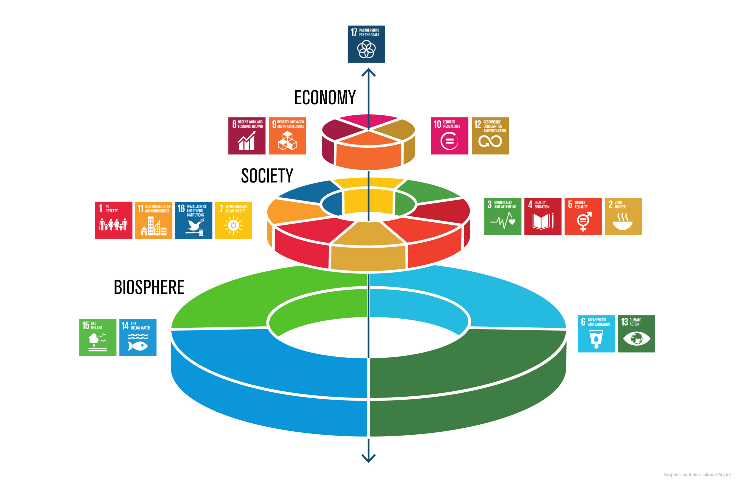
Bridging Ecological Realities: Deep Learning's Promise and Challenges

Anis Ur Rahman

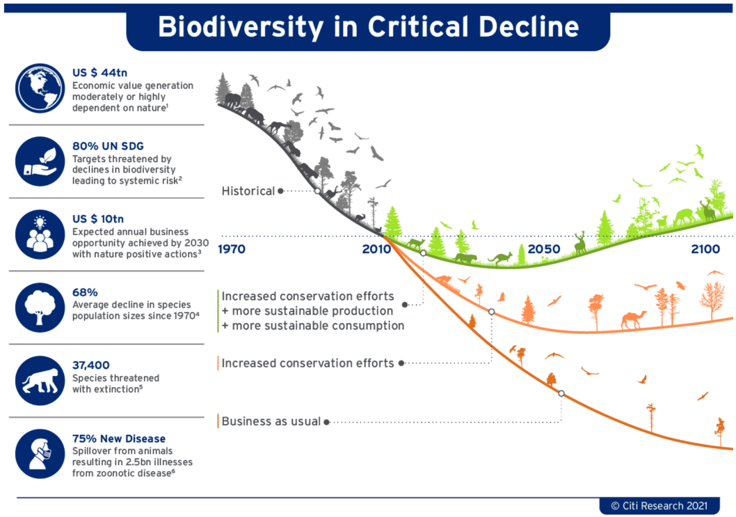
University of Jyväskylä, Finland

Interconnectedness between societal and economic SDGs ensuring a healthy biosphere

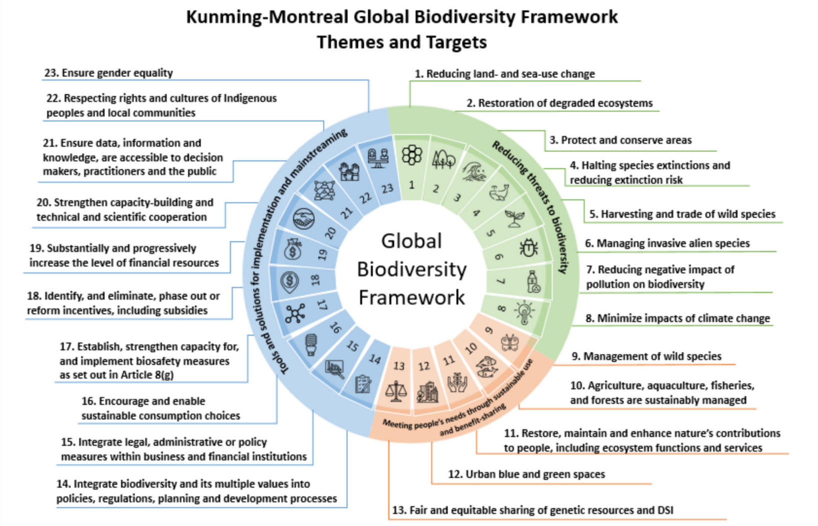


[1] Credit. Stockholm Resilience Center, Stockholm University

Biodiversity policy and science

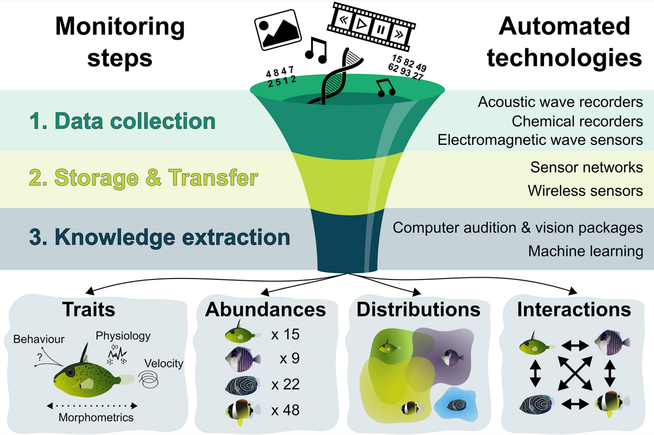


[1]Source: Leclére et al, Nature, 2020



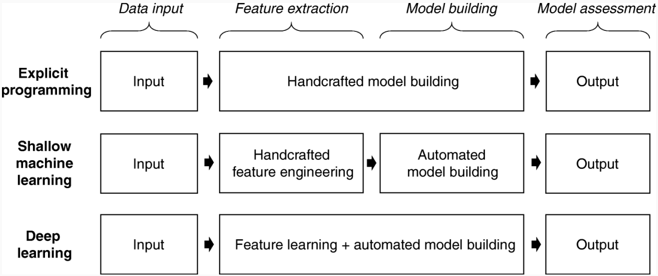
[1] Credit: Environment and Climate Change Canada

What can AI do



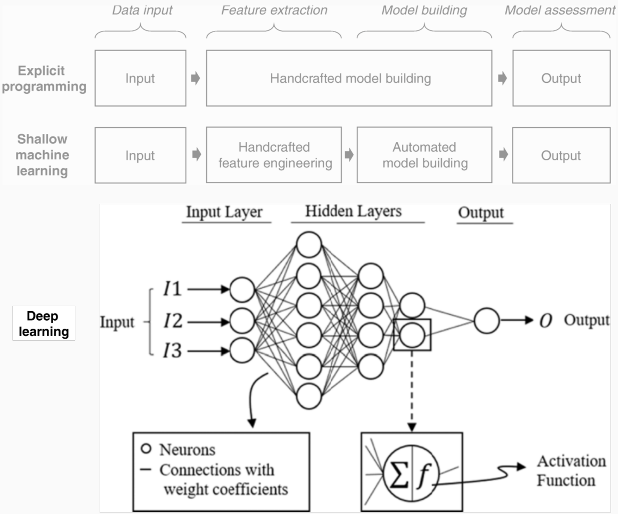
[1]Source. Besson et al. (2022) Ecol. Letters, 25: 2753–2775.

Deep learning fundamentals



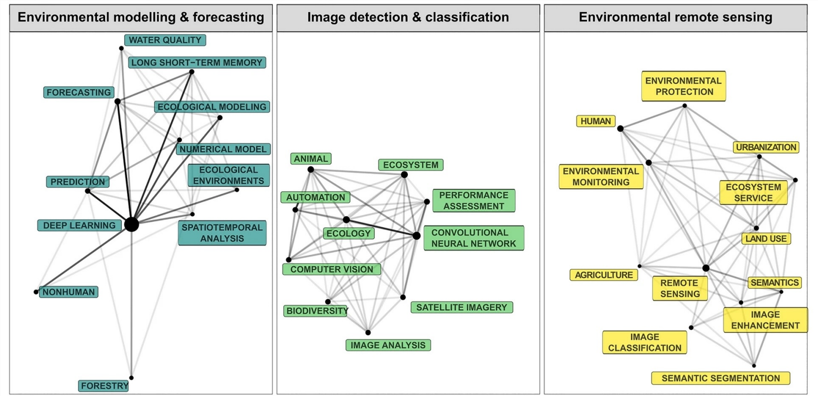
[1]Janiesch et al. (2021) Electron. Markets 31: 685–695.

Deep learning fundamentals



[1] Janiesch et al. (2021) Electron. Markets 31: 685–695.
[2] Source. Hosseiny et al. (2020) Sci. Rep. 10: 8222.

Deep learning fundamentals



[1] Source. Perry et al. (2022) Ecosystems 25: 1700–1718.

Deep learning fundamentals

Several factors lead to the success of deep learning:

$$\theta^*_{\mathcal{A}} = \arg\min_{\theta} \mathbb{E}_{x,y \in \mathcal{D}(X,Y)}[\ell(x, y : \theta)]$$

$$\begin{cases} \mathcal{A}: & \text{optimization algorithm} \\ \theta: & \text{model architecture} \\ \mathcal{D}: & \text{large-scale dataset} \\ \ell: & \text{loss function} \end{cases}$$

Deep learning fundamentals

Several factors lead to the success of deep learning:

$$\theta^*_{\mathcal{A}} = \arg\min_{\theta} \mathbb{E}_{x,y \in \mathcal{D}(X,Y)}[\ell(x, y : \theta)]$$

$$\begin{cases} \mathcal{A}: & \text{optimization algorithm} \\ \theta: & \text{model architecture} \\ \mathcal{D}: & \text{large-scale dataset} \\ \ell: & \text{loss function} \end{cases}$$

Many deep learning studies assume that the dataset follows a balanced class distribution.

Ecological realities

For instance, species are no simple objects to classify; their distribution and abundance present a few challenges for deep learning.

1. Long-tailed dataset issue
2. Scarce data issue
3. Open world problem

Long-tailed dataset issue

Imbalanced distribution where some classes/observations are rare.

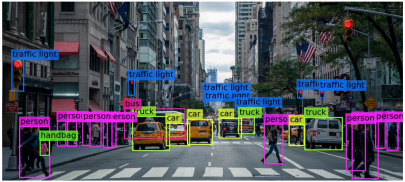
- certain species/populations occur infrequently, leading to skewed distributions.

Challenges with Long-Tailed Ecological Data

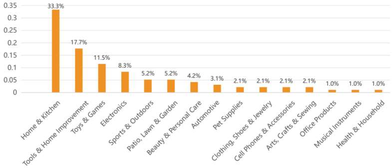
- **Model bias.** Algorithms tend to favor majority classes, neglecting rare observations.
- **Reduced accuracy.** Inadequate representation impacts predictive performance.
- **Misleading conclusions.** Overlooking rare but critical species/populations.

Long-tailed dataset issue

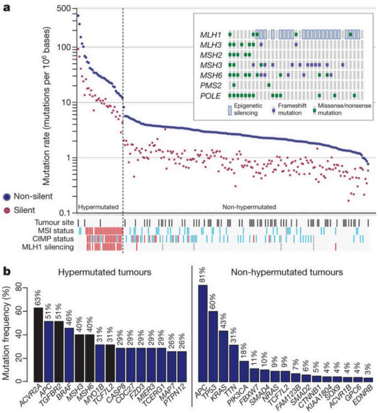
In real applications, training class distribution is often long-tailed.



Self-driving

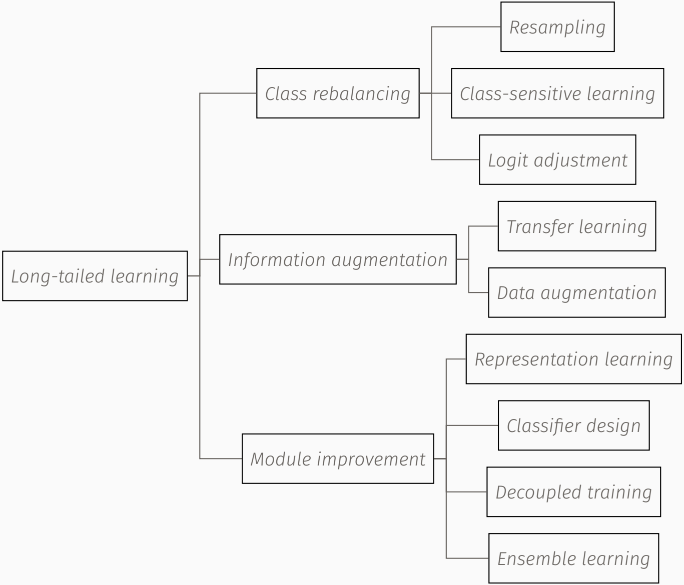


E-commerce (Amazon)



Mutated genes in non-hypermutated CRC

Long-tailed dataset issue: method taxonomy



Long-tailed dataset issue

Class rebalancing, seeking to directly rebalance uneven classes, has three main types:

1. **Re-sampling**
2. **Class-sensitive learning**
3. **Logit adjustment**

Long-tailed dataset issue

Class rebalancing, seeking to directly rebalance uneven classes, has three main types:

1. **Re-sampling** resolves class imbalance by differentially sampling the data from different classes.

$$p_j = \frac{n_j}{\sum_{i=1}^{C} n_i}$$

2. **Class-sensitive learning**
3. **Logit adjustment**

Long-tailed dataset issue

Class rebalancing, seeking to directly rebalance uneven classes, has three main types:

1. **Re-sampling**
2. **Class-sensitive learning** seeks to re-balance classes by adjusting loss values for different classes during training.

$$FL(p) = \begin{cases} -\alpha(1-p)^{\gamma} log(p) & y = 1 \\ -(1-\alpha)p^{\gamma} log(1-p) & otherwise \end{cases}$$

3. **Logit adjustment**

Long-tailed dataset issue

Class rebalancing, seeking to directly rebalance uneven classes, has three main types:

1. **Re-sampling**
2. **Class-sensitive learning**
3. **Logit adjustment** seeks to obtain a large relative margin between classes by post-hoc shifting the model logits via label frequencies.

$$LA(p) = \begin{cases} -log(\sigma(p + \tau * \pi_M) & y = 1 \\ -log(1 - \sigma(p + \tau * \pi_m) & otherwise \end{cases}$$

Long-tailed dataset issue

Information augmentation based methods seek to introduce additional information into model training by:

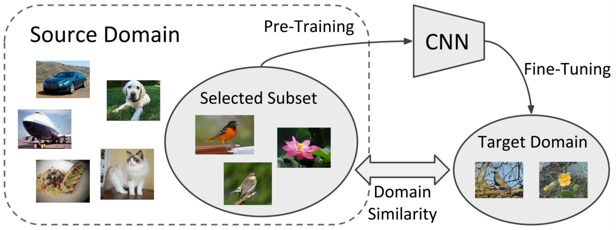
1. **Transfer learning**
2. **Data augmentation**

[1]Source. He et al. (2020) ICLR.

Long-tailed dataset issue

Information augmentation based methods seek to introduce additional information into model training by:

1. **Transfer learning** transfer knowledge from a source domain (e.g., datasets, tasks) to enhance model training on a target domain.



pre-training

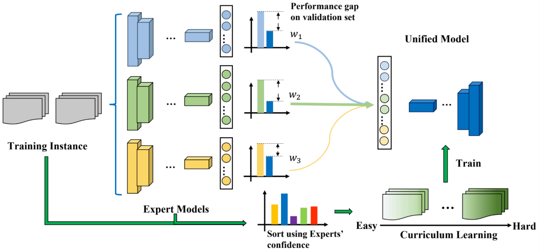
2. **Data augmentation**

[1]Source. He et al. (2020) ICLR.

Long-tailed dataset issue

Information augmentation based methods seek to introduce additional information into model training by:

1. **Transfer learning** transfer knowledge from a source domain (e.g., datasets, tasks) to enhance model training on a target domain.



Knowledge distillation

2. **Data augmentation**

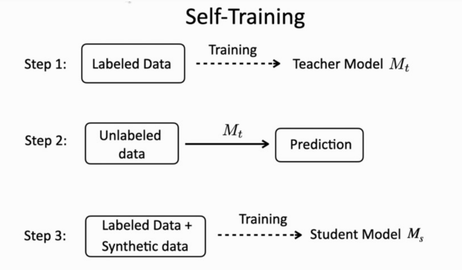
---

[1]Source. He et al. (2020) ICLR.

Long-tailed dataset issue

Information augmentation based methods seek to introduce additional information into model training by:

1. **Transfer learning** transfer knowledge from a source domain (e.g., datasets, tasks) to enhance model training on a target domain.



2. **Data augmentation**

[1]Source. He et al. (2020) ICLR.

Long-tailed dataset issue

Information augmentation based methods seek to introduce additional information into model training by:

1. **Transfer learning**
2. **Data augmentation** pack a set of augmentation techniques to enhance the size and quality of datasets for model training.

[1]Source. He et al. (2020) ICLR.

Long-tailed dataset issue

Module improvement based methods handle long-tailed problem by improving network modules.

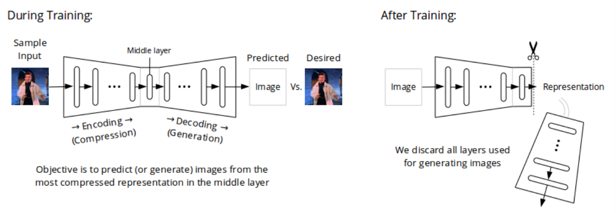
1. **Representation learning**
2. **Classifier design**
3. **Decoupled training**
4. **Ensemble learning**

[1] Source. Li et al. (2022) CVPR.

Long-tailed dataset issue

Module improvement based methods handle long-tailed problem by improving network modules.

1. **Representation learning** improves the feature extractor



2. **Classifier design**
3. **Decoupled training**
4. **Ensemble learning**

[1] Source. Li et al. (2022) CVPR.

Long-tailed dataset issue

Module improvement based methods handle long-tailed problem by improving network modules.

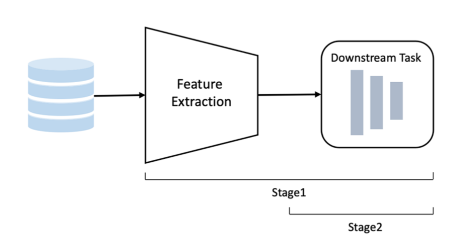
1. **Representation learning**
2. **Classifier design** This category designs various classifiers to handle long-tailed issues
3. **Decoupled training**
4. **Ensemble learning**

[1] Source. Li et al. (2022) CVPR.

Long-tailed dataset issue

Module improvement based methods handle long-tailed problem by improving network modules.

1. **Representation learning**
2. **Classifier design**
3. **Decoupled training** decouples the learning procedure into representation learning and classifier training



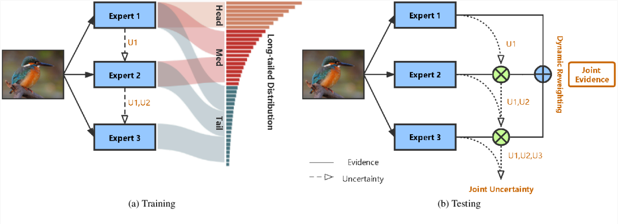
4. **Ensemble learning**

[1]Source. Li et al. (2022) CVPR.

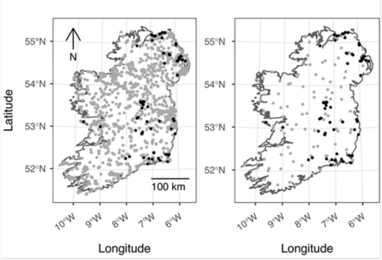
Long-tailed dataset issue

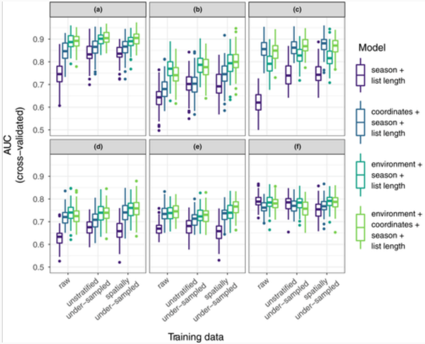
Module improvement based methods handle long-tailed problem by improving network modules.

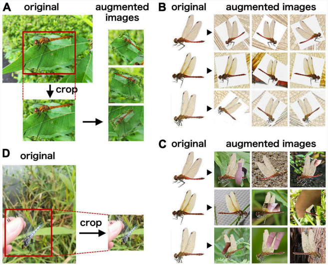
1. **Representation learning**
2. **Classifier design**
3. **Decoupled training**
4. **Ensemble learning** based methods strategically learn multiple network experts to solve long-tailed problems



[1] Source. Li et al. (2022) CVPR.







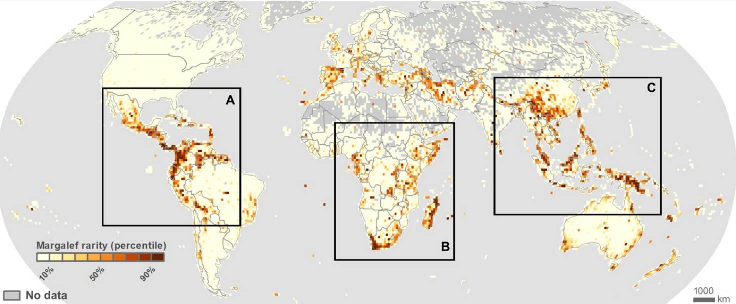
[1]Source (top panel). Gaul et al. (2022) Divers. Distrib. 28 (10):2171-2186

[2]Source (bottom panel). Sun et al. (2021) Front. Ecol. Evol. 9: 1-10

Scarce data issue

Rarity is an intrinsic characteristic of biodiversity, with most communities composed of a large number of rare species.

- observed in species-rich assemblages like coral reef fishes, where most species are demographically rare.
- for deep learning, species rarity implies a lack of training data for a large part of species.



[1]Source. Enquist et al. (2019) Sci. Adv. 5(11): eaaz0414

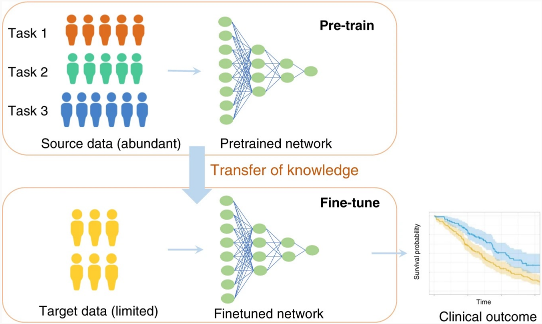
Scarce data issue

1. Meta-training.
2. Metric learning.

[1] Source. Gevaert (2021) British J. Cancer 125: 309–310

Scarce data issue

1. **Meta-training.** Training a model on a variety of tasks or datasets to develop a generalized learning procedure.
   - Enable the model to learn how to learn, acquiring meta-knowledge (generalizable patterns or parameters) for fast adaptation to new tasks.

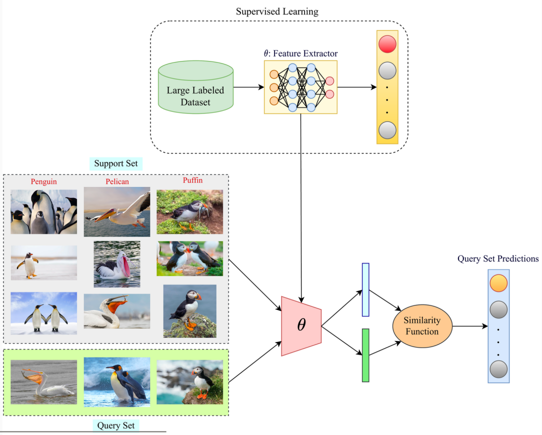


2. **Metric learning.**

[1]Source. Gevaert (2021) British J. Cancer 125: 309–310

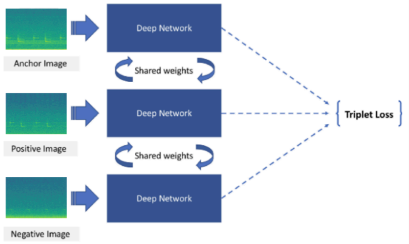
Scarce data issue

1. **Meta-training.**
2. **Metric learning.** to optimize feature representations and improve model performance despite scarce data.
   - learning a distance metric that measures similarity or dissimilarity between instances in a dataset.



[1] Source. Gevaert (2021) British J. Cancer 125: 309–310

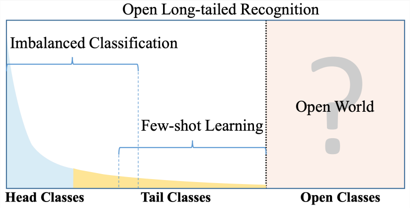
Scarce data issue



	CNN			SNN		
Species	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy	Top-1 Accuracy	Top-3 Accuracy	Top-5 Accuracy
5 – 100 annotated calls (Includes 35 species)	83.25%	90.25%	92.21%	**85.77 %**	**93.19%**	**95.40%**
5 – 20 annotated calls (Includes 13 species)	53.69%	67.89%	75.79%	**73.16%**	**90.00%**	**93.69%**
5 – 10 annotated calls (Includes 7 species)	35.29%	64.71%	82.35%	**60.00%**	**80.00%**	**90.59%**

[1]Source. Zhong et al. (2023) bioRxiv.

Open world problem



Managing the open-world problem in classification involves handling situations where the classifier needs to distinguish between known and unknown classes during inference.

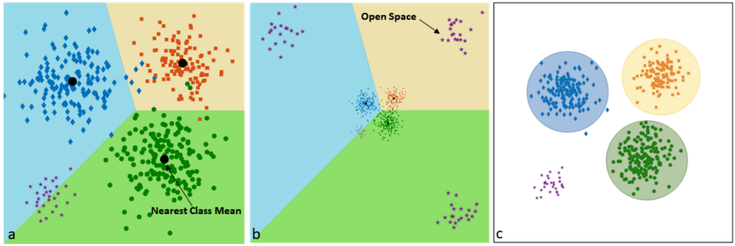
[1]Source. Liu et al. (2019) CVPR.

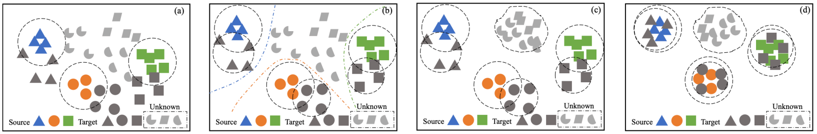
Open world problem

Open set recognition techniques

1. **Thresholding.** Set confidence thresholds to reject samples falling below a certain confidence level, labeling them as unknown or out-of-distribution.
2. **Distance metrics.** Utilize distance-based methods to measure similarity between test samples and known classes. Samples distant from known classes are treated as unknown.
3. **One-class classifiers.** Train models specifically to recognize known classes, ignoring unknown instances during training.
4. **Augmented training data.** Augment training data by generating samples resembling unknown classes or diverse variations within known classes. Techniques like generative models (GANs) or oversampling rare classes can be used.
5. **Representation learning.** Utilize methods that create embedding spaces where known classes cluster together, enabling identification of unknown samples lying outside these clusters.
6. **Active learning.** Human-in-the-loop strategy involve human annotators in the loop to label and incorporate new classes or instances into the classification space.

Example

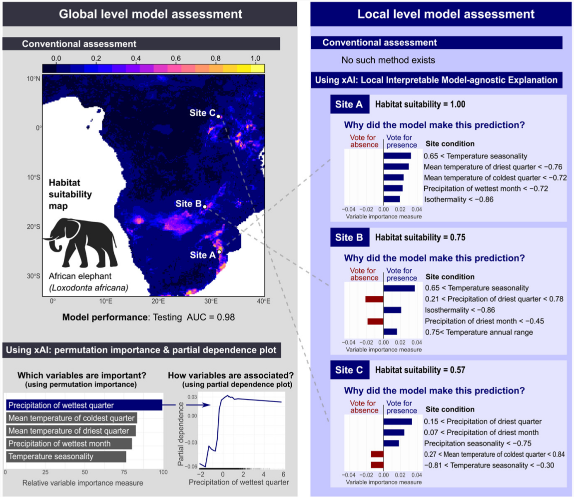




[1]Source. Liu et al. CVPR19.

Model interpretability, explainability and causality

Explainable artificial intelligence (xAI) is the process of understanding how and why a machine learning model makes its predictions.



[1] Source. Ryo et al. (2020) Ecography, 44: 199-205.

Concluding remarks

1. **Complexity of ecological realities**
    - Deep learning presents immense potential for ecological insights but confronts challenges in accommodating the complexity, rarity, and dynamic nature of ecological datasets.
2. **Need for robust adaptation**
    - Collaboration and innovation are pivotal in overcoming these challenges, paving the way for more accurate and robust deep learning applications in ecological studies.

Thank you for your attention!