An efficient and consistent framework for multi-rank taxonomic identification in wildlife images

scientific reports

Check for updates

OPEN

An efficient and consistent framework for multi-rank taxonomic identification in wildlife images

Qianqian Zhang✉, Khandakar Ahmed, Chenhao Xu, Muhammad Imad Khan & Hua Wang

Accurate and scalable taxonomic classification is essential for biodiversity research, supporting systematic species identification across multiple hierarchical ranks. However, current image-based classification methods often fail to enforce taxonomic consistency, a critical limitation that undermines the reliability of their outputs for scientific use. Additionally, field-based biodiversity studies are constrained by limited computational resources and network availability on edge devices. To address these challenges, this paper proposes TaxonomyNet, an ensemble detection model with six independent heads for taxonomic classification, achieving high detection performance across all ranks (mAP: 90.7–99.75%) after training on a dataset of 50 Australian animal species. Furthermore, to resolve the core challenge of prediction inconsistency, we introduce the Weighted Agreement Loss (WAL) metric–a confidence-weighted disagreement measure designed to enforce structural coherence between predicted outputs and a reference taxonomy. Crucially, the application of this consistency-enforcing mechanism enhances hierarchical classification reliability, improving final species-level accuracy by up to 3.87% compared to baseline and recent published domain-specific foundation models, while also demonstrating superior computational efficiency, reducing delay by 22 minutes across 1500 samples and making it highly suitable for deployment on edge devices. This work provides a practical and extensible solution for reliable hierarchical classification in real-world biodiversity monitoring scenarios.

Keywords Image recognition, YOLO, Deep learning, Australian species, Convolutional neural network

Taxonomy classification provides the foundation for understanding biodiversity, enabling the systematic organisation of species based on morphological and evolutionary traits across hierarchical ranks. While molecular techniques such as DNA sequencing offer precise classification, they are often limited by high costs, long processing times, and dependence on laboratory infrastructure, making them less suitable for preliminary screening or broad-scale categorisation in field settings. In contrast, visual identification remains the preferred approach for rapid field-level assessment due to its accessibility and scalability [1]. However, manual taxonomy is increasingly impractical given the volume of data generated by modern monitoring systems, with projects like Snapshot Serengeti [2] producing hundreds of gigabytes of images per season and requiring months of expert effort for annotation [3].

In recent decades, deep learning has advanced image recognition and object detection, achieving remarkable success in identifying individual objects and groups[4–6]. However, automating taxonomic classification introduces unique challenges. The most critical requirement for practical applications in taxonomy is the enforcement of hierarchical consistency: predictions at lower ranks (e.g., species) must be coherent with those at higher ranks (e.g., genus and family)[7]. A model that identifies an animal as a Felis catus (species) but places it in the Canidae (family of dogs) is fundamentally flawed and untrustworthy, regardless of its accuracy on a single rank. Despite advancements, many existing animal identification projects focus on object detection within images, without addressing the hierarchical classification tasks required for taxonomic applications[8–12].

On the other hand, biologists conducting fieldwork in remote environments often face significant constraints. While recent advances in edge computing have enabled partial on-site processing for ecological monitoring,

Institute for Sustainable Industries & Liveable Cities (ISILC), Victoria University, 70/104 Ballarat Rd, Melbourne 3011, Australia. ✉email: qianqian.zhang@vu.edu.au

field-based biologists still face challenges in deploying deep learning models under constrained computational environments and unstable connectivity [13].

To address the above challenges, this paper proposes a multi-rank taxonomic detection and classification model that supports parallel prediction across six hierarchical ranks: phylum, class, order, family, genus, and species. The model integrates multiple rank-specific heads to improve classification accuracy at its corresponding rank. To further improve consistency across ranks, a Weighted Agreement Loss (WAL) metric is proposed to align predicted outputs with a reference taxonomy dictionary while maintaining computational efficiency.

Specifically, the model applies a lightweight post-processing correction mechanism that integrates predictions from all ranks to enforce hierarchical coherence. This design ensures scalability and reliability under resource-constrained conditions, making it suitable for edge deployment in real-world biodiversity monitoring scenarios.

The primary contributions of this work are as follows:

- A novel multi-head detection model concurrently working across different ranks, improves detection and classification accuracy at the hierarchical taxonomic ranking level.
- A biologically confidence-weighted WAL metric helps further improve detection consistency across hierarchical taxonomic rankings while minimising computing burdens.
- Experiments evaluated on a structured dataset with 50 Australian species validate that the proposed model outperforms classic detection models and foundation models in mAP by 13% and 5%, and 882ms in latency.

The rest of this paper is organised as follows. Section Related work reviews prior work on hierarchical taxonomic classification and multi-criteria decision-making strategies. Section Methodology introduces the overall model architecture, followed by the proposed consistency correction mechanism, the construction of a structured taxonomic dictionary, and a discussion on deployment-related considerations. Section Experiments and performance evaluation describes the experimental setup, detection and classification performance comparisons with existing models, and an evaluation of computational efficiency.

Related work

Multiple taxonomic rank identification

Artificial neural networks have made significant advancements in biological identification, expanding research beyond the recognition of individual animals or populations to include the automatic identification of species and their higher taxonomic ranks. Song-Quan and Suhaila evaluated the performance of various Convolutional Neural Network (CNN) models on a custom insect image dataset for three taxonomic ranks, including order, family, and genus [14]. Their findings indicate that a single deep-learning architecture is insufficient to effectively capture the hierarchical relationships among multiple taxonomic ranks, often resulting in inconsistent predictions across levels. In particular, rank-specific characteristics diverge–class imbalance, intra-/inter-class variance, and decision margins differ across levels–making a single loss, thresholding rule, or calibration scheme ill-suited for simultaneous optimisation. This limitation directly motivates the development of multi-head frameworks such as TaxonomyNet, which are explicitly designed to maintain cross-rank consistency. Separately, the study relied on images with plain white backgrounds, a setting uncommon in field conditions, which may limit external validity under cluttered or noisy natural scenes. Addressing both the architectural and data-related challenges is therefore essential for developing robust and practical taxonomic identification systems.

Bjerge et al. proposed a hierarchical classification framework to identify insect species across taxonomic ranks, such as order, family, and species[15]. Their approach utilised convolutional neural networks, including ResNet and EfficientNet. By incorporating a simplified taxonomy and anomaly detection, their method achieved significant improvements in classification accuracy at higher taxonomic ranks. However, evaluations of previously unseen data revealed limited generalisation capabilities, underscoring the need for enhancements in robustness and adaptability.

Recent studies have also explored the integration of natural language and image data for taxonomic classification tasks, leveraging the capabilities of LLMs. Chavez et al. used the Wikimedia animal dataset, which includes annotated images, to assess the effectiveness of the Contrastive Language-Image Pre-training (CLIP) model in multi-rank taxonomic classification[16]. Their evaluation revealed significant variability in model performance across taxonomic ranks, with lower ranks consistently underperforming compared to higher ranks, even after applying data augmentation techniques. This disparity is attributed to factors such as insufficient semantic associations between image and text data, limited availability of annotated datasets, and the challenge of effectively modeling distinct feature distributions across multiple taxonomic ranks within a single architecture. These findings highlight the need for specialised approaches to address the complexities of hierarchical classification.

Multi-criteria decision-making methods

Multi-Criteria Decision-Making (MCDM) methods provide structured approaches for evaluating alternatives based on multiple, potentially conflicting criteria. In the context of classification systems–particularly those involving hierarchical structures such as taxonomic ranks–MCDM frameworks support model selection, prediction reconciliation, and output interpretation when multiple evaluation criteria must be considered concurrently.

MCDM techniques have been widely applied across diverse domains, including engineering [17], environmental management [18], healthcare [19], and energy management [20]. In specialised decision-making systems, expert input is often required to determine criteria weights. However, limitations such as the availability of domain experts and the inherent subjectivity of manually assigned weights present challenges to the robustness and scalability of these methods [21,22]. To address these issues, a range of classical MCDM techniques has been employed, including

the Analytic Hierarchy Process (AHP)[23], the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS)[20], and the Weighted Scoring Model (WSM) [24].

In hierarchical classification, ensuring consistency across predicted taxonomic ranks remains a central challenge. Kosmopoulos et al. identified the shortcomings of flat evaluation metrics and introduced hierarchical-aware performance measures that better align with underlying class structures[25]. Similar concerns have been addressed in ensemble-based models, where prediction confidence is integrated into weighted decision frameworks to improve model interpretability and performance[26].

Recent advances in MCDM research have focused on reducing dependence on expert-defined weights by incorporating outputs generated by machine learning models. This has led to the emergence of adaptive, data-driven MCDM strategies, in which confidence scores and probabilistic estimates are directly embedded in the decision-making process [27,28]. Such methods are particularly valuable in domains such as taxonomic classification, where expert input may be unavailable and inter-rank dependencies are both critical and complex.

The proposed WAL metrics in this study depart from traditional MCDM frameworks and can be characterised as a non-linear, data-adaptive decision model. It extends the principles of multi-criteria ranking by integrating structural consistency and probabilistic confidence scores into a unified optimisation function. The WAL metrics are specifically designed to address the unique constraints of hierarchical taxonomic rank prediction, where biological plausibility and model uncertainty must be jointly considered.

Methodology

This section details the architectural design of the proposed multi-rank detection model, including the Model Overview, the WAL Metrics, Taxonomic Rank Dictionary, and considerations for deployment.

Model overview

The proposed model employs a multi-head architecture to enable parallel classification across six taxonomic ranks as illustrated in Fig. 1.

An input image is first processed by the shared feature extractor, which encodes morphological characteristics such as texture, contour, and spatial structure. The resulting feature map provides a unified representation for all downstream tasks. The task-specific heads include:

- *Species head*: Outputs the predicted species of the input image based on the features extracted by the backbone.
- *Genus head*: Specialised to predict the genus of the image using shared features, operating independently from other heads.
- *Family head*: Focuses on classifying the family to which the detected species belongs, without relying on predictions from other ranks.
- *Order head*: Predicts the order of the specimen, distinguishing broader taxonomic groups in an independent manner.
- *Class head*: Responsible for identifying the class of the specimen using the shared representation.

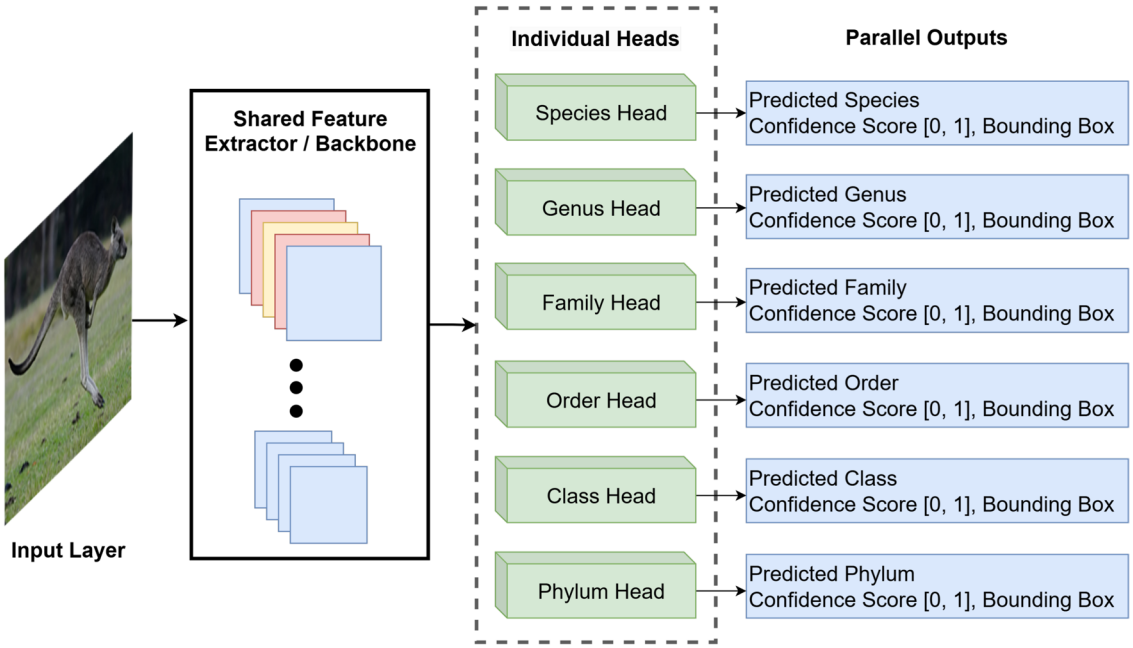

Fig. 1. The architecture of the TaxonomyNet. *Note*: The *Macropus giganteus* (Shaw, 1790) image on the left side of the figure is observed in Australia by Samuel Lee and is licensed under [CC BY 4.0] (http://creativecommons.org/licenses/by/4.0/).

- *Phylum head*: Outputs the predicted phylum, providing the broadest classification of the input image.

Each classification head produces a set of outputs: the predicted category, a confidence score, and the bounding box. Each head functions independently while sharing the same feature representation, allowing parallel predictions across ranks without hierarchical dependency.

All six heads operate concurrently during inference, enabling the model to produce predictions across all taxonomic ranks in parallel. The outputs from this stage are then passed to a post-processing module responsible for evaluating inter-rank consistency.

WAL metrics

Although the multi-rank detection model described in Section Model overview enables simultaneous prediction across six taxonomic ranks, each output head operates independently. This independence may result in predictions that deviate from biologically coherent taxonomic hierarchies. Simple correction strategies, such as inferring higher ranks from species rank predictions or applying majority voting, can introduce systematic bias, particularly in the presence of uncertain or inconsistent outputs. A principled correction mechanism is therefore required to enforce structural consistency across taxonomic ranks while preserving confidence-driven predictions.

Let $\mathcal{R} = \{\text{phylum}, \text{class}, \text{order}, \text{family}, \text{genus}, \text{species}\}$ denote the set of taxonomic ranks. For a given input instance, the model produces rank-wise predictions

$$\hat{r}_i = (\hat{c}_i, \hat{p}_i), \quad \forall i \in \mathcal{R} \tag{1}$$

where $\hat{c}_i$ is the predicted taxon for rank *i*, and $\hat{p}_i \in [0, 1]$ is the associated confidence score.

To reconcile potential inconsistencies, a data-driven correction mechanism termed Weighted Agreement Loss (WAL) is introduced. The method evaluates each taxonomic rank as a potential anchor and selects the anchor whose reconstructed taxonomy exhibits the highest agreement with the original predictions. Let $a \in \mathcal{R}$ denote a candidate anchor. Based on the predicted label $\hat{c}_a$, a full hierarchy $H_a = \{H_a[r] : r \in \mathcal{R}\}$ is reconstructed using a reference taxonomy dictionary $\mathcal{D}$.

The agreement loss for anchor *a* is defined as

$$\mathcal{L}_{\text{WAL}}(a) = \sum_{r \in \mathcal{R}} \delta(\hat{c}_r \neq H_a[r]) \cdot \hat{p}_r \tag{2}$$

where $\delta(\cdot)$ is the indicator function that returns 1 when the predicted label $\hat{c}_r$ differs from the reconstructed label $H_a[r]$, and 0 otherwise. The optimal anchor is selected by minimising the total weighted disagreement:

$$a^* = \arg\min_{a \in \mathcal{R}} \mathcal{L}_{\text{WAL}}(a) \tag{3}$$

When reconstructing a hierarchy from a candidate anchor above the species level (e.g., family), the process does not involve arbitrarily selecting one of the multiple genera or species that may belong to that family. Instead, the reconstruction relies on the taxonomic dictionary, which provides a unique mapping from each species to its corresponding higher ranks. For any anchor, the dictionary is used to generate a complete and internally consistent hierarchy H_a. The indicator function then evaluates disagreement by comparing the model's predicted outputs with this reconstructed hierarchy. This ensures that the reconstruction process is deterministic and avoids ambiguity, even when the anchor is chosen at family, order, or class levels. The hierarchy H_{a^*} derived from the optimal anchor a^* is then adopted as the corrected output. This strategy integrates confidence-weighted consistency with a hierarchical structure, enabling correction decisions that favour both reliability and biological plausibility. Figure 2 illustrates the overall process of the WAL correction mechanism. The WAL-based approach offers a principled alternative to rule-based or anchor-fixed correction schemes, improving alignment with taxonomic conventions in multi-rank classification tasks.

Taxonomic rank dictionary

Predictions for each taxonomic rank are independently produced by their respective classification heads, forming a preliminary hierarchy for each input image. However, because these heads operate in parallel, inter-rank dependencies are not explicitly encoded, which may lead to structural inconsistencies. To evaluate and correct such inconsistencies, a taxonomic rank dictionary was developed as an external reference resource.

The dictionary was constructed by referencing the internationally recognised taxonomic framework and definitions adopted by the Global Biodiversity Information Facility (GBIF) [29]. It consolidates the hierarchical relationships among the 50 species included in the training dataset and their corresponding higher ranks–genus, family, order, class, and phylum. This reference structure provides a verified and standardised taxonomy, ensuring that all consistency checks are grounded in authoritative biological classification rather than model-derived relations.

During post-hoc validation, the taxonomic rank dictionary serves as the ground-truth reference for assessing hierarchical consistency among independently predicted ranks. Each prediction is compared against the corresponding entry in the dictionary to detect and correct structural conflicts, such as mismatched genus–family or class–order assignments. The dictionary functions solely as a verification reference and does not generate, infer, or replace any model predictions. The complete taxonomic rank dictionary is available at https://cutt.ly/qeV7BIz8https://cutt.ly/qeV7BIz8.

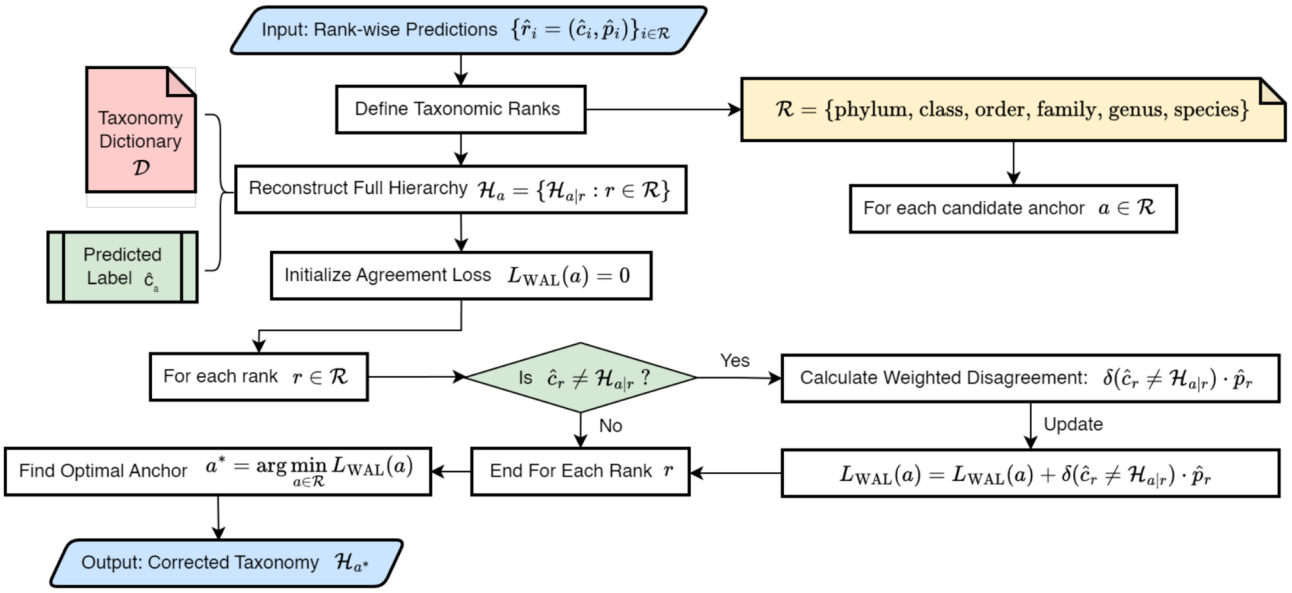

Fig. 2. Workflow of the WAL-based taxonomic correction mechanism.

Further discussion

The proposed model is designed with deployment flexibility, particularly for field-based ecological applications where computational resources and network connectivity are often limited. All inference and correction operations are executed locally, eliminating reliance on external servers or cloud infrastructure. This structure supports real-time decision-making and ensures continued operation in network-disconnected environments.

Edge deployment enables wildlife images captured by camera traps to be processed directly on embedded devices. The model's lightweight design reduces latency and memory requirements, making it suitable for low-power processors commonly used in remote monitoring stations. This configuration minimises bandwidth usage, preserves sensitive data, and allows on-site filtering before selectively transferring results for expert review or long-term analysis.

The system's modularity and low computational overhead allow it to be integrated into a range of field-ready platforms, including solar-powered camera traps, mobile ecological stations, and portable diagnostic kits. These deployment scenarios demonstrate the model's adaptability and practical relevance to biodiversity research conducted under resource-constrained conditions.

Experiments and performance evaluation

Experimental setup

The training and evaluation process is conducted on a high-performance computing system equipped with an NVIDIA A40 GPU with 24,362 MiB of memory. The software environment was configured with Python 3.10.12 and implemented using PyTorch 2.1.1 with CUDA 12.1 for GPU acceleration. The training configuration includes key hyperparameters such as an initial and final learning rate of 0.01, a momentum factor of 0.937, and a weight decay of 0.0005, selected to optimise convergence and generalisation performance.

To evaluate the performance of the proposed model across all taxonomic ranks, standard object detection and classification metrics are adopted. These include Precision (P), Recall (R), Average Precision (AP), mean Average Precision (mAP), and Intersection over Union (IoU). For the overall accuracy reported in subsequent tables, we employ micro-averaged accuracy, which is calculated by dividing the total number of correct predictions by the total number of instances across all classes. This instance-centric metric was chosen as it directly reflects the model's overall real-world performance in identifying individual animals, which is the primary goal of the monitoring application. These metrics collectively assess both the classification and localisation accuracy of the proposed model across all taxonomic ranks.

Detection

The dataset used for training, testing, and validation was derived from a curated collection of 50 Australian animal species [30]. It underwent a rigorous screening process to ensure its quality and relevance, with species selected from the GBIF repository based on their frequency of occurrence records within Australia and their popularity inferred from web search trends, reflecting both ecological abundance and anticipated user interest. For each species, 700 images were selected, resulting in a total of 35,000 annotated instances. A stratified random-sampling script was employed to divide the images into three independent subsets: 600 for training, 70 for testing, and 30 for validation. All images were screened using both metadata and filename checks to remove duplicates or re-uploaded copies, ensuring that each image appeared only once in the dataset. Non-representative samples, such as images containing tracks, feathers, faeces, or partial remains, were excluded during curation, while no selective bias was introduced during sampling. The images were contributed by wildlife camera traps

Model	Precision	Recall	mAP@0.5	mAP@0.5:0.95
Faster R-CNN (ResNet50)	0.4691	0.4549	0.4691	0.2605
Default YOLOv5	0.8565	0.7675	0.7988	0.6885
Default YOLOv9	0.8607	0.7425	0.8040	0.7203
YOLOv9 + Deeper Head	0.8620	0.7520	0.8120	0.7308
YOLOv9 + Deeper Head + Deeper Backbone	0.8600	0.7430	0.8153	0.7364
YOLOv11	0.8910	0.7890	0.8400	0.7650
YOLO-FCE	0.9380	0.8320	0.8774	0.8189

Table 1. Comparison of detection performance across different models.

Taxonomic rank	P (%)	R (%)	mAP50 (%)	mAP75 (%)	mAP95 (%)	mAP50-95 (%)
Species	98.20	87.50	90.80	89.70	72.40	87.50
Genus	97.30	87.50	91.10	89.70	68.40	87.00
Family	99.40	89.10	91.80	91.00	78.90	89.60
Order	98.80	87.10	90.50	89.30	73.60	87.40
Class	99.60	86.10	88.90	88.20	76.80	86.90
Phylum	99.60	81.40	84.30	83.50	72.40	82.30

Table 2. Performance metrics of all taxonomic rank heads.

Taxonomic ranks	No. of instances	No. of incorrect predictions	Accuracy (%)	Avg. Conf. score
Species	1780	161	90.96	0.802562097
Genus	1850	172	90.70	0.781795928
Family	1677	96	94.28	0.84496042
Order	1709	79	95.38	0.832841211
Class	1648	38	97.69	0.862872421
Phylum	1623	4	99.75	0.873138387

Table 3. Validation results of models across six heads.

and public submissions, capturing animals in uncontrolled natural environments. This composition closely mirrors real-world conditions, improving model robustness and ecological validity. The latest version of this dataset is accessible at https://cutt.ly/gwqHHm8D.

In our previous study, the YOLO-FCE model [31] was developed and trained on the training set, which demonstrated strong performance, achieving a mAP:50 of 87.74% and a precision of 93.8%. On the validation set, YOLO-FCE achieved a species identification accuracy of 91.29% with an average confidence score of 0.801, based on a 'top-1' protocol that evaluates only the single highest-confidence prediction per image. This model serves as the foundational architecture for the multi-rank TaxonomyNet proposed in the present work. Table 1 summarises the detection results between YOLO-FCE and several widely adopted object detection architectures, including Faster R-CNN, YOLOv5, various YOLOv9 configurations, and the recently released YOLOv11. Among the evaluated models, YOLO-FCE achieves the highest precision (0.9380) and overall detection performance.

To extend this work, the 50 species were reorganised into five higher taxonomic ranks: phylum, class, order, family, and genus. Species within the same rank were systematically merged to form new categories. For example, species *Litoria fallax* and species *Litoria peronii* were combined into the genus *Litoria*, while genus *Litoria* and genus *Ranoidea* were further merged into the family Pelodryadidae. The dataset was restructured by assigning unique identifiers and updating labels to reflect this hierarchical taxonomy, ensuring consistency and compatibility for training models across different ranks. The detailed distribution of images and labels for each rank is available in the Supplementary File 1, which is also available at https://cutt.ly/LtoOCPBm.

The individual heads were optimised based on reorganised datasets. Table 2 summarises the overall performance of each classification head after 300 epochs of training. Results indicate a general trend of increasing precision at higher taxonomic ranks, while AP and mAP scores gradually decline from species to phylum. This inverse relationship reflects the trade-off between classification granularity and generalisability.

Validation

Validation was performed on an independent dataset comprising 1,500 unseen images. These images were organised by rank and evaluated individually using the corresponding classification head. For each image, the model performed object detection, generated bounding boxes, assigned confidence scores, and produced classification outputs. Table 3 presents the validation results across the six heads. It is important to note that these

metrics are derived from a comprehensive evaluation of all detected instances with a confidence score exceeding 0.5. This multi-instance protocol is more rigorous than the 'top-1' approach and explains why the number of instances reported in the table is greater than the 1,500 images in the validation set. Detailed class-wise accuracy statistics and confidence distributions are provided in Supplementary File 2, which is also available at https://cutt.ly/4toOCM3g.

AP was adopted as the primary evaluation metric due to its ability to integrate both precision and recall, providing a balanced and threshold-independent assessment of model performance. AP is particularly suited to scenarios involving class imbalance or varying object scales, offering a robust measure of both detection accuracy and localisation quality.

Traditional accuracy metrics, which quantify the proportion of correct predictions, proved insufficient for evaluating multi-instance object detection tasks. In images containing multiple species, accuracy failed to account for correct localisation or partial detections. For example, a model could achieve high accuracy by correctly identifying at least one instance while failing to detect others present in the scene. This limitation became evident during validation, where several images annotated with a single species during training were correctly detected with additional, unlabelled species in the scene.

In contrast, AP provides a more comprehensive evaluation by capturing both the spatial precision and class correctness of each detection. To further analyse performance across varying confidence thresholds, AP was computed at multiple ranks (e.g., AP@0.50, AP@0.75, and AP@0.50:0.95). This multi-threshold evaluation reflects the model's ability to balance precision and recall under diverse operating conditions, offering a more nuanced understanding of its detection capabilities in real-world ecological applications.

Efficiency

To evaluate the effectiveness and efficiency of the proposed WAL metrics, a comparative evaluation was conducted between the WAL metric and several locally deployed large-scale language and vision foundation models, focusing on their correction performance and processing efficiency for multi-rank prediction outputs. All experiments presented in this section were conducted using the same independent validation dataset introduced in Section Detection. Identical hardware settings, batch sizes, image resolutions, and confidence thresholds were applied across all models to ensure experimental consistency and fair comparison. In addition, multiple inference runs produced identical predictions, with only minimal variation in total processing time due to routine system-level fluctuations.

Given the constraints of field-deployable systems, emphasis was placed on selecting models that balance computational feasibility with predictive capability. Four widely adopted language models in the 7B–9B parameter range were selected: Mistral[32], Qwen2.5[33], Gemma2[34], and Llama3.1[35]. In addition, two larger models were included as references: Llama3.3 (70B), and a vision-language foundation model capable of direct image input, Llama3.2-vision. To ensure efficient and reproducible inference, all LLM evaluations were conducted using Ollama, a lightweight and optimised framework for the local execution of language models.

Prompt engineering techniques were applied to optimise the instruction format fed into language models. The prompts were iteratively adjusted to align outputs with structured taxonomic expectations, allowing for maximum model adaptation to the multi-rank prediction task.

Table 4 reports the corrected accuracy at each taxonomic rank and the average inference time per image. The results indicate that the WAL metrics achieve higher or comparable accuracy across all ranks. Specifically, species rank accuracy using WAL reached 93.20%, outperforming all foundation models included in the comparison. This represents a substantial enhancement in classification reliability, marking an improvement of up to 3.87% compared to the fastest evaluated LLM (Mistral, 89.33%)." In contrast, WAL metrics require significantly less computational time, approximately 27% of the processing time of the fastest evaluated LLM (Mistral). These results highlight the distinct roles of these technologies: the WAL metric is a lightweight and deterministic corrector, optimised for speed and reliability on edge devices, while LLMs act as generalist reasoners. The experiment was designed to assess if this general reasoning could be a viable alternative, but the findings suggest that for this highly structured correction task, the specialised approach is superior.

Notably, the genus and species rank predictions from larger models (llama3.3) suffered from severe degradation despite extended processing times. The particularly severe performance degradation of the vision-language model Llama3.2-Vision, which directly processes original image inputs, underscores a fundamental task mismatch. VLMs are primarily optimised for open-ended generative tasks, not for the fine-grained, structured-output task of taxonomic predictions.

	Processing time (s)	Phylum (%)	Class (%)	Order (%)	Family (%)	Genus (%)	Species (%)
WAL	494.9899	99.6667	98.2667	96.0667	94.9333	93.6667	93.2000
Mistral	1818.318	99.6667	97.6000	94.1333	94.5333	92.6000	89.3333
Qwen2.5	1849.633	99.6667	98.2000	94.8000	94.0000	91.9333	92.3333
Gemma2	2635.679	99.6667	98.0000	95.2667	94.0667	93.2000	92.7333
Llama3.1	1841.822	99.6000	97.6667	95.0667	93.4667	91.4667	90.9333
Llama3.3	125981.9	99.5333	94.5333	93.5333	93.8667	92.8000	91.0000
Llama3.2-vision	13604.77	95.6667	91.3333	72.4000	51.4667	36.2000	11.4667

Table 4. Comparison of taxonomic prediction accuracy and inference time between WAL and LLM-based methods.

	Processing time (s)	Phylum (%)	Class (%)	Order (%)	Family (%)	Genus (%)	Species (%)
Mistral	5228.704	99.4000	97.4667	94.8667	92.0000	92.8667	92.4000
Qwen2.5	5133.961	99.4000	97.4667	95.1333	93.9333	92.8667	91.9333
Gemma2	6986.031	99.4000	97.4667	95.1333	93.6667	92.8667	92.4000
Llama3.1	4742.822	99.4000	97.4667	94.0000	93.9333	92.7333	92.4000
Llama3.3	150952.9	99.4000	97.4667	95.0000	93.9333	92.8667	92.4000

Table 5. Taxonomic prediction accuracy and inference time of LLM-based methods with enforced knowledge.

Fig. 3. Example prediction results from TaxonomyNet on challenging wildlife images. (**a**) Small targets sample (species rank prediction result for *Dromaius novaehollandiae*). Image credit: *Dromaius novaehollandiae* (Latham, 1790) Observed in Australia by Sean Frey, licensed under CC BY 4.0. (**b**) Extreme angles sample (species rank prediction result for *Phascolarctos cinereus*). Image credit: *Phascolarctos cinereus* (Goldfuss, 1817) Observed in Australia by Lynn Roberts, licensed under CC BY 4.0. (**c**) Noisy backgrounds sample (species rank prediction result for *Pogona barbata*). Image credit: *Pogona barbata* (Cuvier, 1829) Observed in Australia by Geoff Shuetrim, licensed under CC BY 4.0. (**d**) Low-resolution sample (species rank prediction result for *Tursiops aduncus*. Image credit: *Tursiops aduncus* (Ehrenberg, 1833) Observed in Australia by Olivia, licensed under CC BY 4.0.).

To ensure biologically valid outcomes, the taxonomy dictionary introduced in Section Taxonomic Rank Dictionary was incorporated into the prompting process as an enforced knowledge constraint. Table 5 reports the processing time and accuracy for each language model. Notably, the integration of enforced taxonomic knowledge led to improved classification accuracy at the species rank compared to experiments without external constraints. For example, species rank accuracy increased consistently across all models. In contrast, the accuracies of intermediate ranks exhibited slight declines. These trends suggest that, when equipped with a structured taxonomy dictionary, LLMs may prioritise species rank predictions and retrieve corresponding higher ranks through direct dictionary lookup rather than independent validation. Additionally, the incorporation of enforced knowledge significantly increased the processing time of all LLMs. It is also plausible that the LLMs' underperformance was exacerbated by the sparse input provided in this study, which consisted only of a predicted label, bounding box coordinates, and a confidence score. In alternative scenarios where richer contextual information (e.g., habitat descriptions, temporal data) is available, the reasoning capabilities of LLMs might be leveraged more effectively.

Overall, the proposed WAL method outperformed all baseline models in both computational efficiency and taxonomic rank correction accuracy. All complete prompts used for LLM-based methods are available in File 3 and can be accessed online at https://cutt.ly/utoOVtLA. The model's design facilitates efficient inference under constrained computing resources, making it suitable for decentralised deployments in field-based biodiversity monitoring. As a future extension, the development of a lightweight, user-facing platform–such as a mobile-compatible application–would enable real-time species identification in remote environments without reliance on cloud infrastructure.

Discussion

The proposed model demonstrates significant progress in multi-rank species classification, leveraging a dataset comprising 50 Australian animal species and their corresponding higher taxonomic ranks. As illustrated in Fig. 3, the model maintains robust performance across a range of visually challenging scenarios, including images featuring Fig. 3a, small targets, Fig. 3b extreme angles, Fig. 3c noisy backgrounds, and Fig. 3d low resolution. These results highlight the model's adaptability and reliability in practical biodiversity monitoring contexts.

Furthermore, even when species rank identification fails due to encountering an unknown species or severe image degradation, TaxonomyNet can still provide a robust and confident classification at a higher taxonomic rank, such as family or genus. For biodiversity monitoring programmes, this output represents reliable data rather than analytical noise and is substantially more useful than a forced, incorrect species label. In addition, the model's multi-rank output directly serves ecological research questions focused on broader, supra-specific patterns, streamlining data extraction for community-level analyses. Thus, the proposed framework should be viewed not merely as a more accurate classifier but as a more practical and versatile tool designed to meet the nuanced data requirements of modern biodiversity research.

	Phylum (%)	Class (%)	Order (%)	Family (%)	Genus (%)	Species (%)
TaxonomyNet	99.67	98.27	96.07	94.93	93.67	93.20
BioClip	41.20	56.80	37.53	28.00	47.00	53.13

Table 6. prediction accuracy across six taxonomic ranks on validation set.

More recently, the field has seen the emergence of large-scale, vision-language foundation models pre-trained specifically on biological data. These foundation models demonstrate remarkable zero-shot and few-shot capabilities across a vast range of taxa. A notable example is BioCLIP [36], which leverages millions of image-text pairs from sources like iNaturalist to learn generalisable representations of the natural world. However, their primary strength lies in their breadth and generality, which often comes with two trade-offs: substantial computational requirements and potentially lower accuracy on specific, regional datasets compared to a specialised model. Table 6 compared the accuracy across six ranks between BioClip and TaxonomyNet. When evaluated on previously unseen images, TaxonomyNet demonstrated outstanding accuracy.

Despite its effectiveness, the current dataset presents certain limitations. The training distribution exhibits morphological bias, which adversely affects the model's performance when encountering species that deviate significantly from those represented in the training set. Furthermore, the dataset is imbalanced across higher taxonomic ranks, with most species belonging to the phylum Chordata and only minimal representation of other phyla. As a result, the performance reported for class and phylum levels should be considered preliminary. In addition, only three genera in the dataset currently contain more than one species, meaning that results at the genus rank largely overlap with those at the species rank. This constraint reflects the present scope of the dataset and will be addressed in future work through the inclusion of additional species across multiple genera. Although the post-processing WAL correction mechanisms improve hierarchical consistency, these methods remain susceptible to inherited dataset biases. Addressing this issue will require expanding the training data to include a broader range of species from diverse ecosystems and geographic regions, thereby enhancing model generalisability.

Another limitation lies in the current single-target annotation protocol. Each image in the dataset was assigned a single bounding box and species label, even when multiple organisms were present. Consequently, while the model is capable of detecting more than one object per image, it is constrained to report only the primary annotated target. This design choice was made to reduce annotation complexity, as accurately labelling multiple species within a single image demands domain expertise and substantial labour investment. Although multi-target detection remains an important extension, its implementation is expected to require further infrastructure and expert collaboration.

To assess model robustness, a comparative evaluation was conducted against foundation models, including language models for taxonomy-aware correction and vision-language models for direct classification. Results indicate that despite being trained on large-scale corpora, foundation models do not consistently outperform the proposed WAL-based method. This performance gap is particularly evident under constrained input conditions, where foundation models rely solely on sparse textual or structured inputs. Although a small-scale evaluation using 100 samples was conducted on state-of-the-art multimodal models–such as GPT-o1, Gemini2.5, and DeepSeek-r1–their accuracy was comparable to that of smaller LLMs, while requiring significantly more computational resources. These findings suggest that foundation models may benefit from richer, multimodal input pipelines; however, the computational trade-offs pose challenges for deployment in field-based or edge environments.

Conclusion

This study presents a multi-rank detection model for automated taxonomic classification across six hierarchical ranks, integrating rank-specific prediction heads with a unified feature extractor and a post-hoc correction mechanism based on the WAL metrics. Experiments on a curated dataset of 50 Australian animal species demonstrate the model's robustness under diverse visual conditions, with the WAL-based strategy consistently outperforming both standalone predictions and corrections by foundation models, while maintaining lower computational cost. Designed for decentralised deployment, the model supports real-time inference on edge devices without dependence on external infrastructure, making it well-suited for biodiversity monitoring in remote or resource-constrained environments. Future work will explore multi-object detection, unknown species recognition, and mobile deployment to further enhance field applicability.

Data availability

The dataset of 50 species used in this study is publicly available at https://cutt.ly/gwqHHm8D. The datasets used for the higher taxonomic ranks (genus, family, order, class, and phylum) are derived from the same underlying image collection but differ in naming conventions and label structures. These higher-rank datasets can be provided by the corresponding author upon reasonable request.

Received: 21 July 2025; Accepted: 31 December 2025
Published online: 21 January 2026

References

1. Mayr, E. *Systematics and the origin of species, from the viewpoint of a zoologist* (Harvard University Press, 1999).

2. Swanson, A. et al. Data from: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African Savanna. https://doi.org/10.5061/dryad.5pt92(2015).
3. Moallem, G., Pathirage, D. D., Reznick, J., Gallagher, J. & Sari-Sarraf, H. An explainable deep vision system for animal classification and detection in trail-camera images with automatic post-deployment retraining. *Knowl.-Based Syst.***216**, 106815. https://doi.org/10.1016/j.knosys.2021.106815 (2021).
4. Martvel, G. et al. Dog facial landmarks detection and its applications for facial analysis. *Sci. Rep.***15**, 21886. https://doi.org/10.1038/s41598-025-07040-3 (2025).
5. Cao, Z. et al. Semi-automated annotation for video-based beef cattle behavior recognition. *Sci. Rep.***15**, 1–16. https://doi.org/10.1038/s41598-025-01948-6 (2025).
6. Li, R. et al. Open-vocabulary multi-object tracking with domain generalized and temporally adaptive features. *IEEE Trans. Multimedia***27**, 3009–3022. https://doi.org/10.1109/TMM.2025.3557619 (2025).
7. Horton, T. et al. Recommendations for the standardisation of open taxonomic nomenclature for image-based identifications. *Front. Marine Sci.***8**, https://doi.org/10.3389/fmars.2021.620702 (2021).
8. Mulero-Pázmány, M. et al. Addressing significant challenges for animal detection in camera trap images: A novel deep learning-based approach. *Sci. Rep.***15**, 1–18. https://doi.org/10.1038/s41598-025-90249-z (2025).
9. Vidal, M., Wolf, N., Rosenberg, B., Harris, B. P. & Mathis, A. Perspectives on individual animal identification from biology and computer vision. *Integr. Comp. Biol.***61**, 900–916. https://doi.org/10.1093/icb/icab107 (2021).
10. Tang, J., Zhao, Y., Feng, L. & Zhao, W. Contour-based wild animal instance segmentation using a few-shot detector. *Animals***12**, https://doi.org/10.3390/ani12151980 (2022).
11. Roy, A. M., Bhaduri, J., Kumar, T. & Raj, K. Wildect-yolo: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Eco. Inform.***75**, 101919. https://doi.org/10.1016/j.ecoinf.2022.101919 (2023).
12. Simões, F., Bouveyron, C. & Precioso, F. Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. *Eco. Inform.***75**, 102095. https://doi.org/10.1016/j.ecoinf.2023.102095 (2023).
13. Chen, G. et al. Deep reinforcement learning-based cloud-edge offloading for WBANS. *IEEE Trans. Consum. Electron.* 1–1, https://doi.org/10.1109/TCE.2024.3504545 (2024).
14. Ong, S.-Q. & Hamid, S. A. Next generation insect taxonomic classification by comparing different deep learning algorithms. *PLoS ONE***17**, 1–11. https://doi.org/10.1371/journal.pone.0279094 (2022).
15. Bjerge, K. et al. Hierarchical classification of insects with multitask learning and anomaly detection. *Ecol. Inform.***77**, 102278. https://doi.org/10.1016/j.ecoinf.2023.102278 (2023).
16. Chavez, R. K. E., Reynoso, K. G. M., Raquel, C. R. & Naval, P. C. Leveraging large image-caption datasets for multimodal taxon classification. In Nguyen, N. T. et al. (eds.) *Recent Challenges in Intelligent Information and Database Systems*, 13–24, https://doi.org/10.1007/978-981-97-5934-7_2 (Springer Nature Singapore, Singapore, 2024).
17. Zavadskas, E. K., Turskis, Z. & Kildienė, S. State of art surveys of overviews on MCDM/MADM methods. *Technol. Econ. Dev. Econ.***20**, 165–179. https://doi.org/10.3846/20294913.2014.892037 (2014).
18. Huang, I. B., Keisler, J. & Linkov, I. Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends. *Sci. Total Environ.***409**, 3578–3594. https://doi.org/10.1016/j.scitotenv.2011.06.022 (2011).
19. Diaby, V., Campbell, K. & Goeree, R. Multi-criteria decision analysis (MCDA) in health care: A bibliometric analysis. *Op. Res. Health Care***2**, 20–24. https://doi.org/10.1016/j.orhc.2013.03.001 (2013).
20. Musbah, H., Ali, G., Aly, H. H. & Little, T. A. Energy management using multi-criteria decision making and machine learning classification algorithms for intelligent system. *Electric Power Syst. Res.***203**, 107645. https://doi.org/10.1016/j.epsr.2021.107645 (2022).
21. Phulara, S., Kumar, A., Narang, M. & Bisht, K. A novel hybrid grey-BCM approach in multi-criteria decision making: An application in OTT platform. *J. Dec. Anal. Intell. Comput.***4**, 1–15. https://doi.org/10.31181/jdaic10016012024p (2024).
22. Shao, M. et al. A review of multi-criteria decision making applications for renewable energy site selection. *Renew. Energy***157**, 377–403. https://doi.org/10.1016/j.renene.2020.04.137 (2020).
23. Ali, R., Lee, S. & Chung, T. C. Accurate multi-criteria decision making methodology for recommending machine learning algorithm. *Expert Syst. Appl.***71**, 257–278. https://doi.org/10.1016/j.eswa.2016.11.034 (2017).
24. Tzeng, G.-H. & Huang, J.-J. *Multiple attribute decision making: Methods and applications* 1st edn. (A Chapman and Hall book, Taylor & Francis, 2011).
25. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G. & Androutsopoulos, I. Evaluation measures for hierarchical classification: A unified view and novel approaches. *Data Min. Knowl. Disc.***29**, 820–865. https://doi.org/10.1007/s10618-014-0382-x (2015).
26. Utkin, L. V., Konstantinov, A. V., Chukanov, V. S., Kots, M. V. & Meldo, A. A. An adaptive weighted deep forest classifier. https://doi.org/10.48550/arXiv.1901.01334 (2019). arXiv:1901.01334.
27. Hafezalkotob, A., Hafezalkotob, A., Liao, H. & Herrera, F. An overview of multimoora for multi-criteria decision-making: Theory, developments, applications, and challenges. *Inf. Fus.***51**, 145–177. https://doi.org/10.1016/j.inffus.2018.12.002 (2019).
28. Khosravi, K. et al. A comparative assessment of flood susceptibility modeling using multi-criteria decision-making analysis and machine learning methods. *J. Hydrol.***573**, 311–323. https://doi.org/10.1016/j.jhydrol.2019.03.073 (2019).
29. Secretariat, T. G. What is GBIF? https://www.gbif.org/what-is-gbif (2025). Accessed: 24 June 2025.
30. Zhang, Q., Ahmed, K., Sharda, N. & Wang, H. Australian animal species selection and image data collection. In *2023 27th International Conference Information Visualisation (IV)*, 55–63, https://doi.org/10.1109/IV60283.2023.00020 (IEEE, 2023).
31. Zhang, Q., Ahmed, K., Khan, M. I., Wang, H. & Qu, Y. Yolo-FCE: A feature and clustering enhanced object detection model for species classification. *Pattern Recogn.***171**, 112218. https://doi.org/10.1016/j.patcog.2025.112218 (2026).
32. Jiang, A. Q. et al. Mistral 7b, https://doi.org/10.48550/arXiv.2310.06825 (2023). arXiv:2310.06825.
33. Bai, J. et al. Qwen technical report. https://doi.org/10.48550/arXiv.2309.16609 (2023). arXiv:2309.16609.
34. Team, G. et al. Gemma: Open models based on gemini research and technology. https://doi.org/10.48550/arXiv.2403.08295 (2024). arXiv:2403.08295
35. Touvron, H. et al. Llama: Open and efficient foundation language models (2023). arXiv:2302.13971.
36. Stevens, S. et al. Bioclip: A vision foundation model for the tree of life (2024). arXiv:2311.18803.

Acknowledgements

The authors would like to acknowledge Victoria University for supporting the publication of this work. We also thank the Australian Research Data Commons (ARDC) Nectar Research Cloud for providing cloud infrastructure and remote servers, which enabled all machine learning training conducted in this study.

Author contributions

Q.Z. and K.A. conceived the study. K.A., Q.Z., C.X., and M.I.K. contributed to the experimental design and refinement. Q.Z. collected the data, conducted the experiments, and performed data analysis and model development. K.A., C.X., M.I.K., and H.W. provided supervision and critical feedback. All authors reviewed and approved the final manuscript.

Funding
This research received no external funding.

Declarations

Competing interests
We would like to disclose that one of the co-authors of this manuscript, Professor Hua Wang, is a member of the Scientific Reports Editorial Board. However, we confirm that there have been no prior discussions with Professor Wang or any other Editorial Board Member regarding the content or submission of this manuscript. All co-authors, including Professor Wang, have adhered to the journal's policies to ensure a transparent and unbiased submission process.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-34944-x.

Correspondence and requests for materials should be addressed to Q.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.