Understanding Data Augmentation Saturation: A Comparative Study of CNNs and Vision Transformers for Pollen Grain Classification

Júlio César da Silva Soares, Kelson Romulo Teixeira Aires, Juliana do Nascimento Bendini, Luis Felipe Ferreira de Castro, Rodrigo de Melo Souza Veras

[1]Universidade Federal do Piauí (UFPI)

jczars@ufpi.edu.br, kelson@ufpi.edu.br, jbendini@ufpi.edu.br,

luis.castro@ufpi.edu.br, rveras@ufpi.edu.br

Abstract. *Automatic pollen grain classification has advanced with deep learning, but how architecture interacts with progressive data augmentation, especially regarding performance saturation, remains underexplored. We compare DenseNet201, DenseNet201+CBAM, and ViT-B16 on the Cretan Pollen Dataset (CPD1) under progressive augmentation (0–700 synthetic images/class, 10-fold cross-validation). Results show architecture-specific saturation at factors 200 (ViT-B16), 400 (DenseNet201), and 500 (DenseNet201+CBAM), with CNN-based models reaching a peak accuracy of 0.987. CBAM improves stability and delays saturation, offering practical guidance for data-efficient architecture selection in specialized domains such as palynology.*

1. Introduction

Melissopalynology relies on the microscopic analysis of pollen grains to determine the botanical origin of honey. This process is essential for honey quality certification and commercial valuation but remains time-consuming, dependent on specialized expertise, and susceptible to inter-observer variability [Flenley 1968, Tsiknakis et al. 2022, Sevillano and Aznarte 2018].

Recent advances in deep learning have significantly improved automated pollen classification. CNN-based models have achieved high accuracy on benchmark datasets [Astolfi et al. 2020, Tsiknakis et al. 2022], while attention-based approaches [Mahmood et al. 2023] and transformer architectures [Dosovitskiy et al. 2020, Zu et al. 2025] have further expanded the state of the art. However, how these paradigms respond to progressive synthetic data expansion remains largely unexplored.

Data augmentation is widely used to mitigate overfitting in data-constrained domains. However, most studies implicitly assume that increasing the number of synthetic samples consistently improves performance. The existence of architecture-dependent saturation points, where additional synthetic data ceases to provide meaningful gains or even degrades performance, remains poorly understood in palynology.

This work investigates the saturation behavior of three representative architectures: DenseNet201, DenseNet201 enhanced with the Convolutional Block Attention Module (CBAM) [Woo et al. 2018], and ViT-B16. Using the Cretan Pollen Dataset

(CPD1) [Tsiknakis et al. 2022], we evaluate the effects of progressive augmentation on predictive performance, statistical stability, and inter-class discrimination. Section 2 summarizes our research contributions.

The remainder of this paper is organized as follows. Section 3 presents the materials and methods. Section 4 reports the experimental results. Section 5 discusses the findings. Finally, Section 6 concludes the paper and outlines future research directions.

2. Research Contributions

The scientific contributions of this work extend beyond the evaluation of classification accuracy by providing a systematic investigation of the relationship between architectural inductive bias and progressive synthetic data expansion in palynological image classification. Specifically, this study contributes:

1. The identification and characterization of architecture-specific data augmentation saturation points in convolutional and transformer-based models.
2. A controlled experimental comparison between DenseNet201, DenseNet201+CBAM, and ViT-B16 under progressive augmentation regimes and stratified cross-validation.
3. Empirical evidence that attention mechanisms can improve data efficiency, increase statistical stability, and delay performance saturation in specialized image classification tasks.
4. A practical framework for optimizing augmentation strategies and model selection in data-constrained palynological applications.
5. New evidence supporting the hypothesis that Vision Transformers are more sensitive to synthetic noise than convolutional architectures when trained on relatively small and highly specialized datasets.

3. Materials and methods

This section describes the dataset, evaluated architectures, data augmentation protocol, and experimental design adopted to investigate the impact of progressive data augmentation across different deep learning paradigms.

3.1. Dataset and Preprocessing

The Cretan Pollen Dataset (CPD1) [Tsiknakis et al. 2022], containing 4,034 microscopic images from 20 pollen species, was used in all experiments. Images were resized to $224 \times 224 \times 3$ pixels and normalized. No external datasets were employed.

Examples of the dataset are shown in Figure 1.

3.2. Evaluated Architectures

Three architectures representing distinct modeling paradigms were evaluated: DenseNet201 [Huang et al. 2017], DenseNet201 enhanced with the Convolutional Block Attention Module (CBAM) [Woo et al. 2018], and the Vision Transformer ViT-B16 [Dosovitskiy et al. 2020]. DenseNet201 served as the convolutional baseline, CBAM was incorporated to assess the impact of channel and spatial attention mechanisms, and ViT-B16 was selected as a transformer-based alternative relying on global self-attention. For

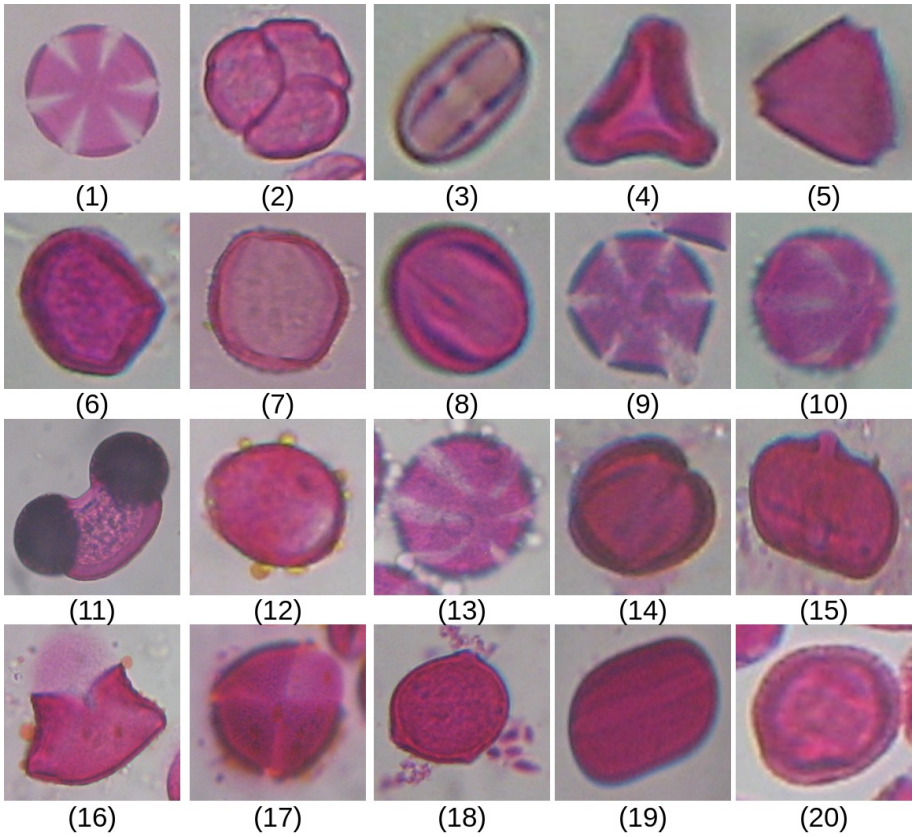


Figure 1. Examples of images from the CPD1 dataset, containing samples from 20 pollen species.

the CBAM-enhanced model, feature refinement is performed through sequential channel and spatial attention:

$$F' = M_c(F) \otimes F \tag{1}$$

$$F'' = M_s(F') \otimes F' \tag{2}$$

where M_c and M_s denote channel and spatial attention maps, respectively.

3.3. Data Augmentation Protocol

To investigate the effect of training data expansion, a progressive data augmentation strategy was adopted. For each augmentation level A_i, up to A_i synthetic images were generated per class:

$$A_i \in \{0, 100, 200, 300, 400, 500, 600, 700\} \tag{3}$$

The augmentation techniques employed are summarized in Table 1. These transformations preserve the semantic structure of pollen grains while increasing variability. All transformations were implemented with the *imgaug* library, applied as a sequential pipeline. Data augmentation was applied exclusively to the training set in order to prevent data leakage into validation and test partitions.

Table 1. Data augmentation techniques utilized.

Method	Parameters	Probability
Gaussian Blur	$\sigma \in [0, 0.3]$	30%
Contrast	$\alpha \in [0.75, 1.25]$	30%
Brightness	$[0.7, 1.3]$	30%
Rotation	$[-25°, 25°]$	100%
Translation (x,y)	$[-0.2, 0.2]$	100%
Horizontal Flip	–	50%
Vertical Flip	–	50%

3.4. Experimental Design

A stratified 10-fold cross-validation protocol was adopted. In each iteration, one fold was reserved for testing, while the remaining folds were split into training (80%) and validation (20%). All architectures were trained using identical data partitions, augmentation procedures, and optimization settings.

3.5. Training Configuration

All architectures were trained under a unified experimental protocol to ensure fair comparison. Models were initialized with ImageNet pretraining and trained for 50 epochs using the Adam optimizer, a batch size of 32, cross-entropy loss, and input images resized to 224×224 pixels.

DenseNet201 and DenseNet201+CBAM employed a learning rate of 10^{-4} and weight decay of 10^{-5}. For ViT-B16, the learning rate was 3×10^{-4} and weight decay was 10^{-4}. The CBAM module used a reduction ratio of 16 and a 7×7 spatial kernel. ViT-B16 employed 16×16 image patches, 768-dimensional embeddings, 12 attention heads, and 12 transformer layers.

3.6. Evaluation and Statistical Analysis

Performance was evaluated using Accuracy, Precision, Recall, F1-score, and Cohen's Kappa, reported as mean $\pm$ standard deviation across folds. Consolidated confusion matrices were generated to analyze recurrent inter-class errors.

Statistical significance was assessed using paired Student's t-tests and repeated-measures ANOVA ($\alpha = 0.05$), enabling comparisons across architectures, augmentation levels, and saturation regimes.

3.7. Use of Generative Artificial Intelligence

Generative AI tools were used exclusively for language revision and text refinement during manuscript preparation. No AI system was involved in the research design, data processing, model development, statistical analysis, interpretation of results, or scientific conclusions. The authors assume full responsibility for the content of this manuscript.

4. Results and Analysis

The experimental results are analyzed according to three central hypotheses: H1 (performance saturation under data augmentation), H2 (statistical stability), and H3 (impact of attention mechanisms on inter-class discrimination).

4.1. Performance Comparison across Architectures

Table 2 summarizes the peak performance metrics achieved by each evaluated architecture. The convolutional paradigm, represented by DenseNet201 and its attention-enhanced variant, consistently outperformed the transformer-based baseline (ViT-B16) on the CPD1 dataset.

Table 2. Peak Performance Metrics (Mean ± Standard Deviation) for each architecture.

Architecture	Aug. Factor	ACC	P	R	F1	K
ViT-B16	200	0.979 ± 0.008	0.980	0.979	0.978	0.976
DenseNet201	400	0.986 ± 0.007	0.986	0.986	0.985	0.984
DenseNet201+CBAM	500	$\mathbf{0.987 \pm 0.005}$	0.987	0.987	0.986	0.985

The DenseNet201+CBAM model achieved the highest overall accuracy (0.987), marginally outperforming the baseline DenseNet201 (0.986) while exhibiting lower variability across folds. Although the absolute performance gain is small, the attention-enhanced architecture demonstrates greater robustness under progressive augmentation regimes.

In terms of computational cost, the average training time per fold at each architecture's saturation point was 24.9 minutes for DenseNet201 (Aug. 400), 26.5 minutes for ViT-B16 (Aug. 200), and 36.4 minutes for DenseNet201+CBAM (Aug. 500), reflecting the added overhead of the attention module.

4.2. Performance under Progressive Augmentation (H1)

Figure 2 illustrates the evolution of accuracy as a function of the augmentation factor.

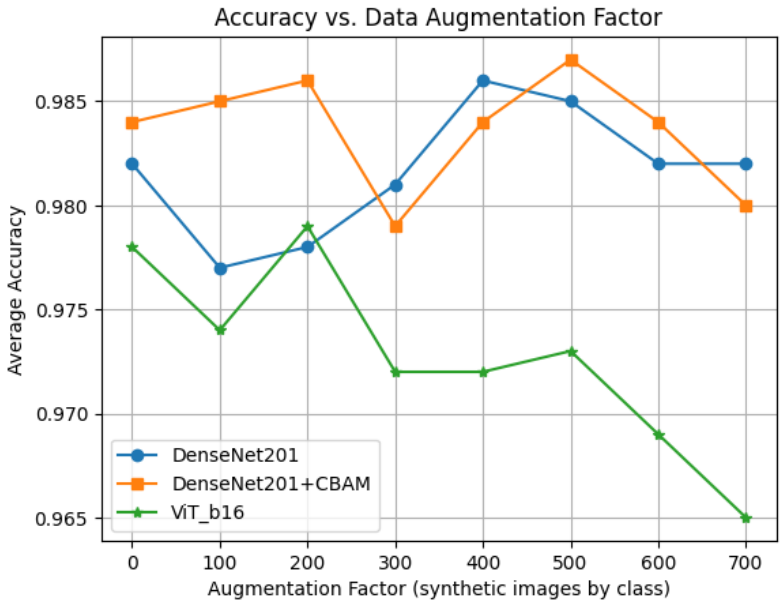


Figure 2. Average accuracy versus augmentation factor. Saturation points are observed for ViT (200), DenseNet (400), and DenseNet+CBAM (500).

All evaluated architectures exhibit a characteristic saturation curve composed of four distinct phases. Initially, a *growth region* is observed, where increasing amounts of synthetic data improve generalization by reducing variance and expanding the effective training distribution. This phase is followed by an *optimal region*, corresponding to the augmentation level that maximizes predictive performance. Beyond this point, a *plateau region* emerges, where additional synthetic samples produce negligible gains. Finally, some architectures exhibit a *degradation region*, in which excessive augmentation introduces statistical redundancy and synthetic noise, leading to reduced performance.

These phases provide empirical evidence that the relationship between data augmentation and classification accuracy is non-linear and strongly dependent on the architectural paradigm. The location and extent of each phase vary across models, reflecting differences in inductive bias and sensitivity to synthetic data.

- **Early saturation (ViT-B16):** Peak performance occurs at augmentation factor 200, followed by a gradual decline, suggesting sensitivity to synthetic noise and limited inductive bias for small datasets.
- **Intermediate saturation (DenseNet201):** The CNN baseline reaches optimal performance at factor 400, benefiting from locality bias and hierarchical feature reuse.
- **Delayed saturation (DenseNet201+CBAM):** The attention mechanism shifts the saturation point to 500, indicating improved robustness to redundant or low-quality synthetic samples.

A comparative inspection of the saturation curves reveals distinct learning dynamics. ViT-B16 reaches its optimum performance earlier, suggesting a greater susceptibility to synthetic noise in limited-data regimes. DenseNet201 benefits from a broader growth region before reaching saturation, while DenseNet201+CBAM exhibits the most extended optimal region, indicating that attention mechanisms improve the effective utilization of augmented samples. These observations reinforce the hypothesis that augmentation efficiency is architecture-dependent rather than universally transferable across deep learning paradigms.

These architecture-specific saturation points provide direct evidence supporting H1. While moderate synthetic expansion improves generalization, excessive augmentation eventually produces diminishing returns and, in some cases, performance degradation. The results indicate that the optimal augmentation level is strongly dependent on the underlying architectural inductive bias.

Table 3. Observed augmentation phases for each architecture.

Architecture	Growth	Optimum	Plateau
ViT-B16	0–200	200	300–700
DenseNet201	0–400	400	500–700
DenseNet201+CBAM	0–500	500	600–700

4.3. Statistical Stability and Variance (H2)

Statistical stability was assessed via standard deviation across the 10 cross-validation folds. Figures 3 and 4 present the distributions for accuracy and F1-score.

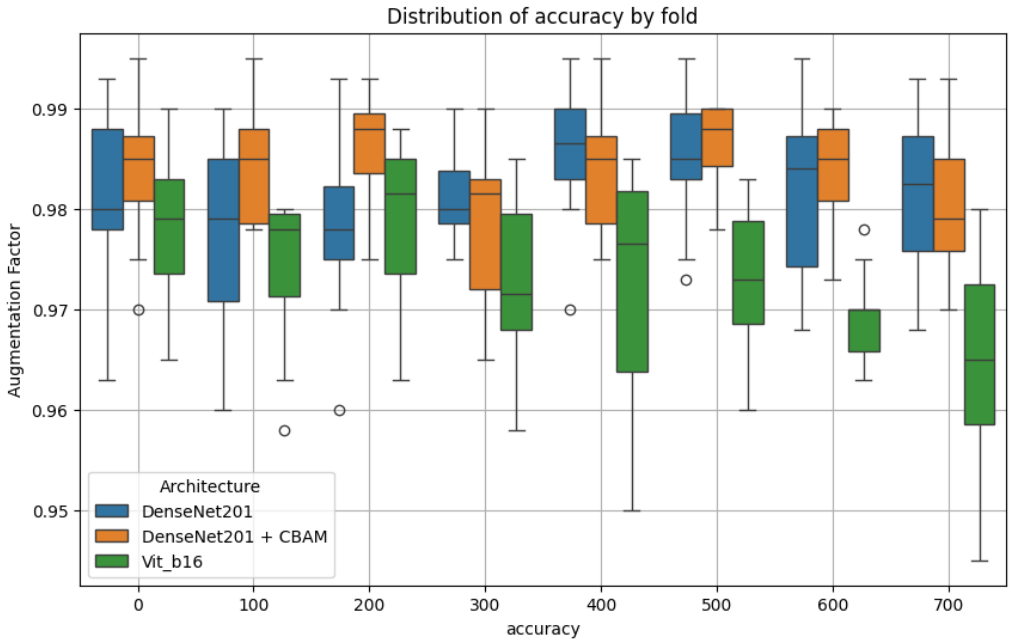


Figure 3. Accuracy distribution across folds for different augmentation factors.

At low augmentation levels (0–200), high variability is observed, reflecting sensitivity to data partitioning. Intermediate levels (300–500) significantly reduce dispersion, indicating improved generalization. However, further increases do not yield consistent gains in stability.

On average, DenseNet201+CBAM consistently exhibits a tendency toward lower fold-to-fold variability, confirming its superior robustness. These findings partially support hypothesis H2.

4.4. Inter-class Confusion Analysis (H3)

The consolidated confusion matrices (Figure 5) reveal that classification errors are primarily associated with morphologically similar classes.

DenseNet201 exhibits higher confusion rates in pairs such as:

- *Sinapis* → *Olea* (4/197 samples)
- *Eucalyptus* → *Myrtus* (4/170 samples)
- *Calicotome* → *Vitis* (5/298 samples)
- *Ceratonia* → *Origanum* (3/100 samples)

The DenseNet201+CBAM model reduces three of these four confusion pairs, with one notable exception:

- *Sinapis* → *Olea* (6/197 samples) — increases relative to the baseline
- *Eucalyptus* → *Myrtus* (3/170 samples) — reduced
- *Calicotome* → *Vitis* (1/298 samples) — substantially reduced
- *Ceratonia* → *Origanum* (1/100 samples) — reduced

At the same time, two additional confusion pairs emerge under CBAM that were not prominent in the baseline: *Origanum* → *Ceratonia* (5/171 samples) and *Satureja* → *Origanum* (4/71 samples).

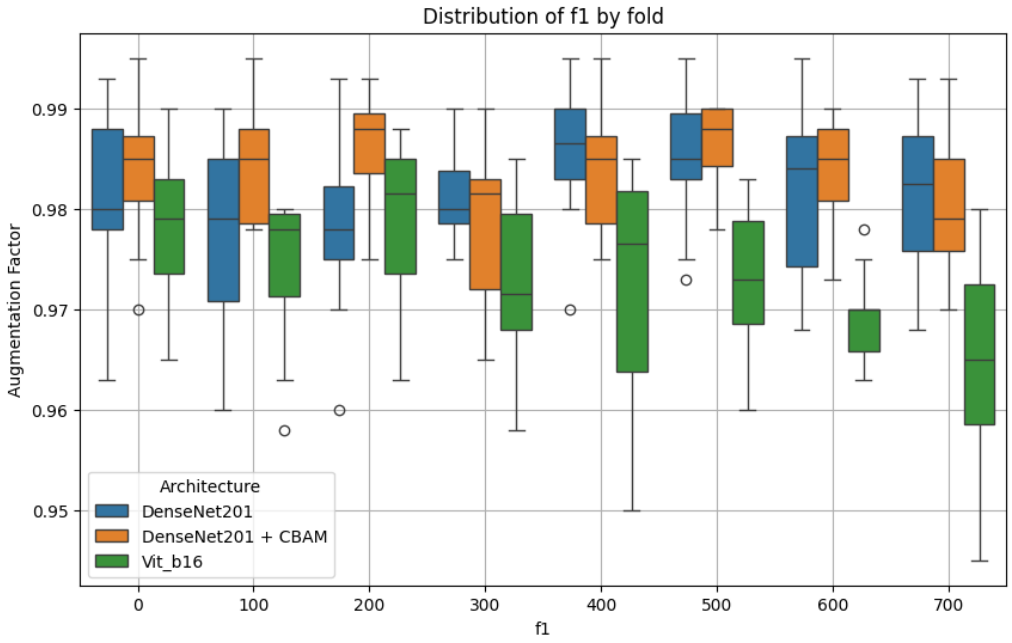


Figure 4. F1-score distribution across folds for different augmentation factors.

Overall, the attention mechanism leads to a more concentrated diagonal distribution for most class pairs, suggesting improved discrimination of subtle local patterns, but the effect is not uniform: confusion is reduced for morphologically related taxa in most cases, yet persists or shifts to new pairs for others (notably Sinapis/Olea and the Origanum/Ceratonia/Satureja group). These results provide partial support for H3. This mixed pattern is consistent with the delayed-rather-than-eliminated saturation behavior observed in the augmentation experiments: attention mechanisms improve feature selectivity for several specific confusions without uniformly resolving morphological ambiguity across all classes.

4.5. Statistical Significance

To rigorously assess performance differences across modeling paradigms and augmentation regimes, statistical analyses were conducted considering cross-validation folds as repeated measures.

A paired Student's t-test ($\alpha = 0.05$) was first applied to compare the baseline **DenseNet201** with the attention-enhanced **DenseNet201+CBAM**. Statistically significant improvements were observed at intermediate augmentation levels, where the attention mechanism most effectively refines feature representations:

- **Aug = 100:** $t(9) = -2.48$, $p = 0.035$, corresponding to a moderate-to-large effect size ($d = -0.78$) and a 95% confidence interval of $[-0.0130, -0.0006]$.
- **Aug = 200:** $t(9) = -3.55$, $p = 0.006$, corresponding to a large effect size ($d = -1.12$) and a 95% confidence interval of $[-0.0128, -0.0028]$.

For higher augmentation levels (Aug $\geq$ 300), differences between the convolutional variants were not statistically significant ($p > 0.25$), indicating convergence as synthetic data redundancy increases.

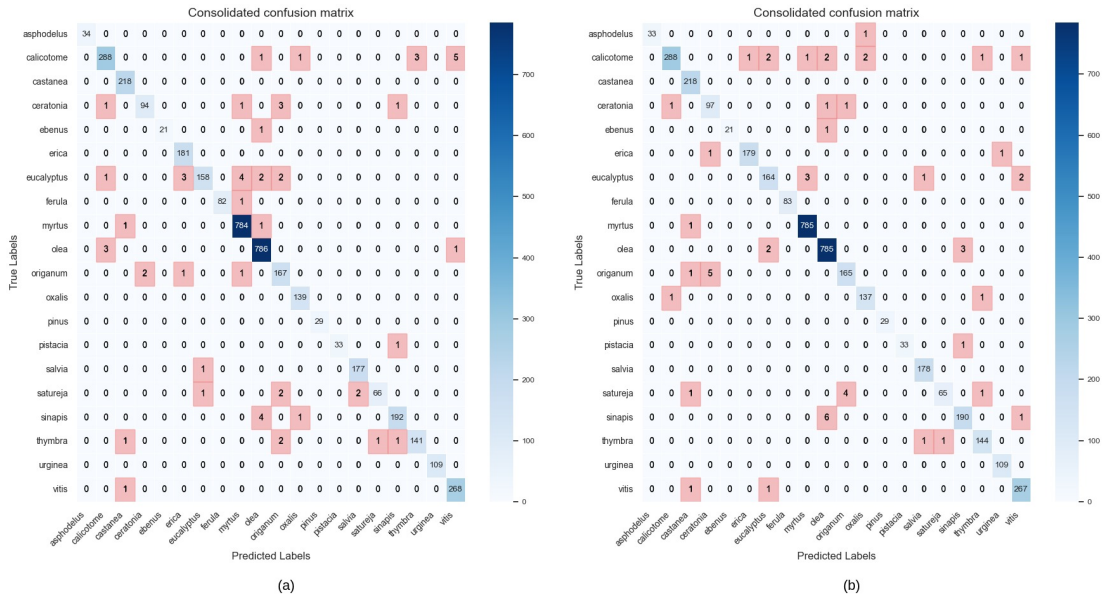


Figure 5. Consolidated confusion matrices for (a) DenseNet201 and (b) DenseNet201+CBAM.

To evaluate the combined effects of augmentation level and attention mechanisms within the convolutional paradigm, a two-way repeated-measures ANOVA was conducted using the fold-wise results obtained from DenseNet201 and DenseNet201+CBAM. Assuming sphericity, the analysis revealed:

- **Significant effect of augmentation factor:** $F(7, 63) = 2.24, p = 0.042$, supporting the performance saturation hypothesis (H1).
- **Significant interaction (architecture $\times$ augmentation):** $F(7, 63) = 2.45, p = 0.027$, indicating that the effectiveness of data expansion depends on the underlying model architecture.
- **Main effect of architecture:** $p = 0.099$, suggesting that while models achieve similar peak accuracies, their learning dynamics and saturation behavior differ.

The observed interaction effect indicates that augmentation does not affect all architectures equally. In particular, ViT-B16 reaches its performance peak earlier than the convolutional architectures, suggesting greater sensitivity to excessive synthetic data.

Overall, these results demonstrate that attention mechanisms (CBAM) provide statistically significant benefits in intermediate data regimes, improving robustness and delaying performance saturation compared to both standard convolutional and transformer-based approaches.

4.6. Hypothesis Synthesis

The results support hypothesis H1, confirming the existence of performance saturation with increasing synthetic data.

Hypothesis H2 is partially supported, as moderate augmentation reduces variability, but excessive augmentation does not provide additional stability.

Hypothesis H3 is also partially supported, since attention mechanisms improve discrimination in specific scenarios but do not guarantee consistent superiority across all conditions.

5. Discussion

The results reveal consistent patterns regarding the interaction between progressive data augmentation and deep learning architectures. The observed saturation behavior and the distinct responses across paradigms provide insights into the roles of inductive bias, data efficiency, and adaptive feature representation.

5.1. Data Augmentation and the Law of Diminishing Returns

A central finding of this work is that data augmentation does not produce monotonic performance improvements. Instead, all evaluated architectures exhibited a clear saturation point, beyond which additional augmented samples led to negligible gains or performance degradation. This phenomenon can be interpreted through the lens of the bias–variance tradeoff.

Initially, data augmentation increases training diversity, reducing variance and improving generalization. However, as synthetic expansion continues, its informational benefit decreases and redundancy becomes dominant. Excessive augmentation may introduce artificial patterns that do not contribute to discriminative learning, indicating that augmentation effectiveness is constrained by the architecture's representational capacity.

5.2. Architectural Inductive Bias and Data Efficiency

The superior performance of DenseNet201 and DenseNet201+CBAM compared to ViT-B16 highlights the importance of inductive bias in specialized scenarios with limited data. CNNs inherently encode spatial locality and translation invariance, properties that are particularly effective for morphological analysis of pollen grains. These built-in priors enable more efficient generalization when training data are scarce.

In contrast, Vision Transformers rely on self-attention mechanisms with minimal image-specific inductive bias and typically require larger datasets to learn spatial relationships effectively. The earlier saturation observed for ViT-B16 suggests that excessive synthetic augmentation introduces non-semantic variations that are more difficult for transformer architectures to filter than for CNNs, reinforcing their well-known dependence on larger training datasets.

5.3. Practical Implications for Training Design

The results of this study offer direct implications for the design of deep learning pipelines in agriculture and botany:

- **DA is not universal:** Increasing augmented data volume should not be treated as a universally beneficial strategy; practitioners must identify architecture-specific optimal points to minimize computational cost.
- **CNNs as a strong baseline:** Convolutional architectures remain the preferred baseline for image classification in constrained data regimes.
- **Attention over Expansion:** Targeted architectural enhancements, such as CBAM, are more effective than massive synthetic data expansions for improving stability and precision.

5.4. Limitations and Future Directions

Despite the contributions, some limitations remain. This study utilized traditional geometric and intensity transformations and was restricted to the CPD1 dataset. Future work should evaluate these patterns across datasets with higher taxonomic diversity and explore generative models (e.g., GANs or Diffusion Models) for more realistic data synthesis. Additionally, implementing interpretability methods like Grad-CAM would help clarify how different architectures weigh augmented features at the layer level.

6. Conclusion

This study demonstrated the interaction between progressive data augmentation and deep learning paradigms in the context of automatic pollen grain classification. By comparing convolutional (DenseNet201), attention-enhanced (DenseNet201+CBAM), and transformer-based (ViT-B16) architectures, we provided empirical evidence of architecture-dependent saturation behavior.

The experimental results lead to the following primary conclusions:

- **Architecture-Dependent Saturation:** Data augmentation did not produce monotonic gains. Distinct saturation points were observed for each architecture: augmentation factor 200 for ViT-B16, 400 for DenseNet201, and 500 for DenseNet201+CBAM. Beyond these points, additional synthetic samples produced marginal gains or slight performance degradation, indicating saturation effects associated with redundancy in the augmented data.
- **Performance Differences Across Paradigms:** DenseNet201 and DenseNet201+CBAM consistently outperformed ViT-B16 on CPD1, suggesting that convolutional inductive biases remain advantageous for pollen classification under limited-data conditions.
- **CBAM as a Robustness Enhancer:** DenseNet201+CBAM achieved the highest accuracy (0.987) and delayed saturation to augmentation factor 500. The attention mechanism also contributed to lower variability across folds, indicating improved robustness under progressive data expansion.
- **ViT Sensitivity to Progressive Augmentation:** ViT-B16 exhibited earlier saturation and lower robustness under extensive synthetic expansion, reinforcing the greater dependence of transformer architectures on data quantity and quality in specialized image classification tasks.

From a methodological standpoint, these findings suggest that practitioners should prioritize targeted architectural enhancements, such as local attention mechanisms, over massive synthetic dataset expansions. Identifying the optimal augmentation factor is critical to balancing computational efficiency and classification accuracy.

Despite the high performance achieved, this work was limited to the CPD1 dataset and traditional geometric/intensity transformations. Future research should explore:

1. The use of generative models (e.g., GANs or Diffusion Models) for more realistic data synthesis.
2. Validation across multiple palynological datasets with higher taxonomic diversity.
3. The application of interpretability methods such as Grad-CAM to visualize how augmentation affects feature saliency in different architectures.

In summary, this research contributes a nuanced understanding of data efficiency in deep learning for palynology, offering a robust framework for designing more effective and stable classification systems in agriculture and botany.

References

Astolfi, G., Gonçalves, A. B., Menezes, G. V., Borges, F. S. B., Astolfi, A. C. M. N., Matsubara, E. T., Alvarez, M., and Pistori, H. (2020). POLLEN73S: An image dataset for pollen grains classification. *Ecological Informatics*, 60:101165.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Flenley, J. (1968). The problem of pollen recognition. *Problems in Picture Interpretation*, pages 141–145.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. (2017). Densely connected convolutional networks. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua:2261–2269.

Mahmood, T., Choi, J., and Park, K. R. (2023). Artificial intelligence-based classification of pollen grains using attention-guided pollen features aggregation network. *Journal of King Saud University - Computer and Information Sciences*, 35(2):740–756.

Sevillano, V. and Aznarte, J. L. (2018). Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. *PLOS ONE*, 13(9):e0201807.

Tsiknakis, N., Savvidaki, E., Manikis, G. C., Gotsiou, P., Remoundou, I., Marias, K., Alissandrakis, E., and Vidakis, N. (2022). Pollen Grain Classification Based on Ensemble Transfer Learning on the Cretan Pollen Dataset. *Plants*, 11(7):919.

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19.

Zu, B., Cao, T., Li, Y., Li, J., Wang, H., and Wang, Q. (2025). RESwinT: enhanced pollen image classification with parallel window transformer and coordinate attention. *The Visual Computer*, 41(7):4975–4990.