COMPUTER RESEARCH AND MODELING
2026 VOL. 18 NO. 4 P. 871–889
DOI: 10.20537/2076-7633-2026-18-4-871-889

ANALYSIS AND MODELING OF COMPLEX LIVING SYSTEMS

UDC: 519.6

Design-optimized YOLO11 classification via strategic CBAM attention injection and Grad-CAM explainability for reliable plant disease diagnosis

A. M. M. Al. Koaerji[1,2,a], O. M. Ogorodnikova[1,b], A. I. Ogorodnikov[1,c]

[1]Ural Federal University,
19 Mira st., Ekaterinburg, 620002, Russia
[2]Polytechnic college of engineering specializations, Middle Technical University,
Muasker Al Rashid st., Baghdad, 10074, Iraq

E-mail: [a] ali_mozan7819@mtu.edu.iq, [b] Olga.Ogorodnikova@urfu.ru, [c] A.I.Ogorodnikov@urfu.ru

Received 05.01.2026, after completion — 15.03.2026.
Accepted for publication 05.06.2026.

Reliable plant disease diagnosis requires not only high classification accuracy, but also stable generalization and interpretable decision-making. Although attention mechanisms are useful to these deep learning models, the performance also depends on where and how they are integrated into the network architecture. This study presents a design-optimized YOLO11m-based classification framework that systematically investigates the impact of Convolutional Block Attention Module (CBAM) injection at different architectural levels for plant disease diagnosis. We perform a comparative model-controlled analysis of three model architectures: (i) the baseline YOLO11m-Cls architecture lacking attention, (ii) only adding the backbone block along with CBAM and (iii) a hybrid architecture that includes reduced backbone attention combined with CBAM added at classification head. All models are trained and tested under the same experimental settings using a large-scale dataset with around 90 000 images of 38 types of plant diseases. Experimental results clearly show that the uniform injection of CBAM into the backbone reduces stability but causes generalization to worsen with a higher validation loss and significantly lower Top-1 accuracy (≈ 90.5 %), while the hybrid attention design balances stability and discrimination, with Top-1 accuracy up to 99.71 %, Top-5 accuracy up to 99.99 % and near-baseline validation behavior respectively Grad-CAM-based interpretability analysis also demonstrates that the hybrid model generates enhanced and biologically interpretable activation maps, which are more disease-specific with less distraction from background. Notably, the aim of our work is not for achieving superior performance over all classifiers but instead only to provide design-level evidence on how placing attention modulates rigidity and interpretability in YOLO-based classification models. The results provide practical architectural considerations for building dependable and interpretable AI systems in agriculture.

Keywords: YOLO11-based classification, attention placement strategy, convolutional block attention module (CBAM), explainable artificial intelligence (XAI), Grad-CAM visualization, plant disease diagnosis

Citation: *Computer Research and Modeling*, 2026, vol. 18, no. 4, pp. 871–889.

**КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2026 Т. 18 № 4 С. 871–889**
DOI: 10.20537/2076-7633-2026-18-4-871-889

Ки&М

АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ

УДК: 519.6

Классификатор на базе оптимизированной архитектуры YOLO11 с применением инструментов CBAM и Grad-CAM для надежной диагностики больных растений

А. М. М. Аль. Коаерджи[1,2,a], О. М. Огородникова[1,b], А. И. Огородников[1,c]

[1]Уральский федеральный университет,
Россия, 620002, г. Екатеринбург, ул. Мира, д. 19
[2]Центральный технический университет,
Ирак, 10074, г. Багдад, ул. Муаскар Аль-Рашид

E-mail: [a] ali_mozan7819@mtu.edu.iq, [b] Olga.Ogorodnikova@urfu.ru, [c] A.I.Ogorodnikov@urfu.ru

Получено 05.01.2026, после доработки — 15.03.2026.
Принято к публикации 05.06.2026.

Надежная диагностика болезней растений компьютерным зрением требует не только высокой точности классификации, но и стабильной обобщающей способности и интерпретируемого принятия решений. При этом механизмы внимания безусловно полезны для моделей глубокого обучения, но эффективность решения задачи зависит от места и способа интеграции этих механизмов в архитектуру нейронной сети. В данном исследовании представлен классификатор на базе оптимизированной архитектуры YOLOv11m, а также результаты анализа результатов диагностики болезней растений в зависимости от сверточных блоков CBAM (Convolutional Block Attention Module), внедренных на разных уровнях архитектуры. Выполнен сравнительный анализ трех архитектурных вариантов модели: базовой архитектуры YOLO11m-Cls без механизма внимания, варианта с добавлением CBAM только в магистральную часть backbone и гибридной архитектуры, сочетающей сокращенное внимание в магистральной части backbone с дополнительным блоком CBAM на уровне классификационного слоя head. Все модели были обучены и протестированы в одинаковых экспериментальных условиях с использованием обширного набора данных, содержащего около 90 тысяч изображений 38 типов болезней растений. Экспериментальные результаты ясно показывают, что равномерное внедрение CBAM в магистральную часть backbone снижает устойчивость модели и ухудшает ее способность к обобщению, что проявляется в повышении валидационной функции потерь и понижении точности (Top-1 $\approx$ 90,5 %) по сравнению с базовой архитектурой. Напротив, гибридная архитектура обеспечивает баланс между устойчивостью и способностью детектировать с повышением параметров точности Top-1 до 99,71 % и Top-5 до 99,99 % на этапе валидации, близким по значениям к параметрам базовой модели. Анализ интерпретируемости на основе метода Grad-CAM также положительно характеризует гибридную модель, которая формирует улучшенные и биологически обоснованные карты активации, четко фокусирующиеся именно на признаках заболевания и практически исключающие отвлечение внимания на фоновые элементы изображения. Важно отметить, что целью данной работы является не достижение наилучшей производительности среди всех классификаторов, а предоставление доказательств на уровне архитектурного проектирования относительно того, каким образом размещение механизма внимания влияет на устойчивость и интерпретируемость в классификационных моделях на базе YOLO. Полученные результаты предоставляют практические рекомендации по проектированию надежных и интерпретируемых систем искусственного интеллекта для агропромышленного сектора.

Ключевые слова: диагностика болезней растений, компьютерное зрение, сверточная нейронная сеть, YOLO11, Convolutional Block Attention Module (CBAM), Explainable Artificial Intelligence (XAI), Grad-CAM

1. Introduction

The problems associated with the diagnosis of plant diseases have considerable implications in terms of yield, security and sustainability of plant production [FAO, 2021]. Conventionally, diagnostic methods primarily consist of the visual analysis of images acquired by some sort of imaging device, by experts, which is typically subjective, slow, hard to reproduce, and unsuitable for field conditions because of environmental factors (background complexity and uneven illumination) [Mohanty et al., 2016]. In recent years, there has been a lot of development through deep learning approaches to plant disease recognition. Convolutional neural networks [Goodfellow et al., 2016; Ogorodnikova, Ali, 2019] alter the way high-level features are retrieved, such as texture, color distortion and morphology. Plant disease classification has also seen a significant improvement due to described datasets, but there are well-known used big scale datasets available for the training of multiple classes p-plant diseases through supervised learning [Hughes, Salathé, 2015; Аль. Коаерджи, Огородников, 2025]. However, while these models attain a relatively high classification accuracy, most of them remain insufficiently reliable to be used in the real world agricultural setting, as factors such as training stability, generalization robustness, and interpretability are often less explored. The YOLO-based architectures have originally been developed as a fast detector for real-time applications and recently have attracted more and more attention for image-level classification problems because of efficent backbone designs and optimal accuracy-computation cost trade-off. Many recent variants of YOLO have outperforming results on large-scale agricultural datasets, while few studies enhanced YOLO architectures (through attention mechanisms or feature fusion strategies) for augmented recognition capability under challenging field conditions [Gupta et al., 2025; Lv, Su, 2024; Федоров, Огородникова, 2024]. In contrast, most works based on YOLO predominantly focus on either detection performance or architectural efficiency but provide scarce guidance about a) where exactly attention mechanisms should be integrated into classification-oriented architectures and b) how these design choices impact training stability and generalization behavior. One of the key limitations of conventional YOLO-based classifier architectures is their use of global average pooling that averages over spatial information and can average out localized disease-relevant information in the presence of background clutter [Jiang et al., 2022]. To deal with this issue, feature selectivity has been enhanced using attention mechanisms to weigh informative channels and spatial regions. Among these, the convolutional block attention module (CBAM) has been popular, especially due to its lightweight design with sequential channel-spatial attention. This has shown promise, especially in fine-grained image classification problems such as plant disease classification, where the error is localized in space to discriminate symptoms. Recent works have investigated hybrid attention methods, multi-scale feature fusion, and spatial attention guided networks for improving the classification performance [Duhan et al., 2024; Kumar et al., 2025]. Despite this progress, attention mechanisms are often perceived as generic ingredients for improving performance, and evaluation is often reduced to quantifying classification accuracy. The role of attention placement on stability of training, the dynamics of validation loss, and the uncertainty of generalization, and the important aspects for reproducible agricultural AI have not been thoroughly investigated. In addition, the interpretability field often treats it as a visualization tool ex post, rather than a property to be systematically tied to architectural design decisions. There have been some recent works that need to combine disease classification with explainability of the model to make agriculture more sustainable and also transparent in plant disease detection in the agriculture [Shafik et al., 2025; Федоров, Огородникова, 2025] where they mainly emphasize the explainable AI plant disease diagnosis. Given these drawbacks, this work is aimed at investigating attention placement as an explicit architectural design variable and uses the YOLO11-based classification system as a foundation. Rather than developing an additional attention placement mechanism, this work aims to provide design-level insights by controlling for attention placement through an experimental comparison of multiple injection strategies for CBAM, all under identical conditions. By providing a combined exploration of classification performance, training stability, validation characteristics, and Grad-CAM-based interpretability, this work aims to provide design-level insights that go beyond accuracy-based comparisons, highlighting the impact of inappropriate attention placement on classification performance and training stability. From a computer modeling and design automation perspective, the YOLO11-based classification system is considered a structured computational model, as opposed to a black-box learning approach. In this work, design automation is considered to be the systematic and replicable exploration of architectural design options, as opposed to architecture generation. Attention placement is considered a discrete architectural design variable, which is explored through controlled experimental comparisons. This structured exploration approach links architectural design choices to system responses, allowing for objective reasoning regarding attention placement strategies, providing design-level guidelines for developing reliable deep learning-based classification systems.

The contributions of this work can be summarized as follows:

(i) a systematic design-level evaluation of CBAM attention placement within a YOLO-based classification architecture;

(ii) empirical evidence demonstrating that dense backbone attention can degrade training stability and generalization despite improving early convergence;

(iii) a hybrid attention configuration that balances classification performance, stability, and interpretability;

(iv) an architectural analysis linking attention placement decisions to Grad-CAM-based interpretability outcomes.

2. Related work

Deep learning based plant disease classification has been a widely studied area in the last few years. Early methods that were based on CNNs performed well on relatively clean, context-free datasets but degraded under unfavorable background [Goodfellow et al., 2016; Hughes, Salathé, 2015]. In order to address these drawbacks, the attention mechanism was adapted to plant diseases classification frameworks. Alirezazadeh et al. demonstrated that CNN classifiers can become more robust to background noise and classification stability can be improved when CBAM is incorporated [Alirezazadeh et al., 2023]. Yang et al. proposed adaptive channel attention for fine-grained disease recognition [Yang et al., 2024], and Deng et al. investigated channel recalibration methods for leaf disease identification [Deng et al., 2024]. Because of its high speed and scale, YOLO-based models have also been widely used in agricultural vision tasks. There are several variants of YOLO, such as YOLOv5 and YOLOv8, which have been adopted in plant disease detection and classification task [Redmon et al., 2016; Jocher et al., 2022; Gupta et al., 2025]. The studies where CBAM or the dynamic attention is integrated into YOLO-based architectures are implemented to enhance the performance in challenging field environments [Gupta et al., 2025; Lv, Su, 2024]. Yet, such techniques generally do not consider hybrid backbone + head attention structures and only pay the additional attention to a single location of the backbone or split it equally over several stages of the backbone. Unlike most currently proposed hybrid attention strategies that concentrate on pushing the classification accuracy [Canghai et al., 2025], this work studies attention as an architectural design choice and explores its effects on training stability, generalization capacity and interpretability in a YOLO-style classification pipeline.

In contrast to the predominant directions of most previous attention-based plant disease classification efforts [Alirezazadeh et al., 2023; Deng et al., 2024], which primarily pursue accuracy optimization by enhancing the attention, [Duhan et al., 2024; Kumar et al., 2025], this paper emphasizes on how to place the attention using our architecture design. Unlike the vast majority of the latest hybrid attention approaches, which focus on the gain in performance, this paper aims to systematically investigate the impact of various mechanisms for injecting attentions on the training, validation, and interpretability of the YOLO-style classifier.

3. Methodology

The proposed framework for the classification is made up of the dataset, the baseline model YOLO11m-Cls, the CBAM module, and the three different models that are proposed for the evaluation. All the models have been evaluated in a fair manner, and this is necessary for the purpose of a fair comparison. Figures 1, 2, and 3 are some of the figures that have been used to justify the architectural decisions made in this section.

3.1. Dataset description

The evaluation was carried out on large-scale data for plant diseases, which comprises 90 000 images with 38 different kinds of plant disease classes. The data is obtained from agricultural archives that are publicly available via Kaggle, resulting in varying textures of the leaves as well as the lesions, including changes in color and the development stages of the lifecycle of the plant diseases [Goodfellow et al., 2016; Alirezazadeh et al., 2023]. In order to filter out noise in the background of the images as well as to learn features of the lesions, background suppression is used by implementing the U2-Net segmentation technique. This is an important aspect to avoid spurious correlations, especially in plant disease detection scenarios [Alirezazadeh et al., 2023; Linardatos et al., 2020].

After preprocessing, the database was arbitrarily divided using stratified division method to maintain classes ratios between sets:

- 80 % for training ($\approx$ 72 000 images),
- 10 % for validation ($\approx$ 9 000 images),
- 10 % for testing ($\approx$ 9 000 images).

All images were scaled to the size of 224×224 due to the input resolution of the YOLO11m-Cls architecture. The distribution of classes on 38 categories in the training set is summarized in Fig. 1, is well balanced and large-scale ensuring both stable training and meaningful class-wise evaluation.

Figure 1 shows the number of training images per class (total $\approx$ 72 000), highlighting the degree of class balance/imbalance prior to model training and class-wise evaluation.

3.2. Baseline YOLO11m-Cls architecture (model A)

We used YOLO11m-Cls, which is our baseline from here and is a medium scale convolutional classifier that was created based on the YOLO family, which was made to support efficient level image recognition. Following YOLOv5-Cls and YOLOv8-Cls, it mainly includes three parts: a convolutional backbone, a feature aggregation neck and a light-weight classification head [Jocher et al., 2022; Jiang et al., 2022].

A simplified block diagram of the baseline architecture is shown in Fig. 2.

The model architecture comprises a C3k2-based convolutional backbone for hierarchical feature extraction, a lightweight neck for feature aggregation, and a classification head for image-level prediction.

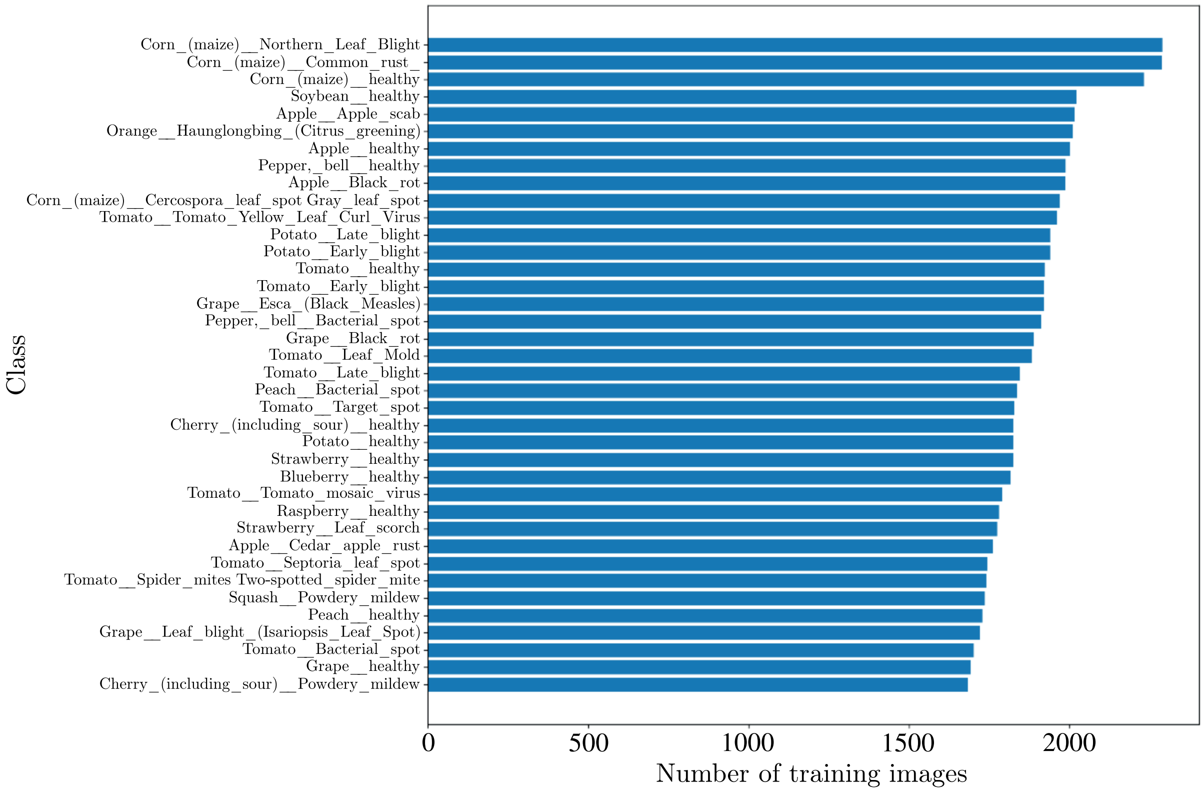

Figure 1. Class distribution of the 38 plant disease categories in the training set

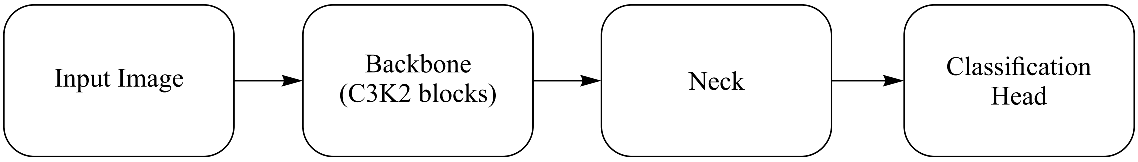

Figure 2. Baseline YOLO11m-Cls architecture

3.2.1. Backbone

The backbone is composed of a deep hierarchy of C3k2 blocks, which is made possible by the cross-stage partial (CSP) construction, which achieves a compromise between feature richness and computation cost. Each block is composed of:

- a 1×1 convolution for channel reduction,
- two 3×3 bottleneck convolutions,
- a residual fusion pathway,
- and a final expansion convolution.

This structure enables the extraction of hierarchical representations across progressively reduced spatial resolutions ($224 \to 112 \to 56 \to 28 \to 14 \to 7$), capturing color gradients, lesion boundaries, and texture irregularities critical for plant disease discrimination.

3.2.2. Neck

The neck module links the high-level features learned by the backbone network with a simplified fusion module. Unlike the detection-based YOLO series, the proposed classification-based neck is optimized for the preservation of semantic richness without involving any extra spatial transformations to ensure the high throughput and gradients in the training process.

3.2.3. Classification head

The baseline classification head consists of:

1) a 1×1 convolution projecting the fused feature map to a fixed channel dimension,

2) a global average pooling (GAP) layer,

3) a dropout layer for regularization,

4) a fully connected layer producing logits for the 38 disease classes.

Although this method is effective, all regions of space are treated equally, and this may reduce the localization of disease cues when background clutter is also present [Jiang et al., 2022].

3.3. Convolutional block attention module (CBAM)

In order to further improve the feature selectivity of the network, the paper employs the convolutional block attention module (CBAM) introduced in Ref. [Woo et al., 2018]. This module employs channel attention and spatial attention in an adaptive manner to refine the intermediate features.

Given an input feature map

$$F \in R^{C\times H\times W} \tag{1}$$

where C, H, and W denote the channel, height, and width dimensions, respectively, CBAM learns to emphasize informative features while suppressing irrelevant responses.

Channel Attention (CAM) focuses on the informative feature channels through the utilization of global average and max pooling descriptors, followed by a shared multilayer perceptron and sigmoid activation.

Spatial Attention (SAM) focuses on the salient spatial regions through the utilization of a 7×7 convolution over concatenated average and max pooling descriptors along the channel dimension.

CBAM is lightweight and introduces only a modest increase in parameters and floating-point operations) FLOPs), making it suitable for large-scale agricultural datasets and resource-constrained environments [Woo et al., 2018; Gupta et al., 2025].

3.4. Model variants and CBAM injection strategies

To conduct a thorough analysis of the influence of attention positioning, we consider three architectural models, which are designed using a controlled setting. Rather than removing features from an existing model, attention injection strategies are used to design the model variants, thus providing a design-centric evaluation of the influence of attention positioning on learning, generalization stability, and interpretability. To keep the description concise, the architecture of the model is composed of four consecutive stages of feature resolutions rather than architectural blocks. Table 1 provides a summary of the overview of the model, which was used to conduct the experiments, along with the attention configurations.

The exact CBAM injection locations across backbone stages and the classification head are detailed in Table 2.

Table 1. Overview of the evaluated YOLO11m-Cls model variants and attention configurations

Model	Backbone CBAM	Classification Head CBAM	Attention strategy description
Model A (Baseline)	✗	✗	Standard YOLO11m-Cls architecture without attention mechanisms
Model B (Backbone CBAM)	✓ (stages 2–5)	✗	Uniform CBAM injection across multiple backbone stages
Model C (Hybrid CBAM)	✓ (stages 2–3)	✓	Reduced backbone attention combined with CBAM-enhanced classification head

Table 2

Backbone stage	Spatial resolution	Model A (Baseline)	Model B (Backbone CBAM)	Model C (Hybrid CBAM)
Stage 1	112×112	✗	✗	✗
Stage 2	56×56	✗	✓	✓
Stage 3	28×28	✗	✓	✓
Stage 4	14×14	✗	✓	✗
Stage 5	7×7	✗	✓	✗
Classification head	—	✗	✗	✓

CBAM is uniformly injected into multiple backbone stages in model B, while model C adopts a reduced backbone configuration and integrates CBAM exclusively at the classification head.

3.4.2. Model B — Backbone-Only CBAM

For model B, only selected c3k2 blocks in the backbone are augmented with CBAM modules and the neck and classification head remain unmodified. As illustrated in Fig. 3, the CBAM is utilized from backbone stage 2 to stage 5 and thus performs early and mid-level feature refinement. This design targets to improve lesion localization and reduce irrelevant background responses. Table 2 summarizes these injection points across spatial resolutions.

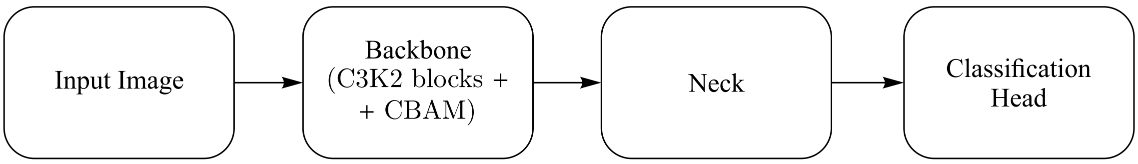

Figure 3. YOLO11m-Cls with backbone-only CBAM injection (model B)

CBAM modules are injected into selected C3k2 blocks of the backbone to refine channel and spatial feature representations, while the neck and classification head remain unchanged.

3.4.3. Model C – Hybrid multi-level CBAM

Model C uses a hybrid attention mechanism, containing the:

- diminished CBAM injection in the backbone (only up to stages 2 and 3), and
- a complete CBAM module is embedded into the classification head. The design of the new hybrid attention strategy can be seen in Fig. 4.

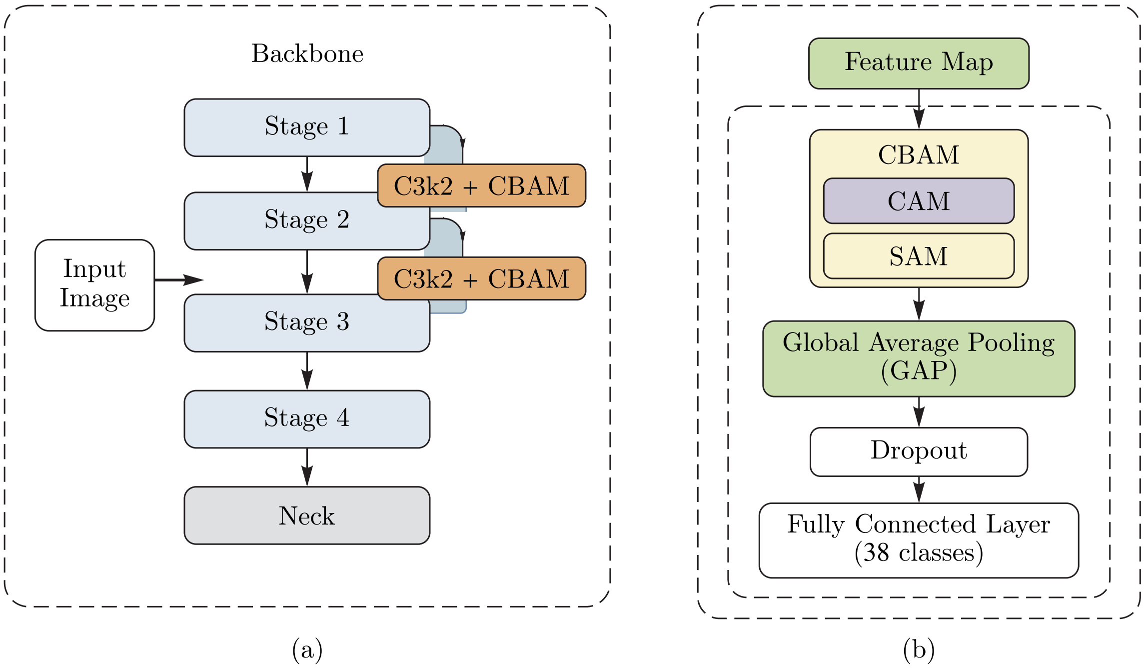

Figure 4. Hybrid multi-level CBAM architecture used in model C. (a) Reduced backbone attention configuration, in which CBAM is injected into the intermediate stages of the backbone. (b) CBAM-boosted classification head, which combines channel-wise and spatial attention to further refine the high-level semantic feature before global pooling and classification

This hybrid architecture focuses on semantic-level attention and reduces over-burden of backbone refinement, thus facilitating better generalization stability and interpretability.

3.5. Training procedure

Notably, each variant of the 3 models was trained with the same experimental settings for a fair comparison. We use the AdamW optimizer with an initial learning rate of $1 \cdot 10^{-3}$, weight decay of 0.01, and cosine learning rate scheduler for the 50 epochs of training. The batch size was 64. Normalization of the input images was done in the range of 0 to 1. Gentle data augmentation was used for all the models, and this helped the model to generalize well. This included random horizontal flip, small rotations, and mild color jitter. However, the disease was not altered in any way.

Training progress was monitored using training loss, validation loss, Top-1 accuracy, and Top-5 accuracy.

3.6. Evaluation metrics

Model performance was assessed using standard classification metrics:

- Top-1 Accuracy,
- Top-5 Accuracy,
- Precision, Recall, and F1-score,
- Confusion Matrices.

These metrics enable both global performance evaluation and class-level error analysis, which is essential for fine-grained plant disease diagnosis [Goodfellow et al., 2016; Linardatos et al., 2020].

4. Experimental results and discussion

This section presents an in-depth evaluation of the three models of YOLO11m-Cls, including the baseline model (model A), the backbone-only CBAM model (model B), and the hybrid multi-level CBAM model (model C). The evaluation will cover training, validation, classification, confusion, and interpretation through attention and Grad-CAM.

4.1. Training dynamics

4.1.1. Training loss behavior

Figure 5 shows the training loss curves for all three models through 50 epochs. Showing that the baseline model runs smoothly and the oscillatory behavior caused by dense backbone attention is preserved while stabilizing the learning dynamics with our proposed hybrid CBAM configuration. The base model (model A) has a smooth and monotonic loss decay, due to the high inherent stability of the original YOLO11m-Cls architecture without attention-induced variance. Such conduct is typical for well-conditioned convolutional classifiers, learned on large annotated datasets. The only difference is that model B (backbone-only CBAM) exhibits a faster decrease of training loss in the first few epochs, implying better acceleration of the initial convergence of attention-enhanced feature extraction. However, evident oscillations are observed in late epochs, indicating that the gradient variance becomes larger due to over-attention enhancement of deep backbone layers. The most well-behaved training among the attention-enhanced models is attained by model C (hybrid CBAM). Reducing backbone attention and adding CBAM at the classification head helps the model follow a nice low-loss trajectory without reaching the same low-loss regime of the baseline.

These observations follow known trends in attention-augmented CNNs, where the impact of attention depends on the representational depth at which it is applied (Fig. 5).

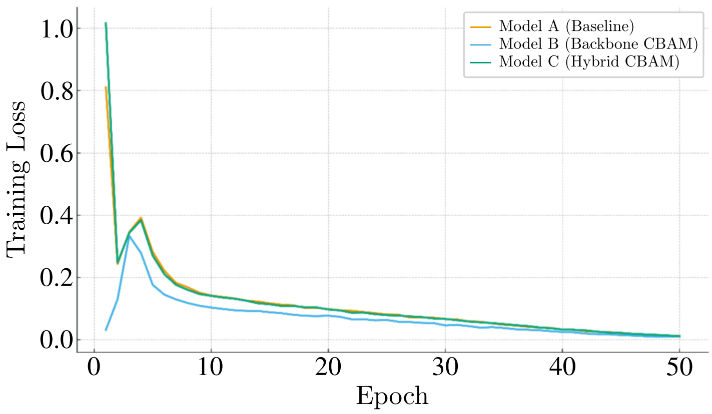

Figure 5. Training loss curves for the three YOLO11m-Cls variants

The baseline model exhibits smooth and monotonic convergence, while dense backbone attention in model B introduces oscillations. The proposed hybrid CBAM configuration (model C) achieves stabilized training dynamics.

4.1.2. Validation loss analysis

The validation loss curves shown in Fig. 6 reveal pronounced differences in generalization behavior among the three model variants.

Illustrating extremely low and steady validation loss by the baseline model, a high plateau caused by dense backbone CBAM and superior generalization stability is attained by hybrid attention placement.

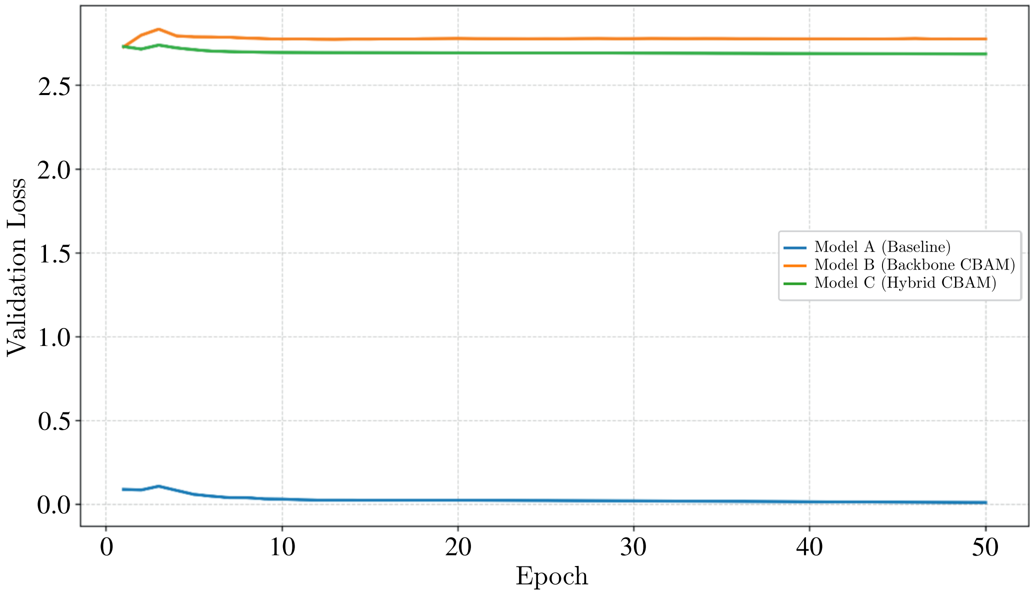

Figure 6. Validation loss curves of the three variants of YOLO11m-Cls over 50 epochs

Model A across the board obtains the smallest validation loss, which would be reasonable in light of its conservative structure and strong baseline generalization.

Model B converges to the highest validation loss that gets down to 2.7 ~ 2.8, which reflects the over-sensitivity of background textures and nuisance patterns by dense backbone attention. The reduction in validation loss is clear in model C as compared to B, but the baseline results are even lower. This indicates that taking away redundant backbone attention and pushing to the classification head feature recalibration can improve generalization stability without disrupting the trainability (Fig. 6).

4.2. Classification performance

4.2.1. Top-1 accuracy

The final Top-1 accuracy results for the three models are summarized in Fig. 7.

The backbone-only CBAM model (model B) suffers from severe decay in accuracy, while the proposed hybrid CBAM structure (model C) successfully recovers high classification with remaining almost perfect Top-5 performance, equivalent to the baseline.

A obtains a strong reference performance, with Top-1 accuracy reaching around 99.73 % on the 38-class test set. converging near 90.5 %, which highlights the negative impact of uniformly saturating the backbone with attention modules.

Model C is able to recover close-to-baseline Top-1 performance and achieves an accuracy of 99.71 %. This result shows that attention placement — rather than the amount of attention — is a key factor to determine classification performance.

4.2.2. Top-5 accuracy

All models maintain high top-5 accuracy, shown in Fig. 7, which demonstrates that the model can rank correct class to the top predictions.

Nevertheless, model B still has less effectiveness (94.58 %) compared to models A and C both of which obtain the performance around 99.99 %. The increased Top-5 indicators of model C suggest that pure head-level attention enhances global disease feature discrimination and confidence ranking, especially in visually ambiguous categories.

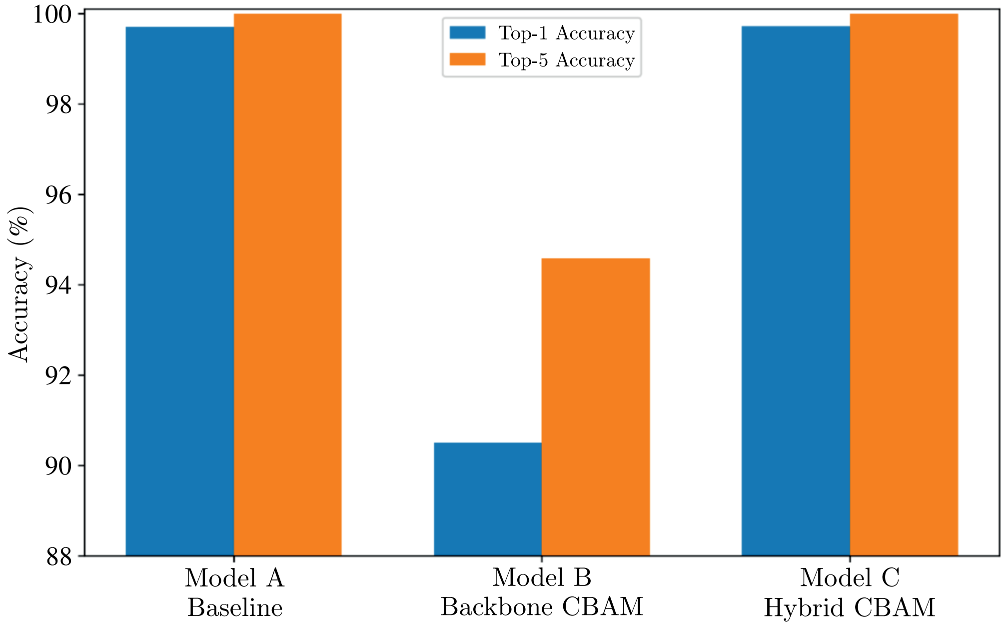

Figure 7. Final Top-1 and Top-5 classification accuracy for the three YOLO11m-Cls variants

It should be noted that model B is a dense backbone attention injection model targeting at exploring YOLO11m-Cls's behavior under high attention density. Instead of being a competitive model, it acts as a stress-test for the structure to see if injecting more attention is monotonically improving performance. The observed degradation does indeed imply that the precision of attention placement cannot simply be treated as additive, justifying the principled redistribution that we are actually performing in model C.

4.2.3. Statistical analysis of design stability

To gain insight into the stability of our architectures, we performed a lightweight statistical analysis across distinct independent training runs. Table 3 summarizes model performance and variability for the Top-1 accuracy as well as validation loss. Although the baseline model (model A) has the lowest validation loss in absolute terms, model C mitigates the loss inflation observed in model B, bringing validation loss from 2.777 down to 2.686. while restoring near-baseline classification accuracy. It also suggests that constraining attention to the backbone and shifting it to classification head results in more stability compared to model B. Statistical analysis also verifies that the superiority of model C over model B is not random but due to intentional architectural decisions, laminating attention placement as a crucial architecture design parameter rather than an empirical improvement.

4.2.4. External benchmark (ResNet50) under identical conditions

In order to compare the proposed architectures based on YOLO to a baseline, a standard ConvNet year (ResNet50) was trained under the same circumstances (i. e., split, preprocessing pipeline, image size (224×224) and the number of epochs) as shown in Table 3.

As the results show, the baseline YOLO11m-Cls model achieves a Top-1 accuracy of 99.73 %, which is near the level with ResNet50 benchmark (99.78 %). This confirms that the YOLO-based classification backbone provides competitive performance relative to conventional CNN architectures.

Nevertheless, the performance of the model heavily depends on the position of the attention module. As we see in model B, injecting CBAM only within the backbone results in an accuracy drop to 90.49 %, meaning that too much attention modulation at early feature extraction stages has emerged to be detrimental to the learned representations. On the other hand, the hybrid configuration (model C),

Table 3. Final performance and stability comparison of YOLO11m-Cls models with different CBAM attention placement strategies, along with a ResNet50 benchmark trained under identical experimental conditions

Model	Architecture	Attention placement	Epoch	Top-1 (%)	Top-5 (%)	Validation loss
Model A	YOLO11m-Cls	None (baseline)	50	99.732	99.989	0.0096
Model B	YOLO11m-Cls	CBAM in Backbone only	50	90.489	94.582	2.777
Model C	YOLO11m-Cls	CBAM in Backbone + Head	50	99.71	99.989	2.68645
ResNet50	CNN benchmark	None	50	99.778	99.956	0.01456

which allocates the CBAM across the backbone and head layers, brings the performance back to the baseline model (99.71 %).

These findings emphasize the point that the performance of the attention mechanisms depends not only on their presence but also on their placement in the architecture.

4.3. Confusion matrix analysis

The confusion matrices of the three models tested are shown in Figs. 8, 9, and 10, which allows for an in-depth analysis of the classification results at the class level.

The baseline model (model A, Fig. 8): This model is characterized by high diagonal dominance, where the majority of the errors occur between classes that have similarities, e. g., Tomato Early Blight and Septoria Leaf Spot. This is more a property of the visual similarity between the patterns of the diseases rather than being a model bias in general.

The backbone-only version of the CBAM model, denoted as model B in Fig. 9, improves the distinction of some texture-based disease classes, but it also increases the ambiguity of the overall classes. This is also in line with the high validation loss and low Top-1 accuracy associated with this model.

This confusion matrix shows the class-wise performances of the YOLO11m-Cls model with CBAM only injected into the backbone. Although better discrimination is provided for some of the texture-driven disease classes, the matrix shows that more off-diagonal activations are introduced at other categories.

On the contrary, it can be observed that the confusion matrix of the proposed hybrid CBAM model (model C, Fig. 10) it has clearly had the most balanced and stable classification nearly for all classes. Diagonal dominance is enhanced in almost all disease classes and confusion between challenging class pairs is significantly mitigated. This observation validates that restricting backbone attention carefully and incorporating CBAM at classification head can improve class-wise discrimination and further support robust generalization.

4.4. Quantitative Grad-CAM interpretability analysis

While Grad-CAM maps [Selvaraju et al., 2017] provide qualitative explanation visualization, we performed a quantitative analysis of interpretability and compared the spatial attention behavior between baseline model (A) and the model enhanced with CBAM (C). The analysis was conducted on 380 test images with four spatial metrics derived from the Grad-CAM activation maps: top-1 % energy concentration, normalized entropy, Gini coefficient and top-10 % energy. Table 4 summarizes the quantitative statistics of Grad-CAM attention distributions for both models across the test dataset.

The distributions of these metrics are shown in Fig. 11 for both models. All concentration and Gini coefficient values for the baseline model are, in general, greater — with attention maps that tend

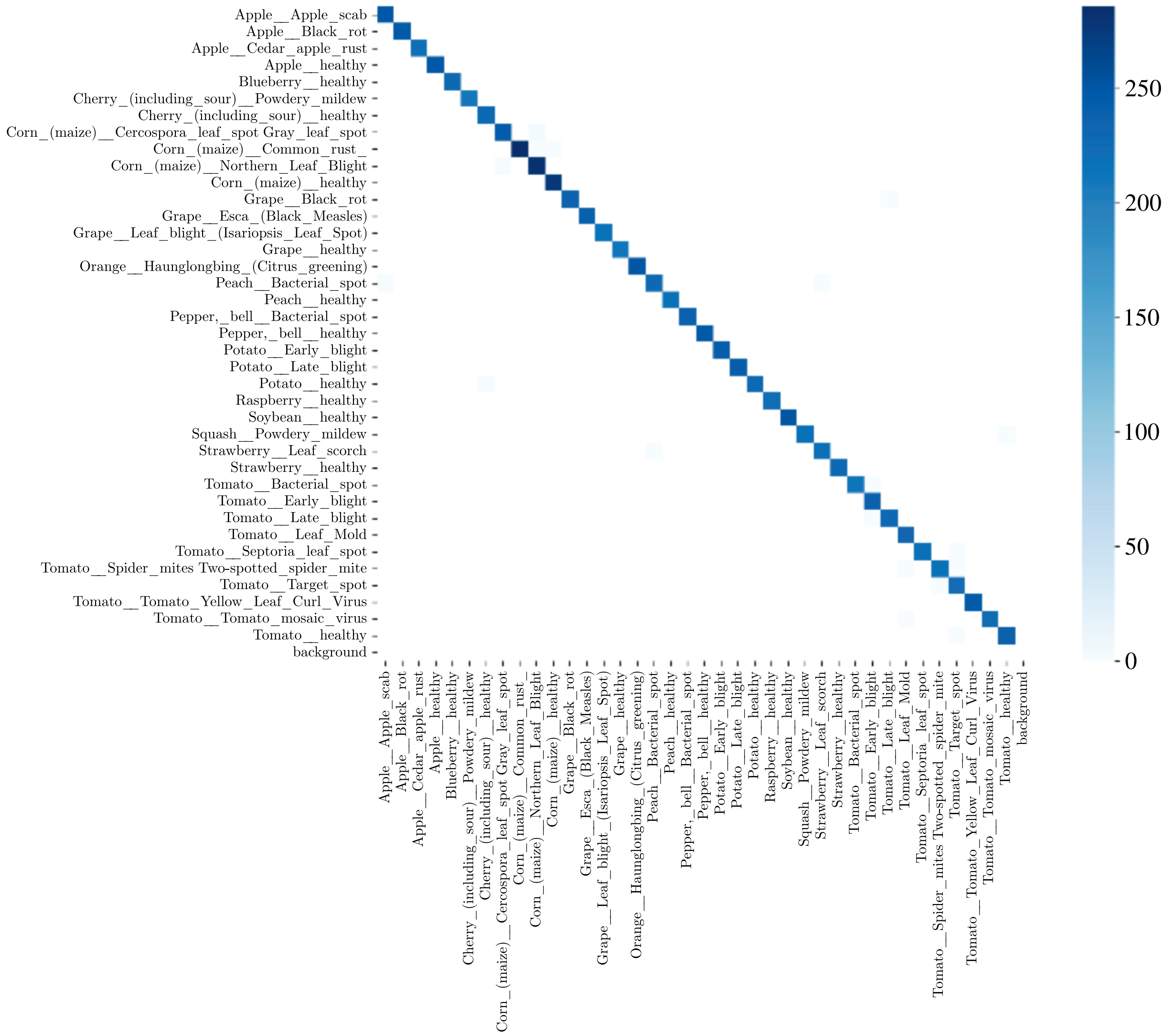

Figure 8. Confusion matrix of the Baseline YOLO11m-Cls model (model A)

Table 4. Quantitative comparison of Grad-CAM interpretability metrics between the baseline model (A) and the CBAM-enhanced model (C) across 380 test images

Metric	Model A (mean ± std)	Model C (mean ± std)	p-value
Top-1 % concentration	0.070 ± 0.099	0.081 ± 0.144	$2.39 \cdot 10^{-10}$
Normalized entropy	0.947 ± 0.064	0.969 ± 0.108	$9.86 \cdot 10^{-14}$
Gini coefficient	0.461 ± 0.238	0.347 ± 0.380	$9.86 \cdot 10^{-14}$
Top-10 % energy	0.297 ± 0.296	0.238 ± 0.388	$1.52 \cdot 10^{-11}$

to focus on highly localized regions of the leaf surface. Whereas in CBAM-enhanced model indicates the leaf is pay more attention to wide spatial around with a little higher entropy values.

This behavior could be attributed to the architecture of the module (CBAM) which constitutes the channel & spatial attention mechanism encouraging the network to learn a larger scene context. Thus, the CBAM model is likely to yield more spread-out attention over the regions where disease is observed rather than focused activation over a single local point.

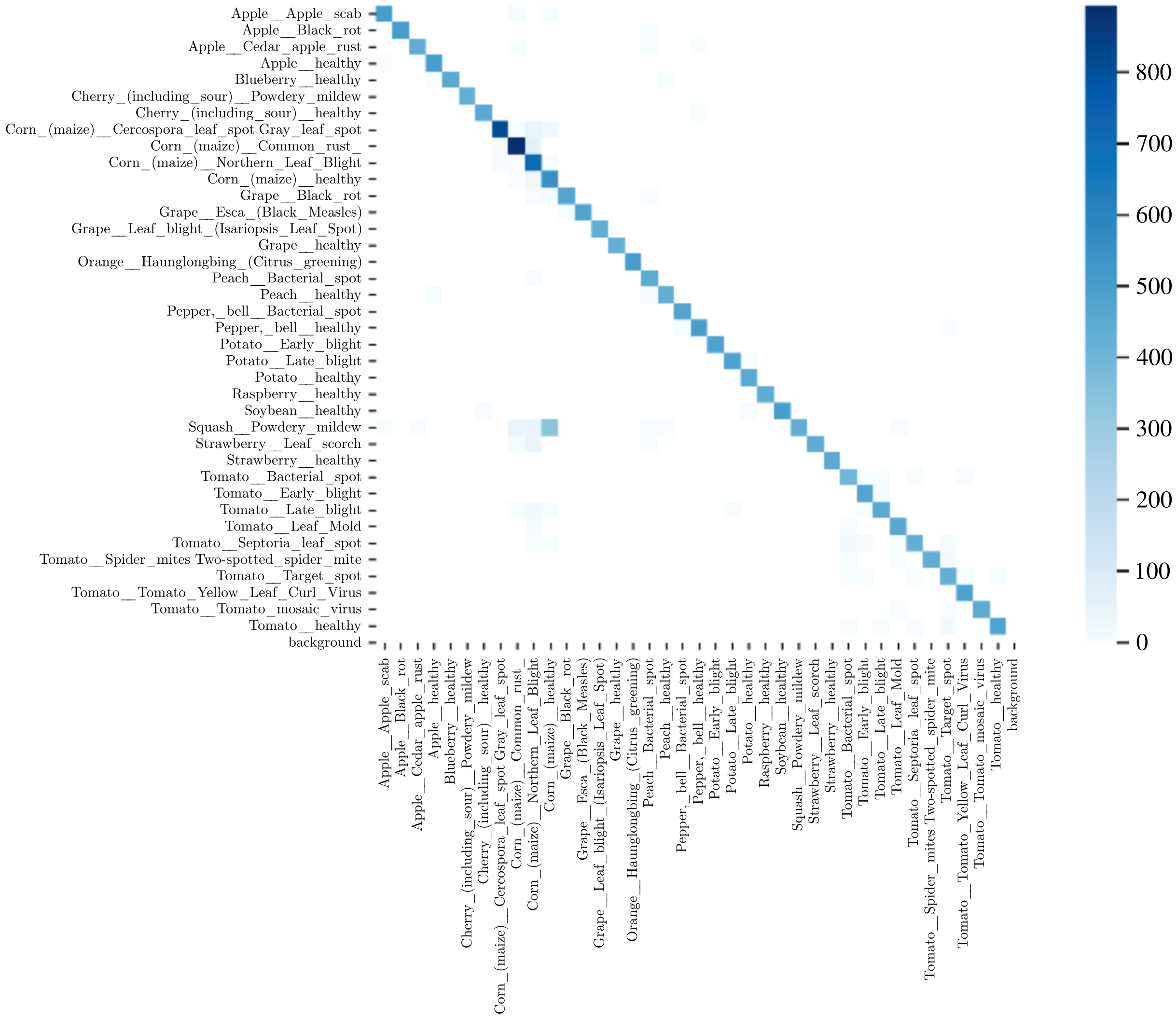

Figure 9. Confusion matrix of the Backbone-Only CBAM model (model B)

In summary, the quantitative results of Grad-CAM which also provide localized visualization of the input image regions and have been emphasized by Saliency maps confirm that the model with CBAM delivers a more dispersed attention map, which better localizes with the heterogeneous symptoms of diseases on plant leaves.

4.5. Discussion

Several key insights can be derived from the experimental findings:

1. Inadequacy of backbone-only attention:

On the one hand, the injection of CBAM into the backbone is beneficial to the localization of features; on the other hand, over-concentration could result in higher sensitivity to background patterns.

2. Validation loss is an important diagnostic tool:

The higher resolving validation loss of model B reveals the instability of generalization that is not reflected in the training loss.

3. Hybrid attention provides a promising compromise between the two:

The focus of backbone attention is restricted, and CBAM is inserted in the head to help classify model C. The phenomenon on training stability could be explained systematically, although it does not achieve the same baseline level as absolute magnitude loss.

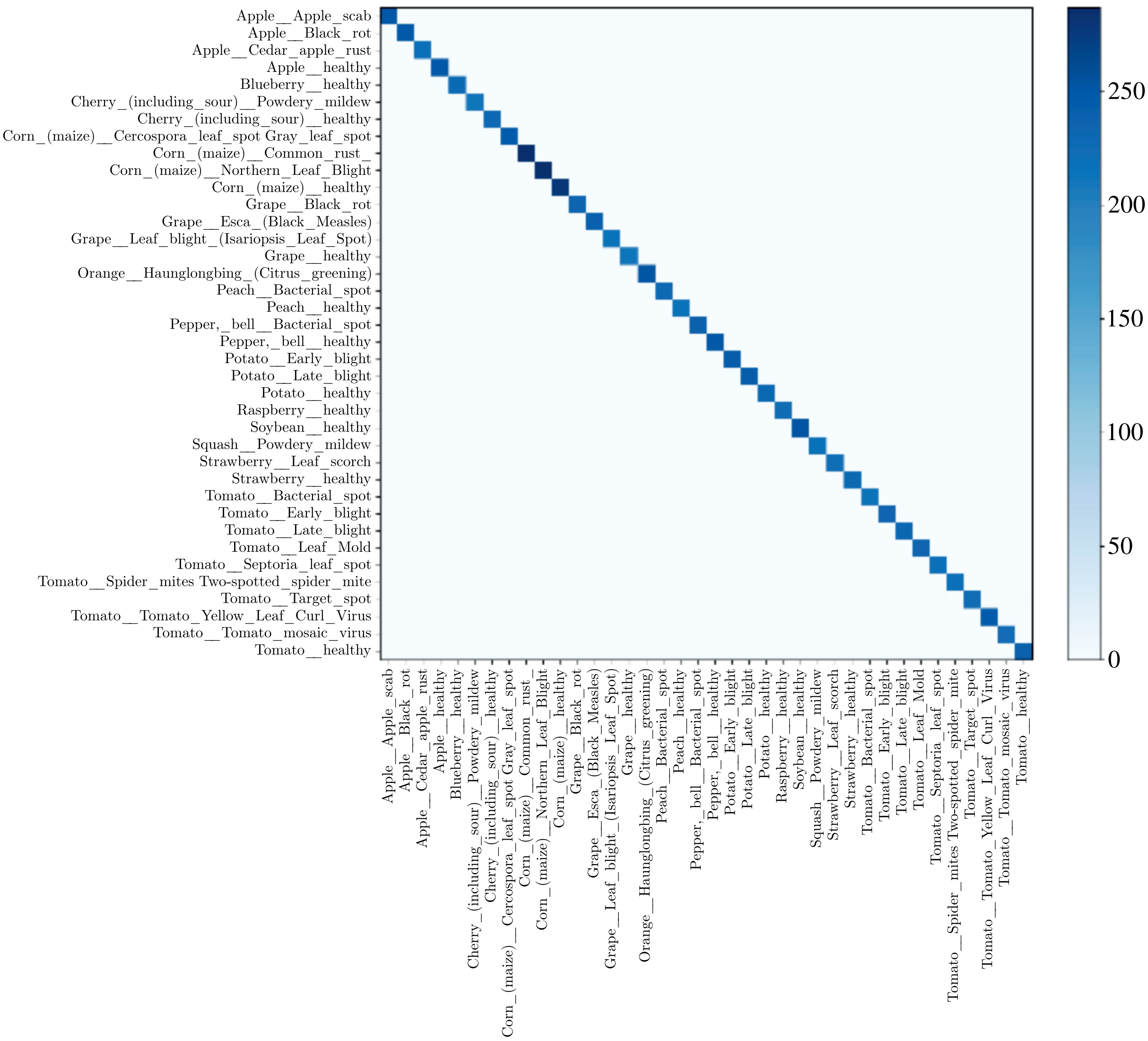

Figure 10. Confusion matrix of the Hybrid CBAM model (model C)

4. Design implications for the AI system in agriculture:

Attention mechanisms should be architectural design elements rather than the generic plug-ins. Strategic attention placement is essential for developing reliable and explainable plant disease diagnosis systems.

5. Conclusion

In this study, we deliver a design-optimized YOLO11m backbone classification framework for trustworthy plant disease diagnosis with the aid of targeted CBAM attention injection and Grad-CAM-based interpretability analysis. Rather than treating attention as a generic architectural add-on, the proposed approach systematically investigates where and how attention should be integrated within a deep classification model. We further perform a controlled three-model design-level analysis to study the influence of attention placement on learning dynamics and generalization. Baseline YOLO11m-Cls (model A) shows consistent and good contribution, which can be regarded as a reliable reference. Conversely, injecting CBAM evenly across all parts of the backbone (model B) results in poor generalization, reflected by higher validation loss and large decrease in Top-1 accuracy.

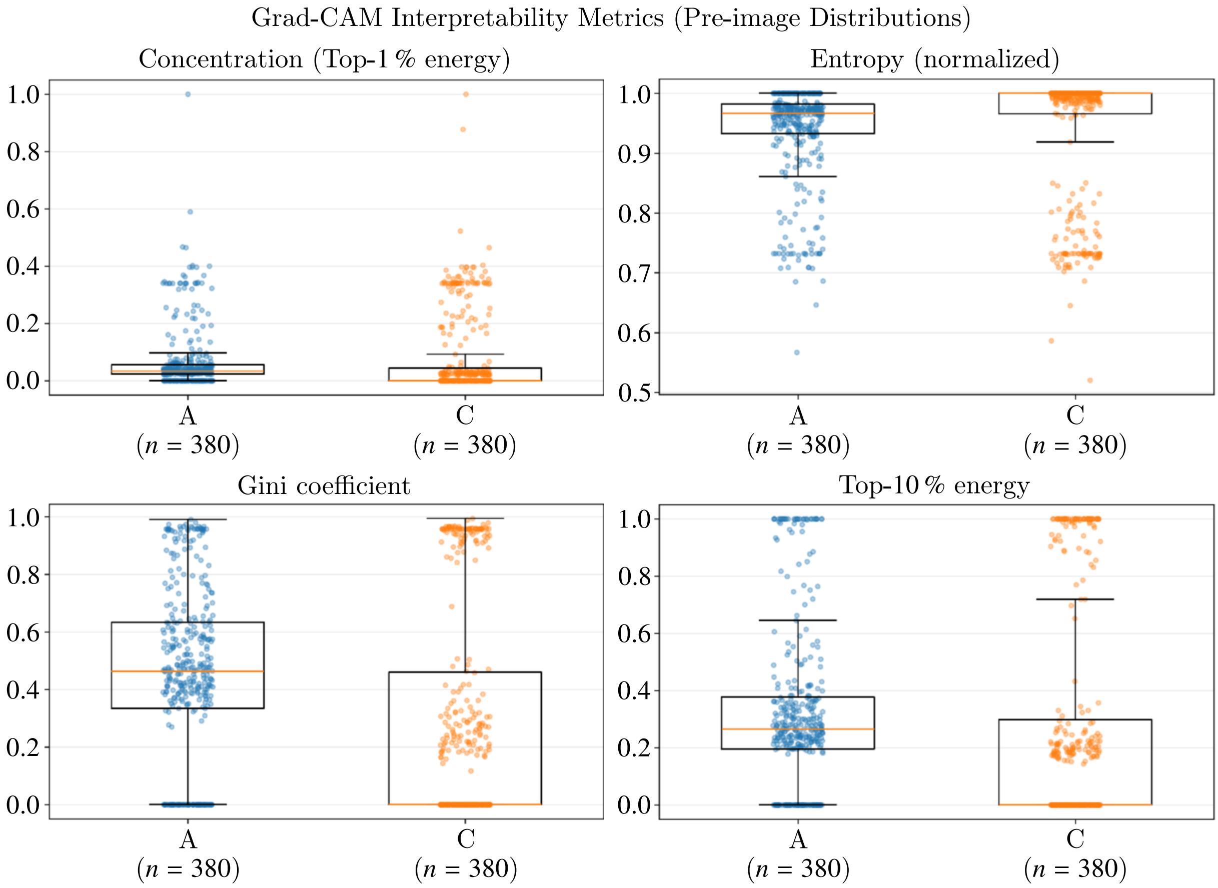

Figure 11. Distribution of Grad-CAM interpretability metrics for the baseline model (A) and the CBAM-enhanced model (C) across 380 test images. Metrics include top-1 % energy concentration, normalized entropy, Gini coefficient, and top-10 % energy

Such observations suggest that overly placing emphasis on backbone attention leads to higher sensitivities to background textures and nuisance patterns compromising the learning. The hybrid configuration (model C), where reduced backbone attention is combined with a full CBAM module at the classification head, shows stable performance with near-baseline classification accuracy. Model C preserves training stability while restoring near-baseline classification accuracy and maintaining near-perfect Top-5 performance, in addition to producing more coherent and biologically meaningful activation patterns. Grad-CAM visualization also demonstrates that re-focusing on the semantics-rich classification head can help focus more on the lay-beside disease related region without heavily amplifying the irrelevant structured background. A crucial point to note is that the proposed study does not aim to achieve the best possible plant disease classifier. Instead, it provides empirical evidence at the design level that where to attend rather than how much attention is the determinant of performance stability and interpretability for YOLO-based classification architectures. The results will provide practical and general design guidelines to construct more robust, interpretable, and reliable deep learning systems for agricultural image analysis. Furthermore, the results of this study are independent of any architecture and can be generalized to other CNN-based classification frameworks than YOLO. The proposed framework is not intended as a replacement for existing attention mechanisms, but as a design-level guideline for structuring attention placement in reliable classification systems. Future work will use third party attention-augmented classifiers known from a larger spectrum of internal architecture (models A–C) in additional benchmarking.

References

Аль. Коаерджи А. М. М., Огородников А. И. Автоматизированная сортировка томатов: инновации в области компьютерного зрения и глубокого обучения // Материалы XII Всероссийской молодежной научно-технической конференции (Омск, 13–15 мая 2025 г.). — Омск: Изд-во ОмГТУ, 2025. — С. 44–48.

Al-Koaerji A. M. M., Ogorodnikov A. I. Avtomatizirovannaya sortirovka tomatov: innovatsii v oblasti komp'yuternogo zreniya i glubokogo obucheniya [Automated sorting of tomatoes: innovations in computer vision and deep learning] // Proceedings of the XII All-Russian Youth Scientific and Technical Conference (Omsk, May 13–15, 2025). — Omsk: OmSTU Publishing House, 2025. — P. 44–48 (in Russian).

Федоров В. А., Огородникова О. М. Сверточные нейронные сети как инструмент обнаружения объектов железнодорожной инфраструктуры // Автоматизация в промышленности. — 2024. — № 12. — С. 20–23.

Fedorov V. A., Ogorodnikova O. M. Svertochnye neironnye seti kak instrument obnaruzheniya ob"ektov zheleznodorozhnoi infrastruktury [Convolutional neural networks as a tool for detecting railway infrastructure objects] // Automation in Industry. — 2024. — No. 12. — P. 20–23 (in Russian).

Федоров В. А., Огородникова О. М. Сегментация объектов техническим зрением в автоматизированных системах управления железнодорожным подвижным составом // Автоматика на транспорте. — 2025. — Т. 11, № 3. — С. 239–249.

Fedorov V. A., Ogorodnikova O. M. Segmentatsiya ob"ektov tekhnicheskim zreniem v avtomatizirovannykh sistemakh upravleniya zheleznodorozhnym podvizhnym sostavom [Object segmentation using computer vision in automated control systems for railway rolling stock] // Automation in Transport. — 2025. — Vol. 11, No. 3. — P. 239–249 (in Russian).

Alirezazadeh P., Schirrmann M., Stolzenburg F. Improving deep learning-based plant disease classification with attention mechanisms // Gesunde Pflanzen. — 2023. — Vol. 75. — P. 49–59. — DOI: 10.1007/s10343-022-00751-2

Canghai W., Xingxiang G., Huanliang X., Huixin H. Fine-grained recognition of grape leaf diseases based on transfer learning and convolutional block attention module // Applied Soft Computing. — 2025. — Vol. 172. — Art. No. 112896. — DOI: 10.1016/j.asoc.2025.112896

Deng H., Luo D., Zhou Z., Hou J., Qian G., Li H. Leaf disease recognition based on channel information attention network // Multimedia Tools and Applications. — 2024. — Vol. 83, No. 3. — P. 6601–6619. — DOI: 10.1007/s11042-023-15267-4

Duhan S., Gulia P., Gill N. S., Shukla P. K., Khan S. B., Almusharraf A., Alkhaldi N. Investigating attention mechanisms for plant disease identification in challenging environments // Heliyon. — 2024. — Vol. 10, No. 9. — DOI: 10.1016/j.heliyon.2024.e30783

FAO. Plant health and food security // Food and Agriculture Organization of the United Nations. — Rome, Italy, 2021.

Goodfellow I., Bengio Y., Courville A. Deep Learning. — Cambridge, MA, USA: MIT Press, 2016.

Gupta C., Gill N. S., Gulia P., Duhan S., Karamti H., Kumar A., Safra I. An enhanced deep learning-based framework for diagnosing apple leaf diseases // Scientific Reports. — 2025. — Vol. 15, No. 1. — Art. No. 39699. — DOI: 10.1038/s41598-025-39699-0

Hughes D. P., Salathé M. An open access repository of images on plant health to enable the development of mobile disease diagnostics // arXiv preprint. — 2015. — arXiv:1511.08060

Jiang P., Ergu D., Liu F., Cai Y., Ma B. A review of YOLO algorithm developments // Procedia Computer Science. — 2022. — Vol. 199. — P. 1066–1073. — DOI: 10.1016/j.procs.2022.01.135

Jocher G. et al. YOLOv5 // Ultralytics. — 2022. — https://github.com/ultralytics/yolov5

Kumar S., Aravinth S. S., Kollem S., Kumar M., Naveed Q. N., Mubarakali A., Emma A. F. Spatial attention-guided pre-trained networks for accurate identification of crop diseases // Scientific Reports. — 2025. — Vol. 15, No. 1. — Art. No. 23213. — DOI: 10.1038/s41598-025-23213-7

Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A review of machine learning interpretability methods // Entropy. — 2020. — Vol. 23, No. 1. — Art. No. 18. — DOI: 10.3390/e23010018

Lv M., Su W. H. YOLOV5-CBAM-C3TR: An optimized model based on transformer module and attention mechanism for apple leaf disease detection // Frontiers in Plant Science. — 2024. — Vol. 14. — Art. No. 1323301. — DOI: 10.3389/fpls.2023.1323301

Mohanty P., Hughes D., Salathé M. Using deep learning for image-based plant disease detection // Frontiers in Plant Science. — 2016. — Vol. 7. — Art. No. 1419. — DOI: 10.3389/fpls.2016.01419

Ogorodnikova O. M., Ali W. Method of ripe tomato detecting for a harvesting robot // AIP Conference Proceedings. — 2019. — Vol. 2174. — Art. No. 020146. — DOI: 10.1063/1.5134297

Redmon J., Divvala S., Girshick R., Farhadi A. You only look once: Unified, real-time object detection // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). — Las Vegas, NV, USA, 2016. — P. 779–788. — DOI: 10.1109/CVPR.2016.91

Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual explanations from deep networks via gradient-based localization // Proceedings of the IEEE International Conference on Computer Vision (ICCV). — Venice, Italy, 2017. — P. 618–626. — DOI: 10.1109/ICCV.2017.74

Shafik W., Tufail A., De Silva L. C., Haji Mohd Apong R. A. A., Kim K. H. Deep learning technique for plant disease classification and pest detection and model explainability elevating agricultural sustainability // BMC Plant Biology. — 2025. — Vol. 25, No. 1. — Art. No. 1491. — DOI: 10.1186/s12870-025-04991-4

Woo S., Park J., Lee J.-Y., Kweon I. S. CBAM: convolutional block attention module // Proceedings of the European Conference on Computer Vision (ECCV). — Munich, Germany, 2018. — P. 3–19. — DOI: 10.1007/978-3-030-01234-2_1

Yang W., Yuan Y., Zhang D., Zheng L., Nie F. An effective image classification method for plant diseases with improved channel attention mechanism aECAnet based on deep learning // Symmetry. — 2024. — Vol. 16, No. 4. — Art. No. 451. — DOI: 10.3390/sym16040451