2025 7th Experiment@ International Conference (expat'25)
September 3-5, 2025, University of the Azores, Faial, Azores, Portugal

# Active Replay for Continual Learning in Plant Species Classification

Dinis Costa[1], Joana Costa[1], Bernardete Ribeiro[1], José Paulo Sousa[2], Rui Pedro Paiva[1], José Rafael Silva[3], Rui Lourenço[3], Catarina Silva[1]
[1]CISUC, Dep. of Informatics Engineering, University of Coimbra, Portugal
[2]Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Portugal
[3]MED, Escola Ciências e Tecnologia, Universidade de Évora, Portugal
{ddcosta,joanamc,bribeiro,ruipedro,catarina}@dei.uc.pt,jps@zoo.uc.pt,{jmsilva,lourenco}@uevora.pt

***Abstract*—Continual Learning (CL) is essential in dynamic environments, such as biodiversity monitoring, due to the continuous emergence of new species or changes in their distribution over time. A major challenge in CL is catastrophic forgetting, where a model loses performance on previously learned classes as it learns new ones. One common approach to mitigate this issue is replay, in which samples from previous tasks are reintroduced during training. We investigate how Active Learning (AL) can improve replay strategies for the classification of plant species in a class-incremental learning setting. Specifically, we compare two AL-based sampling methods: selecting samples with the lowest confidence and those with the highest confidence, with a baseline random sampling strategy. The results show that while all methods achieve similar final accuracy, AL-based strategies significantly reduce catastrophic forgetting. High-confidence sampling reduces forgetting by 11.9 percentage points, while low-confidence sampling achieves a reduction of 9.2 percentage points compared to random sampling. We apply CL in a real-world scenario rather than a benchmark dataset, demonstrating its practical relevance for dynamic and evolving environments. Hence, our approach contributes to biodiversity monitoring, with the potential to support the development of a biodiversity index for tracking species diversity over time.**

***Index Terms*—Continual Learning, Computer Vision, Active Learning, Agriculture 4.0**

## I. Introduction

Continual Learning (CL) plays a crucial role in constantly changing settings, such as ecosystems where new species emerge over time. By enabling adaptive species classification, CL can facilitate biodiversity tracking, helping to monitor species distribution and population changes over time. This, in turn, can support development of biodiversity indexes, providing a quantitative measure of ecosystem dynamics. Therefore, effective monitoring of biodiversity is essential to detect new species and assess their relationship with agricultural production .

Machine Learning (ML) is increasingly used in biodiversity monitoring to classify species through image or sound recognition [1], [2]. Among ML techniques, Convolutional Neural Networks (CNNs) have proven especially effective for image classification tasks. However, when adapting pre-trained CNN models to continuously learn new tasks or recognise new species, these models often suffer catastrophic forgetting [3]. This phenomenon refers to the deterioration or loss of previously learned knowledge when new information is introduced. Addressing this issue in CL would make intelligent systems capable of continuously adapting without sacrificing past performance.

Various strategies exist to improve the efficiency and effectiveness of ML models. One prominent method is Active Learning (AL), which strategically selects the most relevant data samples to train ML models. By doing so, AL reduces the amount of training data required, thus reducing computational energy consumption and lowering annotation efforts.

To mitigate catastrophic forgetting in CL, several techniques have been proposed. A particularly effective strategy is replay [4], which involves retaining and periodically revisiting previously learned data during training for new tasks. However, not all data samples contribute equally to maintaining past knowledge and improving future learning.

A promising approach to improve replay is AL. By prioritising the most relevant data, AL selects the best samples for replay, maximising knowledge retention. Recent works have explored this intersection, including in language tasks using benchmark datasets [5], [6], and in vision tasks using CNNs with selective replay strategies [7]. However, these methods rely heavily on benchmarks and focus mainly on low confidence (uncertain) samples. In this work, we propose using AL to select both low and high confidence samples for replay when incrementally training CNNs to classify plant species from images, enabling the model to adapt to new species while retaining previously learned knowledge.

This paper is structured as follows: Section II introduces the theoretical background on the key topics addressed, providing relevant examples. Section III introduces the methodology, evaluation metrics, and experimental design. Section V presents and discusses the experimental results. Finally, Section VI summarises the main conclusions and provides directions for future research.

## II. Background

### A. Continual Learning

Continual learning focuses on developing models that adapt to new information without forgetting previously acquired knowledge [8]. These adaptations may involve changes in tasks, domains, or classes, requiring the model to learn in-

979-8-3315-7662-2/25/$31.00 ©2025 IEEE

crementally. To address these challenges, various approaches have been proposed, which can be categorized as follows:

- **Regularization strategies** restrict learning to prevent excessive changes in previously acquired knowledge. They achieve this by limiting modifications to key model parameters. Elastic Weight Consolidation (EWC) [9], a notable approach, estimates parameter importance using the Fisher Information Matrix. During new task learning, the model is penalized for altering crucial parameters, preserving past knowledge and reducing catastrophic forgetting.
- **Architectural strategies** modify the model's structure to accommodate new tasks, either by dynamically expanding the model or by assigning specific resources to different tasks. An example is Progressive Neural Networks (PNNs) [10], which add new branches to the network for each task while preserving connections to earlier layers.
- **Replay strategies** mitigate catastrophic forgetting by maintaining access to past data and periodically retraining. These strategies reinforce previously learned knowledge by interleaving past experiences with new ones during training. Replay can be implemented in different ways, such as storing raw data samples or generating synthetic memories through generative models. A widely used replay strategy is Experience Replay [11], which maintains a memory buffer of past samples and reintroduces them during training. This approach, inspired by reinforcement learning, allows the model to retain information from previous tasks and adapt to new ones.

Continual learning can be broadly categorized into [12]:

- **Task Incremental Learning**. The model encounters distinct tasks sequentially, with task identities provided during inference. This allows the model to use task-specific parameters or strategies to mitigate forgetting.
- **Domain Incremental Learning**. The task remains the same, but the data distribution changes over time. The model must adapt to these shifts without explicit knowledge of the domain labels while maintaining performance on previously seen data.
- **Class Incremental Learning (CIL)**. The model learns new classes over time without access to previous data and must classify all classes, old and new, in a unified setting. This is more challenging than Task Incremental Learning, as task labels are unavailable. While regularization methods have been explored, replay-based strategies remain more effective for mitigating catastrophic forgetting [4].

This work addresses a CIL scenario, focussing on replay strategies combined with AL for selective memory retention, aiming to improve efficiency by storing and revisiting only the most informative samples.

### *B. Active Learning*

Active learning is a ML approach that enhances model training by selectively acquiring the most informative data samples. Instead of passively learning from a fixed dataset, AL leverages a partially trained model to evaluate unlabeled data and identify the most relevant samples, as illustrated in Figure 1. These selected samples are then labeled by a human expert or a more robust model and incorporated into the training process. This iterative cycle continues until the model reaches a desired performance level.

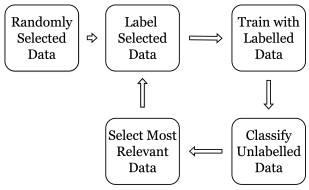



Fig. 1: Representation of Active Learning steps: an initial batch of randomly selected data is labeled, enabling the training of a classifier. The classifier then assigns labels to the remaining unlabeled data, after which the most relevant samples are selected, labeled, and the process is repeated.

Active learning is typically categorized into two main approaches: stream-based and pool-based [13]. In the stream-based approach, data arrives sequentially, and the model must decide in real-time whether to request a label for each instance. In contrast, the pool-based approach starts with a large set of unlabeled samples, $D_u$, and a smaller set of labeled ones, $D_l$, allowing the model to query the most informative samples for labeling.

$$D_u = x_1, x_2, \ldots, x_n, \tag{1}$$

$$D_l = (x'_1, y'_1), (x'_2, y'_2), \ldots, (x'_m, y'_m), \tag{2}$$

where each $y'_i$ represents the label for the corresponding $x'_i$.

Initially, the model trains a classifier $f_0$ using the labeled dataset $D_l$. At each iteration $t = 1, 2, \ldots, T$, the learner employs the previously trained classifier $f_{t-1}$ to select an instance $x_j$ from the unlabeled set $D_u$, queries its label $y_j$, and adds the labeled pair $(x_j, y_j)$ to $D_l$. The updated dataset is then used to train a new classifier $f_t$. The goal is to maximize accuracy while minimizing the number of labeled samples, ensuring that $f_t$ generalizes to unseen data.

In this work, instead of immediately training on classified unlabeled data, we use AL to select the most informative samples for future training iterations. This strategy ensures that revisited data contributes to mitigate catastrophic forgetting, improving long-term model performance.

## III. Methodology

While many CL studies focus on benchmark datasets, this study evaluates replay strategies in real-world scenarios. Specifically, our goal is to develop a model for classification of plant species in a class-incremental setting, where new species may emerge over time or existing species undergo variations. The model must continuously adapt to these changes while preserving previously acquired knowledge, ensuring robustness against catastrophic forgetting.

### A. Proposed approach

Our objective is to minimize catastrophic forgetting [14] while the model learns new classes. To achieve this, we train the model sequentially, where at each iteration $i$, a classifier $f_i$ is trained using data from $s_i$, where $i \in [1, \infty[$. Each subset $s_i$ consists of samples from a specific set of classes $C_i$, such that:

$$s_i = \{x_j \mid x_j \in C_i\}, \tag{3}$$

where $C_i$ represents the set of classes assigned to stream $s_i$. To ensure that classes do not repeat across different streams, we enforce the constraint:

$$s_l \cap C_i = \emptyset, \quad \forall l \neq i. \tag{4}$$

This ensures that each class appears in exactly one subset $s_i$, preventing class overlap across training streams and enforcing a strict-class incremental leaning setup.

*1) Forgetting Measure:* Forgetting refers to the decline in accuracy on previously learned data after training on new data. More formally, it can be quantified by measuring the difference in accuracy for the same data stream across successive training iterations.

To measure Forgetting for $s_i$, we evaluate the classifier's performance on $s_i$ across different iterations. Specifically, for any iteration $j > i$, the forgetting measure for $s_i$ at iteration $j$, denoted as $FM_{j,s_i}$, is defined as:

$$FM_{j,s_i} = \max_{k \leq j} A(f_k, s_i) - A(f_j, s_i), \tag{5}$$

where:

- $A(f_k, s_i)$ represents the accuracy of the classifier $f_k$ on $s_i$ at any past iteration $k \leq j$.
- $\max_{k \leq j} A(f_k, s_i)$ is the highest accuracy achieved in $s_i$ before iteration $j$.
- $A(f_j, s_i)$ is classifier's $f_j$ accuracy in $s_i$ at iteration $j$.

A higher $FM_{j,s_i}$ value indicates greater forgetting, implying a significant decline in the model's performance on $s_i$ after learning new tasks. Our goal is to minimize $FM$ to enhance knowledge retention across training iterations.

*2) Actively Selecting Training Data:* Figure 2 illustrates the process of an iteration $i$ in our proposed approach. To minimize forgetting using replay strategies, in each iteration, we incorporate samples from previous data streams to reinforce past knowledge. Specifically, at each iteration $i$, the classifier $f_i$ is trained not only on $s_i$ but also on selected samples from previous streams $s_j$, where $j < i$.

We investigate the impact of different sampling strategies to select replay samples from $s_j$. In particular, we test two main scenarios: Random Sampling and Active Learning Sampling, where the objective is to identify relevant samples to be revisited in later iterations.

There are several ways to determine the relevant samples. Typically, AL selects samples with the lowest confidence predictions as the most relevant [15], [16]. Since the CNN classifier layer outputs a confidence score for each class, a sample can be considered more relevant than another if the confidence score for its predicted class is lower.

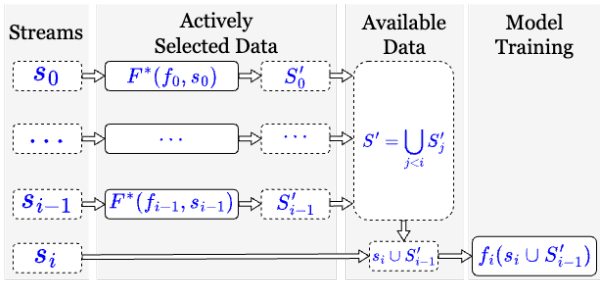



Fig. 2: Illustration of the training process for classifier $f_i$. In each iteration $i$, the model is trained using data from the current dataset $s_i$ along with selected samples from all previous datasets $s_j$ for $j < i$. This replay mechanism helps mitigate catastrophic forgetting by reinforcing past knowledge while learning new tasks.

The most common approach in AL is to select the least confident samples, denoted $F^-(f_i, s_i)$. However, choosing samples with the lowest confidence of $s_i$ can pose challenges when learning $s_{i+1}$, as these samples could be difficult for the model to generalise. To explore this further, we also test high-confidence selection, $F^+(f_i, s_i)$, under the hypothesis that it may affect $FM$. As a baseline, we include random sampling $F^{\text{rand}}(s_i)$, which selects samples uniformly at random.

To formally define these selection strategies:

- **Random sampling**: Selecting a proportion $\rho$ of the samples uniformly at random from $s_i$:

$$F^{\text{rand}}(s_i) = \{x \sim s_i \mid x \text{ random}, |F^{\text{rand}}(s_i)| = \rho \cdot |s_i|\}$$

- **Relevant sampling**:
  - **Low-confidence selection**: Selecting a proportion $\rho$ of samples from $s_i$ with the lowest confidence scores:

$$F^-(s_i) = \{x \in s_i \mid \tau(f_i, x) \leq \tau_\rho\}$$

where $\tau(f_i, x)$ represents the confidence score assigned by the classifier $f_i$ to sample $x$. The threshold $\tau_\rho$ is set such that the total number of selected samples satisfies:

$$|F^-(s_i)| = \rho \cdot |s_i|$$

  - **High-confidence selection**: Selecting a proportion $\rho$ of samples from $s_i$ with the highest confidence scores:

$$F^+(s_i) = \{x \in s_i \mid \tau(f_i, x) \geq \tau_\rho\}$$

where $\tau_\rho$ is set dynamically so that:

$$|F^+(s_i)| = \rho \cdot |s_i|$$

This evaluation allows us to compare how different sampling strategies affect both knowledge retention (minimizing forgetting) and generalization to new tasks in a continual learning setting.

## IV. Experimental Setup

### A. Problem Definition

We aim to develop a model to classify plant species, or at least genus taxonomic level, that may appear in a specific region of Portugal, based on a list of the most common species in the region. However, this list may be updated over time due to the dynamic nature of ecosystems. Given these evolving conditions, CL is a well-suited approach, as it enables the model to adapt to new species as they emerge.

Beyond classification, this approach also contributes to biodiversity tracking, allowing for the continuous monitoring of species diversity. By capturing species variations over time, CL-based classification could aid in the development of a biodiversity index, providing a quantitative measure of ecosystem changes and supporting conservation and agricultural management efforts.

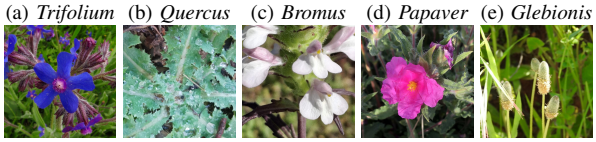
### B. Dataset

From a list of 105 emerging plant species in a region of Portugal, we collected image data from the *iNaturalist* app [17]. As a proof of concept, we first experiment with a small subset of the dataset before scaling to the full dataset. Based on this, we selected the 10 classes with the highest number of images. Furthermore, to ensure a balanced and computationally feasible dataset, the number of images per class was reduced to 50%. The selected classes and their corresponding image counts are as follows:

- *Trifolium* (*716* images)
- *Quercus* (*447* images)
- *Bromus* (*427* images)
- *Papaver* (*375* images)
- *Glebionis* (*311* images)
- *Anchusa* (*300* images)
- *Sonchus* (*298* images)
- *Bellardia* (*298* images)
- *Cistus* (*298* images)
- *Plantago* (*298* images)

The dataset of 3768 images was randomly divided into three subsets: 60% for training, 20% for validation, and 20% for testing. Figure 3 presents sample images from the dataset.

(a) *Trifolium* (b) *Quercus* (c) *Bromus* (d) *Papaver* (e) *Glebionis*



(f) *Anchusa* (g) *Sonchus* (h) *Bellardia* (i) *Cistus* (j) *Plantago*

Fig. 3: Example images for each class in the dataset, that show the visual diversity of plant species used in the study.

### C. Preliminary Experiments

For our study, we use CNNs for image classification. To determine the most effective architecture for our dataset, we conducted preliminary experiments evaluating multiple models. The results showed that RegNet [18], specifically REGNETY-800MF pre-trained on ImageNet, achieved the highest accuracy and was therefore selected as the architecture for our study.

## V. Results and Discussion

Since our main dataset consists of 10 classes, we predefined 10 streams:

$$s = \{s_0, s_1, \ldots, s_9\}$$

Each stream contains exactly one class, and the 10 classes were randomly assigned in these streams.

A model $f_0$ was first trained from scratch on $s_0$ and then incrementally trained on each $s_i$ using the previous model $f_{i-1}$ for every $i > 0$. Each stream was divided into three sets: training, validation, and testing.

Each training iteration lasted for 6 epochs, using a batch size of 16, an image size of 898, and a replay rate of 35% per stream.

After training each $f_i$ on $s_i$, the model was evaluated on all test sets from $s_j$ for $j \leq i$. This allowed us to measure accuracy across previously learned streams and compute the forgetting measure ($FM$) at each iteration for each stream.

To account for randomness in the class distribution between streams, we repeated the experiment 10 times using different random seeds.

The focus of this study was the replay strategy. Therefore, a setup without replay was not conducted. Furthermore, preliminary experiments showed that without any CL strategy, the model failed to retain the knowledge from previous classes, making this setup uninformative for further analysis.

### A. Accuracy Results

TABLE I: Average accuracy results over 10 runs: Performance of each classifier $f_i$ on the test set of each stream $s_i$ using the replay strategy $F^{\text{rand}}$ (**Random sampling**). The highest accuracy per stream is highlighted in **bold**.

| $f_i$ | **Random sampling** ($F^{\text{rand}}$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ |
| $f_0$ | **100.0** | | | | | | | | | |
| $f_1$ | 42.9 | **63.4** | | | | | | | | |
| $f_2$ | 19.5 | 35.3 | **65.8** | | | | | | | |
| $f_3$ | 22.0 | 13.9 | 29.3 | **66.7** | | | | | | |
| $f_4$ | 12.7 | 11.3 | 21.2 | 12.5 | **72.6** | | | | | |
| $f_5$ | 13.9 | 16.2 | 20.4 | 4.4 | 10.0 | **75.7** | | | | |
| $f_6$ | 16.4 | 5.3 | 10.7 | 14.3 | 10.7 | 17.6 | **73.4** | | | |
| $f_7$ | 6.3 | 8.2 | 8.3 | 8.5 | 13.5 | 11.9 | 2.0 | **83.2** | | |
| $f_8$ | 11.2 | 1.5 | 9.7 | 11.0 | 10.0 | 14.8 | 8.1 | 25.8 | **60.6** | |
| $f_9$ | 2.8 | 1.8 | 4.3 | 7.8 | 10.1 | 6.6 | 9.0 | 31.1 | 18.0 | **50.6** |

Table I presents the baseline scenario, where replay samples are selected randomly, while Table II aggregates the results for both AL strategies, where samples are chosen based on

TABLE II: Average accuracy results over 10 runs: Performance of each classifier $f_i$ on the test set of each stream $s_i$ using the replay strategies $F^+$ (**Active Learning with High-Confidence Sampling**) and $F^-$ (**Active Learning with Low-Confidence Sampling**). The highest accuracy per stream is highlighted in **bold**, and background colors indicate relative performance compared to Random Sampling (GREEN means AL performed better, RED means worse).

| $f_i$ | High-Confidence Sampling ($F^+$) | | | | | | | | | | Low-Confidence Sampling ($F^-$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ |
| $f_0$ | **100.0** | | | | | | | | | | **100.0** | | | | | | | | | |
| $f_1$ | 52.6 | **58.5** | | | | | | | | | 58.1 | **50.2** | | | | | | | | |
| $f_2$ | 33.8 | 34.8 | **65.3** | | | | | | | | 31.9 | 19.0 | **72.5** | | | | | | | |
| $f_3$ | 29.5 | 30.6 | 40.1 | **33.8** | | | | | | | 21.9 | 14.2 | 52.3 | **35.7** | | | | | | |
| $f_4$ | 23.1 | 12.0 | 18.6 | 21.8 | **66.1** | | | | | | 9.0 | 11.4 | 21.3 | 9.2 | **70.0** | | | | | |
| $f_5$ | 20.2 | 12.3 | 30.2 | 14.3 | 23.3 | **46.3** | | | | | 11.5 | 9.0 | 20.4 | 4.8 | 25.7 | **48.7** | | | | |
| $f_6$ | 15.8 | 8.6 | 26.6 | 21.5 | 15.3 | 12.8 | **61.4** | | | | 11.5 | 9.7 | 20.4 | 12.6 | 22.2 | 6.3 | **61.7** | | | |
| $f_7$ | 10.7 | 5.6 | 13.6 | 21.4 | 12.7 | 20.8 | 11.3 | **67.1** | | | 19.9 | 3.2 | 10.1 | 9.7 | 21.0 | 1.6 | 14.2 | **66.8** | | |
| $f_8$ | 14.4 | 4.4 | 10.1 | 8.0 | 19.0 | 9.7 | 9.3 | 13.4 | **60.5** | | 14.8 | 1.5 | 7.5 | 10.1 | 18.8 | 5.9 | 11.9 | 16.9 | **58.7** | |
| $f_9$ | 10.8 | 1.7 | 4.2 | 8.0 | 9.7 | 17.0 | 24.2 | 13.1 | 30.0 | **56.4** | 10.2 | 2.0 | 3.5 | 2.9 | 12.1 | 8.8 | 18.6 | 33.6 | 9.1 | **51.6** |

confidence levels. In Table II, background colors indicate the comparison with random sampling: green represents cases where AL resulted in higher accuracy, while red indicates lower accuracy compared to the baseline.

Each table shows the accuracy results for each stream $s_i$ after being trained in each classifier $f_j$ for $j >= i$. As expected, the highest accuracy for each stream $s_i$, highlighted in bold, was achieved immediately after $f_i$ was trained in all scenarios. After each subsequent $f_j$ for $j > i$, performance in the test set of $s_i$ declines, although some knowledge is retained due to the replay mechanism.

Although the highest accuracy achieved in each $s_i$ is generally higher in the random sampling scenario than in the high-confidence scenario, and in almost every $s_i$ of the low-confidence scenario, it is noticeable that there are more green values in both the high-confidence and low-confidence scenarios. This indicates that in later iterations, the accuracy is often higher than in random sampling. In other words, random sampling initially achieves better accuracy, but quickly loses to the AL sampling strategies as training progresses.

Each scenario consists of 55 measured accuracies: 10 for $s_0$, 9 for $s_1$, etc., allowing 55 pairwise comparisons between different replay strategies.

To determine whether there were significant differences between the methods, we performed a paired $t$-test comparing the mean accuracy differences in the 55 comparisons. The results indicate that although high confidence selection outperformed random sampling by an average of 1.8 percentage points and low confidence selection underperformed by 0.5 percentage points, neither difference was statistically significant. At a 95% confidence level with 54 degrees of freedom ($df = 54$), both comparisons failed to reject the null hypothesis.

These results suggest that, while high-confidence selection performed slightly better on average than random sampling, there is no strong statistical evidence to conclude that one replay strategy is significantly superior to another.

### B. Catastrophic Forgetting Results

Table III presents the baseline scenario, in which forgetting ($FM$) is measured when replay samples are selected randomly. Table IV aggregates the results for both AL strategies, where replay samples are selected based on confidence levels.

TABLE III: Forgetting values ($FM$) for **Random Sampling** across 10 runs.

| $f_i$ | Random sampling ($F^{\text{rand}}$) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ |
| $f_0$ | 0.0 | | | | | | | | | |
| $f_1$ | 57.1 | 0.0 | | | | | | | | |
| $f_2$ | 80.5 | 28.0 | 0.0 | | | | | | | |
| $f_3$ | 78.0 | 49.5 | 36.6 | 0.0 | | | | | | |
| $f_4$ | 87.3 | 52.1 | 44.6 | 54.2 | 0.0 | | | | | |
| $f_5$ | 86.1 | 47.2 | 45.4 | 62.3 | 62.6 | 0.0 | | | | |
| $f_6$ | 83.6 | 58.1 | 55.1 | 52.4 | 61.9 | 58.1 | 0.0 | | | |
| $f_7$ | 93.7 | 55.2 | 57.5 | 58.2 | 59.1 | 63.7 | 71.4 | 0.0 | | |
| $f_8$ | 88.8 | 61.9 | 56.2 | 55.7 | 62.6 | 60.9 | 65.3 | 57.4 | 0.0 | |
| $f_9$ | 97.2 | 61.5 | 61.6 | 58.9 | 62.5 | 69.0 | 64.4 | 52.0 | 42.6 | 0.0 |

In Table IV, background colours indicate the comparison with the baseline: Green represents cases where AL resulted in less forgetting, while red indicates greater forgetting compared to random sampling.

Across all scenarios, later iterations consistently introduce forgetting, meaning that the model gradually loses performance in classifying older classes as it learns new ones.

Given the amount of green values in both AL scenarios, it is reasonable to assume that AL sampling helps mitigate forgetting compared to random selection of replay samples.

Specifically, in the high confidence scenario, it outperformed random sampling in 42 cases, underperformed in 3 cases, and matched forgetting in 10 cases. On average, forgetting was reduced by 11.9 percentage points. A paired $t$-test confirmed this intuition by rejecting the null hypothesis at a 95% confidence level with $df = 54$.

Regarding the low-confidence scenario, it outperformed random sampling in 34 cases, underperformed in 11 cases, and matched forgetting in 10 cases. The average reduce in forgetting was 9.2 percentage points across the 55 measurements. The null hypothesis was also rejected under the same conditions.

## VI. Conclusion and Future Work

Continual learning is particularly relevant in dynamic environments, such as species classification in ecosystems. In

TABLE IV: Forgetting values ($FM$) for **Active Learning with High-Confidence Sampling** and **Active Learning with Low-Confidence Sampling** across 10 runs. GREEN values indicate greater forgetting in Random Sampling, while RED values indicate less forgetting in Random Sampling.

| $f_i$ | **High-Confidence Sampling ($F^+$)** | | | | | | | | | | **Low-Confidence Sampling ($F^-$)** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ | $s_0$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ | $s_7$ | $s_8$ | $s_9$ |
| $f_0$ | 0.0 | | | | | | | | | | 0.0 | | | | | | | | | |
| $f_1$ | 47.4 | 0.0 | | | | | | | | | 41.9 | 0.0 | | | | | | | | |
| $f_2$ | 66.2 | 23.8 | 0.0 | | | | | | | | 68.1 | 31.1 | 0.0 | | | | | | | |
| $f_3$ | 70.5 | 27.9 | 25.2 | 0.0 | | | | | | | 78.1 | 35.9 | 20.2 | 0.0 | | | | | | |
| $f_4$ | 76.9 | 46.5 | 46.8 | 12.0 | 0.0 | | | | | | 91.0 | 38.7 | 51.2 | 26.5 | 0.0 | | | | | |
| $f_5$ | 79.8 | 46.2 | 35.1 | 19.5 | 42.8 | 0.0 | | | | | 88.5 | 41.2 | 52.1 | 30.8 | 44.3 | 0.0 | | | | |
| $f_6$ | 84.2 | 49.9 | 38.7 | 12.3 | 50.8 | 33.5 | 0.0 | | | | 88.5 | 40.5 | 52.1 | 23.1 | 47.9 | 42.4 | 0.0 | | | |
| $f_7$ | 89.3 | 53.0 | 51.7 | 12.4 | 53.3 | 25.6 | 50.1 | 0.0 | | | 80.1 | 46.9 | 62.4 | 26.0 | 49.0 | 47.2 | 47.5 | 0.0 | | |
| $f_8$ | 85.6 | 54.1 | 55.2 | 25.8 | 47.1 | 36.6 | 52.1 | 53.6 | 0.0 | | 85.2 | 48.6 | 65.0 | 25.6 | 51.2 | 42.9 | 49.8 | 49.9 | 0.0 | |
| $f_9$ | 89.2 | 56.8 | 61.1 | 25.8 | 56.4 | 29.3 | 37.2 | 53.9 | 30.6 | 0.0 | 89.8 | 48.1 | 69.0 | 32.8 | 57.9 | 40.0 | 43.1 | 33.2 | 49.5 | 0.0 |

this study, we explored replay strategies to mitigate catastrophic forgetting in plant species classification under a class-incremental learning problem setting, where new species are introduced and learned sequentially over time. When a model learns a new class, it experiences catastrophic forgetting, leading to a decline in performance on previously learned classes. To mitigate this, we reintroduce samples from past tasks, which can be selected randomly or based on a specific fitness function.

We used active learning to select the most relevant samples for replay. Specifically, we evaluated two AL strategies: selecting the samples with the lowest confidence and those with the highest confidence as predicted by the model. We then compared these strategies against random sampling for replay.

The results showed no significant difference in model accuracy. However, both AL-based strategies led to a notable reduction in catastrophic forgetting. Specifically, high-confidence sampling reduced forgetting by 11.9 percentage points, while low-confidence sampling achieved a reduction of 9.2 percentage points.

A key strength of this study is its application in a real-world scenario rather than on benchmark datasets. While this presents unique challenges, it also provides valuable insights into the practical implementation of continual learning for dynamic and evolving environments. Moreover, by facilitating the tracking of species over time, this approach supports the development of a biodiversity index, which could help quantify ecosystem changes and inform conservation and agricultural management strategies.

Future work will focus on scaling the experiment to the full 105-class dataset, increasing the number of training epochs, adjusting the replay rate, and refining the methodology. A key direction for improvement is the integration of multimodal learning, where additional data modalities, such as acoustic signals, could complement image-based classification for different living organisms. Other continual learning strategies, as regularization methods, should be explored, as standalone approaches or with replay, to further mitigate forgetting.

## ACKNOWLEDGMENTS

This work was supported by project PEGADA 4.0 (PRR-C05-i03-000099), financed by the PPR - Plano de Recuperação e Resiliência and by national funds through FCT, within the scope of the project CISUC (UID/CEC/00326/2025).

## REFERENCES

[1] D. Axford, F. Sohel, M. Vanderklift, and A. Hodgson, "Collectively advancing deep learning for animal detection in drone imagery: Successes, challenges, and research gaps," *Ecological informatics*, p. 102842, 2024.
[2] B. Van Merriënboer, J. Hamer, V. Dumoulin, E. Triantafillou, and T. Denton, "Birds, bats and beyond: Evaluating generalization in bioacoustics models," *Frontiers in Bird Science*, vol. 3, p. 1369756, 2024.
[3] G. M. van de Ven, N. Soures, and D. Kudithipudi, "Continual learning and catastrophic forgetting," 2024.
[4] G. M. Van de Ven and A. S. Tolias, "Three scenarios for continual learning," *arXiv preprint arXiv:1904.07734*, 2019.
[5] N. Holla, P. Mishra, H. Yannakoudakis, and E. Shutova, "Meta-learning with sparse experience replay for lifelong language learning," *arXiv preprint arXiv:2009.04891*, 2020.
[6] S. Ho, M. Liu, S. Gao, and L. Gao, "Learning to learn for few-shot continual active learning," *Artificial Intelligence Review*, vol. 57, no. 10, p. 280, 2024.
[7] X. Li, B. Tang, and H. Li, "Adaer: An adaptive experience replay approach for continual lifelong learning," *Neurocomputing*, vol. 572, p. 127204, 2024.
[8] L. Wang, X. Zhang, H. Su, and J. Zhu, "A comprehensive survey of continual learning: Theory, method and application," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5362–5383, 2024.
[9] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the National Academy of Sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
[10] A. A. Rusu, N. C. Rabinowitz, G. Desjardins, H. Soyer, J. Kirkpatrick, K. Kavukcuoglu, R. Pascanu, and R. Hadsell, "Progressive neural networks," 2022.
[11] D. Rolnick, A. Ahuja, J. Schwarz, T. Lillicrap, and G. Wayne, "Experience replay for continual learning," in *NeurIPS*, vol. 32, Curran Associates, Inc., 2019.
[12] V. e. a. Lomonaco, "Avalanche: An end-to-end library for continual learning," in *Proceedings of CVPR 2021*, pp. 3600–3610, 2021.
[13] W.-N. Hsu and H.-T. Lin, "Active learning by learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, Feb. 2015.
[14] D. Lopez-Paz and M. A. Ranzato, "Gradient episodic memory for continual learning," in *NeurIPS*, vol. 30, Curran Associates, Inc., 2017.
[15] E. Haussmann, M. Fenzi, K. Chitta, J. Ivanecky, H. Xu, D. Roy, A. Mittel, N. Koumchatzky, C. Farabet, and J. M. Alvarez, "Scalable active learning for object detection," in *2020 IEEE Intelligent Vehicles Symposium (IV)*, pp. 1430–1435, 2020.
[16] C. Silva, D. Costa, J. Costa, and B. Ribeiro, "Data annotation quality in smart farming industry," *Production & Manufacturing Research*, vol. 12, no. 1, p. 2377253, 2024.
[17] J. Nugent, "Inaturalist," *Science Scope*, vol. 41, no. 7, pp. 12–13, 2018.
[18] I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollár, "Designing network design spaces," 2020.