatially informed morphological representation. In parallel, the annotated regions could be expanded beyond teeth and auditory bullae to include additional taxonomically informative structures. Dental features themselves could be hierarchically decomposed into subcomponents, such as the trigonid, talonid, or even individual cusps, to enhance diagnostic resolution. Achieving a truly integrative taxonomic identification system will likely require moving beyond image-based fusion alone. Incorporating more advanced architectures, such as Large Language Models (LLMs), could facilitate the integration of heterogeneous data streams, including morphometric features, genetic sequences, geographic distributions, and ecological traits (Pyron, 2023).

Although the current framework offers high accuracy in classifying known species, it remains constrained within a supervised learning paradigm. Expanding its capabilities to support Novel Category Discovery (NCD) represents a critical advancement (Vaze et al., 2022). This would enable the system to not only classify recognized taxa but also identify morphologically anomalous or previously undescribed specimens, thus shifting its utility from species identification

toward active taxonomic discovery (Badirli et al., 2023).

The automated annotation module developed using YOLOv5x proved to be a highly effective and practical component within the overall classification pipeline. When applied to species identification, automated annotations produced a final accuracy of 93.5%, representing a minor decrease relative to the 95.4% achieved through labor-intensive manual labeling. This marginal trade-off in accuracy was offset by substantial gains in efficiency and scalability. Importantly, most misidentifications occurred between closely related, congeneric species, while genus-level accuracy remained high at 99.6%. Although this level of error warrants caution in formal taxonomic revision, it is well within acceptable bounds for large-scale ecological and macroevolutionary studies, where genus-level resolution is often sufficient. Furthermore, as previously discussed, remaining classification errors could likely be mitigated by incorporating complementary data streams such as DNA sequences and geographic distribution within a multimodal framework.

Performance benchmarking of the automated annotation tool on previously untrained insectivorous mammals further highlighted its potential and its current limitations. While the model proved highly effective at localizing craniodental features, detecting 92% of teeth and auditory bullae, its ability to assign specific positional labels to these structures was limited (precision: 41%; recall: 46%). This performance drop is attributable not to architectural constraints, but to the restricted taxonomic scope of the training dataset, which was based exclusively on talpids. Incorporating a diverse array of annotated specimens from other key mammalian orders (e.g., rodents, bats, shrews), should enable the YOLOv5 model to learn the broader spectrum of dental formulas and morphologies required for high-accuracy, cross-clade generalization.

The broader implications of this automated framework are substantial. Evolutionary and ecological studies frequently rely on extensive specimen datasets, yet progress is often constrained by the laborious manual effort required for accurate morphological identification (Pineda-Munoz et al., 2021; Saarinen & Lister, 2023). By demonstrating robust performance in a morphologically challenging taxonomic group, the proposed approach offers a pathway toward scalable, automated species identification, which will, in turn, facilitate high-throughput morphological workflows and accelerate large-scale, data-driven biodiversity research.

DATA AVAILABILITY

The HISNET-FF data and code are freely available via 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

Z.C., M.B., and K.H. designed and supervised the project. K.H., W.H.Y., Z.Z.C., S.Y.L., S.L., and F.Y.T. collected specimens and prepared skulls for imaging. K.H. and Q.M. photographed the specimens. Z.C., Q.L.T., W.Q.Z., K.H.W., and Z.W.H. performed deep learning analyses. Z.L.Z., Q.Q.L., S.Y.L., X.Y.Z., and S.N.X. conducted data annotation for deep learning. J.J.H. conduct statistical analyses. All authors read and approved the final version of the manuscript.

ACKNOWLEDGMENTS

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 (SI Kawada), Hokkaido University Natural History Museum (T Tsubota), Institute of Ecology and Biological Resources of Vietnam (TS Nguyen), State Museum of Natural History Stuttgart (E Amson, C Leidenroth, S Merker), and University of Montpellier (PH Fabre) for allowing us to examine specimens and take photos of specimens under their curation. We thank J Decher and C Montermann for kindly providing photos of *Galemys*, and T Martin and C Steinweg for kindly providing photos of *Desmana*.

REFERENCES

Agarap AF. 2018. Deep learning using rectified linear units (ReLU). *arXiv*, doi: https://doi.org/10.48550/arXiv.1803.08375.

Al-Qizwini M, Barjasteh I, Al-Qassab H, et al. 2017. Deep learning algorithm for autonomous driving using googlenet. *In*: Proceedings of the 2017 IEEE Intelligent Vehicles Symposium (IV). Los Angeles: IEEE, 89–96.

Badirli S, Picard CJ, Mohler G, et al. 2023. Classifying the unknown: Insect identification with deep hierarchical Bayesian learning. *Methods in Ecology and Evolution*, **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, Hundsdörfer A, Fritz U. 2020. Funding, training, permits—the three big challenges of taxonomy. *Megataxa*, **1**(1): 49–52.

Burgin CJ, Zijlstra JS, Becker MA, et al. 2025. How many mammal species are there now? Updates and trends in taxonomic, nomenclatural, and geographic knowledge. *Journal of Mammalogy*, **106**(5): 1082–1117.

Caci G, Biscaccianti AB, Cistrone L, et al. 2013. Spotting the right spot: computer-aided individual identification of the threatened cerambycid beetle Rosalia alpina. *Journal of Insect Conservation*, **17**(4): 787–795.

Dai YM, Gieseke F, Oehmcke S, et al. 2021. Attentional feature fusion. *In*: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Waikoloa: IEEE, 3559–3568.

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. *Evolutionary Ecology Research*, **4**: 1005–1016.

Gill KS, Gupta R, Malhotra S, et al. 2024. Classification of reptiles and amphibians using transfer learning and deep convolutional neural networks. *In*: Proceedings of the 2024 IEEE 9th International Conference for Convergence in Technology (I2CT). Pune: IEEE, 1–5.

He K, Li AL, 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*, doi: https://doi.org/10.1101/2025.03.06.641773.

Hinton MAC. 1929. MR. M. R. oldfield thomas, F. R. S. *Nature*, **124**(3116): 101–102.

Howard AG, Zhu ML, Chen B, et al. 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. *arXiv*, doi: https://doi.org/10.48550/arXiv.1704.04861.

Iandola F, Moskewicz M, Karayev S, et al. 2014. DenseNet: Implementing efficient convnet descriptor pyramids. *arXiv*, doi: https://doi.org/10.48550/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*. https://doi.org/10.5281/zenodo.7002879.

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. *In*: Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: ACM, 1097–1105.

Kruse R, Mostaghim S, Borgelt C, et al. 2022. Multi-layer perceptrons. *In*: Kruse R, Mostaghim S, Borgelt C, et al. Computational Intelligence: A Methodological Introduction. 3rd ed. Cham: 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. *In*: Proceedings of the 2015 IEEE International Conference on Image Processing (ICIP). Quebec City: IEEE, 452–456.

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

Liu MX, Zhang J, Adeli E, et al. 2019. Joint classification and regression via deep multi-task multi-channel learning for alzheimer's disease diagnosis. *IEEE Transactions on Biomedical Engineering*, **66**(5): 1195–1206.

Lyons SK, Amatangelo KL, Behrensmeyer AK, et al. 2016. Holocene shifts in the assembly of plant and animal communities implicate human impacts. *Nature*, **529**(7584): 80–83.

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*, doi: https://doi.org/10.1101/2020.08.20.259176.

Moritz C, Cicero C. 2004. DNA barcoding: promise and pitfalls. *PLoS Biology*, **2**(10): e354.

Mutanen M, Pretorius E. 2007. Subjective visual evaluation vs. traditional and geometric morphometrics in species delimitation: a comparison of moth genitalia. *Systematic Entomology*, **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 ZL, Huang W, Gu SZ, et al. 2021. Conformer: local features coupling global representations for visual recognition. *In*: Proceedings of the IEEE/CVF International Conference on Computer Vision. Montreal: IEEE, 357–366.

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 of the United States of America*, **118**(2): e1922859118.

Pinho C, Kaliontzopoulou A, Ferreira CA, et al. 2023. Identification of morphologically cryptic species with computer vision models: wall lizards (Squamata: Lacertidae: *Podarcis*) as a case study. *Zoological Journal of the Linnean Society*, **198**(1): 184–201.

Pyron RA. 2023. Unsupervised machine learning for species delimitation, integrative taxonomy, and biodiversity conservation. *Molecular Phylogenetics and Evolution*, **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 DS, Krishna A. 2024. Effi-seg: rethinking efficientNet architecture for real-time semantic segmentation. *In*: Proceedings of the 30th International Conference on Neural Information Processing. Changsha: Springer, 55–68.

Tan MX, Chen B, Pang RM, et al. 2019. MnasNet: Platform-aware neural architecture search for mobile. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach: IEEE, 2815–2823.

Tan MX, Le QV. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. *In*: Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR, 6105–6114.

Targ S, Almeida D, Lyman K. 2016. Resnet in resnet: generalizing residual architectures. *arXiv*, doi: https://doi.org/10.48550/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. *Systematic Biology*, **68**(6): 876–895.

Vaze S, Han K, Vedaldi A, et al. 2022. Generalized category discovery. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 7482–7491.

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 DH, Lu RS, 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. Barcelona: Lynx Nature Books.

Wong TT, Yeh PY. 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. *Systematics and Biodiversity*, **20**(1): 1–14.

Zhai XH, Kolesnikov A, Houlsby N, et al. 2022. Scaling vision transformers. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 1204–1213.

Zhang JJ, Gao Y, Zhang BL, 2025. A deep learning lightweight model for real-time captive macaque facial recognition based on an improved YOLOX model. *Zoological research, 46*(2), 339.

Zhang XY, Zhou XY, Lin MX, et al. 2018. ShuffleNet: an extremely efficient convolutional neural network for mobile devices. *In*: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 6848–6856.

Zhao ZY, Lu YY, Tong YJ, et al. 2023. PENet: a phenotype encoding network for automatic extraction and representation of morphological discriminative features. *Methods in Ecology and Evolution*, **14**(12): 3035–3046.