

# Overview of FungiCLEF 2025: Few-Shot Classification With Rare Fungi Species

Klara Janouskova<sup>1</sup>, Jiří Matas<sup>1</sup> and Lukas Pícek<sup>2,3,\*</sup>

<sup>1</sup>Visual Recognition Group, Faculty of Electrical Engineering, Czech Technical University in Prague, Czech Republic

<sup>2</sup>Department of Cybernetics, Faculty of Applied Sciences, University of West Bohemia, Czech Republic

<sup>3</sup>Inria, LIRMM, University of Montpellier, France

## Abstract

FungiCLEF 2025, the 4<sup>th</sup> edition of the FungiCLEF challenge, was organized as part of the LifeCLEF and the FGVC workshops. This year's edition targeted few-shot classification of rare fungi species. Participants were tasked with identifying species from multimodal observations, including images, structured metadata, and environmental data. The data was collected through citizen science and underwent expert-based labeling. Building upon the FungiTastic dataset, FungiCLEF 2025 emphasized real-world constraints such as limited training samples, high intra-class variability, fine-grained inter-class similarities, and distribution shift. The competition attracted 74 teams, with the leading submissions demonstrating significant gains over the provided baselines, showcasing the potential of pretrained vision transformers, contrastive learning, and ensemble techniques. This overview summarizes the challenge setup, dataset, baselines, participant strategies, and key findings, and outlines directions for future work. The winning team achieved a top-5 accuracy of 78.9%, outperforming baselines by over 52%.

## Keywords

LifeCLEF, FungiCLEF, fine-grained, classification, multi-modal, fungi, species, machine learning, computer vision

## 1. Introduction

Accurate recognition of fungi species is important for biodiversity monitoring, ecological research, and early detection of invasive or toxic species [1, 2, 3, 4]. Given the vast number of species and their morphological variability [5], automated tools have become essential for both experts and citizen scientists in the identification process. Many fungal species are known from only a few observations [3, 6]. This makes it difficult to obtain large, well-annotated image datasets that are required for training conventional machine learning models. Few-shot learning [7, 8] is a machine-learning paradigm that targets learning from only a few labeled examples per class. This better fits real-world conditions of fungi recognition, where collecting and labeling sufficient data for each species is often expensive, time-consuming, or infeasible.

Fine-grained visual recognition of fungi presents several additional challenges, including high intra-class variability (Figure 1), subtle inter-class differences (Figure 2), and complex backgrounds. The inter-class visual appearance similarities are common in mycology, as many fungal species closely resemble one another. Consider the following species from Figure 1: *Psathyrella fragrans*, known for its distinctive fragrant aroma; *Psathyrella citerinii*, which often exhibits a unique color palette; *Psathyrella sphagnicola*, typically associated with specific mossy habitats; *Entoloma favrei*, characterized by its unique cap and gill structure; and *Psathyrella orbiculari*, recognized for its rounded fruiting body. While their visual characteristics may blend into one another, precise identification is crucial in understanding their ecological roles and interactions within their environments [4].

The morphological similarity often makes visual identification extremely challenging, as such subtle differences may be difficult to spot without expert knowledge. These issues are amplified in few-shot fungi classification due to variations in lighting, growth stages, and environmental context, often captured by non-expert users [6, 9]. In many cases, the most reliable means of distinguishing

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

\*Corresponding author.

✉ lukaspicek@gmail.com (L. Pícek)

ORCID 0000-0002-0191-7510 (K. Janouskova); 0000-0003-0863-4844 (J. Matas); 0000-0002-6041-9722 (L. Pícek)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).



**Figure 1: Intra-class visual appearance changes.** A *Comatricha alta* specimen exhibits a notable progression in color and shape during its developmental stages. From initial pale yellow-brown through reddish hues to deep, dark brown and black at maturity.



**Figure 2: Inter-class visual appearance similarities.** Some species resemble each other. In many cases, only DNA analysis or microscopic images of spores allow correct identification. From left to right: *Psathyrella fragrans*, *Psathyrella citerinii*, *Psathyrella sphagnicola*, *Entoloma favrei*, *Psathyrella orbiculari*.

these species lies in comprehensive and time-consuming techniques such as DNA sequencing or the examination of microscopic structures like spores.

To address these challenges, FungiCLEF [10, 11, 12] has been organized annually, now in its 4<sup>th</sup> edition, with the goal of advancing the state of the art in fine-grained recognition, fostering collaboration between machine learning researchers and domain experts, raising public awareness about fungal biodiversity, and promoting real-world impact through applied ML solutions. This year, the focus is on few-shot fine-grained recognition, which is especially important in biological datasets, where species exhibit a long-tail distribution and many classes have very limited annotated data. In the context of species prediction, and fungi in particular, advances in few-shot recognition support tasks such as biodiversity assessment, ecological monitoring, and the development of tools for citizen scientists, where collecting large training datasets is often infeasible [2, 3, 4, 13].

The competition attracted 74 teams, of which 6 submitted working notes detailing their solutions. The remainder of this overview outlines the challenge, including its organization, data, and evaluation protocol. This is followed by the overall results and a summary of the submitted working notes. The conclusions highlight key findings from the competition and they suggest directions for future work.

## 2. Challenge Description

The FungiCLEF 2025 competition, hosted on Kaggle as part of the LifeCLEF [14, 15] and FGVC workshops, focused on few-shot classification of rare fungi species with multimodal data. Participants were asked to build models to recognize species from very limited training examples (<5) taken from expert-verified records in the Atlas of Danish Fungi and reflecting real-world biodiversity monitoring scenarios. Submissions were CSV files matching observation IDs to predicted labels, and evaluation was based on top-5 accuracy to account for the difficulty of distinguishing visually similar species.

The challenging nature of the task is demonstrated by the results of the few-shot baselines presented in Table 1, clearly showing the impact of the number of training samples: while for species with 4

**Table 1**

Few-shot classification accuracy (%) of BioCLIP-based methods (from [6]) as a function of the number of training observations per class. A consistent increase in performance with increased training samples can be observed for both baselines. Note: each training example may include multiple photographs, so the single observation results for the NN and centroid methods can be different.

| classifier        | training observations per class |       |       |       |
|-------------------|---------------------------------|-------|-------|-------|
|                   | 1                               | 2     | 3     | 4     |
| Centroid          | 12.69                           | 23.48 | 32.86 | 40.26 |
| Nearest Neighbour | 12.84                           | 21.33 | 30.02 | 29.39 |

training observations, the baselines achieve top-1 accuracies of 40% and 29%, respectively. For species with one training observation, the performance drops dramatically to only 13% for both of the methods.

## 2.1. Evaluation Protocol

To account for the difficulty of distinguishing visually similar fungi species, we used top-5 accuracy as the primary evaluation metric for the FungiCLEF 2025 challenge. It measures the percentage of test samples where the correct label is among the model’s top-5 predicted labels. Top-5 accuracy allows to account for reasonable alternatives, reflecting real-world scenarios where multiple plausible predictions can be useful, e.g. as a shortlist for experts. More generally, the standard *top-k accuracy* metric is defined as:

$$\text{top-}k = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(y_i \in \hat{Y}_i^k),$$

where  $N$  is the total number of test samples,  $y_i$  is the true label for the  $i$ -th sample,  $\hat{Y}_i^k$  is the set of top  $k$  predicted labels for the  $i$ -th sample, and  $\mathbf{1}(\cdot)$  is the indicator function, returning 1 if the condition is true and 0 otherwise.

While the top-5 accuracy determines the official leaderboard rankings, we also report *top-k accuracies* for a range of different cutoff thresholds ( $k = 1, 2, \dots, 10$ ) to provide a more detailed evaluation of models’ performance.

## 2.2. Baselines

We provided a [Kaggle starter notebook](#) that implements two few-shot classification baselines from [6] with the objective to support participants and reduce entry barriers. The baselines rely on BioCLIP [16] embeddings and FAISS [17] for efficient nearest-neighbor search. The notebook includes all core components: data loading, pre-processing, feature extraction and storage, classification on top of precomputed features and evaluation. It enables participants to run reproducible experiments and extend the code with minimal computational and coding overhead. It encourages broad participation across both machine learning and biodiversity research communities and beyond by offering a functional and modular starting point.

The first baseline is a **centroid-based classifier**, which computes a prototype (mean embedding) for each class using the BioCLIP representations of the training data. In inference, each query image is embedded and then classified by assigning it to the nearest class centroid using cosine similarity.

The second baseline is a **nearest neighbors (NN)** classifier using FAISS [17], an efficient similarity search library optimized for high-dimensional vector spaces. Test samples are embedded and compared against all training embeddings, and the label is assigned based on the label of the nearest neighbor. FAISS accelerates the nearest neighbor search and thus supports scalable evaluation on larger datasets.

## 2.3. Timeline

The FungiCLEF 2025 competition launched on March 7, 2025, around one week earlier than last year. It was open to submissions for about 10 weeks until May 19. The competition was hosted on [Kaggle](#) and promoted through [LifeCLEF](#) [14, 15] and [FGVC](#), as well as on social media.

## 2.4. Working Notes

Participants were strongly encouraged to submit both their code and a detailed technical report (Working Notes) to ensure their results be fully reproducible. The working notes provide an in-depth analysis of the techniques employed, including hyperparameter tuning, model ensembling, and loss function selection, offering valuable insights into the development of the method for fungal image classification. The Working Notes underwent a thorough review; two experts with strong publication records in Computer Vision and Machine Learning provided detailed feedback. The review process was primarily designed to guarantee reproducibility and maintain quality standards. The review was single-blind, allowing participants to respond with up to two rebuttals to address any concerns raised by the reviewers.

**New in 2025: AI-Assisted Reviewing.** To explore whether large language models can improve review consistency and coverage, we introduced a third, *automatic* reviewer based on ChatGPT. Each Working Note therefore received an additional LLM-generated review alongside the two expert reviews. The ChatGPT review was produced with the prompt shown below, constraining the model to the workshop’s emphasis on reproducibility and clarity.

### Prompt used for the ChatGPT reviewer:

You are acting as a peer reviewer for a scientific workshop in computer science: LifeCLEF 2025, part of CLEF. The workshop focuses on biodiversity informatics challenges involving machine learning, computer vision, and related techniques. The emphasis of the submissions is on reproducibility, careful experimental evaluation, and thoughtful analysis, rather than purely on novelty. Please carefully read the following paper submission and write a professional, constructive review. Your review should include the following sections:

1. Summary:
  - \* Briefly summarize the task, methods, datasets, and key findings.
  - \* State clearly what problem the authors are addressing, which LifeCLEF challenge it pertains to, and what their main contributions are.
2. Strengths: List the strong aspects of the paper, such as:
  - \* Reproducibility (e.g., availability of code, data, clear methodology)
  - \* Careful experimental design
  - \* Well-performed ablation studies or error analyses
  - \* Insightful discussions of results
  - \* Clarity of writing and presentation
3. Weaknesses / Areas for Improvement: Identify any weaknesses or limitations, such as:
  - \* Missing details that would prevent reproduction
  - \* Lack of ablations or sensitivity analyses
  - \* Incomplete or unclear description of the method
  - \* Insufficient discussion or interpretation of the results
  - \* Missing comparison to appropriate baselines
4. Detailed Comments:
  - \* Provide actionable, constructive feedback that the authors can use to improve their paper.
  - \* You may point out specific sections, figures, or tables that need clarification, expansion, or correction.
  - \* Comment on both scientific and presentational aspects.
5. Overall Evaluation: Please provide your overall recommendation, choosing one of:
  - \* Strong Accept | Accept | Weak Accept | Borderline | Weak Reject | Reject | Strong Reject

### Important Reviewing Guidelines:

- \* Focus on scientific rigor, reproducibility, and clarity rather than novelty alone.
- \* Do not hallucinate or infer information not present in the submission.
- \* Be neutral, unbiased, and professional.
- \* If some required information is missing, state it explicitly.






**Figure 3:** FungiTastic-FS. Photos in the dataset are of very rare species. Some do not look like typical mushrooms.

## 2.5. Dataset

The FungiCLEF 2025 competition is based on the **FungiTastic-FS** dataset, a curated few-shot subset of the FungiTastic benchmark [6]. It consists of selected, expert-verified observations of fungi species submitted to the Atlas of Danish Fungi. Each observation includes one or more photographs of the same specimen. It is enriched with structured metadata, including location, timestamp, substrate, habitat, toxicity status, GPS coordinates, elevation, land cover classification, biogeographical zone, and other relevant details. In addition to photographs, labels and metadata, the dataset includes automatically generated captions describing the specimen’s visual characteristics, created using the Malmo-7B VL model [18]. See Figure 4 for an example observation. For more info, please see the data source paper [6].

The FungiCLEF 2025 dataset specifically targets *rare and under-recorded species*, each with between 1 and 4 training observations. These species represent approximately 20% of all labeled instances in the broader FungiTastic dataset. The challenge emphasizes fine-grained classification in a realistic low-data regime. Importantly, many fungi in this dataset deviate from typical mushroom morphology, as shown in Figure 3. The dataset is divided into the following subsets:

- **Training set** with 7,819 images across 4,293 observations, covering 2,427 species.
- **Validation set** with 2,285 images across 1,099 observations, covering 570 species.
- **Test set** with 1,911 images across 999 observations, covering 567 species.

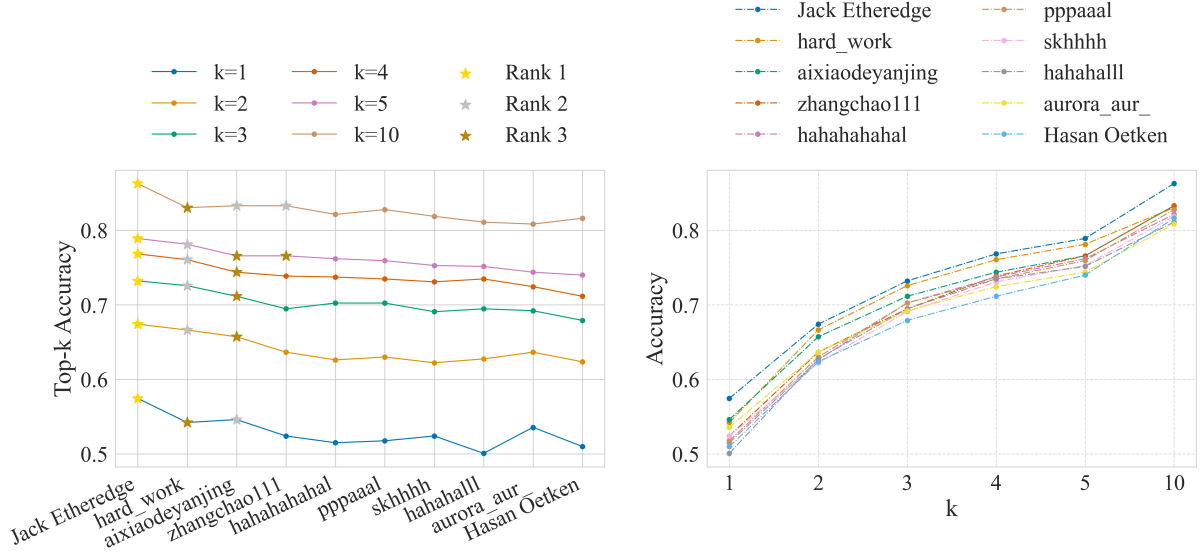
| Photographs   | Caption  | Labels  | Metadata  |
|---|--|---|---|
|  | <p>The image shows a cluster of yellowish-orange fungi growing on the fallen tree trunk. These fungi have a distinctive appearance with several notable features: <b>Shape:</b> The fungi have a rounded, cap-like shape typical of many bracket fungi or polypores ... <b>Texture:</b> The surface of the fungi appears smooth and slightly glossy, indicating a moist environment. This texture is common in many species of fungi that thrive in damp conditions. <b>Size:</b> While it's difficult to determine exact sizes without a reference scale, the fungi seem to be relatively small compared to the tree trunk. They cover a significant portion of the visible area on the fallen log, suggesting they are mature but not overly large. <b>Colour:</b> The fungi display a vibrant yellowish-orange hue, ...</p> | <p><b>Kingdom:</b> Protozoa<br/><b>Phylum:</b> Mycetoza<br/><b>Class:</b> Myxomycetes<br/><b>Order:</b> Physarales<br/><b>Family:</b> Physaraceae<br/><b>Genus:</b> Fuligo<br/><b>Species:</b> F. luteonitens</p> | <p><b>Toxic:</b> No<br/><b>Elevation:</b> 87 m<br/><b>Substrate:</b> Dead wood<br/><b>Date:</b> 1. 1. 2022<br/><b>Location:</b> [55.59,12.35]<br/><b>Land cover:</b> Evergreen needleleaf forests<br/><b>Bio region:</b> Atlantic<br/><b>Habitat:</b> Thorny scrublandd</p> |

**Figure 4: FungiCLEF observations.** Each includes one or more photos of a single specimen, a Malmo-7B generated [18] caption for each photo, expert-verified taxon labels, and observation metadata.

## 3. Challenge Results

This section provides an analysis of the overall results. We compare teams based on their submission with the best top-5 accuracy on the private test set. The winning submission achieved an impressive top-5 accuracy of 78.9%. The runner-up achieved 78.1%, followed closely by the 3<sup>rd</sup> submission at 76.6%.

Figure 5 plots the top- $k$  accuracies,  $k \in \{1, 2, 3, 4, 5, 10\}$ , of the 10 highest-ranked submissions on the private leaderboard. The winning submission consistently outperforms all others across all the metrics (values of  $k$ ). Notably, it leads by a significant margin in top-1 and top-6-top-10 (with  $k \in \{6, 7, 8, 9\}$  not displayed in the plot) accuracy, with much narrower gaps observed for intermediate  $k$  values. It can be further observed that the leaderboard rankings would change if a different metric (i. e. top-1) was



**Figure 5: Private Leaderboard results of the top 10 submissions.** (left) Performance of the teams across different top-k accuracy metrics. Teams ranking 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> for each metric are highlighted. (right) Evolution of the top-k accuracy metric as  $k$  increases for each team.

chosen. The second and third place submissions, for example, would shift under top-1 or top-10 metrics. This suggests that the choice of metric has a strong influence on leaderboard outcomes. Some teams may have specifically optimized for top-5, while others aimed for more general performance across different  $k$  values.

**Baseline Performance.** We implemented two baselines, described in Subsection 2.2: a *Centroid* prototype-based method and a *Nearest Neighbor (NN)* classifier, both based on image-level features only. The Centroid baseline achieved an accuracy of 33.2% on the public test set and 26.7% on the private test set. The NN baseline performed slightly worse, with 28.3% and 24.7% on the public and private test sets, respectively. In comparison, the winning team’s approach outperformed the stronger *Centroid* baseline by more than 52 percentage points.

## 4. Participants and Methods

Out of 74 participating teams, 53 outperformed both baselines, and six teams submitted working notes detailing their approach, including the winning team and the runner-up. Each team’s method and its performance on both public and private leaderboards are listed in Table 2. All working note submissions utilize embeddings from strong pre-trained vision transformers, including DINOv2 [19], BEIT [20], BioCLIP [16], SigLIP [21], and SAM [22]. While all teams use similar building blocks, their approaches differ in the choice of the backbone, the way embeddings are used (e.g., projection heads, contrastive losses), and whether or how additional modalities are incorporated.

Notably, the top two teams, **Jack Etheredge** and **hard\_work**, rely solely on the image modality, outperforming the next-best working-note method by more than 20%. This highlights that there is still significant headroom in optimizing image-based pipelines, particularly when using strong data augmentations and ensemble strategies.

Several teams show that when integrated effectively, multimodal information can bring substantial improvements. For example, **Yang Tuan Anh** improved performance from 30% to 46.2% by incorporating textual captions and metadata. **I2C-UHU-Pegasus** achieved a 7.5% gain from modeling ecological context. In contrast, **DS@GT LifeCLEF** saw no improvement from metadata, indicating that the effectiveness of multimodal integration depends heavily on implementation details.

Specialized domain encoders such as BioCLIP consistently outperform general-purpose alternatives like SigLIP. Class-balanced sampling, Mixup, and careful prototype construction also contribute to performance, each providing gains of 4–7%. Finally, generative and large language models (e.g., Gemini, GPT-4.1 Mini, Mistral) performed poorly compared to vision-only approaches, suggesting that they currently lack the fine-grained visual grounding needed for this task.

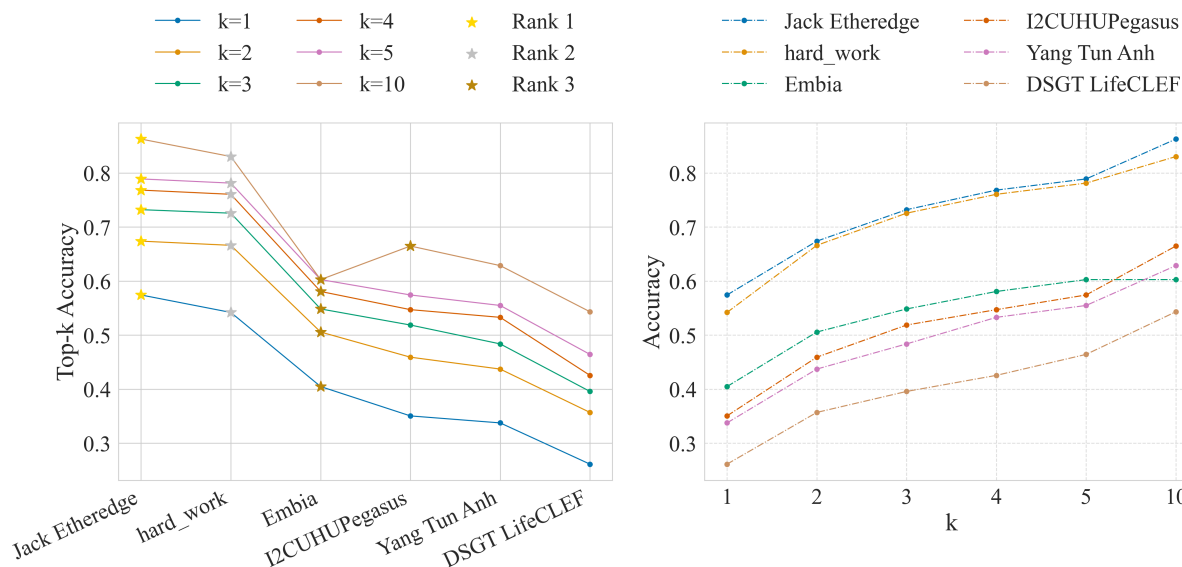
Several submissions also point to a domain shift between validation and hidden test sets. In one case, an ensemble hurt validation performance but improved test performance, highlighting the limitations of current validation splits and the need for better modeling of real-world distributional shifts.

Figure 6 shows the top- $k$  accuracies on the private leaderboard for all the teams that submitted a working note. Up to  $k = 5$ , rankings remain stable. For  $k > 5$ , team Embia’s performance plateaus, resulting in a lower rank at  $k = 10$ . This is because their submission only included predictions up to top-5, preventing any gains beyond that point. In contrast, all the other teams see improved performance as  $k$  increases.

**Table 2**

**Performance of provided baselines and methods described in the submitted working notes.** All methods consistently outperformed baselines by a considerable margin, with a noticeable performance gap between higher- and lower-ranked submissions.

| Rank | Team            | Method core                      | Top-5 (%) [public] | Top-5 (%) [private] |
|------|-----------------|----------------------------------|--------------------|---------------------|
| 1    | Jack Etheredge  | Aug+Proj+Ensemble                | 81.0               | 78.9                |
| 2    | hard_work       | Contrastive Transformer          | 78.3               | 78.1                |
| 17   | Embia           | Context Fusion                   | 63.3               | 60.3                |
| 22   | I2C-UHU-Pegasus | Multimodal Proto                 | 58.0               | 57.4                |
| 26   | Yang Tuan Anh   | BioCLIP ProtoNet                 | 57.1               | 55.5                |
| 35   | DS@GT LifeCLEF  | Transfer + Mixup                 | 50.9               | 46.4                |
| -    | <i>baseline</i> | <i>centroid-based classifier</i> | 33.2               | 26.7                |
| -    | <i>baseline</i> | <i>nearest neighbors</i>         | 28.3               | 24.7                |



**Figure 6: Private Leaderboard results of all teams that submitted a working note.** (left) Performance of the teams across different top- $k$  accuracy metrics. Teams ranking 1<sup>st</sup>, 2<sup>nd</sup> and 3<sup>rd</sup> for each metric are highlighted. (right) Evolution of the top- $k$  accuracy metric as  $k$  increases for each team.

## 4.1. Working Notes

Team **Jack Etheredge** [23] (Top1) developed an ensemble of prototypical networks that integrates embeddings from several pretrained models (e.g., SAM [22], BEIT [20], DINOv2 [19]). For each image, multiple geometric augmentations are applied to increase variability, both during training and testing. The embeddings from these models are concatenated and passed through a two-layer projection network to enhance class separation. Predictions are made using cosine similarity between observation-level and class-level prototype embeddings. An ensemble of multiple independently trained pipelines ensures stable and robust performance. Test-time data augmentation (specifically five-crop, horizontal flip, and 90° rotations) is a key component; removing it leads to a performance drop from 63% to 37.3%. The two-layer projection network is also critical: without it, accuracy drops from 63% to 54.6%. Optimal results require the use of both cross-entropy and InfoNCE losses [24], which together improve the quality of the projected embeddings. Individually, BEIT and DINOv2 achieve top-1 accuracies in the range of 58–60%, while SAM performs poorly at 11.7%; however, including SAM embeddings still boosts ensemble performance, suggesting they provide valuable complementary information.

Team **hard\_work** [25] (Top2) designed a supervised contrastive learning approach that combines DINOv2 features with a customized Transformer model. The approach helped to create a rich feature space that better captures subtle visual differences between species, even with limited data. A specially designed supervised contrastive loss helps the model learn more distinct class representations by efficiently organizing positive and negative training samples. Experiments revealed that contrastive learning benefits significantly from larger batch sizes, with a minimum of 256 needed for good performance. While increasing the batch size beyond this point provides some improvement, the gains are marginal.

Team **Embia** [26] (Top17) focused on enhancing prototype quality for few-shot classification. They used BioCLIP embeddings and prototypical networks, carefully constructing class prototypes by averaging embeddings from both the training and validation datasets. This approach makes the prototypes more representative, especially for species with few samples, resulting in over 30 percentage points improvement in top-5 accuracy compared to the baseline. A comparison between domain-specific BioCLIP and the general-purpose SigLIP embeddings shows a clear advantage for BioCLIP, with accuracies of 61.1% and 53.5%, respectively. Further, extending the training set with validation images increases performance from 61.1% to 64.2%.

Team **I2C-UHU-Pegasus** [27] (Top22) developed a multimodal pipeline that integrates visual features from fine-tuned BioCLIP and DINOv2 with structured textual descriptions, ecological metadata, and hierarchical taxonomic information. They fine-tuned the BioCLIP model while keeping most of its layers frozen, adding a specialized multimodal classifier that weighs visual and textual information. Ecological context modeling significantly improved performance, and the ensemble combines multiple sources of information for better rare species recognition. Ablation studies show that ecological context delivers the largest single gain, adding 7.5% to top-1 accuracy. Addressing class imbalance contributes a further 5.0%, while fine-tuning the visual backbone yields an additional 4.0%. A failure analysis reveals a small set of species that are always misclassified (100% error), with most errors occurring between taxa of the same genus or family. Finally, DINOv2-based ensemble experiments highlight a domain shift: the ensemble lowers accuracy on the validation set but improves it on the hidden test set, underscoring the challenges posed by evolving ecological data.

Team **Yang Tuan Anh** [28] (Top26) introduced a multimodal few-shot learning system that merges image embeddings from BioCLIP, SigLIP, and DINOv2 with metadata and automatically extracted textual features. The training involves two stages: first, supervised pretraining of the multimodal encoder; second, few-shot fine-tuning using a prototypical network under episodic training. They also use an observation-level re-ranking method to consolidate predictions across multiple images of the



same observation. Compared to the image-only performance of 30%, adding textual captions improves accuracy to 44.2% while including metadata boosts it to 45.8%. Using all three modalities together results in 46.2% demonstrating their complementarity. The single-stage supervised pretrained model achieves 46.2% the two-stage fine-tuned model reaches 50.7% and combining both with observation-level re-ranking leads to 55.5%.

Team **DS@GT LifeCLEF** [29] (Top35) combine transformer-based embeddings (e.g., DINOv2, PlantCLEF [30]), class-balanced sampling, Mixup, and a linear classifier. While generative AI and multi-objective loss were explored, the best results were achieved with vision-only, domain-pretrained embeddings and class-balanced training strategies. Interestingly, experiments with advanced models, such as Gemini, Mistral, and ChatGPT, led to negative results. Their results show the dominance of tailored, domain-specific solutions, where the proposed method with Mixup and class weighting scores 45.4% while the best-performing generative model, Gemini 2.5 Flash, reaches a significantly lower accuracy of 13.6%. OpenAI GPT-4.1 Mini and MistralAI Mistral Medium 3 lagged behind even further with 6.2% and 3.1%, respectively.

## 5. Conclusions

The paper presented an overview and results evaluation of the 4<sup>th</sup> edition of the FungiCLEF 2025 challenge, organized in conjunction with the CLEF LifeCLEF lab [14] and CVPR-FGVC11 — The 11th Workshop on Fine-Grained Visual Categorization, held within the CVPR conference. FungiCLEF 2025 built upon previous editions [10, 11, 12], with a continued focus on the challenging task of few-shot fungi species recognition. In this year’s edition, Participants were tasked with classifying observations of species with a limited number of training samples from the FungiTastic dataset using multimodal inputs such as images, machine-generated captions, and environmental metadata.

In contrast to previous years, the 2025 edition removed constraints on the model size and it reverted to an open competition format. This change led to increased participation of more than 70 teams. It sparked a diverse range of methodological innovations across vision-only, multimodal, and metric-based approaches. Several top-performing teams demonstrated that well-optimized image-only pipelines can still achieve state-of-the-art performance, while others explored the potential of structured multimodal fusion strategies. The main outcomes from this year’s evaluation are:

- **A highly-optimized vision-only method wins.** The top-performing solutions in the challenge demonstrate that carefully tuned strong vision-only pipelines, built on pretrained transformers and enhanced with effective data augmentation and ensemble strategies, can outperform more complex multimodal systems. For example, the winning solution combined embeddings from several pretrained models with strong augmentations and a simple projection network.
- **Multimodality is promising but tricky to get right.** Teams that integrated metadata and text saw modest to strong improvements, but only with carefully designed pipelines. Straightforward fusion, even with strong general-purpose models, lagged far behind structured text prompts, learned weighting between modalities, or explicit taxonomic hierarchies. Simply adding more modalities is not enough.
- **Prototypes over neighbours.** Prototypical networks remained a popular choice, with teams refining the basic prototype idea using multiple embeddings, reranking, or ensemble voting. Even without gradient-based fine-tuning, class averaging plus a smart similarity function can go a long way in these settings, and nearest prototypes were preferred over nearest-neighbor approaches.
- **Contrastive learning is finally pulling its weight.** Supervised contrastive loss showed clear gains this year. Especially when using transformer heads or class-specific projection layers, this training objective encouraged better separation in embedding space, crucial when fine-grained differences are subtle and data is sparse.

- **Ensembles help, but at what cost?** Top teams improved their performance by combining predictions from multiple embedding models, data augmentation variants, or even entirely separate pipelines. Ensembling helps to smooth noise and improve generalization, proving once again to be a good strategy for a competition. However, these gains come with a steep price in computational overhead, which, in practice, is often not worth it.
- **Generative models? Not yet.** Some teams tried large vision-language models for zero-shot inference via prompting. While creative, the results fell far short of specialized models and more handcrafted approaches. For now, it seems generative multimodal models still lack the resolution and structure to handle fine-grained multimodal biological classification.
- **Data imbalance.** As in previous editions, rare species dominated the dataset. The teams reported that performance dropped sharply for species with fewer observations. Mixup, class-aware sampling, focal loss, and contrastive training all helped.

**Directions for future work.** An important direction is to investigate whether the strong performance of vision-only systems can be further enhanced through principled multimodal integration. This includes exploring unified multimodal pretraining that jointly leverages visual, ecological, taxonomic, and textual inputs. Robustness to domain shift remains a challenge. Promising future exploration strategies include test-time adaptation, distribution-aware ensembling, and synthetic data augmentation to better generalize across ecological and seasonal variation. Additionally, prototype refinement during inference and uncertainty-guided active learning may support more reliable recognition of rare or ambiguous species.

## 6. Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checks and ChatGPT for improving clarity and rewording sentences. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

## Acknowledgments

This research was supported by the Technology Agency of the Czech Republic, project No. SS73020004. We extend our sincere gratitude to the mycologists from the Danish Mycological Society, particularly Jacob Heilmann-Clausen, Thomas Læssøe, Thomas Stjernegaard Jeppesen, Tobias Guldberg Frøslev, Ulrik Söchting, and Jens Henrik Petersen, for their contributions and expertise. We also thank the dedicated citizen scientists whose data and efforts have been instrumental to this competition.

## References

- [1] D. L. Hawksworth, R. Lücking, Fungal diversity revisited: 2.2 to 3.8 million species, *Microbiology spectrum* 5 (2017) 10–1128.
- [2] P. Stephenson, M. C. Londoño-Murcia, P. A. Borges, L. Claassens, H. Frisch-Nwakanma, N. Ling, S. McMullan-Fisher, J. J. Meeuwig, K. M. M. Unter, J. L. Walls, et al., Measuring the impact of conservation: The growing importance of monitoring fauna, flora and fungi, *Diversity* 14 (2022) 824.
- [3] L. A. Lofgren, J. E. Stajich, Fungal biodiversity and conservation mycology in light of new technology, big data, and changing attitudes, *Current Biology* 31 (2021) R1312–R1325.
- [4] S. D. Warnasuriya, D. Udayanga, D. S. Manamgoda, C. Biles, Fungi as environmental bioindicators, *Science of The Total Environment* 892 (2023) 164583.
- [5] T. Niskanen, R. Lücking, A. Dahlberg, E. Gaya, L. M. Suz, V. Mikryukov, K. Liimatainen, I. Druzhina, J. R. Westrip, G. M. Mueller, et al., Pushing the frontiers of biodiversity research: Unveiling

the global diversity, distribution, and conservation of fungi, *Annual review of Environment and resources* 48 (2023) 149–176.

- [6] L. Pícek, K. Janoušková, V. Cermák, J. Matas, Fungitastic: A multi-modal dataset and benchmark for image categorization, in: *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, 2025, pp. 2046–2056.
- [7] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni, Generalizing from a few examples: A survey on few-shot learning, *ACM computing surveys (csur)* 53 (2020) 1–34.
- [8] Y. Song, T. Wang, P. Cai, S. K. Mondal, J. P. Sahoo, A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities, *ACM Computing Surveys* 55 (2023) 1–40.
- [9] L. Pícek, M. Šulc, J. Matas, T. S. Jeppesen, J. Heilmann-Clausen, T. Læssøe, T. Frøslev, Danish fungi 2020-not just another image recognition dataset, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1525–1535.
- [10] L. Pícek, M. Šulc, J. Heilmann-Clausen, J. Matas, Overview of FungiCLEF 2022: Fungi recognition as an open set classification problem, in: *Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, 2022.
- [11] L. Pícek, M. Šulc, J. Heilmann-Clausen, J. Matas, Overview of FungiCLEF 2023: Fungi recognition beyond 0-1 cost, in: *Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum*, 2023.
- [12] L. Pícek, M. Šulc, J. Matas, Overview of FungiCLEF 2024: Revisiting fungi species recognition beyond 0-1 cost, in: *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, 2024.
- [13] M. Lotfian, J. Ingensand, M. A. Brovelli, The partnership of citizen science and machine learning: benefits, risks, and future challenges for engagement, data collection, and data quality, *Sustainability* 13 (2021) 8087.
- [14] L. Pícek, S. Kahl, H. Goëau, L. Adam, T. Larcher, C. Leblanc, M. Servajean, K. Janoušková, J. Matas, V. Čermák, K. Papafitsoros, R. Planqué, W.-P. Vellinga, H. Klinck, T. Denton, J. S. Cañas, G. Martellucci, F. Vinatier, P. Bonnet, A. Joly, Overview of lifeclef 2025: Challenges on species presence prediction and identification, and individual animal identification, in: *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, Springer, 2025.
- [15] A. Joly, L. Pícek, S. Kahl, H. Goëau, L. Adam, C. Botella, M. Servajean, D. Marcos, C. Leblanc, T. Larcher, et al., Lifeclef 2025 teaser: Challenges on species presence prediction and identification, and individual animal identification, in: *European Conference on Information Retrieval*, Springer, 2025, pp. 373–381.
- [16] S. Stevens, J. Wu, M. J. Thompson, E. G. Campolongo, C. H. Song, D. E. Carlyn, L. Dong, W. M. Dahdul, C. Stewart, T. Berger-Wolf, et al., Bioclip: A vision foundation model for the tree of life, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 19412–19424.
- [17] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, H. Jégou, The faiss library (2024). [arXiv:2401.08281](https://arxiv.org/abs/2401.08281).
- [18] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al., Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, in: *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 91–104.
- [19] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, et al., Dinov2: Learning robust visual features without supervision, *arXiv preprint arXiv:2304.07193* (2023).
- [20] H. Bao, L. Dong, F. Wei, Beit: Bert pre-training of image transformers, *arXiv preprint arXiv:2106.08254* (2021).
- [21] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, et al., Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, *arXiv preprint arXiv:2502.14786* (2025).
- [22] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg,

- W.-Y. Lo, et al., Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
- [23] J. Etheredge, Few-shot fungi classification with prototypical networks using multiple pretrained embedding models, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
  - [24] A. v. d. Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv preprint arXiv:1807.03748 (2018).
  - [25] L. Lu, H. Yang, S. Li, F. Liu, P. Chen, W. Ma, Few-shot fine-grained classification of fungi species using contrastive representation learning, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
  - [26] A. Traore, Éric Hervet, A. Couturier, Improving fungi prototype representations for few-shot classification, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
  - [27] F. C. García, V. P. Álvarez, J. M. Vázquez, M. G. García, I2c-uhu-pegasus at fungiclef 2025: Multi-modal pipeline for rare fungal species classification using fine-tuned vlms and ecological context, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
  - [28] T.-A. Yang, M.-Q. Nguyen, Mushroom for improvement: Prototypical few-shot learning with multimodal fungal features, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
  - [29] J. K. Tam, M. Gustineli, A. Miyaguchi, Transfer learning and mixup for fine-grained few-shot fungi classification, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.
  - [30] H. Goëau, J.-C. Lombardo, A. Affouard, V. Espitalier, P. Bonnet, A. Joly, PlantCLEF 2024 pretrained models on the flora of the south western Europe based on a subset of Pl@ntNet collaborative images and a ViT base patch 14 dinoV2, 2024. doi:10.5281/zenodo.10848263.