I2C-UHU-PEGASUS at FungiCLEF 2025: Multimodal Pipeline for Rare Fungal Species Classification Using Fine-Tuned VLMs and Ecological Context

Notebook for the LifeCLEF Lab at CLEF 2025

Fernando Carrillo García[1], Victoria Pachón Álvarez[1], Jacinto Mata Vázquez[1] and Manuel Guerrero García[1]

[1]*University of Huelva, Andalusia, Spain*

Abstract

Automatic identification of rare fungal species represents one of the most complex challenges in computational mycology and biodiversity conservation. Analysis of collections like the Atlas of Danish Fungi reveals that approximately 20% of verified observations correspond to poorly documented species, highlighting the critical need for systems capable of accurately identifying these underrepresented taxa. The scarcity of labeled samples prevents inclusion of rare species in conventional training sets, severely limiting traditional AI approaches. This research was conducted within the FungiCLEF 2025 framework, an international challenge focused on automatic fungal species classification with particular emphasis on rare species identification. Our methodology combines Vision-Language Models (VLMs) with advanced transfer learning and few-shot learning techniques, integrating multimodal fine-tuning of BioCLIP, multimodal ensemble with DINOv2, probabilistic ecological context modeling, and comprehensive textual description analysis. The results demonstrate a Recall@5 of 0.57438 on the test set, achieving 22nd position among 74 participating teams in FungiCLEF 2025, demonstrating the effectiveness of multimodal integration for few-shot scenarios.

Keywords

VLMs, Few-shot learning, Fungal classification, Rare species, FungiCLEF, Multimodal AI, Deep Learning, Transfer Learning

1. Introduction

Fungal species identification represents a historically complex field where automated solutions are crucial to support experts and researchers in biodiversity conservation efforts. The challenge is particularly acute for rare species, which are often underrepresented in training datasets yet critical for ecological understanding and conservation planning. With an estimated 2.2 to 3.8 million fungal species worldwide [1], most remaining largely undocumented, the development of accurate automated identification systems becomes essential for accelerating biodiversity research and conservation initiatives.

Unlike other biological domains with extensive labeled datasets, mycology presents a substantially different reality characterized by extreme data scarcity. According to data from the FungiTastic dataset [2] used in the FungiCLEF 2025 few-shot recognition challenge, 84.6% of fungal species are represented by five or fewer samples in the training set. This distribution reflects the real-world scenario where rare species are significantly underrepresented in available datasets, creating a critical bottleneck for traditional machine learning methods that require substantial amounts of labeled data per class.

Conventional supervised learning methods face particular difficulties in this context, as they typically demand extensive training examples to achieve reasonable performance. This limitation is especially problematic in biodiversity applications, where taxonomic experts are scarce, field collection is challenging, and the cost of obtaining high-quality annotations is prohibitive. Furthermore, the morphological similarity between closely related species and the high intraspecific variability within species compound the difficulty of accurate identification.

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

fernando.carrillo051@alu.uhu.es (F. C. García); vpachon@dti.uhu.es (V. P. Álvarez); mata@dti.uhu.es (J. M. Vázquez); manuel.guerrero790@alu.uhu.es (M. G. García)

The FungiCLEF 2025 competition [3], part of the LifeCLEF 2025 lab [4], specifically targets this challenge by focusing on few-shot learning scenarios where models must generalize to rare species with minimal training examples. This setup closely mirrors real-world applications where new species discoveries or rarely encountered taxa must be identified based on very limited reference material. Recent advances in Vision-Language Models (VLMs) and multimodal learning present new opportunities to address these limitations by leveraging multiple information sources simultaneously.

Our work addresses these challenges through a comprehensive multimodal fine-tuning strategy that systematically adapts pre-trained Vision-Language Models to the specific requirements of fungal classification. By this approach, we refer to a systematic methodology that jointly optimizes visual and textual representations by fine-tuning both the visual encoder and text encoder of pre-trained VLMs while incorporating structured prompts that combine morphological descriptions, ecological metadata, and taxonomic hierarchies. This methodology differs from traditional fine-tuning by explicitly incorporating domain-specific textual knowledge during the adaptation process, enabling the model to leverage both visual features and structured biological knowledge simultaneously.

This work presents three fundamental contributions to the field of biodiversity informatics and few-shot learning:

1. We demonstrate how to effectively adapt pre-trained models from general biological domains to specific mycology tasks through systematic multimodal fine-tuning that combines visual features with structured textual prompts and ecological context.
2. We propose a multi-source integration framework that systematically combines visual, textual, ecological, and hierarchical taxonomic information using probabilistic modeling and ensemble strategies.
3. We experimentally validate that our methodology achieves reasonable performance in extreme data scarcity scenarios, particularly for species with fewer than five available examples, demonstrating the effectiveness of multimodal learning for biodiversity applications.

2. Related Work

Recent advances in artificial intelligence have transformed biological species classification, particularly through the integration of deep learning models and multimodal approaches. We examine key developments in few-shot learning, vision-language models, and their applications to biodiversity challenges to establish the theoretical foundation for our work.

Biological species classification using artificial intelligence techniques has experienced significant advances with the adoption of deep learning models [5]. Traditional approaches have relied heavily on convolutional neural networks trained on large datasets, but these methods face limitations when applied to long-tailed distributions typical of biological data. The challenge becomes particularly pronounced in biodiversity applications where Zipfian distributions are common, with a few species having abundant samples while the majority remain severely underrepresented.

In the few-shot learning context, various approaches have been developed to address scenarios with limited training data. Prototypical Networks [6] learn to compute prototypes for each class and classify based on distances to these prototypes, while Model-Agnostic Meta-Learning (MAML) [7] learns initialization parameters that can be quickly adapted to new tasks with minimal examples. These approaches have shown promise in computer vision tasks but require careful adaptation for biological applications due to the domain-specific challenges involved.

In FungiCLEF competitions, successful solutions have explored various architectural innovations and training strategies tailored to the challenges of fungal identification. Recent work has combined architectures like MetaFormer [8], implemented specialized loss functions such as Seesaw Loss for long-tailed distributions [9], and explored ensemble approaches [10]. More recently, Chiu et al. [11] demonstrated the effectiveness of self-supervised models like DINOv2 [12] for feature extraction in fungal classification tasks, highlighting the value of general-purpose visual representations in biological domains.

Vision-Language Models (VLMs) represent a paradigm shift in multimodal learning, integrating both image and text processing capabilities in a unified representation space. These models, trained using contrastive techniques on large datasets of image-text pairs, have emerged as powerful tools in biodiversity applications due to their ability to overcome traditional supervised learning limitations by leveraging textual descriptions and metadata. The key advantage of VLMs lies in their capacity to understand relationships between visual features and textual descriptions, enabling zero-shot and few-shot learning capabilities.

BioCLIP [13], one of the base models used in our work, adapts the CLIP architecture specifically to the biological context, significantly improving performance in organism classification tasks across the Tree of Life. BioCLIP consists of a ViT-B/16-based visual encoder and an autoregressive text encoder, trained jointly on TreeOfLife-10M, a dataset spanning over 450,000 taxa with approximately 10 million images. This specialized training allows BioCLIP to capture taxonomic hierarchical relationships inherent to biology, consistently outperforming general domain models by 17-20% in fine-grained biological classification tasks.

DINOv2 [12] —another key component of our system— stands out for its ability to learn robust visual representations without supervision. DINOv2 employs knowledge distillation and contrastive techniques with a Vision Transformer-based architecture, generating high-quality visual embeddings that encode rich semantic information even for classes not seen during training. This characteristic makes it particularly valuable as a complement to domain-specific models like BioCLIP [13], especially in few-shot scenarios where visual diversity is limited.

Unlike previous works that focused mainly on traditional CNN architectures or self-supervised models for visual feature extraction, our approach explores the comprehensive use of multimodal capabilities of VLMs applied to fungal biodiversity, developing a systematic framework that combines specific multimodal fine-tuning, probabilistic ecological context modeling, and hierarchical taxonomic analysis.

3. Dataset and Evaluation Metrics

The FungiCLEF 2025 challenge utilizes the FungiTastic dataset, which presents unique characteristics that make it particularly suitable for evaluating few-shot learning approaches in biological classification. Understanding these dataset properties and the evaluation framework is essential for interpreting our experimental results.

The FungiCLEF 2025 dataset is based on FungiTastic [2] and exhibits the following characteristics:

- **Training**: 4,293 observations distributed across 2,427 species.
- **Validation**: 1,099 observations across 570 species.
- **Images**: Available in multiple resolutions (300p, 500p, 720p, full size).
- **Rich metadata**: Complete taxonomic hierarchy, habitat information, substrate, biogeographic region, and temporal data.
- **Textual descriptions**: Each observation includes a natural language description generated by Malmo-7b VLM.

The challenge requires Top-10 predictions and the extreme long-tail distribution (84.6% of species with five or fewer examples) creates an ideal scenario for evaluating few-shot learning techniques.

3.1. Evaluation Metric: Recall@5

The metric used in FungiCLEF 2025 is Recall@K, which evaluates the percentage of cases where the correct class is found among the k most probable predictions:

$$\text{Recall@}k(y, \hat{Y}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(y_i \in \hat{Y}_i^k)$$

This metric, unlike precision which measures the percentage of exact matches, allows for a more realistic evaluation where a prediction is considered correct if the true class appears among the top k predictions made by the model.

4. Methodology

Our approach combines multiple complementary strategies to address the fundamental challenges of rare species identification through a comprehensive multimodal pipeline. The integration of vision-language models, ecological context, and taxonomic knowledge forms the core of our system designed to handle extreme data scarcity scenarios.

4.1. Multimodal Pipeline Overview

The developed system implements a comprehensive pipeline that leverages multiple information sources for rare fungal species classification. The system architecture combines a multimodal feature extraction backbone (primarily fine-tuned BioCLIP [13]) with: (1) domain-specific data augmentation for fungi, (2) caption processing and prompt structuring, (3) ecological context modeling, (4) hierarchical taxonomic analysis, and (5) efficient search through optimized HNSW [14] indices.

The proposed methodology starts from the premise that visual information alone is insufficient for precise identification of rare species with limited training data. Therefore, the pipeline systematically integrates textual information (morphological descriptions), ecological context (habitat, substrate, biogeographic region) and hierarchical taxonomic knowledge to enrich learned representations.

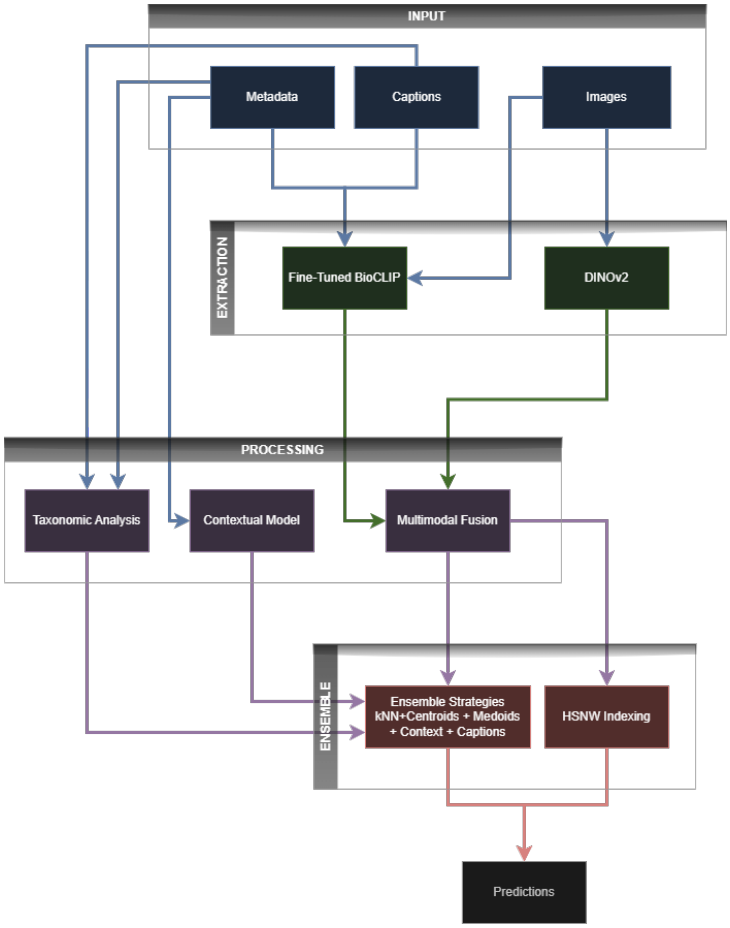

Figure 1: Proposed multimodal pipeline divided into five main stages.

As illustrated in Figure 1, the pipeline follows a systematic approach where input modalities (images, metadata, and descriptions) are processed through specialized extraction modules before being integrated in the processing layer, culminating in ensemble-based predictions through optimized indexing strategies.

4.2. Multimodal Fine-tuning of BioCLIP

Our approach to adapting BioCLIP [13] for fungal classification leverages systematic multimodal fine-tuning that preserves pre-trained knowledge while adapting to domain-specific patterns.

The system's core component is the multimodal fine-tuning of BioCLIP, specifically designed to leverage both visual and textual information available in the dataset. The fine-tuned model architecture introduces a multimodal classifier that operates on the pre-trained BioCLIP model. The BioCLIP backbone remains largely frozen, except for the last 6 transformer layers, all normalization layers, and the visual projection layer.

The multimodal classifier implements a weighted fusion strategy:

$$\mathbf{f}_{fused} = \beta \cdot \mathbf{f}_{image} + (1 - \beta) \cdot \mathbf{f}_{text} \tag{1}$$

where $\beta = 0.75$ was optimized empirically, favoring visual information (75%) while retaining a significant contribution from textual context (25%).

The multimodal classifier implements a robust three-layer architecture:

$$\begin{aligned} \mathbf{h}_1 &= \text{Dropout}(0.3, \text{GELU}(\text{LayerNorm}(\text{Linear}(\mathbf{f}_{fused}, 2048)))) \\ \mathbf{h}_2 &= \text{Dropout}(0.4, \text{GELU}(\text{LayerNorm}(\text{Linear}(\mathbf{h}_1, 1024)))) \\ \mathbf{o} &= \text{Linear}(\mathbf{h}_2, \text{num\_classes}) \end{aligned} \tag{2}$$

This deeper architecture enables more effective learning of multimodal representations. Training uses Focal Loss [15] with alpha=0.3, gamma=1.5 and label smoothing=0.1 to handle extreme class imbalance. Fine-tuning uses AdamW [16] with differentiated learning rates (3e-5 for backbone, 2e-4 for adapter) and cosine annealing with warmup for 10 epochs.

4.3. Multimodal Feature Extraction

The feature extraction process combines visual and textual information at multiple resolutions and modalities to create robust representations for few-shot learning scenarios.

The system extracts both pure image embeddings and fused multimodal embeddings. For each image, the system processes multiple resolutions with specific weights favoring higher resolutions:

$$\mathbf{e}_{obs} = \sum_{r \in R} w_r \cdot \mathbf{e}_r \tag{3}$$

where $w_{300p} = 0.4$, $w_{500p} = 0.9$, $w_{720p} = 1.3$, $w_{fullsize} = 1.5$. These weights were determined experimentally, starting from standard values reported in the literature and optimized for the mycological domain.

Structured prompts are generated systematically:

```
"Identify this fungal species.
Description: [detailed morphological description].
Ecological context --- habitat: [habitat]; substrate: [substrate];
region: [region]; collected in: [month].
Taxonomic information: genus: [genus], family: [family],
order: [order], class: [class], phylum: [phylum]."
```

This structure allows the model to leverage both visual information and multiple relevant context sources for the mycological domain.

4.4. Multimodal Ensemble Architecture

The ensemble strategy combines BioCLIP [13] and DINOv2 [12] features to improve classification robustness, particularly important in few-shot scenarios where individual models may struggle with limited training data.

The system implements an ensemble combining fine-tuned BioCLIP with DINOv2 to improve the robustness of learned representations. Ensemble weights (BioCLIP: 1.4, DINOv2: 1.2) were optimized through systematic experimentation, selecting the configuration that maximized Recall@5 on the validation set:

$$\mathbf{e}_{combined} = w_{BioCLIP} \cdot \mathbf{e}_{BioCLIP} + w_{DINOv2} \cdot \mathbf{e}_{DINOv2} \tag{4}$$

This ensemble leverages the domain-specific knowledge of BioCLIP while benefiting from DINOv2's general-purpose and resilient visual representations, which are particularly valuable for distinguishing morphologically similar species.

4.5. Ecological Context Integration

Our probabilistic approach incorporates ecological metadata into the classification process, leveraging the strong ecological dependencies exhibited by fungal species to improve identification accuracy.

Fungal species often exhibit strong ecological dependencies, which can provide valuable signals for classification. We developed a probabilistic model across four dimensions: habitat types (31 categories), substrate types (30 categories), biogeographic regions (7 main categories), and temporal patterns (12 monthly distributions).

For each ecological factor, conditional probabilities are calculated using Laplace smoothing [17] to handle the long-tail distribution:

$$P(\text{species}|\text{context}) = \frac{\text{count} + \alpha}{\text{total} + \alpha \times N_{classes}} \tag{5}$$

where $\alpha = 0.1$ is used to smooth rare combinations and $N_{classes}$ represents the total number of species.

During prediction, these probabilities are applied as multiplicative boost factors. The complete ecological context integration is formalized as:

$$P_{eco}(\text{species}|\text{context}) = \prod_{c \in \{\text{habitat, substrate, region, month}\}} P(\text{species}|c)^{w_c} \tag{6}$$

where the context weights were optimized empirically: $w_{habitat} = 0.45$, $w_{substrate} = 0.35$, $w_{region} = 0.12$, $w_{month} = 0.25$.

4.6. Textual Description Processing

Our approach extracts and utilizes morphological information from textual descriptions, enabling the system to leverage expert knowledge encoded in natural language descriptions.

The morphological descriptions represent a rich source of information, requiring comprehensive processing. The feature extractor identifies terminology specific to fungal morphology across multiple categories: 16 color terms with location-specific descriptors, 16 shape descriptors for key structures, and 14 texture characteristics. Feature similarity is calculated using a weighted Jaccard index for each category, with higher weights for taxonomic information (0.7-0.8) compared to morphological descriptors (0.2-0.35).

4.7. Hierarchical Taxonomic Classification

We leverage taxonomic hierarchy information for improved classification accuracy, exploiting the well-defined biological classification structure to enhance species-level identification.

Biological classification follows a well-defined hierarchy, providing valuable structural information. We implemented an approach that exploits these relationships through taxonomic consensus voting and verification of hierarchical consistency. Mappings are created between species and their taxonomic classification at five levels: genus, family, order, class, phylum. The weights, optimized through systematic experimentation, are as follows: genus (0.45), family (0.25), order (0.12), class (0.04), phylum (0.01).

Vote aggregations are built in which each species prediction contributes votes to all hierarchical levels, with taxonomic boost mechanisms providing enhanced accuracy for taxonomically challenging groups.

4.8. Optimized Multi-Strategy Fusion

The final integration strategy combines multiple prediction approaches using optimized search algorithms to create a robust ensemble system capable of handling the challenges of few-shot learning.

The final prediction combines five complementary strategies with weights specifically optimized for each strategy in the context of multimodal fusion:

- Multimodal k-NN (weight 0.5) - provides the main foundation through multimodal embedding space search
- Centroid similarity (weight 0.3) - distances to class prototypes
- Medoid similarity (weight 0.2) - most representative example of each class
- Metadata matching (weight 0.45) - ecological context similarity
- Description similarity (weight 0.5) - evaluates similarity between textual morphological descriptions

Search is performed using FAISS [18] with HNSW [14] indexing, specifically optimized for handling multimodal embeddings: M=160, efConstruction=400, efSearch=400, with k_neighbors=50.

4.9. Class Imbalance Handling

We address the extreme class imbalance present in the FungiCLEF 2025 dataset through specialized techniques beyond traditional resampling approaches.

The FungiCLEF 2025 dataset exhibits a high degree of class imbalance where 84.6% of species have five or fewer samples. We addressed this challenge without resorting to traditional oversampling or undersampling techniques, which are inadequate due to extreme data scarcity.

4.9.1. Adaptive Weighted Sampling

To ensure adequate representation of rare species in training batches and prevent the model from being dominated by abundant classes, we implemented adaptive weighted sampling. This approach assigns higher sampling probabilities to underrepresented species, maintaining a balanced learning process across the extreme class imbalance present in the dataset.

We implemented adaptive weighted sampling:

$$w_i = \max(0.05, (\frac{\text{max\_count}}{n_i})^{1/3}) \tag{7}$$

The exponent of 1/3 (the cubic root) provides a moderate compensation that ensures abundant species remain represented during training.

4.9.2. Specific Data Augmentation

These transformations generate realistic visual variations while preserving key morphological traits essential for species identification: random rotations (±20°), horizontal/vertical flips, photometric variations (±30%), and grayscale conversion (10% probability).

For rare species, data augmentation is enhanced through frequency-based boosting to compensate for limited training samples. We developed a custom rare species boost technique that adaptively increases prediction weights based on ecological context and taxonomic hierarchy similarity:

$$\text{boost}_{rare} = 1.0 + \beta \times \left(1.0 - \frac{f_{class}}{t_{rare}}\right) \tag{8}$$

where f_{class} is the class frequency, $t_{rare} = 15$ is the threshold to consider a species rare, and $\beta = 0.5$ is the boost factor. Test-time augmentation applies seven transformations while maintaining consistency with the associated textual prompts.

5. Experiments and Results

We evaluate our multimodal pipeline through comprehensive experiments on the FungiCLEF 2025 challenge, analyzing both overall performance and individual component contributions to understand the effectiveness of our approach for rare species classification.

5.1. Overall Performance

The final model achieves a Recall@5 of 0.57438 on the FungiCLEF 2025 private test set, resulting in 22nd place out of 74 teams. Experiments were conducted using a 200-sample subset from the FungiCLEF 2025 validation set for ablation studies and component analysis.

5.2. Component Analysis

As shown in Table 1, the ablation study reveals interesting results. Multimodal fine-tuning provides a solid foundation (+4.0%), while ecological context emerges as the strongest individual contributor (+7.5%). Notably, the DINOv2 ensemble shows modest regression on validation (-3.1%), but it performs better on the full test dataset, as the increased visual diversity benefits general-purpose visual features.

Table 1
Ablation study of pipeline components on validation set

Component	Recall@5	Improvement
Base BioCLIP (frozen)	0.6200	baseline
+ Multimodal Fine-tuning	0.6450	+4.0%
+ DINOv2 Ensemble	0.6250	-3.1%
+ Ecological Context	0.6430	+2.9%
+ Description Processing	0.6550	+1.9%
+ Multi-Strategy Fusion	0.6700	+2.3%
Total Improvement (Val)		**+8.1%**
Projected Test Performance		**+9-10%**

The results demonstrate pipeline resilience, where individual component setbacks are compensated by complementary strategies. The ecological context provides meaningful recovery (+2.9%), indicating that habitat and substrate metadata contain highly informative co-occurrence patterns. Multi-strategy fusion achieves the strongest cumulative effect (+2.3%), validating the synergistic benefits of combining multiple information sources.

Table 2 shows that the rare species boost provides a meaningful improvement (5.0%) for species with limited training data (5 training examples), while strategy fusion achieves progressive enhancements, with the combined approach yielding an 8.1% improvement over the baseline k-NN method.

Table 2
Performance on rare species and strategy fusion

Method/Strategy	Recall@5	Improvement
Rare Species (5 training observations):		
No boost	0.6200	baseline
Rare species boost	0.6510	+5.0%
Strategy Fusion:		
k-NN only	0.6200	baseline
+ Centroid + Medoid	0.6448	+4.0%
+ Metadata + Descriptions	0.6700	+8.1%

5.3. Validation vs. Test Performance Analysis

The validation results reveal important insights about component behavior across different dataset distributions. While the DINOv2 ensemble shows modest regression on the 200-sample validation subset (-3.1%), our analysis reveals that DINOv2 outperforms BioCLIP in 18% of individual cases (36 out of 200 validation samples), suggesting potentially stronger performance on the larger, more diverse test dataset. This discrepancy between validation and test performance highlights the importance of complementary feature representations in few-shot scenarios, where the limited validation subset may not fully capture the diversity present in the complete test distribution.

5.4. Competitive Analysis

Our system achieved 22nd place out of 74 teams, placing our approach within the top 30% of participants and above the competition median (0.4489), demonstrating the effectiveness of our multimodal approach while highlighting areas for improvement compared to top-performing methods.

Table 3
FungiCLEF 2025 competition results

Position	Team	Recall@5
1	Jack Etheredge	0.78913
2	hard_work	0.78137
3	aixiaodeyanjing	0.76584
15	Bodhisatta Maiti	0.61707
20	ChestnutKurisu	0.58732
22	**I2C-UHU-Pegasus (Ours)**	**0.57438**
25	creazy555	0.56274
50	mao_mao_chonga	0.33505

6. Error Analysis

Understanding the failure modes of our multimodal approach provides valuable insights into the fundamental challenges of automatic fungal species identification and guides future research directions. We examine both systematic taxonomic patterns and specific morphological factors that influence classification accuracy.

To understand the residual limitations of our approach, we conducted a detailed error analysis of the final pipeline using the validation set. This analysis focused on identifying persistent error patterns that highlight persistent challenges in automatic fungal species identification.

6.1. Most Problematic Species

Our analysis identified several species that were consistently misclassified (100% error rate), as presented in Table 4. Notably, all samples from species belonging to the genera *Diaporthe* and *Plagiostoma* were misclassified, indicating fundamental difficulties in distinguishing these taxa.

Table 4
Species with highest error rates in the validation set

Species (ID)	Genus	Family	Error	Error Top-5
2373	Uromyces	Pucciniaceae	1.00	1.00
722	Discostroma	Sporocadaceae	1.00	1.00
931	Golovinomyces	Erysiphaceae	1.00	0.93
1738	Plagiostoma	Gnomoniaceae	1.00	0.86
1739	Plagiostoma	Gnomoniaceae	1.00	1.00
1744	Pleomassaria	Pleomassariaceae	1.00	1.00

The species listed in Table 4 represent the most challenging cases for automatic classification, with perfect error rates indicating that current multimodal approaches struggle with these particular taxa, highlighting the utility of taxonomic boosting strategies for enhancing species-level classification.

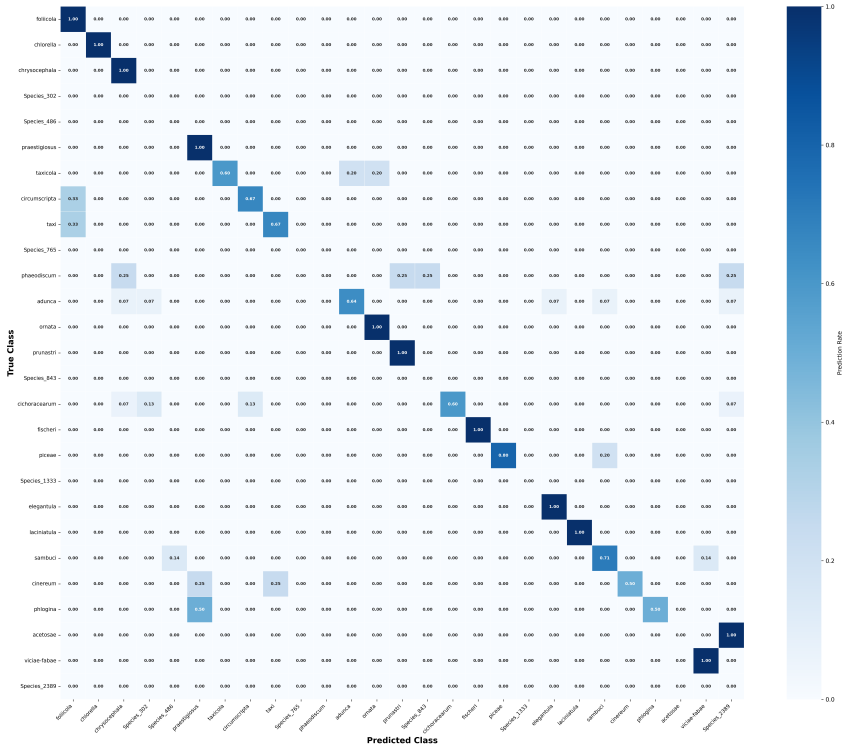

Figure 2: Confusion matrix of most frequently confused classes

Confusion patterns reveal frequent misclassifications between species within the same genus or family, as shown in Figure 2. This structure indicates that confusions tend to occur mainly between taxonomically related species, reflecting that the model learns to distinguish broad taxonomic groups but has difficulties differentiating species within the same clade.

6.2. Taxonomic Error Patterns

Error rates exhibit clear taxonomic patterns, with certain fungal groups consistently more difficult to classify, as shown in Table 5.

Table 5
Error rates by taxonomic level

Taxonomic Level	Highest Error Rate	Lowest Error Rate
Genus	*Diaporthe* (100%), *Plagiostoma* (100%)	*Puccinia* (44%), *Cortinarius* (47%)
Family	Diaporthaceae (100%), Gnomoniaceae (100%)	Russulaceae (30%), Parmeliaceae (30%)
Order	Sordariales (92%), Amphisphaeriales (88%)	Russulales (27%), Boletales (33%)

These patterns suggest that taxonomic relationships strongly influence model performance. As evidenced in Table 5, orders like Russulales and Boletales, which contain species with more distinctive morphological characteristics, are classified with much higher accuracy than orders containing species exhibiting subtle morphological distinctions.

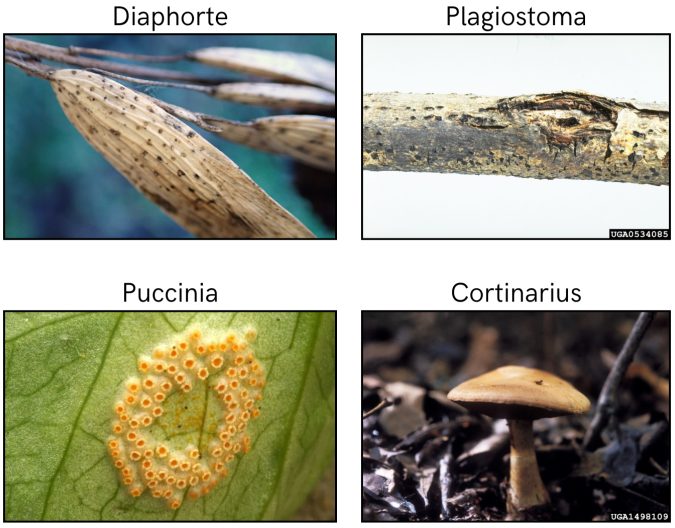

Figure 3: Contrasting classification performance across fungal genera. Top: Diaporthe and Plagiostoma showing microscopic features embedded in plant substrates. Bottom: Puccinia and Cortinarius with distinctive macroscopic morphology suitable for visual classification.

Figure 3 illustrates the contrasting classification performance across different fungal genera, highlighting the morphological factors that determine model success or failure. The top row shows the most challenging genera for automatic classification: Diaporthe and Plagiostoma species exhibit atypical fungal morphologies characterized by extremely small fruiting bodies (perithecia) embedded within plant substrates. These genera present minute, dark structures that are difficult to distinguish even for human experts, with microscopic diagnostic features and similar ecological preferences for woody substrates creating fundamental challenges for visual feature extraction.

In contrast, the bottom row demonstrates genera with significantly better classification performance. Puccinia species display distinctive orange circular patterns with high visual contrast against plant tissue, providing unique colorimetric and geometric features that facilitate accurate identification. Cortinarius species exemplify conventional mushroom morphology with clearly defined caps, stems, and gills that offer multiple discriminative visual characteristics. These morphological differences explain the dramatic performance gap: while Diaporthe and Plagiostoma achieve 100% error rates, Puccinia and Cortinarius maintain relatively low error rates of 44% and 47% respectively, demonstrating that our multimodal approach performs effectively when distinctive visual features are present at macroscopic scales.

6.3. Implications for Model Design

These findings influenced our model design in several ways: (1) incorporation of hierarchical taxonomic information to leverage better performance at higher taxonomic levels, (2) implementation of taxonomic boost mechanisms that improved accuracy for difficult groups, (3) specific boost techniques for the most problematic species, and (4) adaptive multimodal fusion giving more weight to discriminatory features in genera with high error rates. This error analysis provides valuable information about current system limitations and guides future improvements, especially in handling difficult genera like *Diaporthe* and *Plagiostoma*.

7. Conclusions

The multimodal approach developed in this work demonstrates the potential of integrating vision-language models with domain-specific knowledge for addressing challenging biodiversity classification tasks. Our results provide important insights for the broader application of AI systems in biological conservation and species identification.

The developed multimodal ensemble approach demonstrates reasonable performance in fungal classification by systematically integrating BioCLIP-specific fine-tuning, multimodal ensemble with DINOv2 [12], probabilistic ecological context modeling, and comprehensive textual description analysis, contributing valuable insights for multimodal learning in biodiversity applications.

The success of our approach highlights several key lessons for biological classification. Domain-specific pre-training provides crucial advantages over general models, although careful fine-tuning is still required to adapt these models to specific tasks. Multimodal integration significantly improves performance, particularly when systematically incorporating biological knowledge through ecological and taxonomic modeling.

Few-shot learning techniques prove essential for handling the long-tail distribution characteristic of biological datasets, with the combination of test-time augmentation, rare species boost, and ensemble methods providing robust performance across the full range of species in the dataset. However, significant challenges remain in distinguishing closely related species, particularly within genera like *Diaporthe* and *Plagiostoma*, indicating the need for more sophisticated approaches.

The methodology demonstrates that Vision-Language Models, when appropriately adapted through domain-specific fine-tuning and enriched with ecological, textual and taxonomic knowledge, can effectively address the challenge of rare species classification in biological domains. While our approach achieves reasonable performance compared to the median, the substantial gap with top-performing methods indicates significant opportunities for improvement.

Regarding computational considerations, our pipeline requires approximately 2.5 hours for training the complete system on a single A100 GPU, with inference time of approximately 0.8 seconds per sample when processing the multimodal ensemble. While this computational overhead is manageable for research applications, practical deployment would benefit from model compression and optimization techniques to reduce inference time and resource requirements.

Future work should focus on several key areas: (1) expanding the ecological context modeling to include more detailed habitat relationships and seasonal patterns, (2) exploring advanced few-shot learning techniques specifically designed for extreme class imbalance scenarios, (3) investigating the integration of phylogenetic information to better distinguish closely related species, and (4) developing more efficient architectures that maintain performance while reducing computational requirements for practical deployment. Additionally, the integration of citizen science data and active learning approaches could help address data scarcity for rare species.

The source code and trained models are available at https://github.com/cgarciafernando/fungiclef-2025-tfg to facilitate reproducibility and enable further research in multimodal biodiversity informatics.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude (Anthropic) to improve writing style, translate text, and generate scripts for creating tables and visualizations based on the authors' experimental pipeline code and results. After using these tools, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] D. L. Hawksworth, R. Lücking, Fungal diversity revisited: 2.2 to 3.8 million species, Microbiology Spectrum (2017).

[2] L. Picek, K. Janoušková, V. Cermak, J. Matas, Fungitastic: A multi-modal dataset and benchmark for image categorization, arXiv preprint arXiv:2408.13632 (2025). URL: https://arxiv.org/abs/2408.13632.

[3] K. Janouskova, J. Matas, L. Picek, Overview of FungiCLEF 2025: Few-shot classification with rare fungi species, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, 2025.

[4] L. Picek, 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.

[5] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature (2015).

[6] J. Snell, K. Swersky, R. Zemel, Prototypical networks for few-shot learning, in: Advances in Neural Information Processing Systems, 2017.

[7] C. Finn, P. Abbeel, S. Levine, Model-agnostic meta-learning for fast adaptation of deep networks, in: Proceedings of the 34th International Conference on Machine Learning, 2017.

[8] Z. Xiong, Y. Ruan, Y. Hu, Y. Zhang, Y. Zhu, S. Guo, B. Han, 1st place solution for fungiclef 2022 competition: Fine-grained open-set fungi recognition, in: CLEF Working Notes, 2022.

[9] F. Hu, P. Wang, Y. Li, C. Duan, Z. Zhu, Y. Li, X.-S. Wei, A deep learning based solution to fungiclef2023, in: CLEF Working Notes, 2023.

[10] K. Desingu, et al., Fungiclef: Deep-learning for the visual classification of fungi species using network ensembles, in: CLEF Working Notes, 2022.

[11] C. Chiu, M. Heil, T. Kim, A. Miyaguchi, Fine-grained classification for poisonous fungi identification with transfer learning, arXiv preprint arXiv:2407.07492 (2024). URL: https://arxiv.org/abs/2407.07492.

[12] 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). URL: https://arxiv.org/abs/2304.07193.

[13] 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, W.-L. Chao, Y. Su, Bioclip: A vision foundation model for the tree of life, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.

[14] Y. A. Malkov, D. A. Yashunin, Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs, IEEE transactions on pattern analysis and machine intelligence 42 (2018) 824–836.

[15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.

[16] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2017.

[17] S. F. Chen, J. Goodman, Good-turing frequency estimation for feature selection, Computer Speech & Language 13 (2006) 61–77.
[18] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2019) 535–547.