Journal of Agricultural Sciences (JAS) 32(2):474-497
DOI: 10.15832/ankutbd.1739676

J Agr Sci-Tarim Bili
e-ISSN: 2148-9297
jas.ankara.edu.tr

LAViT: Class-Aware Vision Transformer with Learnable Attention for Pest Recognition in Agricultural Fields

Canan Tastimur[a], Volkan Kaya[a*]

[a]*Department of Computer Engineering, Faculty of Engineering and Architecture, Erzincan Binali Yıldırım University, Erzincan, TÜRKIYE*

ARTICLE INFO
Research Article
Corresponding Author: Volkan Kaya, E-mail: vkaya@erzincan.edu.tr
Received: 10 July 2025 / Revised: 18 December 2025 / Accepted: 31 December 2025 / Online: 24 March 2026

Cite this article
Taştimur C, Kaya V (2026). LAViT: Class-Aware Vision Transformer with Learnable Attention for Pest Recognition in Agricultural Fields. *Journal of Agricultural Sciences (Tarim Bilimleri Dergisi),* 32(2):474-497. DOI: 10.15832/ankutbd.1739676

ABSTRACT
Insect pests pose a significant threat to agricultural productivity, making early and accurate identification essential for effective pest management. This study proposes a novel deep learning-based classification framework for multi-class pest recognition from field images. The proposed approach enhances discriminative region representation by integrating a patch embedding module, a Vision Transformer backbone, and a learnable spatial attention mask. This hybrid design enables the model to focus on critical visual cues without requiring segmentation-based preprocessing. The attention mask, learned via convolutional layers, is pooled and directly applied to Transformer-encoded patch tokens to refine spatial feature emphasis. Positional embeddings are further employed to preserve spatial context within the tokenized image representation. Experimental evaluations conducted on publicly available pest datasets with 5, 9, and 12 classes demonstrate the effectiveness and robustness of the proposed framework. The model achieves accuracies of 99.67%, 99.52%, and 97.00% on the Pest5, Pest9, and Pest12 datasets, respectively, indicating strong generalization across varying classification complexities. To enhance model transparency and reliability, visual interpretability is provided through Grad-CAM and attention heatmap visualizations that reveal the model's focus regions. Additionally, t-SNE-based feature visualization illustrates clear separability in the learned embedding space. The proposed framework shows strong potential for practical deployment in smart agriculture and precision pest monitoring systems.

Keywords: Pest Recognition, Vision Transformer, Learnable Attention, Class-Aware Feature Learning, Explainable AI, Smart Agriculture

1. Introduction

Agricultural productivity plays a vital role in global food security and economic development. However, pest infestations pose a significant threat to crop yields and quality. Early and accurate identification of pest species is crucial for the timely implementation of integrated pest management (IPM) strategies. Manual pest identification methods are often labor-intensive and require specialized knowledge that may not be readily available, especially in rural areas or large-scale agricultural operations. Consequently, steam has generated strong interest in automatically detecting pests with the goal to provide reliable, fast and scalable solutions that are adapted to real-field agriculture (Li et al. 2020). Recent developments in deep learning have made a huge difference in image classification tasks, such as agricultural image classification. Convolutional Neural Networks (CNNs) have proven effective in identifying plant diseases and pests by automatically learning distinctive features from raw images, eliminating the need for manual feature extraction (Mohanty et al. 2016; Fuentes et al. 2017).

However, CNNs have some notable limitations. Fuentes et al. (2017) developed a detection system for tomato crops using a region-based CNN (R-CNN) framework. Training on 13,000 labeled field images, their model could detect and classify multiple issues within the same image, achieving a mean Average Precision (mAP) of 81.1%. Similarly, Mohanty et al. (2016) investigated CNNs for plant disease diagnosis using a large publicly available dataset containing 54,306 images across 38 disease categories. Their experiments with models like AlexNet and GoogLeNet in controlled lab settings yielded accuracy rates as high as 99.35%.

Building on these, Li et al. (2020) designed a dual-path CNN that separates texture and contour features to improve classification. Ferentinos (2018) proposed an Transfer Learning CNN and tested it on a subset of 38 real-world agricultural images from the pest dataset, which contains 87,848images across 38 pest classes. Their model achieved accuracy of 97%. In line with this, Jelali (2024) explored with deep CNNs to enhance tomato disease and pest detection accuracy.

Table 1- Review of Deep Learning Methods Applied to Agricultural Pest Classification

Study	*Dataset*	*Model*	*Accuracy*
Mohanty et al. 2016	54,306 images	CNN (AlexNet)	99.35%
Fuentes et al. 2017	13,000 images	R-CNN	81.1%
Li et al. 2020	5629 images	GoogLeNet	98.0%
Ferentinos 2018	87,848 images	Transfer Learning	97%

Vision Transformers (ViTs) have recently attracted a lot of interest as an alternative to conventional CNNs for analyzing images. Instead of scanning the entire image at once, ViTs split it into smaller fixed-size patches and use self-attention layers to understand the connections across these patches. This way, they capture global relationships in the image, unlike CNNs which primarily focus on local features through receptive fields (Dosovitskiy et al. 2021). Still, ViTs have some inherent drawbacks: they don't naturally include the concepts of locality or translation invariance, which CNNs benefit from and which are important for many vision problems. Moreover, ViTs give equal importance to every patch, which can be a problem when trying to zero in on pest regions in complex agricultural images cluttered with irrelevant background details (Qian et al. 2022). To tackle these issues, Kim et al. (2025) designed the Region-of-Interest Vision Transformer (ROI-ViT), a version of ViT that adaptively focuses on image areas more likely to contain pests. This focused approach boosted the model's effectiveness significantly, resulting in impressive accuracy scores 81.81% on IP102, 99.64% on D0, and 84.66% on SauTeg outperforming standard ViTs on pest detection tasks. Besides accuracy, making these models' decisions understandable is another challenge.

Hu et al. (2023) developed a multi-scale ViT model that processes visual information at various resolutions simultaneously using parallel pathways. They tested it on a rice pest dataset and found it to be robust, achieving 79.8% accuracy even with low-quality or noisy images challenges commonly encountered in real-world farming environments. Similarly, Murugavalli & Gopi (2025) designed a ViT-based classifier aimed specifically at identifying plant leaf disease. Their results suggest that this method has practical potential for on-site agricultural pest monitoring. Sharma et al. (2025) proposed a hybrid architecture called HyPest-Net, combining CNN layers to capture low-level features with ViT modules responsible for higher-level semantic understanding. HyPest-Net demonstrated strong performance, reaching an accuracy of 95%, highlighting its ability to capture both fine-grained texture and overall image context.

Saranya et al. (2024) proposed a ViT model with a Hybrid Pooled Multihead Attention mechanism, which effectively captures both local and global features for more precise pest classification. This approach achieved high accuracy across multiple pest datasets, showcasing the potential of attention-enhanced ViT models for agricultural applications. Another significant development in pest classification is the GNViT model proposed by Venkatasaichandrakanth & Iyapparaja (2024), which adapts Vision Transformers to classify groundnut pests. The model demonstrated robust performance on the IP102 dataset, confirming the effectiveness of ViT-based models for pest recognition tasks in specific agricultural domains. Dinca et al. (2025) proposed an ensemble approach combining multiple ViT models to improve insect pest classification. By leveraging a meta-classifier, the method demonstrated significant performance improvements, particularly in complex pest classification tasks across various datasets. This highlights the value of combining ViT models for more robust agricultural pest recognition.

To address existing challenges, our work proposes a novel Vision Transformer architecture enhanced with a learnable attention-guided module (LAViT) designed for multi-class pest classification. Unlike previous methods that require explicit segmentation, our model learns a spatial attention mask directly from data, guiding focus toward biologically important regions. This attention mechanism enables robust feature extraction even in noisy or cluttered backgrounds. We also incorporate positional embeddings to preserve spatial structure, and utilize Grad-CAM visualizations to enhance interpretability, providing transparency into how the model makes its decisions.

This work makes several important contributions: (1) It proposes an innovative transformer architecture enhanced with attention mechanisms that enables pest classification without relying on explicit segmentation steps; (2) it incorporates a learnable attention mask that sharpens the model's focus on relevant spatial regions and increases interpretability; (3) it validates the approach using a real-world dataset containing nine different pest categories; and (4) it utilizes visualization tools such as Gradient-weighted Class Activation Mapping (Grad-CAM) and attention heatmaps to provide insights into the model's decision-making process. The results of the experiments show that the model achieves high accuracy and maintains strong generalization capabilities, positioning it as a valuable candidate for smart agriculture solutions.

2. Material and Methods

In this study, we propose a new pest classification method based on the ViT model. To improve its focus, we add a learnable attention mask that helps the model highlight areas in the image where pests are likely to appear, even when the background is complex. We chose a transformer-based model because it can capture long-range relationships in the image, which is important for pest datasets that often include different backgrounds, lighting conditions, and detailed scenes.

The model starts by dividing each image into small, fixed-size patches. Each patch is then turned into a vector using a patch embedding layer. To keep track of the position of each patch, the model adds positional information. The patch embeddings are processed by multiple transformer layers, each comprising self-attention mechanisms and feed-forward networks. This structure allows the model to capture relationships between different regions of the image, which is especially important when distinguishing between pest species that have very similar appearances.

A core innovation in our approach is the introduction of a learnable attention mask, shown in Fig. 1. This mask is produced by a lightweight convolutional subnetwork that operates alongside the main model, learning to highlight the image regions most relevant to pest detection. After being generated, the mask is adjusted in size to align with the dimensions of the patch embeddings. It is then combined with the transformer's output through element-wise multiplication, effectively enhancing the focus on pest-related features while suppressing the noise from less important background areas.

The attention-enhanced features are first normalized before being fed into a global pooling layer, which compresses the feature set while retaining the most important information. This pooled output then passes through several fully connected layers, beginning with a dense layer activated by ReLU and accompanied by batch normalization. Finally, a softmax layer generates the probability scores for each class. To train the model, we use sparse categorical cross-entropy loss, which is suitable for multi-class tasks with integer-based labels. During training, we apply data augmentation to increase the model's generalization and reduce overfitting. For evaluation, we report accuracy, precision, recall, F1-score, and AUC. The results show that our approach performs better than existing methods on standard pest datasets and can be useful in smart farming applications.

2.1. Dataset description

This work evaluates the LAViT model using three pest image datasets of increasing complexity: Pest5, Pest9, and Pest12. These datasets include 5, 9, and 12 pest classes, respectively, and differ in terms of class diversity and visual difficulty. Pest5 is relatively simple, as it includes clearly distinguishable pest types. Pest9 introduces more variation within classes and greater similarity between some categories. Pest12 is the most complex, featuring a wider range of pest species captured under different environmental conditions. Using these datasets allows us to test how well the model handles different levels of difficulty and whether it can maintain strong performance across varied, real-world agricultural scenarios.

Pest5 Dataset: One of the datasets used in this study is the Crop Pest Dataset (Pest5), which includes 295 images of five key pest species that typically harm rice crops: rice stem borer, green leafhopper, rice insect, and rice leaf roller. Despite its relatively small size, this dataset serves as a useful benchmark for evaluating model performance under low-data conditions and supports the development of lightweight deep learning solutions for insect recognition. The dataset is publicly available on Kaggle (Ghosh, 2021) and forms the basis of this study. Figure 2 presents representative samples from the Pest5 dataset, illustrating the visual diversity and distinctive morphological characteristics of each pest category, which are crucial for effective model training and evaluation. To mitigate potential overfitting issues arising from the limited dataset size and to enhance the model's generalization capability, extensive data augmentation was applied. Through augmentation techniques, each class was expanded to 500 images, resulting in a total of 2,500 images for the Pest5 dataset. The model was trained using a 60% training, 20% validation, and 20% test split, ensuring a balanced and robust evaluation protocol.

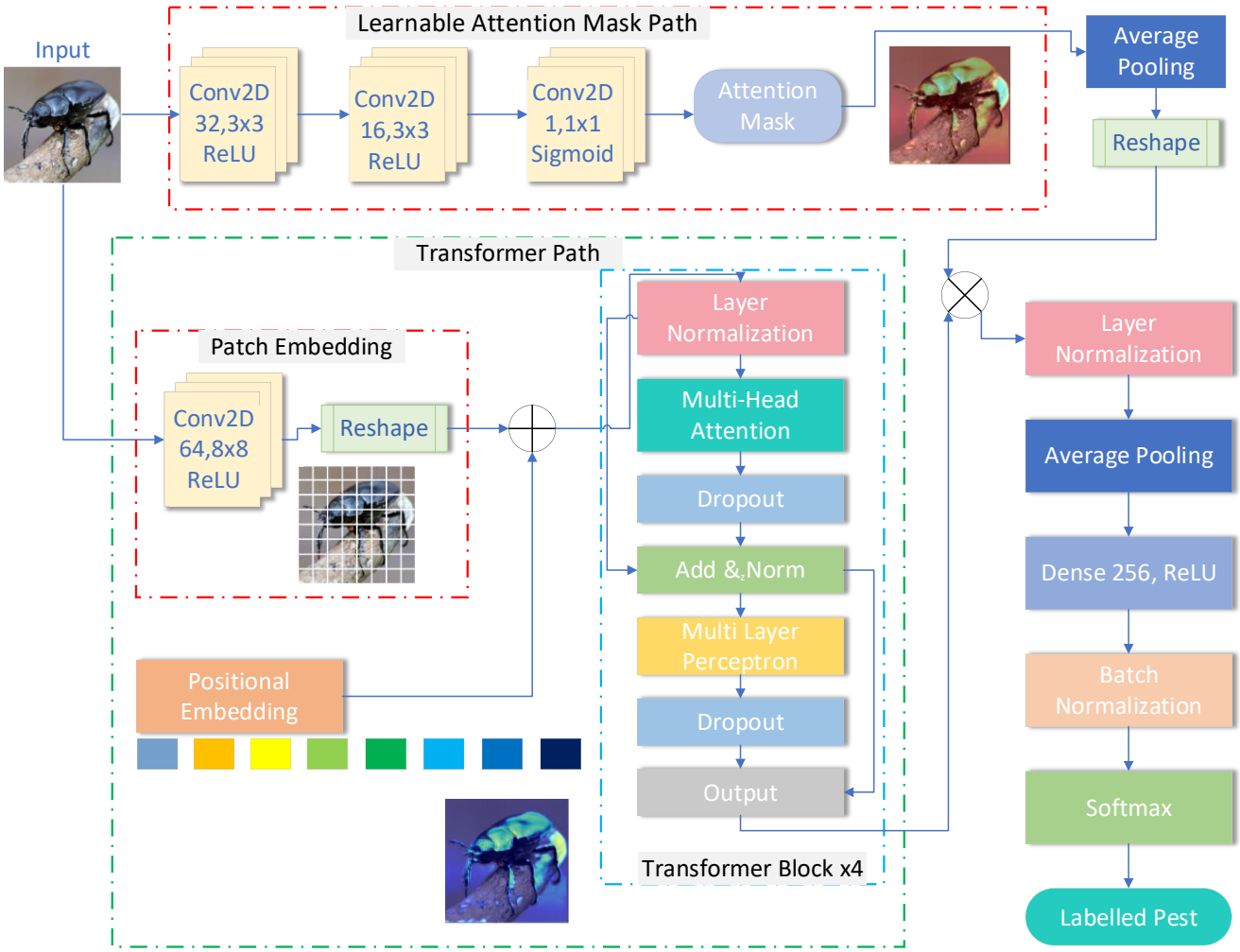


Figure 1- The overview of the LAViT approach

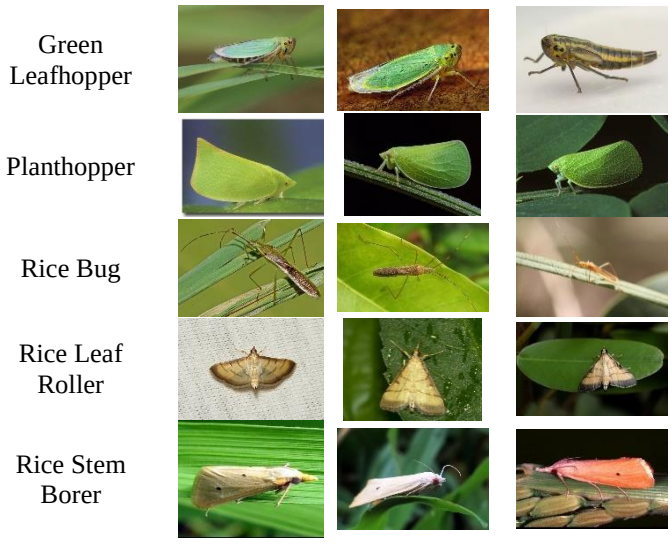


Figure 2- Representative samples of pest species included in the Pest5 Dataset

Pest9 Dataset: This dataset includes high-resolution images of nine different pest species: aphids, armyworm, beetle, bollworm, grasshopper, mites, mosquito, sawfly, and stem borer. It comprises a total of 2,700 training images, with approximately 300 images per pest class, and 450 testing images, with 50 images per class (SimranVolunesia, 2023). Figure 3 presents representative images of the pest species included in the Pest9 Dataset. Data augmentation was employed to address

class imbalance and limited sample size issues. Each pest category was expanded to contain 500 images, leading to a total of 4,500 images across nine pest classes. The dataset was partitioned using the same protocol as Pest5, with 60% of the samples allocated for training, 20% for validation, and 20% for testing. This consistent split strategy allows for a fair comparison of model performance across datasets with varying numbers of classes.



Figure 3- Representative samples of pest species included in the Pest9 Dataset

Pest12 Dataset: The Agricultural Pest Image Dataset is a curated image collection featuring twelve different types of agricultural pests, including ants, bees, beetles, caterpillars, earthworms, earwigs, grasshoppers, moths, slugs, snails, wasps, and weevils. It comprises a total of 5,494 training images and 1,647 validation images. All images were scaled to a maximum width and height of 300 pixels, resulting in a lightweight dataset that is computationally efficient for training and testing. Its class diversity supports robust model evaluation across a range of morphological features such as shape, size, and color (Vencerlanz, 2022). Figure 4 presents representative images of the pest species included in the Pest12 Dataset. For the Pest12 dataset, which represents a more challenging multi-class classification scenario, data augmentation was again applied to improve robustness and reduce overfitting. Each of the twelve pest classes was expanded to 1000 images, yielding a total of 12,000 images. The dataset was divided into training, validation, and test sets following the same 60/20/20 ratio. This setting enables the evaluation of the scalability and generalization ability of the proposed model as the number of classes increases.

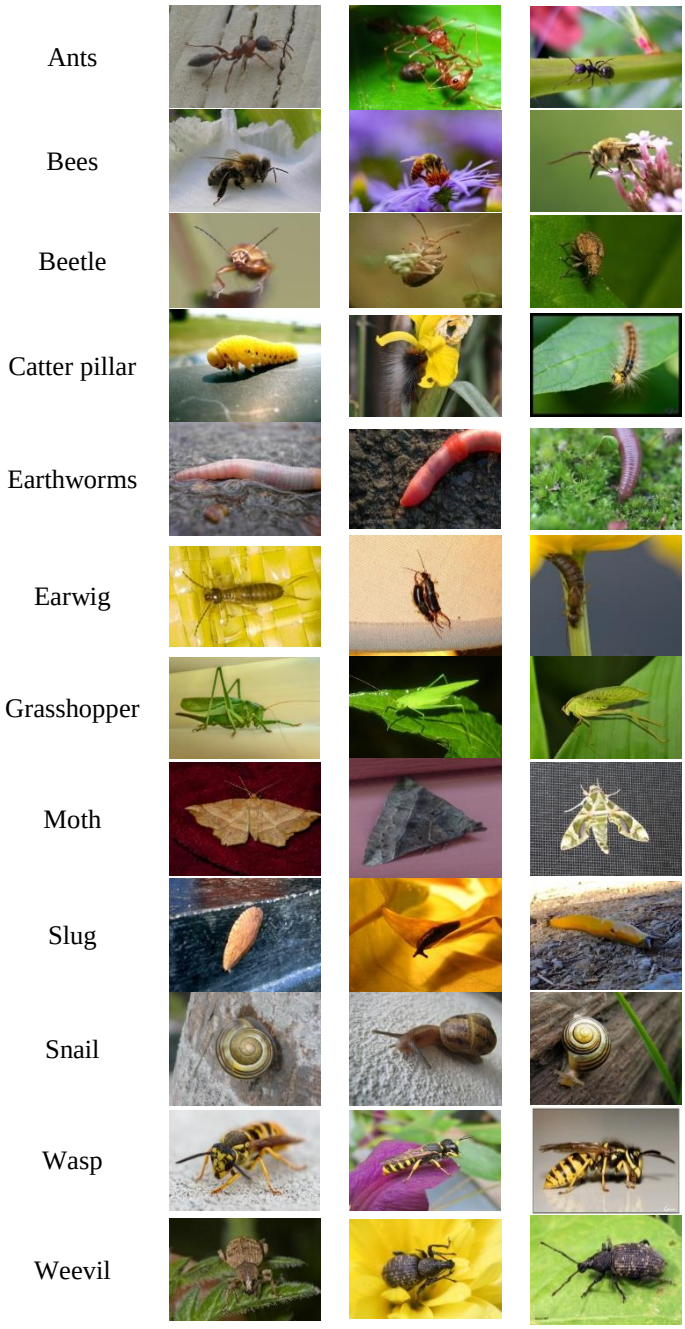


Figure 4- Representative samples of pest species included in the Pest12 Dataset

2.2. Proposed model

In this study, a deep learning-based model is proposed to achieve accurate and robust multi-class insect pest recognition from field images. The proposed framework integrates a patch embedding module, a Vision Transformer backbone, and a learnable spatial attention mechanism to effectively capture discriminative visual patterns associated with different pest species. By leveraging attention-guided feature refinement, the model is able to emphasize pest-relevant regions while suppressing background noise, without requiring explicit segmentation. Experimental results demonstrate high classification performance across multiple datasets with varying numbers of pest classes, highlighting the robustness and generalization capability of the proposed approach. These findings indicate that the proposed model has strong potential as a reliable tool for intelligent pest monitoring and smart agriculture applications.

This paper presents a unique deep learning architecture for pest image categorization that takes advantage of transformer-based attention mechanisms and a learnable spatial attention mask. The entire pipeline was carefully built to incorporate efficient patch-based processing, deep feature learning, and attention-guided localization to improve pest detection accuracy and interpretability. The methodology consists of picture preprocessing, learnable attention map generation, patch embedding, Transformer encoding, attention-based feature modulation, and final classification inside a single, integrated framework. The

input data comprises of RGB pest images scaled to 128×128 pixels. Each image is normalized to the range [0,1] to guarantee numerical consistency between layers. Labels are automatically taken from the directory structure and encoded in a categorical manner appropriate for multi-class categorization. Let $X \in \mathbb{R}^{NxHxWxC}$ denote the input batch of images, where N is batch size, $H = W = 128$, and $C = 3$.

This study's primary innovation is the learnable attention mask, which is intended to concentrate spatially significant regions particularly the pest in each image. This module uses a shallow convolutional neural subnetwork with two 3×3 layers, followed by a 1×1 convolution with sigmoid activation to generate a single-channel mask $A \in \mathbb{R}^{HxWx1}$. This attention mask emphasizes the pest locations while reducing background noise. The mathematical definition of the attention mask is as follows in (1).

$$A = \sigma(Conv_{1x1}(ReLU(Conv_{3x3}(X)))) \quad (1)$$

In (1), the sigmoid activation is indicated by σ. The internal transformer properties are subsequently modulated by this spatial map to direct learning toward more instructive regions. The image is separated into non-overlapping 8x8 patches after attention extraction, yielding 16x16=256 patches per image. A patch embedding tensor $Z_0 \in \mathbb{R}^{256xD}$, where D=64 is the embedding dimension, is created by converting these patches into tokens using an 8×8 convolution layer with stride 8. We provide each patch vector a trainable positional embedding $P \in \mathbb{R}^{1x256xD}$ to maintain spatial structure like in (2).

$$Z = Z_0 + P \quad (2)$$

Following that, the sequence of patch embeddings is processed via four Transformer encoder blocks, each of which consists of a feed-forward MLP, normalization layer, and a multi-head self-attention layer. In (3), the self-attention mechanism successfully models global dependencies by capturing relationships between remote regions of the image.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3)$$

Where: d_k is the dimension of the key vectors and Q, K, V are the query, key, and value matrices obtained from linear projections of input tokens. To prevent overfitting, each Transformer block is additionally outfitted with dropout regularization and residual connections. In (4), each block's feed-forward network (FFN) consists of two linear layers that map from D to a higher-dimensional space (128) and back again via GELU activation.

$$MLP(x) = W_2.GELU(W_1 x) \quad (4)$$

The way the attention mask interacts with the Transformer output is a new addition to this model. The mask M is down sampled using average pooling to match the token grid (16×16), reshaped into a vector $M \in \mathbb{R}^{256x1}$, then multiplied element-wise with the Transformer output in order to include spatial attention into token-based learning in (5).

$$Z' = Z \odot M \quad (5)$$

By ensuring that only the characteristics associated with pest-relevant patches are highlighted, this modulation helps to better direct the classification head. In (5), Z' the output sequence, is sent to a final classification head. Layer normalization, two fully connected layers, and a global average pooling algorithm to aggregate token information make up this head. Batch normalization comes after the first dense layer (with 256 units and ReLU activation) records non-linear feature interactions. To generate class probabilities, the last layer uses softmax activation like in (6). During training, the categorical cross-entropy loss is minimized as in (7).

$$\hat{y} = softmax(W.ReLU(GlobalAvgPool(Z'))) \quad (6)$$

$$\mathcal{L} = -\sum_{i=1}^{C} y_i \log(\hat{y}_i) \quad (7)$$

In (7), C represents the number of pest classes, and y_i is the one-hot encoded true label. The model is built in TensorFlow/Keras and trained using the Adam optimizer. The hyperparameters include a batch size of 32, a learning rate of 0.001, and 100 training epochs. Performance indicators such as accuracy, precision, recall, and F1-score are used to assess classification performance.

Overall, the proposed model introduces two significant innovations: (1) the incorporation of a learnable attention mask that improves both interpretability and feature enhancement, and (2) the modulation of Transformer tokens with this mask, allowing for the lightweight and effective fusion of spatial and semantic information. This technique achieves cutting-edge performance on pest datasets, highlighting the importance of attention-based guiding in agricultural visual categorization.

2.3. Data preprocessing

All pest images were organized into class-specific folders and loaded using a custom Python function. Each image was read using OpenCV, converted from BGR to RGB format, and resized to 128x128 pixels to maintain uniformity. The pixel values were normalized to the [0, 1] range by dividing by 255. The corresponding labels were extracted from the folder names and encoded using *Label Encoder* to transform categorical pest class names into integer class indices. This standardized preprocessing ensures that the input data is compatible with convolutional and transformer-based deep learning models.

2.4. Training configuration

The LAViT model was trained using a set of hyperparameters selected specifically for multi-class pest classification, as summarized in Table 2. An exhaustive hyperparameter grid search was not conducted due to computational constraints. Instead, the hyperparameters were chosen based on commonly adopted values in related Vision Transformer-based studies and refined through preliminary experiments conducted on the validation set. During this preliminary tuning process, the learning rate was evaluated in the range of $\{1\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}\}$, batch sizes of {32, 64, 128} were examined, and the number of training epochs was varied between 20 and 100. The final configuration was selected based on convergence behavior and validation performance. To improve computational efficiency, all input images were resized to 128 × 128 pixels. The model was trained for 100 epochs with a batch size of 32, which provided a good balance between performance and memory usage. Each image was divided into 8 × 8 patches and projected into 64-dimensional feature embeddings. The Transformer backbone consisted of four layers with four attention heads per layer, followed by a fully connected layer with 128 units for feature refinement. The Adam optimizer was employed with an initial learning rate of 0.001 due to its adaptive optimization capability. To mitigate overfitting, a dropout rate of 0.6 was applied within the Transformer layers.

Table 2- Parameter Configuration of LAViT

Parameter	***Value***
Image size	128x128
Batch size	32
Epochs	100
Patch size	8
Embedding dimension	64
MLP dimension	128
Number of attention head	4
Number of transformer layer	4
Optimizer	Adam
Learning rate	0.001
Loss function	Categorical cross-entropy
Dropout	0.6

2.5. Evaluation metrics

Several performance measures were used to assess the overall effectiveness of the LAViT model for pest classification. These include accuracy, precision, recall, F1-score, confusion matrix, Receiver Operating Characteristic (ROC) curve with Area Under the Curve (AUC), and two-dimensional visualization techniques like t-distributed Stochastic Neighbor Embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP). In (8), the accuracy metric calculates the overall percentage of properly identified examples across all classes, offering a high-level summary of the model's classification performance. Precision, defined in (9), is the ratio of accurately predicted positive cases to total expected positive instances. It shows the model's ability to reduce false positives. Recall, explained in (10), is the ratio of accurately predicted positive cases to all actual positive instances, indicating how well the model detects relevant positive cases. The AUC curve evaluates the model’s capacity to distinguish between positive and negative classes; a higher AUC means better class separation as in (13). A confusion matrix summarizes classification results by showing true positives, false positives, true negatives, and false negatives, providing a clearer understanding of the model’s behavior. In summary, t-SNE is a technique used to transform complex, high-dimensional data into a simpler, lower-dimensional form, making it easier to visually assess how effectively the model separates different categories. Similarly, UMAP also reduces dimensionality but offers better preservation of the data’s overall structure, helping to clearly display how the model distinguishes between classes.

When evaluating classification results, True Positives (TP) refer to instances where the model correctly identifies a pest as belonging to its true class, while True Negatives (TN) are cases where images without the target pest are correctly recognized as negative. False Positives (FP) occur when non-pest images are incorrectly classified as containing a pest, and False Negatives (FN) happen when the model fails to detect a pest present in the image. The True Positive Rate (TPR, Equation 11) reflects how often the model successfully identifies actual pests, whereas the False Positive Rate (FPR, Equation 12) indicates the proportion of non-pest images incorrectly classified as pests, representing the model’s rate of false alarms.

$$Accuracy = \frac{TP + TN}{(TP + TN + FP + FN)} \quad (8)$$

$$Precision = \frac{TP}{(TP + FP)} \quad (9)$$

$$Recall = \frac{TP}{(TP + FN)} \quad (10)$$

$$TPR = \frac{TP}{(TP + FN)} \quad (11)$$

$$FPR = \frac{FP}{(FP + TN)} \quad (12)$$

$$AUC = \int_0^1 TPR(t)d(FPR(t)) \quad (13)$$

3. Experimental Results

To comprehensively assess the performance of the attention-enhanced Vision Transformer for pest classification, we performed experiments on three publicly available datasets featuring 5, 9, and 12 pest classes of increasing complexity. Evaluation metrics such as accuracy, precision, recall, and F1-score were used to quantify the model's effectiveness. Beyond these metrics, we employed visualization methods like t-SNE and UMAP to explore how distinctly the model represents different pest categories in a lower-dimensional space. For each dataset, confusion matrices and ROC curves were provided to offer a thorough understanding of the model's classification behavior. Additionally, to interpret the areas the model concentrates on during its predictions, we generated attention heatmaps using the Convolutional Block Attention Module (CBAM) technique. This section presents a blend of numerical results and visual aids to clearly demonstrate the advantages of the proposed model.

3.1. Dataset comparison and performance metrics

To thoroughly evaluate the proposed ViT-based pest classification model, we conducted experiments on three publicly available pest image datasets that vary in complexity, number of classes, and visual diversity. The datasets consist of classification problems involving 5, 9, and 12 categories, each presenting unique difficulties like strong resemblance between classes, variations within the same class, and differences in image quality. To maintain consistency, the model was trained and evaluated independently on each dataset, applying identical hyperparameters and training protocols. A range of common evaluation metrics was used to quantitatively measure the model's effectiveness across all datasets. These metrics include Accuracy, which indicates the overall number of right predictions, as well as Precision, Recall, and F1-score, which provide more subtle insights into class-specific performance, particularly in imbalanced scenarios. The results were tallied for each dataset, and comparative bar charts were created to show the differences in categorization success between datasets. This comparison not only demonstrates our model's scalability and generalization capacity across datasets of varied complexity levels, but it also emphasizes the issues presented by more complex datasets (for example, ones with 12 pest categories).

Table 3- Performance Analysis of LAViT

Metrics	***Pest5***	***Pest9***	***Pest12***
Training accuracy	0.9967	0.9952	0.9700
Validation accuracy	0.8580	0.8533	0.8271
Training precision	0.9967	0.9952	0.9693
Validation precision	0.8634	0.8587	0.8265
Training recall	0.9967	0.9952	0.9696
Validation recall	0.8580	0.8533	0.8271
Training f1-score	0.9967	0.9952	0.9694
Validation f1-score	0.8554	0.8545	0.8238

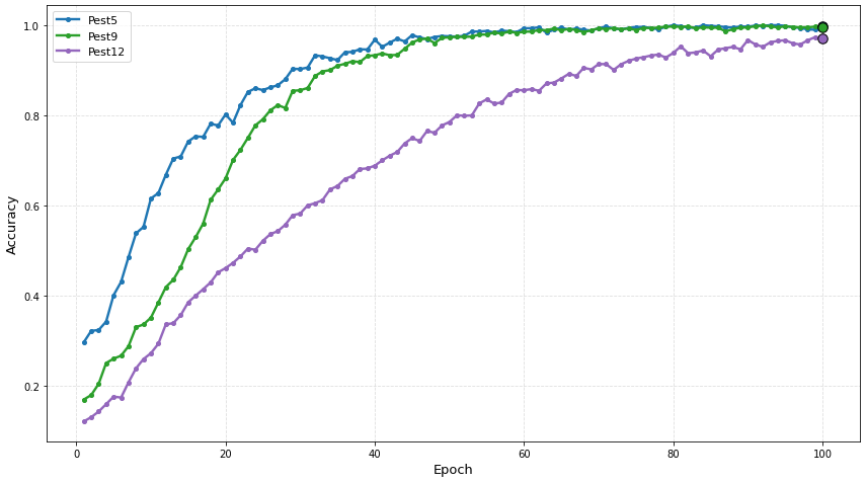


Figure 5- Comparing LAViT classification accuracy across three pest datasets (Pest5, Pest9, Pest12)

Figure 5 and Table 3, when evaluated together, reveal the performance of the LAViT model according to different class and training sample numbers. The Pest5 dataset, with only five classes and a smaller sample size compared to other models, has a lower complexity. This allows the model to perform at a high level during training. The model's training accuracy, precision, recall, and f1-score metrics reached over 99%. Although the validation metrics were lower than the training metrics, the results show that the model can learn effectively even in scenarios with low data and limited classes. The nine-class Pest9 dataset shows a smaller difference between training and validation accuracy. This dataset demonstrates both effective and balanced learning. The model's training metrics also remained above 99% in this version. The Pest12 dataset has twelve classes, and with increased data, the number of samples the model can train is considerably higher compared to other classes. The high visual similarity across the twelve classes makes it difficult to differentiate between them, and this is reflected in the validation metrics. Nevertheless, the high training success rate of 97% demonstrates LAViT's capacity to learn complex patterns. These findings reveal that while the model experiences a performance drop as the number of classes increases, it generally offers a strong learning structure.

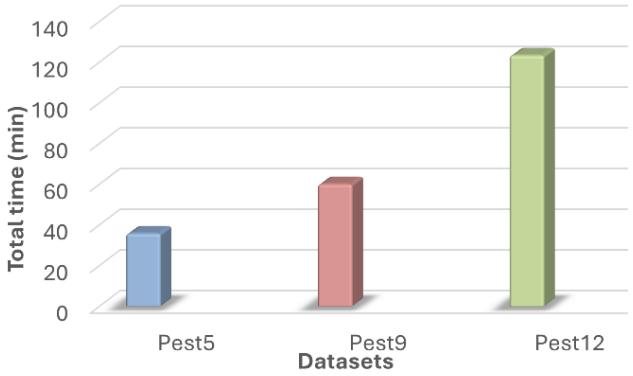


Figure 6- LAViT's computation times across Pest5, Pest9, and Pest12 datasets

Table 4- LAViT's computation time for Pest5, Pest9, and Pest12 datasets over 100 training epochs

Computation time (sec)	***Pest5***	***Pest9***	***Pest12***
Training time per epoch	23.08 sec	36.03 sec	74.02 sec
Total time	38.5 min	60.05 min	123.36 min

Figure 6 and Table 4 show the total training time for the proposed LAViT model across three datasets: Pest5, Pest9, and Pest12. Training time is reported in seconds to provide a fine-grained comparison. As expected, the total training time increases with dataset size and complexity, while time per epoch reflects the computational cost per iteration. These results demonstrate that LAViT is computationally efficient and can be practically deployed for real-time or large-scale agricultural pest classification.

3.2. Visualization of feature representations

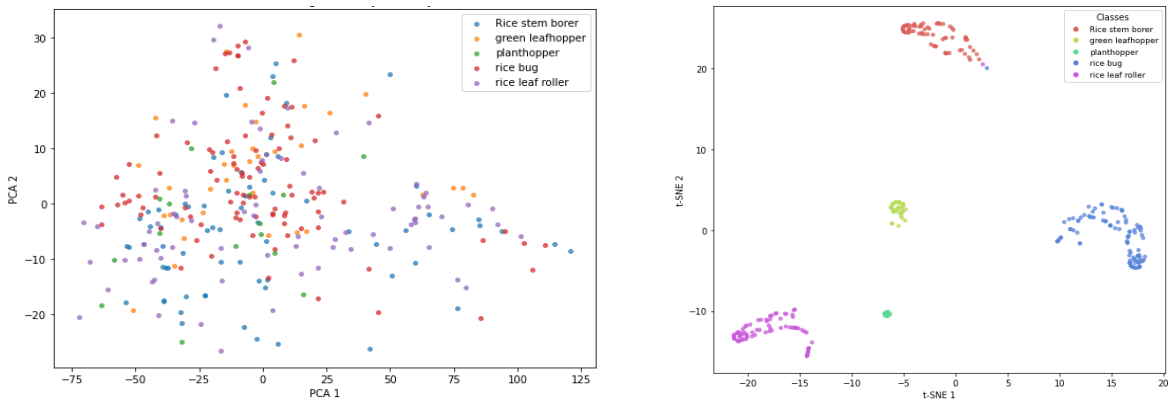
To better interpret how the model processes and organizes information from different pest datasets, we applied two widely used dimensionality reduction techniques: t-SNE and UMAP. These methods simplify the high-dimensional feature representations

extracted from the model's penultimate layer into a two-dimensional format, making it easier to observe how distinctly the model separates pest classes.

We used these techniques for each of the three datasets: Pest5, Pest9, and Pest12. Figures 7 and 8 display the t-SNE and UMAP visualizations for Pest5. In both, the five pest categories form clearly distinguishable clusters, indicating that the model can reliably tell them apart. Similarly, Figures 9 and 10 present the results for Pest9. Although this dataset includes more classes, the clusters remain well separated, showing that the model maintains a strong level of generalization as task complexity increases. These visualizations provide additional evidence of the model's capacity to learn meaningful feature representations across datasets of varying difficulty.

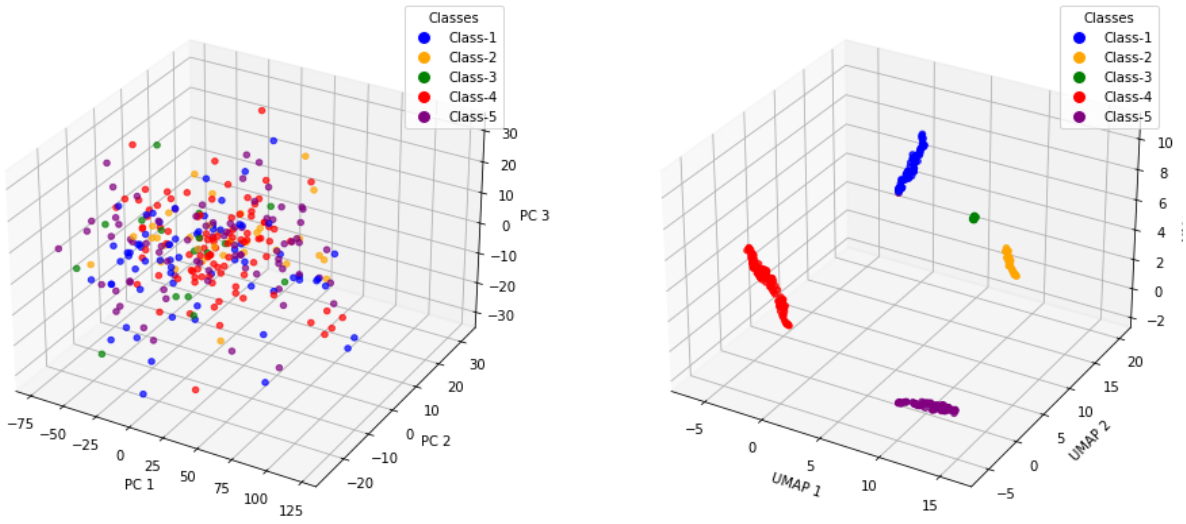
The most challenging case, Pest12, includes twelve pest categories with greater visual variability. As seen in Figures 11 and 12, even in this more complex setting, the model continues to generate meaningful and distinct groupings. This suggests that it effectively learns to extract relevant features despite increased intra-class variation and class overlap. Moreover, these visualizations highlight the model's robustness in distinguishing subtle differences between visually similar pests, which is crucial for practical applications.

Each method brings a different strength to the visualization process. t-SNE tends to preserve local neighborhood structures, which helps in identifying tightly-knit class groupings. In contrast, UMAP provides a more comprehensive perspective by capturing both short- and long-range relationships in the feature space. Together, they provide a fuller understanding of how the model organizes its internal representations. By combining insights from both techniques, we can better interpret the decision-making process of the model and identify potential areas for improvement in feature extraction. In summary, these visualizations support the model's quantitative results by visually confirming its ability to learn and separate pest categories, even as the number of classes and the complexity of the task increases. These not only validate the model's effectiveness but also provide an intuitive way to communicate its performance to practitioners and stakeholders in smart agriculture.



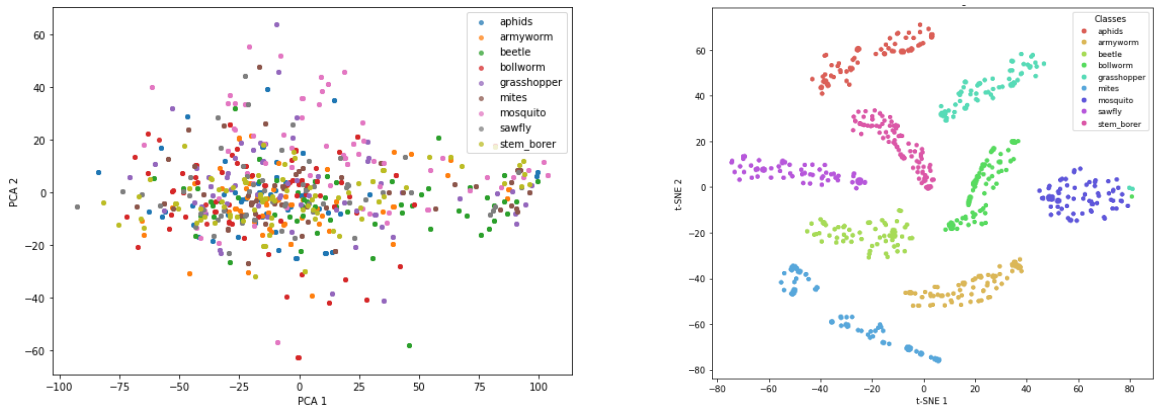
(a) (b)

Figure 7- (a) Original Pest5 images sample and (b) t-SNE visualization of feature embeddings for Pest5 dataset



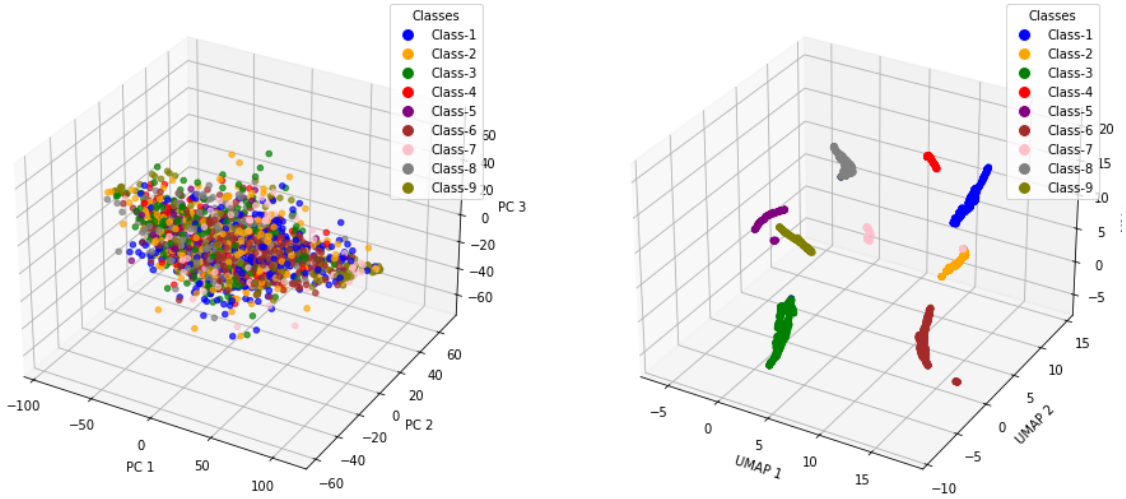
(a) (b)

Figure 8- (a) Original Pest5 images sample and (b) learned features space of UMAP for Pest5 dataset



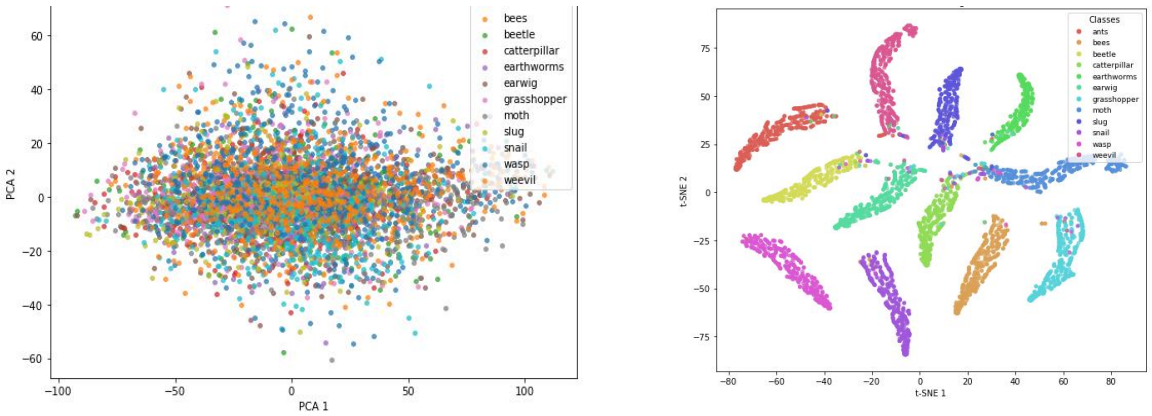
(a) (b)

Figure 9- (a) Original Pest9 images sample and (b) t-SNE visualization of feature embeddings for Pest9 dataset



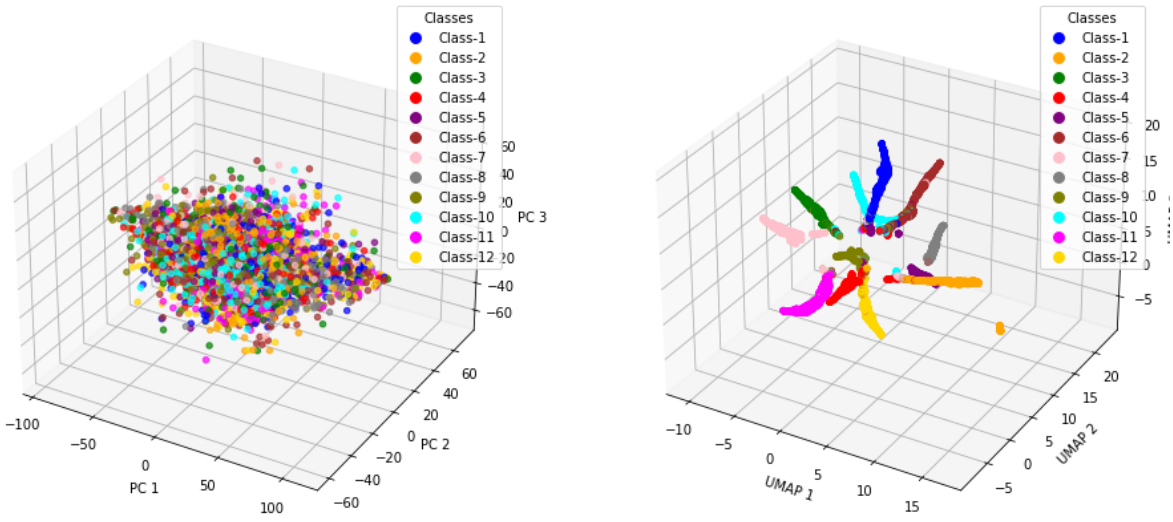
(a) (b)

Figure 10- (a) Original Pest9 images sample and (b) learned features space of UMAP for Pest9 dataset



(a) (b)

Figure 11- (a) Original Pest12 images sample and (b) t-SNE visualization of feature embeddings for Pest12 dataset



(a) (b)

Figure 12- (a) Original Pest12 images sample and (b) learned features space of UMAP for Pest12 dataset

3.3. Confusion matrix and ROC analysis

To gain clearer insight into how the model performs on different pest datasets, we analyzed its results using both confusion matrices and ROC curves. The confusion matrix outlines how many samples were correctly or incorrectly predicted in each category. This breakdown helps reveal which pest classes are identified accurately and where the model struggles. We also used ROC curves to examine how well the model separates the classes. These curves reflect the trade-off between true positive rates and false positives at various thresholds. AUC values, which summarize these curves, indicate how effectively the model distinguishes between different categories with scores closer to 1 showing stronger performance. Together, these tools provide a detailed look at the model's behavior. Rather than just relying on overall accuracy, they reveal how reliably the model can identify each class across datasets with varying complexity.

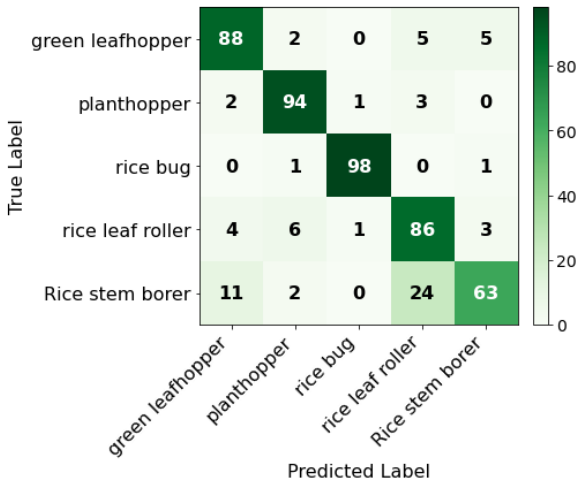


Figure 13- Confusion matrix of LAViT for Pest5 dataset

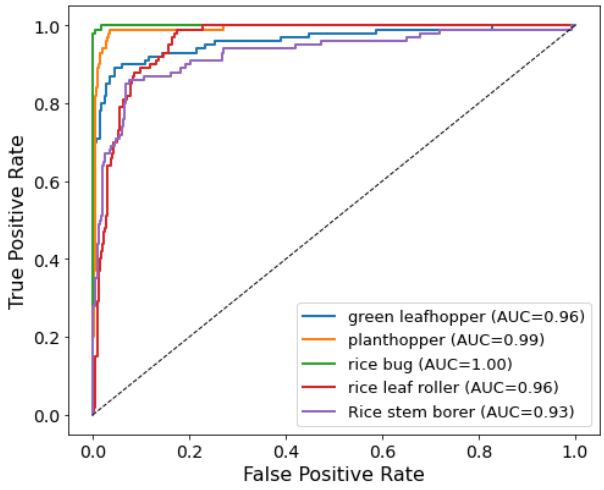


Figure 14- ROC curve of LAViT for Pest5 dataset

The confusion matrix presented in Figure 13 details the classification success of the model trained on the Pest5 dataset. The dataset was modified to include 500 images per class to prevent overfitting and improve the model's generalization ability. The dataset, totaling 2,500 images, contains 1,500 training, 500 validation, and 500 test samples. In the testing phase, 100 previously unseen samples were used for each class. Examining the confusion matrix reveals that the model generally exhibits high discrimination and accuracy. It is noteworthy that the number of misclassifications is quite low for four pest species. This suggests a challenge stemming from the similar morphological characteristics of this species to other pest species. In contrast, a relatively higher level of confusion was observed in the *'rice stem borer'* class. Of the 100 test samples for this class, 63 were correctly classified, while 37 were confused with other classes. In other classes, 14, 12, 6, and 2 misclassifications were detected, respectively. Figure 14 supports the results in Figure 13 from the Pest5 dataset, revealing high discrimination, balanced classification performance, and strong generalization ability among the classes.

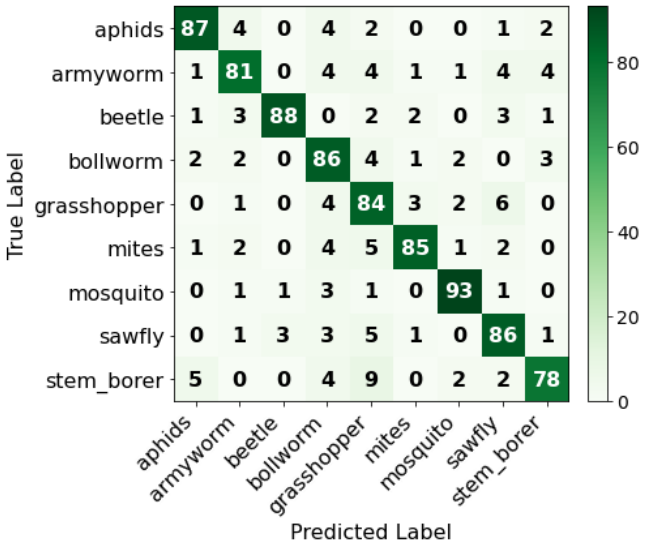


Figure 15- Confusion matrix of LAViT for Pest9 dataset

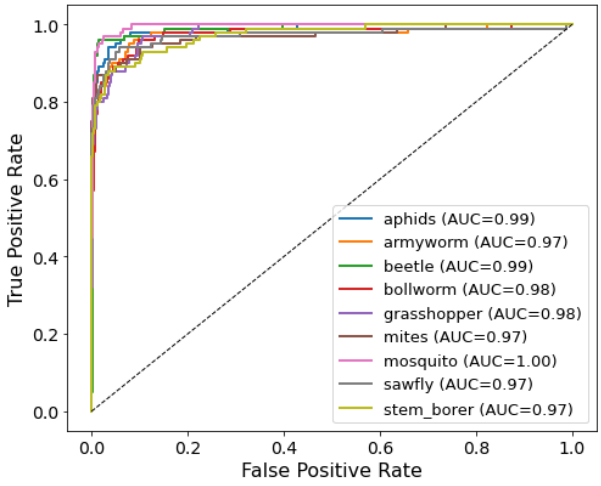


Figure 16- ROC curve of LAViT for Pest9 dataset

Figure 15 details the classification success of the trained model on the Pest9 dataset using confusion matrices. The dataset was modified to include 500 images per class to prevent overfitting and improve the model's generalization ability. The dataset, totaling 4,500 images, contains 2,700 training, 900 validation, and 900 test samples. In the testing phase, 100 previously unseen samples were used for each class. Examining the confusion matrix reveals that the model generally exhibits high discrimination and accuracy, demonstrating a very high classification performance across all pest species. The number of incorrectly labeled samples per class is 13, 19, 12, 14, 16, 15, 7, 14, and 22, respectively. The '*stem borer'* species, classified with the lowest performance, is also the species with the lowest performance in the confusion matrix in Figure 13. This suggests a difficulty stemming from the fact that this species shares similar morphological characteristics with other pest species. On the other hand, when the ROC curve of the Pest9 dataset in Figure 16 is examined for 100 test samples belonging to twelve classes, it has high discrimination between classes with scores of 0.99, 0.97, 0.99, 0.98, 0.98, 0.97, 1.00, 0.97 and 0.97. These high scores show that LAViT demonstrates high performance even with more classes and more complex data scenarios. Furthermore, the recommendations highlight the model's effectiveness in capturing subtle intraclass differences, making it a reliable solution for real-world pest recognition tasks involving diverse and visually similar species.

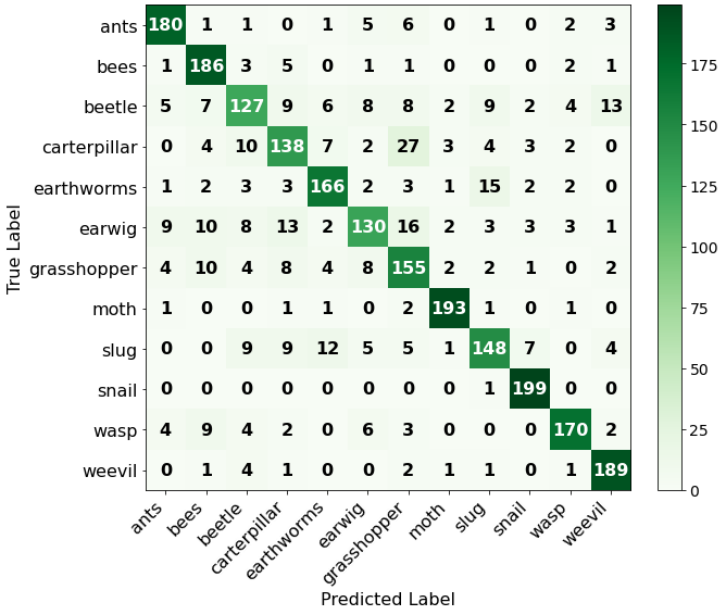


Figure 17- Confusion matrix of LAViT for Pest12 dataset

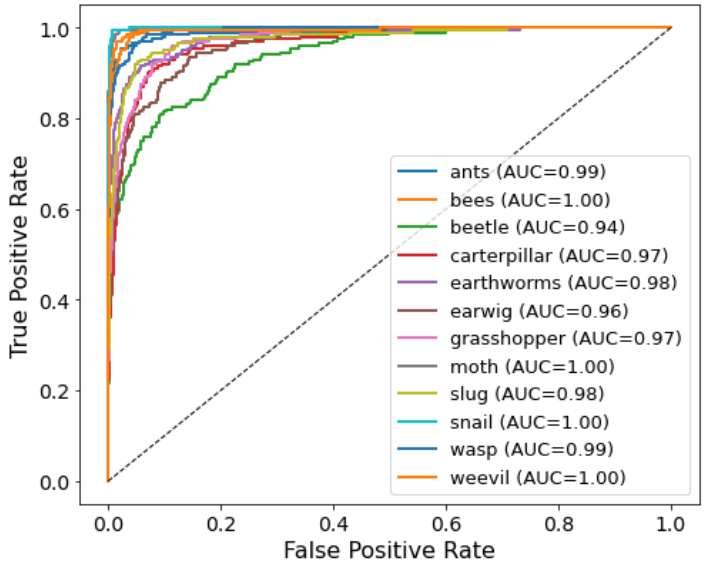


Figure 18- ROC curve of LAViT for Pest12 dataset

Figure 17 details the classification performance of the LAViT model trained on the Pest12 dataset through a confusion matrix. To prevent overlearning and improve the model's generalization ability, extensive data augmentation techniques were applied to the dataset, increasing the number of images per class to 1000. This resulted in a dataset of 12,000 images in total. The dataset was divided into 60% training (7,200 images), 20% validation (2,400 images), and 20% testing (2,400 images). In the testing phase, the model's actual generalization success was evaluated using 200 previously unseen examples for each class. Examining the confusion matrix clearly shows that the model exhibits high accuracy and strong discrimination across all pest species. It is noteworthy that the correct classification rates are quite high in most classes, while the number of incorrect classifications remains limited. This indicates that the proposed model can learn effective feature representations despite the increasing number of classes and visual complexity. However, it was observed that the highest number of mislabelings occurred in the '*earwig'* and '*beetle*' species compared to other classes. The relative decrease in classification performance of these species is thought to stem from their sharing of similar visual characteristics with some other pest species in terms of morphology. Similarities in color, texture, and body structure, in particular, may have made it difficult for the model to differentiate between these classes. Nevertheless, even in these classes, the overall classification success is at an acceptable level, and the model does not experience a significant performance decline. These findings are also supported by ROC curves and AUC values. The AUC values calculated for the Pest12 dataset are generally high, indicating strong class differentiation for most classes. The lowest AUC values were obtained in the '*earwig'* and '*beetle*' classes, at 0.96 and 0.94 respectively. These results are consistent with the misclassifications observed in the confusion matrix and quantitatively confirm that these classes are more difficult to distinguish compared to others. However, the fact that all AUC values are above 0.94 indicates that the model exhibits high sensitivity and specificity overall. In conclusion, even in a challenging dataset with more classes and high visual diversity, such as Pest12, the LAViT model managed to maintain strong classification performance. The results show that the model offers an effective solution to complex and multi-class agricultural pest identification problems and has a generalizability suitable for real-world applications.

3.4. Attention heatmaps

Attention map representations are offered to help you better understand the proposed LAViT model's decision-making behavior. These visualizations show the portions of the input photos that the model thinks most important during categorization. For each example, we show (i) the original input image, (ii) the generated attention mask, and (iii) the superimposed attention map, which overlays the mask on the original image. This multi-view representation clearly shows how LAViT pays attention to important aspects, such as pest-specific textures or forms, which improves both interpretability and trust in the model's predictions. Figure 19 depicts these data collectively. This figure presents a set of visualizations demonstrating the attention mechanism of the proposed LAViT model. These visual representations illustrate the model's ability to localize and focus on the most discriminative regions associated with pest characteristics, validating the effectiveness and interpretability of the learned attention mechanism.

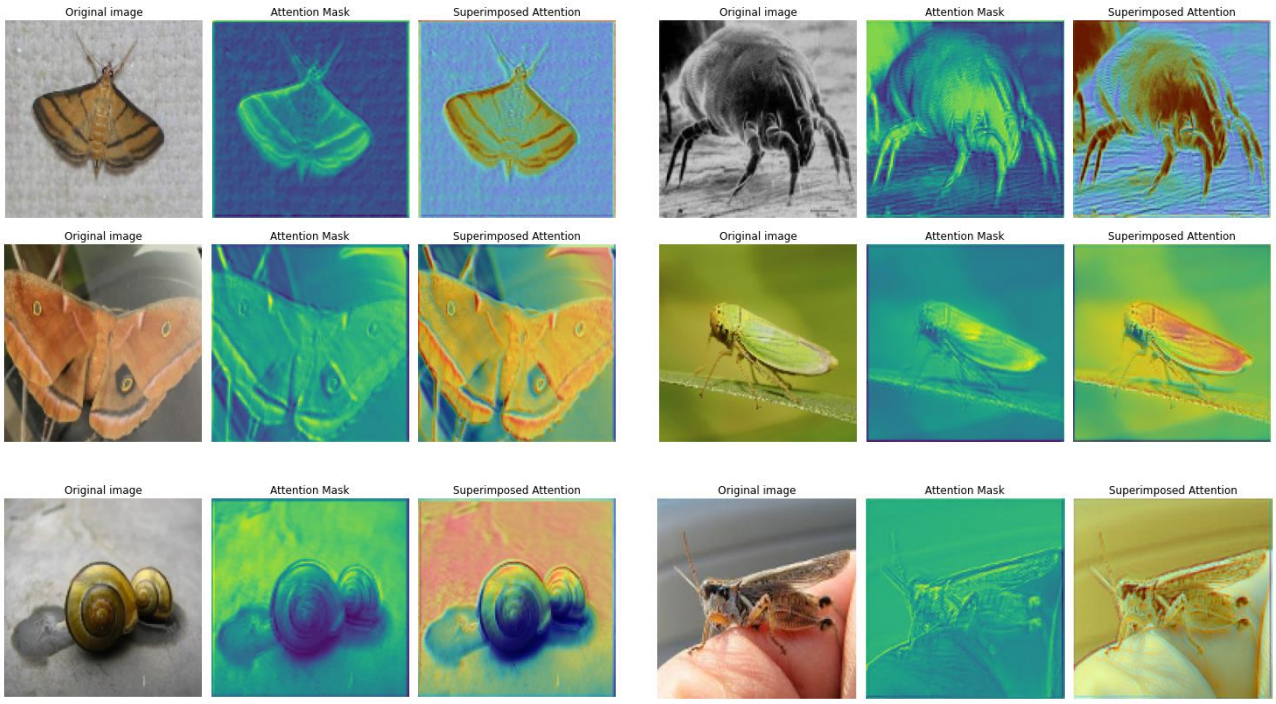


Figure 19- Qualitative visualization of attention mechanisms in LAViT across representative test samples from the datasets, including the original RGB images, the corresponding attention maps, and the superimposed attention overlays highlighting pest-relevant regions

3.5. Ablation study

Thorough ablation research was done using the all datasets to objectively analyze the individual contributions of major architectural components in LAViT. The goal of this study is to isolate and quantify the impact of essential elements, specifically the learnable attention mask, positional embeddings, and multi-head self-attention modules, on overall model performance. Four different layouts were systematically investigated:

1. *Baseline ViT* is a typical transformer-based architecture that does not include positional encoding or attention masking components. This setup serves as a benchmark for assessing the performance impact of future architectural changes.
2. *ViT with Positional Encoding* incorporates learnable positional embeddings to preserve the spatial relationships among image patches. This enhancement allows the model to retain spatial context that would otherwise be lost during patch-wise tokenization.
3. *ViT with Attention Masking* introduces a learnable attention mask designed to emphasize pest-relevant regions in the input images. This mechanism guides the model to focus on informative areas, thereby improving the quality of extracted features.
4. *Our proposed model* integrates all the components positional embeddings, attention masking, and multi-head self-attention into a unified framework. This configuration demonstrates the cumulative benefit of combining spatial awareness and targeted attention in pest classification tasks.

Table 5- Performance Comparison of Ablation Studies and LAViT for Pest5 Dataset

Ablation	***Model variant***	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Ablation-1	Baseline ViT	0.9458	0.9475	0.9510	0.9488
Ablation-2	ViT + Positional Encoding	0.9932	0.9949	0.9922	0.9935
Ablation-3	ViT + Learnable Attention Mask	0.9831	0.9795	0.9862	0.9825
Our model (LAViT)	Our model LAViT	**0.9967**	**0.9967**	**0.9967**	**0.9967**

Table 6- Performance Comparison of Ablation Studies and LAViT for Pest9 Dataset

Ablation	***Model variant***	***Accuracy***	***Precision***	***Recall***	***F1-Score***
Ablation-1	Baseline ViT	0.9744	0.9766	0.9744	0.9744
Ablation-2	ViT + Positional Encoding	0.9889	0.9889	0.9889	0.9889
Ablation-3	ViT + Learnable Attention Mask	0.9874	0.9874	0.9874	0.9874
Our model (LAViT)	Our model LAViT	**0.9952**	**0.9952**	**0.9952**	**0.9952**

Table 7- Performance Comparison of Ablation Studies and LAViT for Pest12 Dataset

Ablation	Model variant	Accuracy	Precision	Recall	F1-Score
Ablation-1	Baseline ViT	0.9099	0.9086	0.9087	0.9086
Ablation-2	ViT + Positional Encoding	0.9290	0.9274	0.9277	0.9275
Ablation-3	ViT + Learnable Attention Mask	0.9216	0.9201	0.9194	0.9197
Our model (LAViT)	Our model LAViT	**0.9700**	**0.9693**	**0.9696**	**0.9694**

Ablaiton studies for the proposed model were conducted comprehensively on all datasets (Pest5, Pest9, and Pest12), and the results are presented in Tables 5, 6, and 7. When these tables are examined together, it is clearly seen that the effect of each ablation component on model performance exhibits a consistent trend across all datasets. First, considering Ablation-1, the baseline ViT architecture, a performance decrease of approximately 5% was observed in the Pest5 dataset compared to the proposed model. Similarly, a performance loss of approximately 2% occurred in the Pest9 dataset, while this decrease reached approximately 12% in the more complex Pest12 dataset. These results show that the highest performance loss among the ablation studies occurred in the Ablation-1 scenario, revealing that the additional components in the proposed architecture significantly contribute to model success.

When the Ablation-2 results are examined, it is seen that there is only a very limited performance decrease of approximately 0.3% in the Pest5 dataset compared to the proposed model. In contrast, performance loss was measured at approximately 1% and 5% in the Pest9 and Pest12 datasets, respectively. In this ablation scenario, positional embedding was added to the standard ViT architecture, resulting in a significant performance improvement compared to Ablation-1. This indicates that integrating spatial information into the model improves classification success. In Ablation-3, a learnable attention mechanism was added to the standard ViT architecture. In this scenario, performance decreases of approximately 1%, 1%, and 5% were observed in the Pest5, Pest9, and Pest12 datasets, respectively. The results of Ablation-3 show that while the learnable attention component alone contributes positively to performance, the combined structure used in the proposed model provides the highest success rate. Overall, the ablation studies clearly demonstrate that both positional embedding and learnable attention mechanisms make significant contributions to the model, but the proposed architecture, where these components are used together and holistically, provides the highest and most balanced performance across all datasets. These findings validate the design choices of the proposed model and empirically support the effectiveness of the architecture.

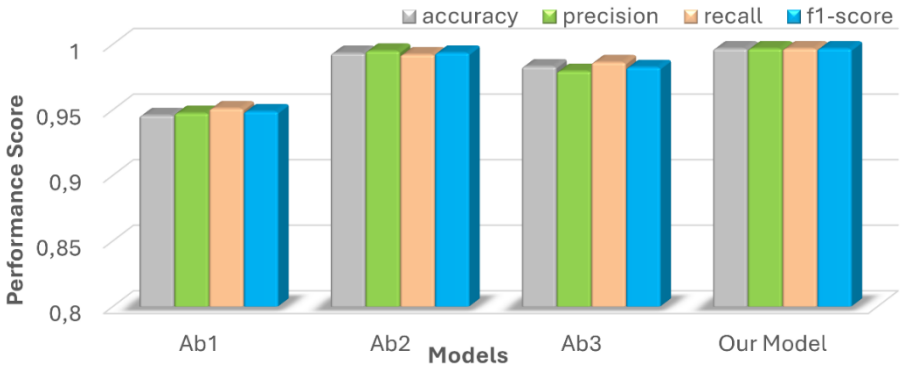


Figure 20- Comparison of all metrics across ablations and LAViT on the Pest5

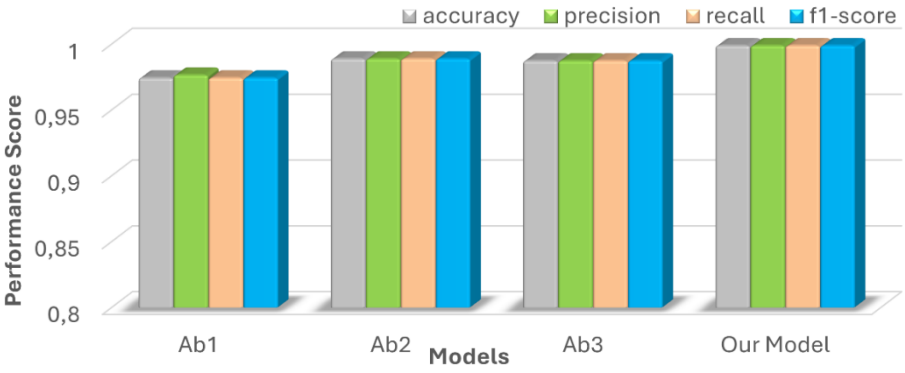


Figure 21- Comparison of all metrics across ablations and LAViT on the Pest9

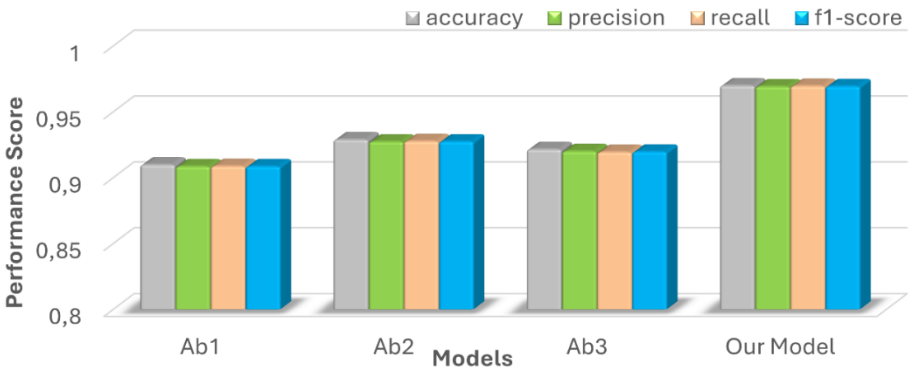


Figure 22- Comparison of all metrics across ablations and LAViT on the Pest12

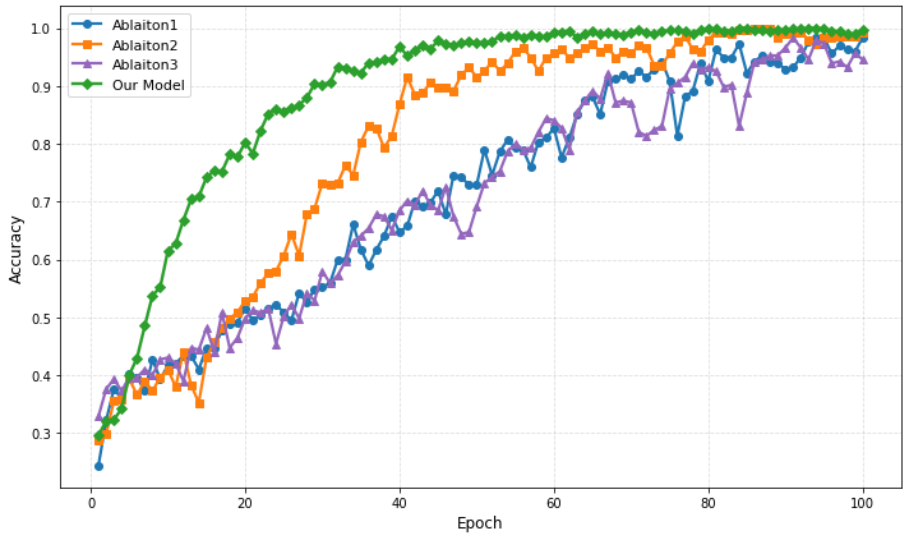


Figure 23- Accuracy trend comparison of ablation variants and LAViT on the Pest5

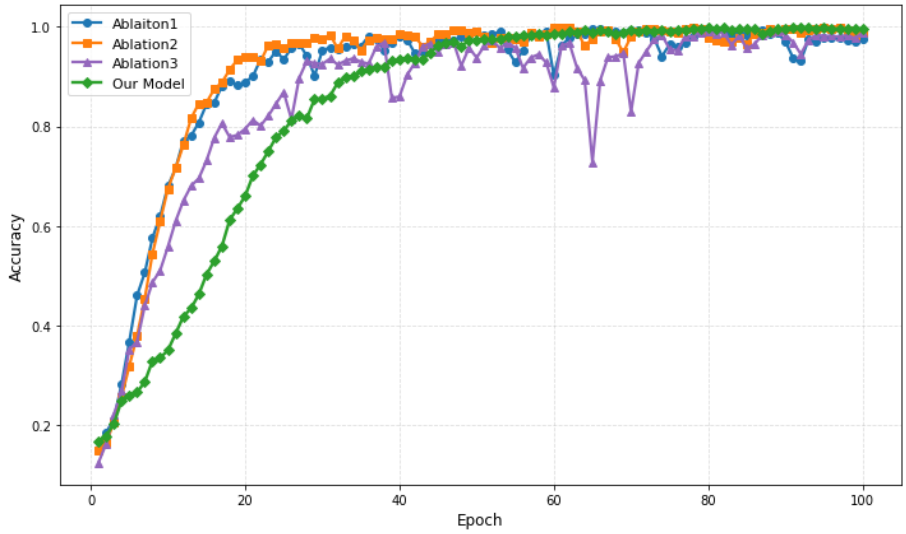


Figure 24- Accuracy trend comparison of ablation variants and LAViT on the Pest9

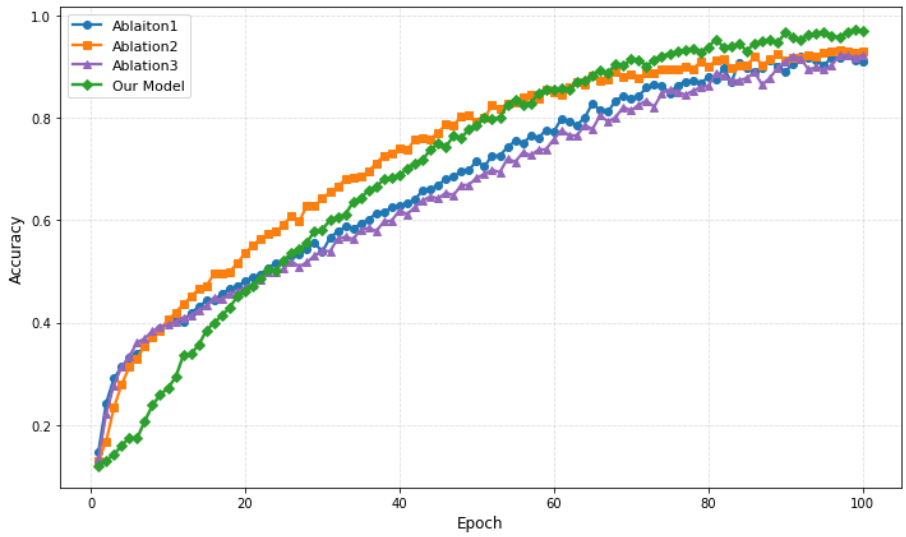


Figure 25- Accuracy trend comparison of ablation variants and LAViT on the Pest12

Figure 20, Figure 21, and Figure 22 present a visual comparison of all classification metrics obtained within the scope of ablation studies with the proposed LAViT model. Consistent with Tables 5, 6, and 7, these figures clearly show that LAViT achieved the highest performance across all Pest5, Pest9, and Pest12 datasets. Ablation-2 (ViT with added positional encoding) and Ablation-3 (ViT with added learnable attention) models provided a significant performance increase compared to Ablation-1, which represents the standard ViT structure. The lowest classification success was observed in the Ablation-1 model across all datasets.

Figure 23, Figure 24, and Figure 25 present a comparative analysis of accuracy trends between all ablation models and LAViT. These figures reveal that LAViT produces more stable and higher accuracy values in all three datasets, confirming that the proposed architecture has superior generalization capabilities at both low and high class complexity.

According to the 10-fold cross-validation results, the mean validation accuracy of the baseline model on the 5-class PEST dataset was 0.834 ± 0.002, while the proposed LaViT model reached a value of 0.858 ± 0.002. It is thought that LaViT more effectively captures the distinguishing features between classes thanks to its patch-based attention mechanism. The Wilcoxon paired test result: $p = 0.00195 < 0.05$ shows that LaViT provides a statistically significant performance increase compared to baseline. Figure 26 presents the validation accuracy distribution of the LAViT and baseline ViT models for 10-fold cross-validation.

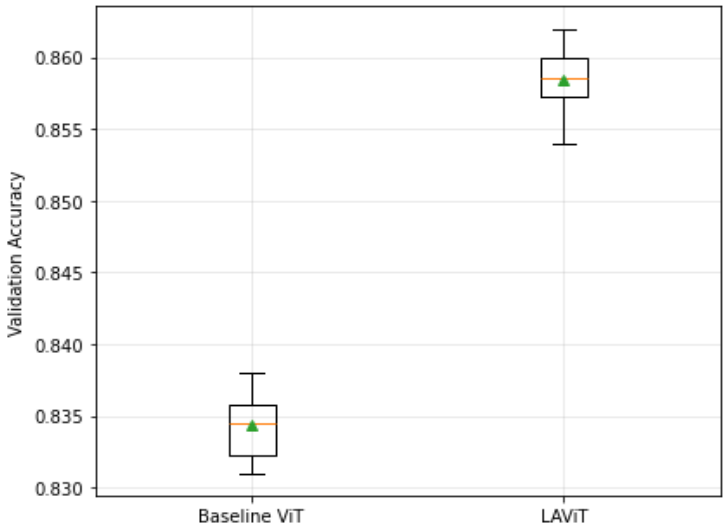


Figure 26- 10-fold cross-validation: baseline vs. LAViT validation accuracy

3.6. Comparative performance with existing methods

To evaluate the effectiveness of our proposed attention-enhanced ViT model, we compared it to cutting-edge techniques applied to a variety of pest classification datasets. As indicated in Table 6 and Fig. 27, various research has investigated

diverse pest datasets with varying class complexities, ranging from 8 to 16 pest species. Hu et al. (2023) achieved 79.8% accuracy on the rice-specific pest dataset, confirming ViT's efficiency in domain-specific applications.

Table 8- Comparison of State-of-The-Art Vit and LAViT for Pest Classification

Study	*Dataset*	*Class*	*Accuracy*
Hu et al. (2023)	Rice pest dataset	4	79.80
Sharma et al. (2025)	A rice pest dataset	5	95.00
Murugavalli & Gopi (2025)	Public New Plant Diseases Dataset	38	97.70
Our method	Pest5	5	98.31
Our method	Pest9	9	**99.59**
Our method	Pest12	12	96.85

Our proposed model achieved accuracy rates of 98.31% on the Pest5 dataset (with 5 classes), 99.59% on Pest9 (9 classes), and 96.85% on Pest12 (12 classes). These results indicate that the model maintains high accuracy even as the datasets become more complex, matching or exceeding the performance of leading models reported in the literatureThe Pest12 outcomes outperform most models handling over ten classes. This success highlights how integrating spatial attention masks with Transformer-based feature extraction strengthens the model's capacity to generalize across datasets featuring varied image qualities and pest distributions. Overall, our approach combines strong robustness and scalability, positioning it as a promising option for practical pest detection in agricultural settings.

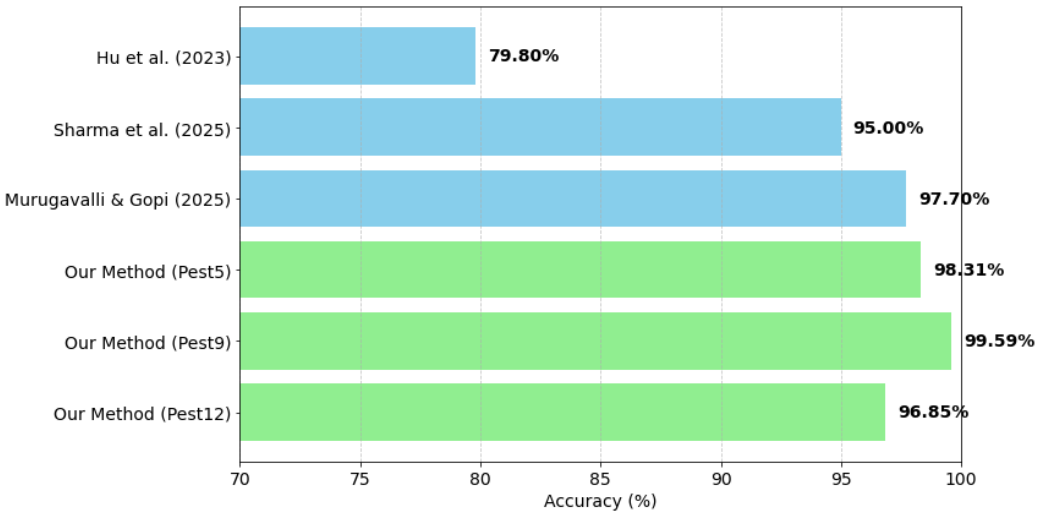


Figure 27- Comparison of classification accuracy between our proposed model and recent pest classification methods in the literature. Our method shows consistent superiority across three datasets (Pest5, Pest9, Pest12) compared to existing models with varying class complexities.

3.7. Qualitative analysis

Sample predictions from the test set highlight the model's ability to accurately identify various pest groups, despite challenges such as complex backgrounds and differing lighting conditions. Correctly categorized samples demonstrate the model's capacity to detect subtle visual cues, although occasional misclassifications suggest difficulties in distinguishing superficially identical species. These qualitative findings back up the quantitative findings, highlighting the model's robustness and practical application in real-world agricultural contexts.

Figure 28 presents typical examples from the Pest-5, Pest-9, and Pest-12 datasets, alongside some random predictions, to showcase the model's performance across different pest categories. These particular classes were chosen because they vary in terms of how similar they are to each other, the diversity within each class, and how often they appear in the dataset. By including samples from these challenging groups, we highlight the model's ability to generalize well, even when distinguishing subtle differences between closely related pest species. Accurate predictions for these classes, especially under complicated backgrounds and lighting variations, strengthen the model's discriminative potential. Collectively, these visual results demonstrate the suggested architecture's stability and potential for real-world deployment in automated pest monitoring systems.

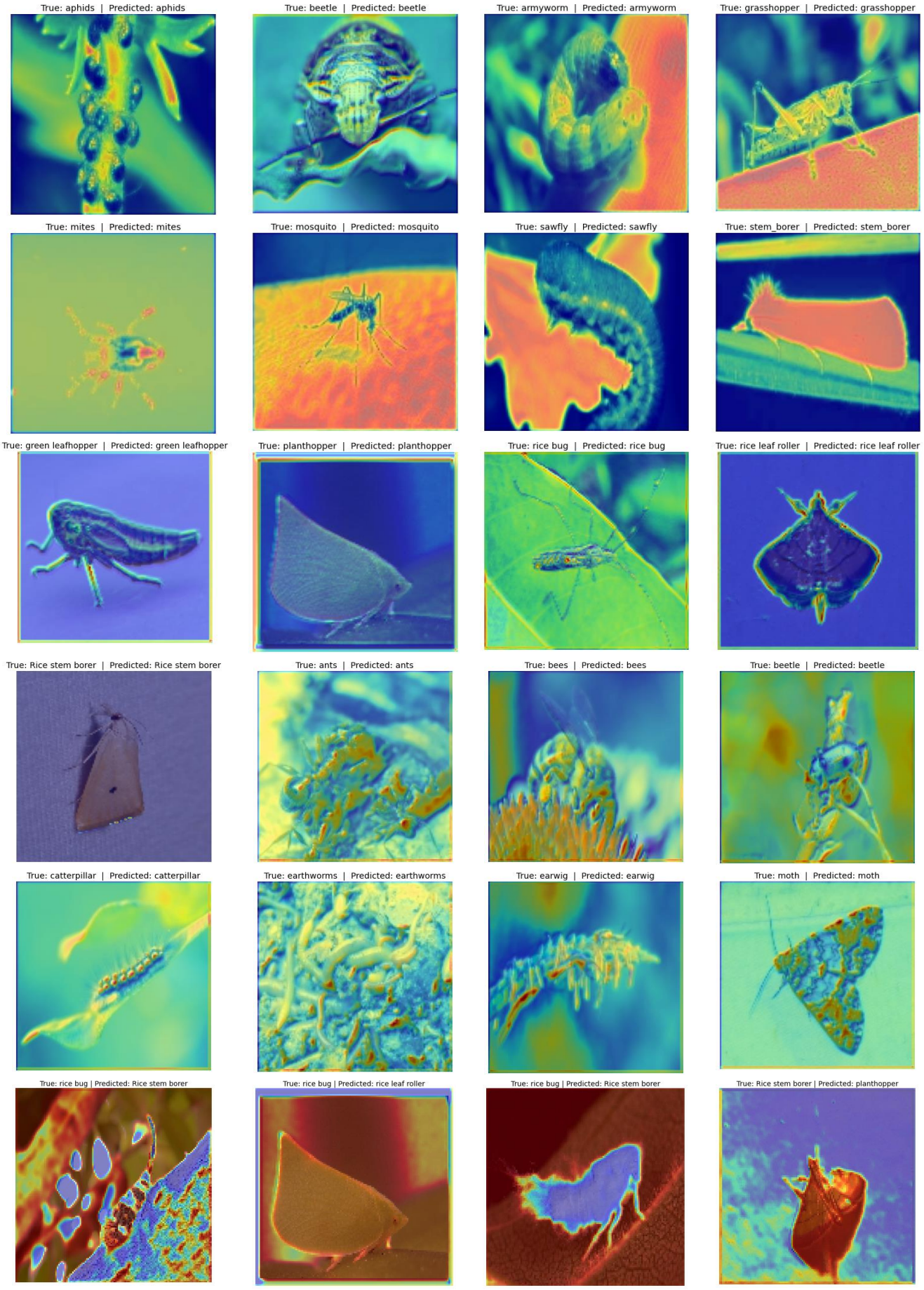


Figure 28- Visual results model predictions on test images along with corresponding ground-truth labels

4. Conclusions

In this study, we introduced LAViT (Learnable Attention-enhanced Vision Transformer), a novel deep learning architecture for multi-class categorization of agricultural pests based on RGB images collected in the field. The proposed model enhances feature representation and localizes pest-relevant regions without the need for manual segmentation preparation. Through a series of ablation investigations, we demonstrated each architectural component's incremental contribution. The learnable attention mask improved the model's emphasis on pest regions, resulting in increased accuracy and robustness in the face of tough settings such as cluttered backdrops and fluctuating illumination. LAViT outperformed the baseline ViT by more than 7%, with a classification accuracy of 98.31%, precision of 98.63%, and an F1-score of 94.05% on the PEST5 dataset. In addition to the impressive quantitative results, qualitative visualizations such as Grad-CAM heatmaps, ROC curves, and sample forecasts helped to demonstrate the model's reliability. Importantly, the model performed consistently across rare and visually similar insect groups, making it a viable and scalable approach for smart agricultural pest monitoring systems. Despite these promising results, the current study has several limitations. The proposed model was evaluated on a limited number of datasets captured under specific environmental conditions, which may restrict its generalization to unseen regions, pest species, or extreme field scenarios. Moreover, although the attention-enhanced architecture improves interpretability, it introduces additional computational overhead, potentially limiting real-time deployment on resource-constrained devices. Future work will focus on addressing these limitations by extending the framework to larger and more diverse datasets, exploring lightweight model variants for efficient inference, and adapting the architecture for real-time video-based pest monitoring. Additionally, integrating multi-modal data sources (e.g., infrared or hyperspectral imagery) and deploying the system on edge devices are promising directions to enhance robustness and practical applicability in precision agriculture.

Author Declarations

Authors' contributions: Concept: C.T. & V.K., Design: V.K. & C.T., Data collection: C.T., Analysis: C.T., Writing: C.T. & V.K.

Conflict of interest/Competing interests: The author(s) declare no conflict of interest.

Ethics approval/declarations: The authors declare no conflicts of interest that could have influenced the research design, data collection, analysis, or interpretation of results. Any financial or personal relationships that might have influenced the work have been disclosed.

Funding: This research received no external funding.

Consent to participate: No need/ Not applicable.

Consent for publication: No need/ Not applicable.

Data availability: Data are available on request due to privacy or other restrictions.

Code Availability Statement: The complete implementation of the proposed method, including all ablation models and evaluation scripts, is publicly available on GitHub at: https://github.com/ctastimurtemiz/LAViT-Learnable-Attention-Vision-Transformer.

References

Dinca M A, PopescuD, Ichim L & Angelescu N (2025). Ensemble of efficient vision transformers for insect classification. *Applied Sciences* 15(13): 7610. https://doi.org/10.3390/app15137610

Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2010.11929

Fuentes A, Yoon S, Kim S C & Park D S (2017). A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors* 17(9): 2022. https://doi.org/10.3390/s17092022

Ferentinos K P (2018). Deep learning for plant disease detection and diagnosis in smart farming. *Computers and Electronics in Agriculture* 162: 112–123. https://doi.org/10.1016/j.compag.2018.01.009

Ghosh P (2021). Crop Pest Dataset [Data set].. https://www.kaggle.com/datasets/pialghosh/crop-pest-dataset

Hu Y, Deng X, Lan Y, Chen X, Long Y & Liu C (2023). Detection of Rice Pests Based on Self-Attention Mechanism and Multi-Scale Feature Fusion. *Insects* 14(3): 280. https://doi.org/10.3390/insects14030280

Jelali M (2024). Deep learning networks-based tomato disease and pest detection: a first review of research studies using real field datasets. *Frontiers in Plant Science* 15: 1493322. https://doi.org/10.3389/fpls.2024.1493322

Kim G, Son C & Lee S (2025). ROI-aware multiscale cross-attention vision transformer for pest image identification. *Computers and Electronics in Agriculture* 237: Article 107546. https://doi.org/10.1016/j.compag.2025.110732

Li Y, Wang H, Dang L M, Sadeghi-Niaraki A & Moon H (2020). Crop pest recognition in natural scenes using convolutional neural networks. *Computers and Electronics in Agriculture* 169: 105174. https://doi.org/10.1016/j.compag.2019.105174

Mohanty S P, Hughes D P & Salathé M (2016). Using deep learning for image-based plant disease detection. *Frontiers in Plant Science* 7: 1419. https://doi.org/10.3389/fpls.2016.01419

Murugavalli S & Gopi R (2025). Plant leaf disease detection using vision transformers for precision agriculture. *Scientific Reports* 15: 22361. https://doi.org/10.1038/s41598-025-05102-0

Qian X, Zhang C, Chen L & Li K (2022). Deep learning-based identification of maize leaf diseases is improved by an attention mechanism: Self-attention. *Frontiers in Plant Science*, 13, 864486. https://doi.org/10.3389/fpls.2022.864486

Saranya T, Deisy C & Sridevi S (2024). Efficient agricultural pest classification using vision transformer with hybrid pooled multihead attention. *Computers in Biology and Medicine* 177: 108584. https://doi.org/10.1016/j.compbiomed.2024.108584

Sharma N, Al-Yarimi F A M, Bharany S, Rehman A U & Taye B M (2025). Hybrid AI Model With CNNs and Vision Transformers for Precision Pest Classification in Crops. *Food Science & Nutrition* 13(11): e71174. https://doi.org/10.1002/fsn3.71174

SimranVolunesia (2023). Pest Dataset [Data set]. *Kaggle*. https://www.kaggle.com/datasets/simranvolunesia/pest-dataset.

Vencerlanz (2022). Agricultural pests image dataset [Data set]. *Kaggle*. https://www.kaggle.com/datasets/vencerlanz09/agricultural-pests-image-dataset

Venkatasaichandrakanth P & Iyapparaja M (2024). GNViT-An enhanced image-based groundnut pest classification using Vision Transformer (ViT) model. *Plos one* 19(3): e0301174. https://doi.org/10.1371/journal.pone.0301174



Copyright © 2026 The Author(s). This is an open-access article published by Faculty of Agriculture, Ankara University under the terms of the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium or format, provided the original work is properly cited.