MOSAICO: Hierarchical Classification of Mosquito Species

in collaboration with Guido Gigante (ISS) and Alberto Tubito (ISS)

Advanced Machine Learning for Physics's final project

Mattia Liguoro (2207422)

Academic Year 2024/2025

SAPIENZA
UNIVERSITÀ DI ROMA

Table of Contents

Motivation

- Mosquitoes are primary vectors of diseases (malaria, dengue, Zika, etc).
- Accurate species monitoring is crucial for early prevention.
- Manual classification is:
  - Expensive
  - Time-consuming
  - Dependent on highly trained entomologists
- AI-based automatic classification offers a scalable alternative.

Example of a MOSAICO grid

Project Goals (MOSAICO System)

- Build an automatic hierarchical classifier:
  - Reliable **Genus classification** always.
  - **Species classification** only when confident.
  - Open-set recognition of unknown species.
- Support entomologists in field surveys.
- Enable continuous data collection for long-term improvements.
- Deployable on portable acquisition devices.

Table of Contents

2 Preprocessing Pipeline (Detection, Cropping, Augmentation)

Biological and Data Acquisition Variability

2 Preprocessing Pipeline (Detection, Cropping, Augmentation)

Biological Variability

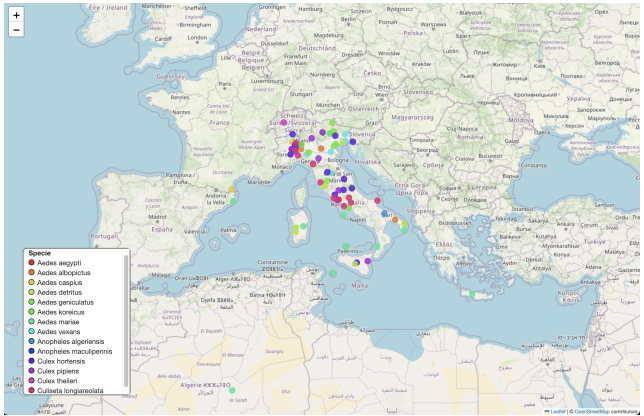
- ∼10,000 mosquitoes annotated by expert entomologists.
- Specimens from:
  - Field collections
  - Controlled laboratory colonies
- 15 species across 4 genera: *Aedes*, *Anopheles*, *Culex*, *Culiseta*.
- Geographic variability across multiple regions.
- High intra-species morphological variability.

Data acquisition variability

- Specimen damage during capture.
- Image acquisition inconsistencies:
  - Illumination, background, camera settings.
- Strong class imbalance:
  - Some species: thousands of images.
  - Rare species: few dozen samples.

Geographical Distribution of Mosquito Samples

2 Preprocessing Pipeline (Detection, Cropping, Augmentation)



Screenshot of the interactive map of mosquito sample locations (Folium). Zoomed in on Italy

Preprocessing Pipeline

2 Preprocessing Pipeline (Detection, Cropping, Augmentation)

- **YOLO Object Detection:**
  - Detects bounding boxes over grid images.
  - Automatically extracts individual mosquito crops.
- **Data Augmentation:**
  - **Geometric augmentations:**
    - Random rotations, flips, shifts
    - Random zoom, cropping, perspective warping
  - **Photometric augmentations:**
    - Gamma correction, white–balance shifts
    - Brightness/contrast/hue/saturation changes
- **Degradation:**
  - Gaussian noise, motion blur, JPEG artifacts
  - Dropout
- **Why so aggressive?**
  - Avoid shortcut learning (illumination/background)
  - Encourage invariance to domain shifts
  - Force the model to learn shape-based features

Data Augmentation Examples

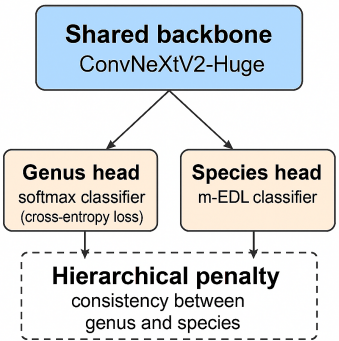
2 Preprocessing Pipeline (Detection, Cropping, Augmentation)

Table of Contents

Hierarchical Multitask Structure

3 Model and architecture

- **Shared backbone:** ConvNeXtV2-Huge.
- **Two heads:**
  - Genus head: softmax classifier (weighted cross entropy criterion).
  - Species head: m-EDL classifier.
- **Hierarchical penalty:** consistency between genus and species.



Scheme of multitask hierarchical model

Why ConvNeXtV2-Huge?

3 Model and architecture

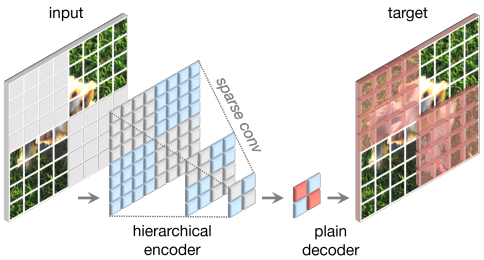
facebookresearch/
ConvNeXt-V2

Code release for ConvNeXt V2 model

- **State-of-the-art visual backbone:** Combines the inductive biases of CNNs with architectural insights from vision transformers.
- **Large capacity (660M parameters):** Allows extraction of fine-grained morphological patterns critical for species-level discrimination.
- **Hierarchical feature maps:** Multi-scale representation is essential for modeling both local and global features.
- **High transferability:** Pretrained on ImageNet-22K: strong generalization even with limited mosquito-specific data.

Why ConvNeXtV2-Huge?

3 Model and architecture



Fully Convolutional Masked Autoencoder Framework

Type	Backbone	size	#param	FLOPS	Val acc.
Conv	Efficient V2-XL	480^2	208M	94.0G	87.3
	ConvNeXt V1-XL	384^2	350M	179.0G	87.8
Hybrid	CoAtNet-4	512^2	275M	360.9G	88.1
	MaxViT-XL	384^2	475M	293.7G	88.5
	MaxViT-XL	512^2	475M	535.2G	88.7
Trans	MViTV2-H	384^2	667M	388.5G	88.6
	MViTV2-H	512^2	667M	763.5G	88.8
Conv	ConvNeXt V2-H	384^2	659M	337.9G	88.7
	ConvNeXt V2-H	512^2	659M	600.7G	**88.9**

Comparison of state-of-the-art visual backbones on ImageNet-1K test set

Table of Contents

4 Theoretical Foundations: Evidential Deep Learning (EDL)

Why Evidential Deep Learning (EDL)?

- **Limitations of softmax:**
  - Softmax outputs a probability vector, but the highest probability does *not* always mean the prediction is reliable.
  - Tends to be **overconfident**, even for ambiguous or out-of-distribution samples.
  - No direct measure of uncertainty, making error analysis and rejection difficult.
- **EDL: Uncertainty-aware prediction**
  - Predicts a **Dirichlet distribution** over class probabilities, not just a point estimate.
  - Each class k is assigned an **evidence** value $e_k \geq 0$.
  - Dirichlet parameters:

$$\alpha_k = e_k + 1 \qquad \text{(evidence parameterization)}$$

  - High evidence = confident and sharp prediction; low evidence = flat/uniform and uncertain prediction.
  - Naturally enables **open-set rejection**: when evidence is low for all classes, the model can defer to "unknown".
- **Uncertainty:**
  - once computed the evidences e_k and the "strength of the evidence" S $= \sum_{i=1}^{K} \alpha_i$, we can also compute the uncertainty as $u = K/S$

Bayesian Risk Minimization in EDL

- **Classical training:**
  - Standard networks minimize cross-entropy on softmax outputs.
  - This ignores predictive uncertainty, leading to overfitting and over-confidence.
- **Bayesian framework:**
  - EDL minimizes the expected loss under the Dirichlet predictive distribution **(Bayes risk)**.
  - Squared-error Bayes risk for a single sample i:

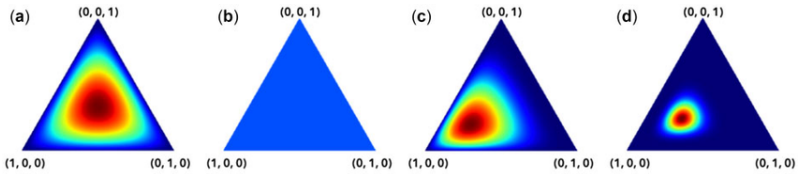
$$\mathcal{L}_i(\boldsymbol{\alpha_i}, \mathbf{y_i}) = \mathbb{E}_{\mathbf{p_i} \sim \mathrm{Dir}(\boldsymbol{\alpha_i})}[\|\mathbf{y_i} - \mathbf{p_i}\|^2]$$

$$\mathcal{L}_i(\boldsymbol{\alpha}_i, \mathbf{y}_i) = \sum_{j=1}^{K} \Big(y_{ij} - \mathbb{E}[p_{ij}]\Big)^2 + \mathrm{Var}\big(p_{ij}\big) = \sum_{j=1}^{K} \Big(y_{ij} - \hat{p}_{ij}\Big)^2 + \frac{\hat{p}_{ij}\big(1 - \hat{p}_{ij}\big)}{S_i + 1},$$

$$S_i = \sum_{j=1}^{K} \alpha_{ij}, \qquad \hat{p}_{ij} = \frac{\alpha_{ij}}{S_i}.$$

- **Result:**
  - More reliable, calibrated predictions.
  - Safer open-set behaviour: uncertain cases can be flagged for review.

Dirichlet Distribution on the Simplex



Example a Dirichlet distribution for K=3

- Dirichlet density defined over $(K-1)$-dimensional simplex: $\Delta^{K-1} = \left\{ \vec{p} \in \mathbb{R}^K \,\middle|\, p_i \geq 0,\ \sum_{i=1}^{K} p_i = 1 \right\}$

Table of Contents

5 m-EDL: modified Evidential Deep Learning

m-EDL: Extending EDL with an Unknown Class

5 m-EDL: modified Evidential Deep Learning

- **Motivation:** Standard EDL assumes all inputs belong to one of K known classes.
- **Challenge:** In real-world data (e.g. wild mosquito images), we may encounter *unknown species*.
- **Solution:** m-EDL introduces an additional class u (unknown), extending:

$$\mathbf{e}_i^+ = (e_{i1}, \ldots, e_{iK}, e_{iu}) \Rightarrow \boldsymbol{\alpha}_i^+ = \mathbf{e}_i^+ + 1$$

(it can be shown that $e_u = K$)

- Predictive mean and strength of evidence:

$$\mathbb{E}[p_{ij^+}] = \frac{\alpha_{ij^+}}{S_i^+}, \quad S_i^+ = \sum_{j^+=1}^{K+1} \alpha_{ij^+}$$

- **Key idea:** If model is uncertain, it allocates more probability mass to the unknown class.

m-EDL Loss and Integration in MOSAICO

5 m-EDL: modified Evidential Deep Learning

- **m-EDL Loss Function:**

$$\mathcal{L}_{\text{species}}^{\text{m-EDL}}(\Theta) = \sum_{j^+} \left(\left[y_{ij^+} - \mathbb{E}[p_{ij^+}] \right]^2 + \text{Var}[p_{ij^+}] \right)$$

- **Hierarchical architecture:**
  - Genus head: cross-entropy over 4 genera
  - Species head: m-EDL over $K + 1$ species (including "unknown")
- **Hierarchy loss:**

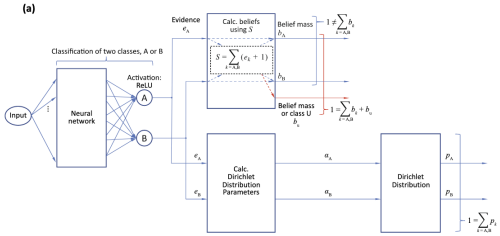
$$\mathcal{L}_{\text{hierarchy}} = \exp^D - 1, \quad D = \begin{cases} 0 & \text{if hierarchically consistent} \\ 1 & \text{otherwise} \end{cases}$$

- **Total loss:**

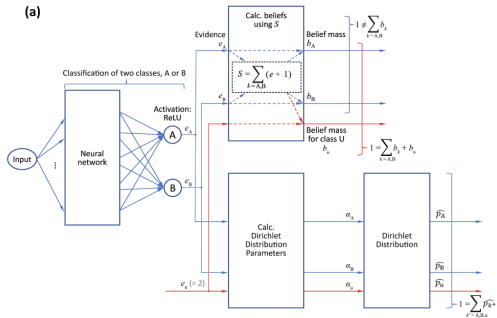
$$\mathcal{L}_{\text{total}} = \alpha \cdot (\mathcal{L}_{\text{genus}} + \mathcal{L}_{\text{species}}^{\text{m-EDL}}) + (1 - \alpha) \cdot \mathcal{L}_{\text{hierarchy}}$$

Comparison: EDL vs m-EDL Architectures

5 m-EDL: modified Evidential Deep Learning



Standard EDL architecture (closed-set)



m-EDL architecture for $K = 2$ + unknown (open-set)

Table of Contents

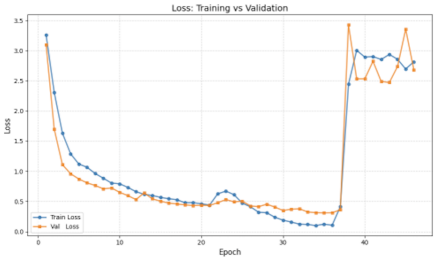
6 Training Optimization and Fine-Tuning Strategy

Fine-tuning Pipeline

6 Training Optimization and Fine-Tuning Strategy

- **Three-phase fine-tuning:**
  - **Phase 1:** Only classification heads are trained (20 total epochs, with 5 warm-up epochs).
  - **Phase 2:** Last stage of the backbone is unfrozen (15 total epochs, with 3 warm-up epochs).
  - **Phase 3:** Full model fine-tuning with discriminative LR (30 total epochs, with 3 warm-up epochs).
- **Layer-wise learning rates:**
  - Larger LR for classification heads.
  - Smaller LR for deeper backbone layers.

- **Motivation:** improves stability and prevents catastrophic forgetting.



Optimization and Training Stability

6 Training Optimization and Fine-Tuning Strategy

- **Optimizer:** `AdamW`
  - Better generalization than SGD on imbalanced data.
  - Decoupled weight decay for stability in large models.
- **Learning rate schedule:**
  - Linear warmup at the beginning of each phase
  - Cosine annealing after LR.
- **Mixed precision training (AMP):**
  - Reduced GPU memory and faster convergence.
  - Stabilized with dynamic gradient scaling.

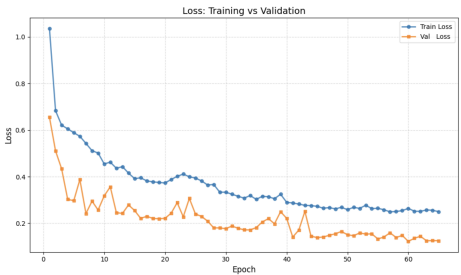
Table of Contents

7 Training and Validation Analysis

Training Loss and Validation Loss

7 Training and Validation Analysis

- The two curves (train and validation) remain well aligned throughout.
- This indicates good generalization and the absence of overfitting.
- The loss steadily decreased thanks to learning rate annealing and progressive unfreezing.

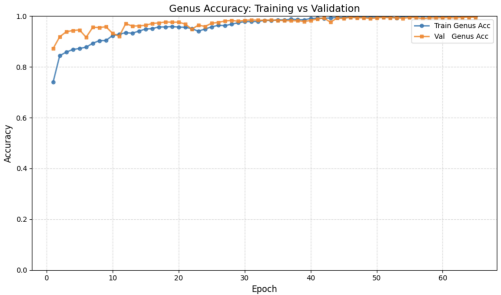


*Train (blue) and Validation (Orange) loss over the epochs

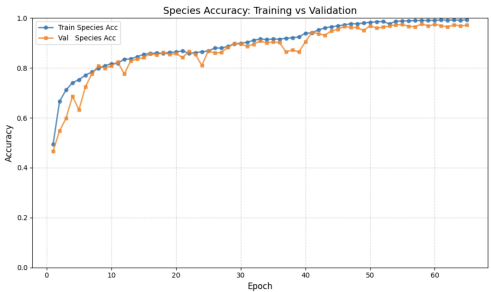
Genus and Species Accuracy

7 Training and Validation Analysis

- The model achieved very high accuracy on genus prediction, consistently above 99% on validation data.
- Species-level accuracy was more variable due to class imbalance, but reached up to 96% at peak.



*Genus accuracy vs. epochs



*Species accuracy vs. epochs

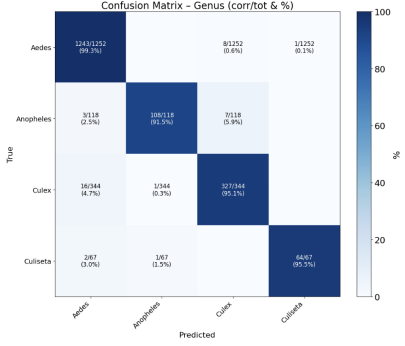
Table of Contents

8 Results and Considerations

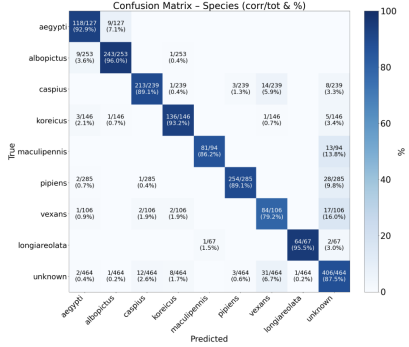
Classification Results: Genus and Species Accuracy

8 Results and Considerations

- The model reaches **98% accuracy** at the genus level.
- Species-level accuracy peaks at **91%**, with some drops for species that are rare, damaged or captured under conditions significantly different from the majority of the dataset.
- These results confirm the value of hierarchical modeling.



Confusion matrix – Genus

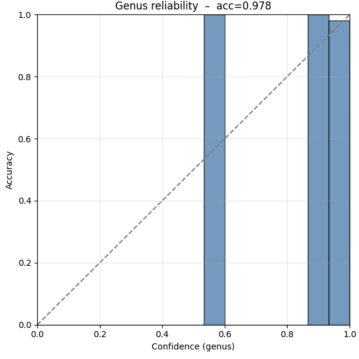


Confusion matrix – Species

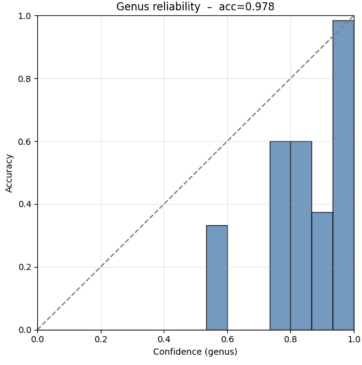
Model Calibration and Reliability for Genus classification

8 Results and Considerations

We evaluated predictive uncertainty using the **Expected Calibration Error (ECE)** and **Reliability diagram**. The genus classifier was slightly **overconfident** before calibration. After **temperature scaling**, predicted confidence aligns more closely with actual accuracy.



Before scaling: ***T = 1.00***, ***ECE = 0.013***



After scaling: ***T = 1.598***, ***ECE = 0.021***

More on Unknown Class Detection: FP, FN and Hierarchical Consistency

8 Results and Considerations

- The model correctly identifies most unknown samples, achieving an accuracy of **87.5%** on the unknown class.
- A final hierarchical check is applied via the discriminator on both genus and species predictions: during inference, **20 out of 479** unknown predictions were reassigned using this mechanism.
- However, **73 known specimens were misclassified as unknown** (false positives), often belonging to species with lower confidence or genus-species mismatches.
- Conversely, **58 unknown samples were incorrectly assigned to known classes** (false negatives), typically when the model exhibited high but misplaced confidence.

Table of Contents

9 Future work

Next Steps

9 Future work

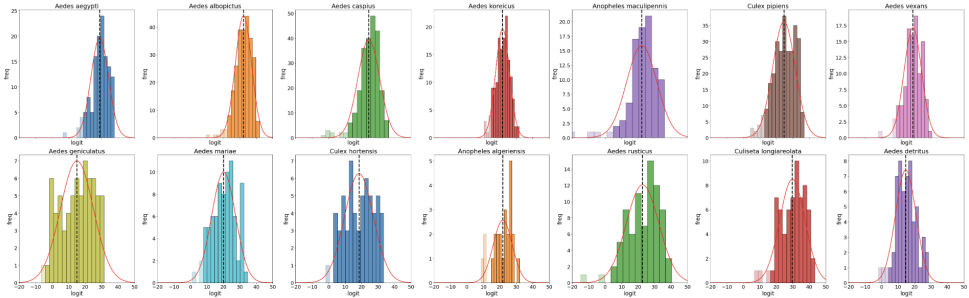
Future improvements:

- Expand the dataset with expert-reviewed relabeling.
- Transition to fully discriminative models as more labeled data becomes available.
- Investigate inverse focal loss and post-hoc calibration for both genus and species heads, aiming for sharper and more reliable predictions.
- Analyze logit distributions to define real-time rejection thresholds based on model confidence.

Analyzing Logit Distributions for Rejection Thresholds

9 Future work

We analyze the distribution of species logits to determine a rejection threshold. By modeling the distribution of correct predictions, we define a threshold at $t = \mu - 2\sigma$. Bins are color-coded: **colored bars** correspond to predictions accepted as known species; **pale bars** indicate samples rejected as unknown.



Logit distribution of known-class predictions with threshold-based rejection at $t = \mu - 2\sigma$

Computational Resources

9 Future work

This work was made possible thanks to access to the **Leonardo supercomputer** at **CINECA**, equipped with thousands of **NVIDIA A100 GPUs**. All model training and large-scale inference were conducted on this high-performance infrastructure.

HPC Leonardo (CINECA) – used for training and inference

Thank you for your attention!

Questions?

MOSAICO

Mosquito Artificial Intelligence Control