Explainable Vision Transformers for Candida Species Identification from Brightfield Microscopy

Rodrigo Sá*, Luís H. M. Torres*, Catarina Pimentel† and Bernardete Ribeiro*
*Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal
Emails: rodrigosa@student.dei.uc.pt, luistorres@dei.uc.pt, bribeiro@dei.uc.pt
†Yeast Molecular Biology Lab, Nova University Lisbon, Lisbon, Portugal
Email: pimentel@itqb.unl.pt

***Abstract*—The high mortality of invasive fungal infections necessitates rapid, automated differentiation between *Candida albicans* and *Candida glabrata* to overcome the bottlenecks of manual microscopy. We propose a classification framework using the SimpleViT architecture, trained on a novel dataset of 3,731 brightfield microscopy patches derived from 60 slide preparations. To prevent data leakage, we enforced a strict biological partition between training and testing replicates. Results show that aggregating predictions from the patch level (84.39%) to the slide level significantly improves decision stability, achieving 92.86% balanced accuracy on a held-out test set. A rigorous explainable (XAI) analysis using Uniform Manifold Approximation and Projection (UMAP) and Attention Rollout exposes two diverging error regimes. In contrast to the entropic signal loss characterizing false negatives, the predominant false positives exhibit the intense attentional focality of correct predictions. This indicates the model suffers not from stochastic noise, but from confident misidentification of ambiguous yeast-like textures. We conclude that while global attention effectively localizes cellular targets, resolving fine-grained phenotypic mimicry requires architectural enhancements for stronger local feature modeling.**

***Index Terms*—Fungal Infections, Vision Transformers, Candida Species, Explainable AI, Computational Biology, Deep Learning.**

I. Introduction

THE epidemiology of invasive fungal infections (IFIs) presents a critical clinical challenge, with mortality burdens exceeding 1.5 million cases per year [1]. Within the nosocomial landscape, the World Health Organization designates *Candida* as a priority pathogen. A critical diagnostic bottleneck lies in differentiating the ubiquitous *Candida albicans* from non-albicans species, specifically *Candida glabrata*.

Taxonomic discrimination is confounded by the disparate phenotypic traits of both species. Whereas *C. albicans* leverages phenotypic plasticity, transitioning between yeast and hyphal forms, to facilitate tissue invasion, *C. glabrata* is a strictly blastoconidial, haploid yeast lacking this dimorphic switching capacity. Consequently, accurate identification is therapeutically decisive, as *C. glabrata* frequently exhibits intrinsic or acquired resistance to azole-based antifungals, necessitating divergent treatment protocols [2].

Current diagnostic standards remain suboptimal. Blood cultures suffer from diagnostic inefficiency, characterized by delayed turnaround times and low sensitivity ($\approx$40%) for invasive candidiasis [1]. Conversely, direct microscopic examination reduces latency but is hindered by inter-observer variability due to morphological ambiguities, specifically subtle variations in blastospore size and budding patterns. Similarly, advanced proteomic methods like MALDI-TOF MS face operational constraints in mold identification by the physical embedding of hyphae within the culture medium, leading to subsequent spectral quality degradation [3].

Deep Learning (DL) has transformed computational biology, with Convolutional Neural Networks (CNNs) emerging as the de facto standard. However, CNNs are architecturally constrained by an inductive bias towards local spatial processing [4]; the receptive field of a neuron is restricted to a local sub-region, expanding only as data propagates deeper into the network hierarchy. This limitation is critical in fungal cytology, where diagnostic features such as dispersed spores or hyphal fragments are often scattered without immediate adjacency. By processing images as patch sequences, Vision Transformers (ViTs) inherently maintain a global contextual scope, allowing the modeling of long-range dependencies [5]. Despite this potential, clinical adoption is stalled by the scarcity of high-quality, non-leaky datasets and the opaque decision-making of deep networks.

This paper addresses three core research questions (RQs) by testing their corresponding hypotheses. We hypothesize that: (RQ1) a strictly curated dataset minimizes background artifacts to improve model robustness; (RQ2) SimpleViT's achieves high accuracy with global attention capturing long-range morphological features essential for fungal differentiation; and (RQ3) the model results show localized attention explained by biologically relevant patterns. These are evaluated through three key contributions:

1) **Dataset Curation:** Presentation of a pipeline for the acquisition and processing of fungal microscopy images.
2) **SimpleViT Evaluation:** The implementation of a Vision Transformer [6] to capture morphological intricacies and achieve high performance in fungal image classification.
3) **Explainability Analysis:** Explainable AI (XAI) techniques, including Attention Rollout and latent space topology analysis (UMAP) to verify that the model's predictions are driven by local features.

To facilitate further research and reproducibility, the source code and the curated dataset are publicly available.[1]

[1]https://github.com/Rodrigo2003-PT/simpleViT_fungi

II. Related Work

The automation of fungal diagnosis relies heavily on the **DeFungi** benchmark [7], constructed from direct examination of KOH-stained samples. While initial transfer learning baselines yielded ≈85% accuracy [7], subsequent studies utilizing dense architectures have reported metrics exceeding 90% [8]–[10]. However, these metrics are frequently confounded by inconsistent protocols. The risk of such validation-dependent selection was explicitly quantified by Sultana et al. [11], who observed a severe generalization gap in EfficientNetB0 from 95.82% (validation) to 42.07% (test).

Regarding architecture, Vision Transformers (ViTs) theoretically benefit fungal classification by capturing global morphological dependencies, yet they struggle to generalize in data-scarce regimes. Standalone transformer baselines frequently underperform relative to hybrid architectures that inject convolutional inductive biases or require aggressive synthetic augmentation [12]. Ahmed and Haque [13] observed that a standalone ViT achieved only ≈83% accuracy on DeFungi, necessitating a hybrid ViT-ResNet50 architecture to reach competitive metrics (90.13%). Similarly, hierarchical designs like Swin Transformers have demonstrated high sensitivity to data volume, with Gümüş [14] reporting a performance leap from 82.3% to 98.36% only after heavy augmentation. While recent hybrid models have achieved 95.57% accuracy (Swin-DenseNet121) under strict 5-fold cross-validation [15], the prevalence of disparate balancing strategies across these studies continues to confound comparative analysis.

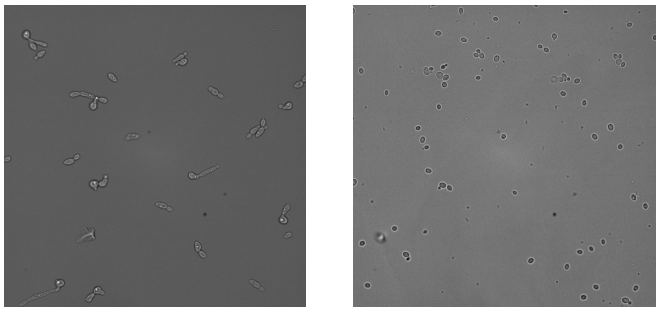
High classification metrics do not guarantee biological validity, as deep models may exploit non-causal shortcuts such as staining artifacts [16]. To mitigate this, segmentation-guided attribution has been used to verify model focus on cellular structures [17], while clinical trials confirm that visual explanations (e.g., Grad-CAM) significantly reduce false positives by enabling human verification of errors [18]. Furthermore, aggregation techniques utilizing superpixels have been proposed to enhance the perceptual fidelity of these pixel-wise attribution maps [19].

III. Dataset Curation

A contribution of this work is the curation of a high-fidelity microscopic dataset designed explicitly to overcome the limitations of existing benchmarks described in Section II. The dataset construction enforces strict biological isolation to ensure that subsequent model evaluations reflect genuine generalization rather than artifact memorization.

A. Acquisition Protocol

Data from reference strains of *Candida albicans* and *Candida glabrata* (ATCC) were acquired in collaboration with the Yeast Molecular Biology Lab (Nova University Lisbon). Cultures were grown in Synthetic Complete (SC) medium (pH 6.5) at 35°C and standardized to an optical density (OD600) of 3.0. Despite equal biomass, morphological differences yield distinct cellular concentrations: 3.6×10^7 cells/mL for *C. albicans* and 6×10^7 cells/mL for *C. glabrata*.



(a) *C. albicans* (b) *C. glabrata*

Fig. 1. Raw brightfield microscopy acquisitions (1200×1200 px). Images are displayed with standard contrast stretching. (Pixel spacing: $0.175\,\mu$m).

To prepare for imaging, 5 µL aliquots were mounted using a 1.7% agarose pad. This establishes the physical slide as the primary *independent biological replicate*, effectively encapsulating inter-sample confounders like agar thickness and refractive index variations. Conversely, each Field of View (FOV) serves as a *dependent observational unit*.

Images were acquired via brightfield illumination (10% LED) on a Zeiss Axio Observer with a Plan-Apochromat 63x/1.4 Oil objective. A Prime 95B monochrome camera (12-bit sensor) captured random FOVs (62 images/slide) with a fixed focus per slide (pixel spacing: 0.175 µm). The higher inter-slide variance in *C. albicans* reflects its lower cellular concentration and phenotypic plasticity, representing expected clinical heterogeneity rather than sampling artifacts.

B. Statistics

The dataset comprises $N = 60$ unique biological slides yielding 3,731 microscopic patches. As detailed in Table I, the partitioning preserves a held-out test cohort of 20% of the biological replicates. The acquisition density reflects real-world variability, with *C. albicans* averaging 62.3 ± 35.2 images/slide and *C. glabrata* 62.0 ± 15.7 images/slide.

TABLE I
Dataset Distribution by Class and Partition

Class	Slides	Train Patches	Test Patches	Total
C. albicans	33	1616	440	2056
C. glabrata	27	1294	381	1675
Total	**60**	**2910**	**821**	**3731**

C. Preprocessing

To prevent inflated metrics caused by intra-slide correlation [20], all images originating from a single biological replicate (Slide ID) were assigned exclusively to either the training or testing partition. Class balance was optimized via a randomized search ($s = 42$) minimizing the KL-divergence between slide assignments ($\delta < 0.02$). Raw 12-bit sensor data ($I_{raw} \in [0, 4095]$) was linearly mapped to $[0, 1]$ and resampled

to 384×384 pixels using Lanczos interpolation to preserve high-frequency morphological details [21].

IV. Proposed Model

We adopt a streamlined Vision Transformer (SimpleViT) [6] optimized for data-efficient binary classification. This architecture refines the standard ViT [5] by replacing learnable positional embeddings and class tokens with fixed spatial encodings and global pooling.

A. Architecture Configuration

The input microscopy images, denoted as $x \in \mathbb{R}^{H \times W \times C}$, are processed at a resolution of $H = W = 384$ pixels with a single grayscale channel ($C = 1$). The image is serialized into a sequence of non-overlapping patches of size $P \times P$ (where $P = 16$), resulting in a sequence length of $N = 576$ tokens.

The implementation followed the *vit_pytorch* instantiation of the SimpleViT architecture [22]. This variant injects spatial awareness by using *fixed 2D sinusoidal positional embeddings* (E_{sincos}), computed analytically from the patch coordinate grid [6]. The patch projection layer employs a *dual-normalization* strategy to bound signal variance at initialization:

$$z_0 = \text{LN}(\text{Linear}(\text{LN}(x_{patches}))) + E_{sincos} \quad (1)$$

The encoder is compact ($L = 8$ layers, $D = 256$), utilizing 6 parallel attention heads and a Feed-Forward expansion ratio of 2 ($D_{MLP} = 512$).

Deviating from the canonical ViT architecture [5], the model eliminates the learnable class token (x_{class}) by implementing *Global Average Pooling (GAP)*, computing the mean vector of the output patch sequence z_L. The final class logits y are computed via $y = \text{Linear}(\text{Mean}(z_L))$, where z_L denotes the encoder output. This mechanism forces the model to distribute semantic attention across the patch sequence, thereby facilitating the extraction of high-fidelity attention maps.

V. Experimental Setup and Results

This section details the training protocol and provides a rigorous assessment of the model's diagnostic capabilities, focusing on generalization stability and error analysis.

A. Implementation Details

To guarantee biological independence, the training cohort (S_{train}) was partitioned via a stratified $k = 5$ Repeated Cross-Validation scheme, ensuring *atomic slide separation* between folds. Reproducibility was guaranteed via a hierarchical deterministic seeding protocol ($S_{base} = 42$). To ensure representative sampling, partition seeds were algorithmically selected to enforce class stratification across k folds, thus eliminating variance arising from random data splitting. Input normalization statistics (μ_j, σ_j) were computed strictly within training folds to prevent statistical leakage.

Models were optimized using AdamW (learning rate 3×10^{-4}, weight decay 0.05) with a cosine decay schedule over 100 epochs (batch size 64, 5 warmup epochs). To enforce morphological invariance, we applied random horizontal/vertical flips, rotations ($\theta \in [-30°, 30°]$), and cropping (scale $\in [0.9, 1.0]$). Following SimpleViT principles [6], explicit dropout was replaced by conservative MixUp regularization ($\alpha = 0.2, p = 0.5$) [23]. This low-alpha regime smooths the decision boundary without obscuring fine morphological details or highlighting underrepresented artifacts. In addition, literature shows that this augmentation strategy also effectively prevents overfitting in data-scarce biological datasets.

Model selection followed a lexicographic protocol prioritizing slide-level Balanced Accuracy (BA), with validation loss serving as a tie-breaker. To filter transient fluctuations, the stopping condition relied on an Exponential Moving Average (smoothing $\alpha = 0.3$) of the primary metric signal, with a patience counter reset every time the smoothed signal exceeds the minimum threshold ($S_t > S_{best} + \delta_{min}$). The final deployment model was retrained on the full S_{train} cohort for a duration determined by the median optimal epoch across folds.

B. Cross Validation

1) Optimization Dynamics: Given the heterogeneous stopping epochs (T_f) induced by early stopping, validation trajectories were synchronized using Last Observation Carried Forward (LOCF) imputation. As shown in Fig. 2, the initial optimization phase (epochs 0–28) induced rapid training loss reduction. Although the generalization gap ($\Delta\mathcal{L}_t = \mathcal{L}_{val,t} - \mathcal{L}_{train,t}$) peaked at $\Delta\mathcal{L}_{max} \approx 0.23$, it subsequently compressed to ≈ 0.11 as validation loss stabilized ($\mathcal{L}_{val} \approx 0.43$). The low inter-fold variability ($\pm 1\sigma$, shaded) confirms consistent convergence without overfitting.

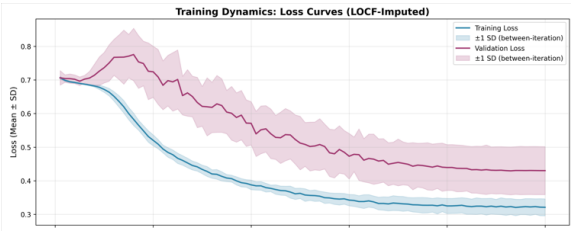


Fig. 2. Aggregate learning dynamics ($N = 30$; Shading: $\pm 1\sigma$).

2) Quantitative Performance: Table II details the diagnostic performance, with statistics derived from per-fold best-epoch summaries and aggregated via intra-iteration averages to account for data overlap [24]. Uncertainty is quantified via descriptive 95% confidence intervals (t_{N-1}).

Transitioning from patch-level to slide-level diagnosis yields consistent improvements. Notably, the *Matthews Correlation Coefficient (MCC)* exhibits a 26.8% gain ($0.656 \rightarrow 0.832$) upon aggregation. Since MCC is highly sensitive to confusion matrix symmetry, this sharp increase indicates that mean-pooling effectively attenuates local textural noise/ambiguity, resulting in a robust specimen-level signal. The stability of this process is further evidenced by a low Coefficient of Variation ($CV \approx 4.0\%$) across scales.

TABLE II
CROSS-VALIDATION PERFORMANCE (MEAN ± STD) [95% CI]

Metric	Patch-Level	Slide-Level
Balanced Acc.	0.8287 ± 0.0330 $[0.816, 0.841]$	$\mathbf{0.9117 \pm 0.0374}$ $[0.898, 0.926]$
MCC	0.6562 ± 0.0666 $[0.631, 0.681]$	$\mathbf{0.8319 \pm 0.0715}$ $[0.805, 0.859]$
F1-Score	0.8249 ± 0.0350 $[0.812, 0.838]$	$\mathbf{0.8992 \pm 0.0490}$ $[0.881, 0.918]$
AUC	0.9067 ± 0.0359 $[0.893, 0.920]$	$\mathbf{0.9618 \pm 0.0340}$ $[0.949, 0.975]$

TABLE III
PERFORMANCE METRICS ON HELD-OUT TEST SET (95% CI)

Metric	Patch-Level	Slide-Level
Bal. Acc.	0.8439 $[0.745, 0.897]$	**0.9286** $[0.786, 1.000]$
MCC	0.7025 $[0.516, 0.816]$	**0.8452** $[0.529, 1.000]$
F1 (W)	0.8326 $[0.729, 0.926]$	**0.9172** $[0.733, 1.000]$
AUC	0.9422 $[0.847, 0.985]$	**0.9714** $[0.852, 1.000]$

C. Held-Out Test Set Evaluation

Generalization was assessed on an independent cohort comprising 821 patches across 12 slides (7 *C. albicans*, 5 *C. glabrata*).

To account for the limited sample size ($N_{slides} = 12$) and quantify uncertainty, we computed 95% Confidence Intervals (CI) via non-parametric bootstrapping ($B = 1000$). We utilized cluster-robust resampling for patches to preserve intra-slide correlations [25], and grouped singleton resampling for slide-level predictions. Replicates resulting in mono-class distributions were excluded to ensure metric calculability.

1) Performance and Error Analysis: As shown in Table III, the model achieves a patch-level Balanced Accuracy of 0.8439. However, the substantial width of the confidence intervals underscores the impact of *inter-slide variability* on patch-level estimation. The mean-pooling aggregation successfully filters local noise, yielding consistent improvements across all metrics; notably, the slide-level Balanced Accuracy reaches $\mathbf{0.9286}$. The upper CI bounds frequently saturate at 1.0, a consequence of the discrete resampling in small cohorts, potentially omitting hard examples.

A granular analysis of the decision boundary (Table IV) reveals a sensitivity-specificity trade-off. At the patch level, the model prioritizes sensitivity for the target class *C. glabrata* (Recall 0.976), missing only 9 instances. However, this incurs a 28.9% False Positive Rate (127 misclassified *C. albicans* patches), causing a sharp precision disparity. While *C. albicans* predictions are highly reliable (Precision 0.972), *C. glabrata* precision drops to 0.745 due to noise contamination.

The slide-level probabilistic aggregation effectively neutralizes this local specificity deficit. As shown in Table Vb, the accumulation of evidence suppresses sporadic patch errors provided they remain the minority. Consequently, the system achieves 100% detection of *C. glabrata* slides and misclassifies only a single *C. albicans* specimen. This confirms that the model's high local sensitivity translates into effective specimen-level diagnosis, acting as a reliable screening filter.

TABLE IV
CONFUSION MATRIX ANALYSIS

		Predicted	
		Alb	*Gla*
Actual	*Alb*	**313**	127
	Gla	9	**372**

(a) Patch-Level ($N = 821$)

		Predicted	
		Alb	*Gla*
Actual	*Alb*	**6**	1
	Gla	0	**5**

(b) Slide-Level ($N = 12$)

VI. EXPLAINABLE ANALYSIS

To investigate the precision deficit observed in the *C. glabrata* class (Section V-C), we conducted a forensic analysis of the model's internal representations. This section tests the hypothesis that the architecture induces a systematic bias, mapping ambiguous features from the decision boundary into the target class.

A. Latent Manifold Topology

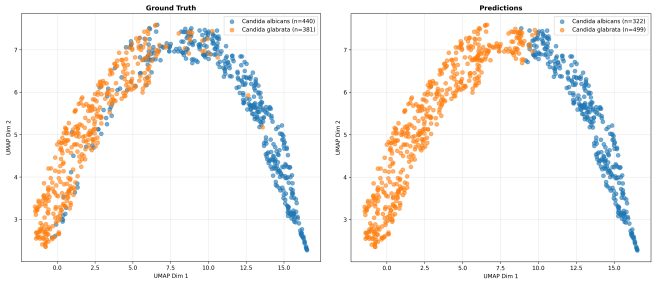
We extracted high-dimensional feature embeddings ($\mathbf{z} \in \mathbb{R}^D$) from the penultimate layer and projected them into $\mathbb{R}^2$ space using UMAP [26] ($N_{neighbors} = 15, min\_dist = 0.1$).

The resulting geometry (Fig. 3) is characterized by a "curvilinear gradient" where biological traits blend seamlessly, rather than segregating into disjoint clusters. The preservation of the structure of the local neighborhood is confirmed by a high Trustworthiness score ($T(k = 5) \approx 0.97$).

Geometrically, the manifold is bipartite: the distal tails contain high-confidence canonical samples, while the central apex represents a zone of feature intersection. Comparing the Ground Truth (Fig. 3a) with the Prediction Manifold (Fig. 3b) reveals the model deals with the inherent biological entanglement at the apex by imposing a monolithic assignment to *C. glabrata*.

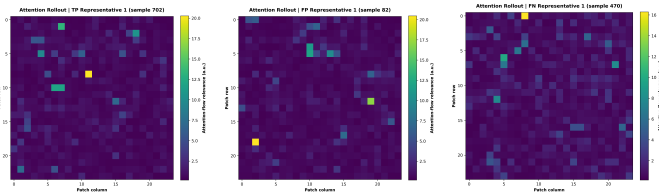
Quantitatively, the model imposes a *topological contraction*, evidenced by a predicted compactness score of 0.62 relative to the ground truth baseline of 0.47. This geometric compression suggests the network minimizes decision uncertainty by absorbing the transitional manifold region into the positive class, implicitly biasing the decision boundary toward a higher recall.

To verify that these errors are systematic rather than stochastic, we performed a Class-Stratified Permutation Test ($B = 1000$ iterations). The analysis rejected the null hypothesis of independence ($p < 0.001$), revealing that the observed error variance across slides ($\sigma^2_{obs} \approx 0.069$) is $3.5\times$ higher than the random expectation, indicating that errors are not uniformly distributed but structurally bound to specific biological samples.



(a) Ground Truth (b) Predicted Labels

Fig. 3. Latent Manifold Topology. (a) Ground Truth labels reveal a central "confounded region" where classes naturally overlap. (b) Predicted Labels show the model artificially resolving this ambiguity by assigning the entire region to *C. glabrata*, creating a denser but less accurate cluster.



(a) True Positive (b) False Positive (c) False Negative

Fig. 4. Attention Rollout Topology. (a) TP and (b) FP exhibit similar high-focality "hotspots," suggesting the model hallucinates salient features in negative samples (mimicry). (c) FN displays diffuse entropy, indicating signal failure.

B. Attention Dynamics

To elucidate intra-sample feature attribution, we adapted Attention Rollout [27] for the global average pooling architecture, defining patch relevance as the total accumulated flow across all query positions. The propagation algorithm incorporates residual signal preservation ($0.5\mathbf{A} + 0.5\mathbf{I}$) and a heuristic pruning of the bottom 90% of attention links. Maps were quantified via Gini Coefficient (G, measuring sparsity) and Spatial Entropy (H, measuring diffuseness). To ensure representativeness, visualizations depict the rank-1 exemplar minimizing the Euclidean distance to the group's median Gini score.

The topological analysis identifies two distinct regimes:

1) **Morphological Mimicry:** False Positives defy the expectation of entropic confusion, exhibiting high mean focality ($\mu_G \approx 0.460$) functionally equivalent to True Positives ($\mu_G \approx 0.465$). Although a Welch's t-test yields $p \approx 0.027$, the negligible effect size ($d \approx 0.22$) confirms that the model treats these samples as confident positives. This suggests the network hallucinates salient features by over-indexing on "yeast-like" local textures within *C. albicans*, failing to integrate the broader absence of hyphal structures. However, definitive validation of this causal alignment necessitates granular region-level annotations [28] to strictly map attention hotspots to ground-truth biological semantics.
2) **Signal Attenuation:** Conversely, False Negatives diverge via high spatial entropy ($\mu_H \approx 5.94$). Despite the limited sample size ($N = 9$), the topology reveals fragmented attention masses embedded in noise. This dispersion supports the hypothesis that Global Average Pooling acts as a bottleneck, attenuating the relevant signal when the ratio of target-to-background variance falls below a learnable threshold.

Summary: We conclude that the primary failure mode is not stochastic noise, but a structural over-interpretation of ambiguous textures. The model maintains high precision on canonical samples but minimizes entropy in the manifold's "confounded region" by forcing uncertain phenotypes into the positive class, effectively prioritizing recall over granular separability.

VII. Conclusion and Future Work

This study establishes a rigorous baseline for AI-driven fungal diagnostics by introducing a novel dataset grounded in real clinical laboratory regimes. The results validate that *slide-level probabilistic aggregation* effectively neutralizes sporadic patch level errors, verifying that clinical diagnosis benefits from accumulating evidence across multiple fields of view. However, the explainable analysis highlights that standard Vision Transformers struggle with dispersed attention and fail to separate distinct morphologies in ambiguous latent regions. These findings indicate that purely global attention mechanisms may be insufficient for fine-grained fungal taxonomy; consequently, comparative evaluations with models that inherently capture local texture (e.g., DenseNet121) will be actively pursued.

Furthermore, the current dataset size is naturally constrained by the resource-intensive biological acquisition protocol. Expanding this cohort to further validate generalization remains a primary objective. Subsequent research will also explore hybrid architectures that restore local inductive biases within the global transformer model. For instance, replacing linear flattening with convolutional tokenization [29] is hypothesized to better preserve high-frequency textural cues. Furthermore, locality-constrained mechanisms (e.g., LSA) [30] present a viable strategy to counteract attention entropy by penalizing self-token dominance. Complementing this, spatial input augmentation (e.g., Shifted Patch Tokenization) [30] may prevent the discretization of continuous hyphal structures across grid partitions. Collectively, these architectural changes could improve the model's understanding of biological features, increasing the reliability of clinically verifiable decision support.

Clinically, the model's bias toward high sensitivity (recall) for azole-resistant *C. glabrata* supports its utility as a critical safety screening tool. In this context, the risk of therapeutic failure from a false negative (misclassifying *C. glabrata* as *C. albicans*) far outweighs the operational burden of verifying false positives (flagging *C. albicans* as *C. glabrata*). Consequently, the system functions as a risk-averse triage mechanism, prioritizing the interception of high-risk pathogens for expert confirmation.

Funding

This work is funded by national funds through the Portuguese Recovery and Resilience Plan (PRR) through project C645008882-00000055, Center for Responsible AI.

Acknowledgments

This work was supported by the Center for Responsible AI, and also by the FCT - Foundation for Science and Technology, I.P., within the scope of the research unit UID/00326 - Centre for Informatics and Systems of the University of Coimbra, and through the Research Grant with Ref. 2024.01388.BD.

Generative AI tools (Gemini) and automated grammar checkers (Grammarly) were used solely for linguistic refinement and editorial assistance. All scientific content, data analysis, and conclusions remain the full responsibility of the authors.

References

[1] D. W. Denning, "Global incidence and mortality of severe fungal disease," *The Lancet Infectious Diseases*, vol. 24, no. 7, pp. e428–e438, 2024. [Online]. Available: https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(23)00692-8/fulltext

[2] M. Katsipoulaki, M. H. T. Stappers, D. Malavia-Jones, S. Brunke, B. Hube, and N. A. R. Gow, "Candida albicans and Candida glabrata: global priority pathogens," *Microbiology and Molecular Biology Reviews*, vol. 88, no. 2, pp. e00 021–23, 2024. [Online]. Available: https://journals.asm.org/doi/10.1128/mmbr.00021-23

[3] J. Pemán and A. Ruiz-Gaitán, "Diagnosing invasive fungal infections in the laboratory today: It's all good news?" *Revista Iberoamericana de Micología*, vol. 42, pp. 1–14, 2025.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[5] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[6] L. Beyer, X. Zhai, and A. Kolesnikov, "Better plain ViT baselines for ImageNet-1k," *arXiv preprint arXiv:2205.01580*, 2022. [Online]. Available: https://arxiv.org/abs/2205.01580

[7] C. J. P. Sopo, S. Gheisari, and F. Hajati, "DeFungi: Direct mycological examination of microscopic fungi images," *arXiv preprint arXiv:2109.07322*, 2021. [Online]. Available: https://arxiv.org/abs/2109.07322

[8] J. Bhimavarapu and S. V. Movva, "Fungi classification: Enhancing diagnosis using deep learning," in *2nd World Conference on Communication & Computing (WCONF)*, Raipur, India, 2024.

[9] S. Asadi Amiri and S. F. Mohammady, "VGG19-DeFungi: A novel approach for direct fungal infection detection using VGG19 and microscopic images," *Journal of Artificial Intelligence and Data Mining (JAIDM)*, vol. 12, no. 2, 2024.

[10] H. Mohamed, A. E. Nagib, and M. Hany, "Comparative study of microscopic fungii classification using transfer learning models," in *2024 Intelligent Methods, Systems, and Applications (IMSA)*, Cairo, Egypt, 2024.

[11] S. Sultana, S. B. S. Omit, M. H. Imam, M. S. Hossain, N. Nahar, and A. Hanip, "Microscopy-based fungi species classification using a deep learning ensemble approach," in *2025 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, Chattogram, Bangladesh, 2025.

[12] M. G. Anggara, A. Hindasyah, J. M. Amda, and A. A. Waskita, "Fine tuning Swin transformer based pretrained model for microscopic fungi images classification," in *2024 International Conference on Computer, Control, Informatics and its Applications (IC3INA)*, Pamulang, Indonesia, 2024.

[13] S. I. Ahmed and A. H. M. O. Haque, "Microscopic fungi classification using vision transformer guided by transfer learning approach," in *26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2023.

[14] A. Gümüş, "Classification of microscopic fungi images using vision transformers for enhanced detection of fungal infections," *Turkish Journal of Nature and Science (TJNS)*, vol. 13, no. 1, 2024.

[15] P. Seth *et al.*, "Hybrid deep learning models for microscopic fungi image classification," *Preprint/Recent Literature*, 2024.

[16] A. Singh, S. Sengupta, and V. Lakshminarayanan, "Explainable deep learning models in medical image analysis," *Journal of Imaging*, vol. 6, no. 6, p. 52, 2020.

[17] J. D. Guthrie, S. A. Shankarnarayan, and D. A. Charlebois, "Peering inside the black box: Explainable AI to interpret advanced computer vision fungal pathogen prediction," *bioRxiv*, 2025.

[18] F. Xu, L. Jiang, W. He, G. Huang, Y. Hong, F. Tang, J. Lv, Y. Lin, Y. Qin, R. Lan, X. Pan, S. Zeng, M. Li, Q. Chen, and N. Tang, "The clinical value of explainable deep learning for diagnosing fungal keratitis using in vivo confocal microscopy images," *Frontiers in Medicine*, vol. 8, p. 797616, 2021.

[19] A. Kallipolitis, P. Yfantis, and I. Maglogiannis, "Improving explainability results of convolutional neural networks in microscopy images," *Neural Computing and Applications*, vol. 35, pp. 21 535–21 553, 2023.

[20] N. Bussola, V. Maggio, A. Marcolini, G. Jurman, and C. Furlanello, "AI slipping on tiles: data leakage in digital pathology," *arXiv preprint arXiv:1909.06539*, 2020. [Online]. Available: https://arxiv.org/abs/1909.06539

[21] C. E. Duchon, "Lanczos filtering in one and two dimensions," *Journal of Applied Meteorology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[22] P. Wang, "vit-pytorch: Implementation of vision transformers in pytorch," https://github.com/lucidrains/vit-pytorch, 2020, accessed: 2026-01-27.

[23] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations (ICLR)*, 2018. [Online]. Available: https://openreview.net/forum?id=r1Ddp1-Rb

[24] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," *Journal of Machine Learning Research*, vol. 5, pp. 1089–1105, 2003. [Online]. Available: https://www.jmlr.org/papers/volume5/grandvalet04a/grandvalet04a.pdf

[25] F. L. Huang, "Using cluster bootstrapping to analyze nested data with a few clusters," *Educational and Psychological Measurement*, vol. 78, no. 2, pp. 297–318, 2018.

[26] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2020, version 3. [Online]. Available: https://arxiv.org/abs/1802.03426

[27] S. Abnar and W. Zuidema, "Quantifying attention flow in transformers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4190–4197. [Online]. Available: https://arxiv.org/abs/2005.00928

[28] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 782–791. [Online]. Available: https://arxiv.org/abs/2012.09838

[29] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," *arXiv preprint arXiv:2104.05704*, 2021. [Online]. Available: https://arxiv.org/abs/2104.05704

[30] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021. [Online]. Available: https://arxiv.org/abs/2112.13492