Computers in Biology and Medicine 177 (2024) 108584

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Efficient agricultural pest classification using vision transformer with hybrid pooled multihead attention

T. Saranya *, C. Deisy, S. Sridevi

Thiagarajar College of Engineering, Madurai, India

ARTICLE INFO

Keywords:
Convolutional neural networks (CNN)
Vision transformer
Hybrid pooled multihead attention (HPMA)
And pest classification

ABSTRACT

Accurate pest classification plays a pivotal role in modern agriculture for effective pest management, ensuring crop health and productivity. While Convolutional Neural Networks (CNNs) have been widely used for classification, their limited ability to capture both local and global information hinders precise pest identification. In contrast, vision transformers have shown promise in capturing global dependencies and enhancing classification performance. However, the traditional attention mechanism employed in vision transformers, which uses the same query (Q), key (K), and value (V), overlooks spatial relationships between patches, limiting the model's capacity to capture fine-grained details and long-range dependencies in the image. To address these limitations, this study presents a novel approach, termed Hybrid Pooled Multihead Attention (HPMA), for superior pest classification that outperforms both CNN models and vision transformers. The HPMA model integrates hybrid pooling techniques and modifies the attention mechanism to effectively capture local and global features within images. By emphasizing discriminative features and suppressing irrelevant information, the HPMA model achieves heightened robustness and generalization capabilities. The model is trained and tested on a newly built dataset consisting of 10 pest classes, achieving a remarkable accuracy of 98 %. Furthermore, the proposed HPMA model is validated on two benchmark datasets and achieves accuracies of 98 % and 95 %, demonstrating its effectiveness across diverse pest datasets. The results and ablation study of the proposed model contribute to exceptional performance in accurate pest classification. This tackles agricultural pest challenges and enables prompt pest control to reduce crop losses.

1. Introduction

Accurate pest classification is necessary in modern agriculture to manage pests and maintain crop health and yield. However, it might be challenging to appropriately characterize pest species due to their complexity and variety [1,2]. Conventional approaches such as Convolutional Neural Networks (CNN) have been widely applied for classification applications [3,4]. Nonetheless, there is considerable discussion about whether CNNs or vision transformers are more efficient at obtaining the local and global information required for accurate pest classification [5–7]. CNN models have demonstrated extraordinary performance in a range of computer vision applications, including image classification, because to their proficiency in capturing local data through convolutional layers and pooling operations. Nevertheless, CNNs often struggle to accurately identify long-range connections and global context in images, which are critical for accurate pest categorization. In contrast, vision transformers have emerged as a potential substitute for image classification tasks [8]. They show tremendous potential for discovering distant relationships and enhancing classification performance by utilizing self-attention mechanisms to detect global dependencies and contextual information in images [9]. Conventional attention systems struggle to recognize complicated patterns and have limited expressiveness, discrimination, and attention capacity.

To overcome these shortcomings, this work presents a novel method for classifying pests that performs better than traditional vision transformers and CNN models. The proposed Hybrid Pooled Multihead Attention (HPMA) architecture employs a modification of the standard attention mechanism to extract the query (Q), key (K), and value (V) representations. The HPMA model effectively captures both local and global information inside the image. This is achieved by means of hybrid pooling techniques and the extraction of keys (K) and values (V) from the concatenation of both max and mean pooled representations. By focusing on relevant regions of the image, the HPMA attention mechanism recognizes relationships and enables the model to interpret the

* Corresponding author.
E-mail addresses: saranshakthi09@gmail.com, tsaranya@student.tce.edu (T. Saranya), cdcse@tce.edu (C. Deisy), sridevi@tce.edu (S. Sridevi).

https://doi.org/10.1016/j.compbiomed.2024.108584
Received 16 November 2023; Received in revised form 21 March 2024; Accepted 8 May 2024
Available online 13 May 2024

context and interactions between different visual components [10]. This adaptability allows the model to focus on discriminative features. It also helps suppress irrelevant or noisy information, thereby improving overall robustness and generalization capability [11]. presents a method for identifying crop pests using CNNs and MobileNet [12]. explores the use of MMTL-IPCAC in conjunction with CLAHE, NASNet, MGWO, and XGBoost for insect identification. Using DL models like Unet and ResNet [13], addresses yield losses and achieves a 93.54 % classification rate. In contrast, this work presents HPMA, which outperforms conventional models and advances agricultural pest classification.

To evaluate the efficacy of the proposed model, it was trained and tested on a newly built dataset comprising 10 distinct pest classes. The results clearly demonstrate that the HPMA model outperforms traditional CNN models and vision transformers in accurately classifying pests, reconciling the controversy surrounding these two approaches. Additionally, the efficiency of the HPMA model was further tested on larger datasets such as IP102 and xie1, confirming its effectiveness and versatility across diverse pest datasets.

1.1. Motivation for the work

- Addressing pest-related challenges: The motivation behind this research is to tackle the challenges posed by pests in crop fields, especially in Tamil Nadu, where they can cause a detrimental impact on crop yield and quality, resulting in significant economic losses for farmers.
- Improving the accuracy and efficiency of pest classification: Accurate pest classification is essential for timely pest control and effective mitigation of crop losses. By improving the accuracy and efficiency of pest classification, this research aims to provide valuable support for agricultural practices.

1.2. Our contributions

- This research significantly contributes to the field by introducing the Hybrid Pooled Multihead Attention (HPMA) architecture, a novel approach for pest classification. HPMA innovatively modifies the traditional self-attention mechanism by incorporating hybrid pooling techniques to extract query, key, and value information from embedded features. This approach effectively captures both local and global information, enhances long-range dependency learning, and adapts to various levels of abstraction.
- In addition to this technical innovation, the study encompasses the creation of a newly collected dataset featuring 10 pest classes sourced from Tamil Nadu crop fields and Integrated Pest Management (IPM) images.
- To further validate the model's robustness and versatility, its efficiency and performance were rigorously assessed on two additional datasets, xie1 and IP102, demonstrating its effectiveness across a diverse range of pest datasets.
- Furthermore, the HPMA model has been rigorously evaluated and has demonstrated superior performance when compared to traditional attention mechanisms, CNN models, and vision transformers.
- The evaluation includes comprehensive results and discussions, along with an ablation study that provides further evidence of the effectiveness and innovation presented by this novel approach. This comprehensive set of contributions aims to advance the state of the art in pest classification technology and provide valuable support to agricultural practices.

The remaining part of this work is organized as follows: Section 2 presents the related work in the field of pest classification. In Section 3, the dataset and the proposed model's architecture are discussed. Section 4 analyzes and discusses the experimental results. Section 5 evaluates the effectiveness and robustness of the proposed approach through an ablation study. Finally, Section 6 concludes the paper, summarizing the key findings and contributions.

2. Related study

[14]explored DL's role in smart pest monitoring, emphasizing insect classification and detection using field images. It outlines DL frameworks across image acquisition, preprocessing, and modeling, aiming to promote diverse SPM applications [15]. developed a high-resolution model with a deep recursive residual network, enhancing recall for low pixel resolution [16]. created a hybrid model with Inception-ResNet-v2 for diseases and pest identification [17]. introduced three-stage frameworks to automate lesion identification in coffee trees, utilizing Mask-RCNN, UNet, PSPNet, and ResNet for segmentation and pest classification.

[18]found that the VGG-16 model with the SGDm algorithm had the highest detection accuracy and F1 score of 98.33 % and 98.36 %, respectively, during the pest outbreak stage. In the third stage, the AlexNet model with the SGDm algorithm achieved the highest detection accuracy and F1 score of 99.33 % and 99.34 %, respectively, demonstrating effective pest management in agricultural products [19]. proposed an accurate pest classification method, using MMAlNet for multi-scale features and DenseNet Vision Transformer for enhanced accuracy. The ensemble of MMAlNet and DNVT achieves high classification accuracies (99.89 % and 74.20 %) on D0 and IP102 datasets, surpassing existing methods and demonstrating practical value [20]. presented a novel DL approach for identifying mung bean pests and diseases, vital for Indian food security. Utilizing CNN and transfer learning, the smartphone-based model attains an impressive 93.65 % accuracy in detecting six diseases and four pests, offering an efficient solution for quick detection [21]. presented PCNet, a lightweight CNN-based pest classification method with attention for accurate insect recognition on mobile devices. Achieving 98.4 % accuracy on a self-built dataset, PCNet outperforms classic and lightweight CNN models with only 20.7 M parameters, making it suitable for real-time pest recognition on resource-constrained mobile devices [22]. tackled agriculture's annual yield losses from pests using DL models like Unet and ResNet, achieving a high 93.54 % classification rate. This accuracy is crucial for effective pest management and avoiding misidentification [23]. developed a four-step pest detection framework using an attention module fused with a deep ResNet, sampling-balanced RPN, adaptive RoI selection, and Faster RCNN for classification and location of smaller multi-pests.

[24]proposed a method for identifying pests and diseases of grape leaves, combining Transformer and Ghost-convolution networks. The Ghost Enlightened Transformer model (GeT) achieves 98.14 % accuracy, surpassing competitors in speed and efficiency [25]. presented a CNN-based approach for crop pest detection with adaptive feature fusion and augmentation, outperforming methods like Cascade R–CNN, Dynamic R–CNN, SSD, FPN, and RetinaNet on the AgriPest21 dataset with an accuracy of 77.0 % [26]. creates a dataset of 5135 whitefly-attacked leaf images and proposes a classification method based on the Compact Convolutional Transformer (CCT). The CCT-based model outperforms MobileNet, ResNet152v2, and VGG-16, achieving an accuracy of 97.2 %.

[27]proposed a self-supervised transformer-based pre-training approach for classifying agricultural pests and diseases. It outperformed advanced methods, obtaining accuracy of 99.93 %, 76.99 %, and 74.69 % on the CPB, Plant Village, and IP102 datasets, respectively. The technique used feature relationship conditional filtering (FRCF) and latent semantic masking auto-encoder (LSMAE) to enhance feature learning and classification performance over CNN-based approaches [28]. presented a novel method for identifying insect pests that combines transformer architecture and convolutional blocks. The model outperforms advance techniques on the IP102 benchmark dataset, achieving classification accuracies of 74.897 % at the size of 224x224 and 75.583 % at the size of 480x480, respectively. The long-tailed

distribution problem in pest datasets was addressed by the DL integration architecture that was presented by Ref. [29]. ConvNeXt and Swin Transformer models were combined at the feature level in the suggested method, which enhanced pest classification performance overall. Experimental results on three datasets demonstrated superior accuracy, outperforming state-of-the-art techniques by 76.1 % for IP102, 98.5 % for d0, and 92.3 % for the insect dataset.

[30]introduced a two-stage model called the ODP-Transformer, which combined methods for creating image captions with those for classifying pest images. By including actions for observation, description, and prediction, the model replicated the diagnostic procedure used by experts in agriculture. In the task of classifying images of pests, the ODP-Transformer outperformed six other techniques with a higher accuracy (12.91 %) than frequently used CNN models. For creating image captions, the ODP-Transformer outperformed six other techniques by increasing the indicators of Bleu1, CiderD, and Rouge by 1.62, 8.08, and 1.08, respectively.

For accurate pest location and identification in field images [31], developed a CNN design, combining channel attention and spatial mechanisms. The proposed dataset, the Li's dataset, and the IP102 dataset, the model's classification accuracy was 96.78 %, 96.50 %, and 73.29 %, respectively. It was outperforming conventional CNN models and earlier attention-related DL models. Improving the pest recognition, multimodal fine-grained transformer (MMFGT) model was proposed by Ref. [32]. The MMFGT combined learning of self-supervised, recognition of fine-grained and joint multimodal data will improve accuracy of recognition. The MMFGT outperformed other models in the experiments, compared to the latest DINO method by 5.92 % in the baseline and achieving recognition accuracy of up to 98.12 %.

Inferences from related studies: In the domain of agricultural pest management, accurate classification of pests is essential for effective mitigation strategies and crop protection. However, traditional self-attention mechanisms in transformers exhibit limitations when it comes to capturing both local and global information in vision-based tasks, such as pest classification. The primary limitation lies in their proficiency at capturing long-range dependencies while struggling with fine-grained details and local features in images. Conventional attention mechanisms operate on the entire input feature map without distinguishing the significance of different regions. Consequently, important fine-grained features crucial for accurate classification may not receive sufficient attention. In tasks like pest classification, where both local and global features play a critical role, this limitation can lead to suboptimal performance. The effect of this limitation is evident in the model's ability to accurately classify pests. It fails to effectively discern and prioritize crucial fine-grained features within input images. As a result, the model may misclassify images or struggle to distinguish visually similar pest types, ultimately leading to reduced classification accuracy and reliability.

To address these gaps, researchers have explored various approaches. Traditional methods, such as Convolutional Neural Networks (CNNs), excel at capturing local features but often struggle to integrate essential global context for precise classification. Recent advancements in deep learning, particularly vision transformers, show promise in capturing long-range dependencies and global information. However, there still remains a significant research gap in effectively integrating both local and global features to enhance pest classification models.

Addressing the Research Gap with HPMA: In response to this gap, the Hybrid Pooled Multihead Attention (HPMA) architecture is introduced, offering a novel solution to bridge the divide between local and global features. By modifying the conventional attention mechanism and leveraging hybrid pooling techniques, HPMA adeptly captures both local and global information within pest images. Unlike conventional CNNs, HPMA dynamically adapts to varying levels of abstraction, facilitating a comprehensive understanding of context and interactions among image components. Furthermore, HPMA's adaptability enables it to prioritize discriminative features while suppressing noise, thereby enhancing overall robustness and generalization capability in pest classification tasks. Through exhaustive experimentation and evaluation on diverse datasets, including a newly constructed dataset and established benchmarks, HPMA demonstrates superior performance. This superiority is evident when comparing it to traditional CNN models and vision transformers. The innovative architecture not only addresses the shortcomings of existing approaches but also contributes significantly to advancing pest classification technology in agriculture. This comprehensive understanding of the research landscape underscores the significance of HPMA in filling the identified research gap and propelling the field of agricultural pest management forward.

Table 1 offers a comparative analysis of pest classification methodologies utilizing CNN and Transformer models, presenting key performance metrics and ablation studies sourced from relevant literature. The inclusion of "Ours" signifies our proposed approach for pest classification, with the tick mark indicating the work done on that aspect.

3. Materials and methods used

This section covers the dataset, the proposed model, including patch creation, patch and positional embedding, Hybrid Pooled Multihead Attention (HPMA), and the classifier head.

3.1. Dataset collection

The paper presents a meticulously curated pest dataset aimed at evaluating the effectiveness of the proposed Hybrid Pooled Multihead Attention (HPMA) model for pest classification. The dataset consists of 2405 pests distributed across 10 classes and was collected in the agricultural field Ramanathapuram district of Tamil Nadu using a mobile camera and pi camera. This ensured a diverse representation of pest populations across various crop fields, including cotton, brinjal, rice, coconut, and peanut. During data collection, significant variability in the dataset counts was observed, reflecting the inherent diversity of pest populations in agricultural environments. To address potential class imbalances and ensure robust representation, the dataset was augmented by sourcing additional images from reputable Integrated Pest Management (IPM) sources (available at https://www.ipmimages.org/Collections/SubColls.cfm). IPM serves as the authentic source for pest and disease images, and since most of the pest datasets were sourced from IPM, such as [33]. This process not only enhanced the diversity of the dataset but also mitigated any bias towards simplicity. It achieved this by enriching it with images from external sources, crucial for capturing the complexity of pest populations in real-world scenarios.

From these crop fields, 43 % of the pest images were collected, representing 12 different types of pests. However, to streamline the dataset and focus on the most prevalent pests, the decision was made to exclude two types of pests that were present in single-digit counts. Therefore, the final dataset consists of 10 pest types. The selection of the 10 target pest types was guided by extensive consultation with Madurai Agricultural College experts, field surveys, and observations. This ensured that the chosen classes accurately represented the most commonly encountered and agriculturally significant pests in the surveyed fields.

Moreover, crop-specific factors were considered, as certain pests may have a greater impact on particular crops. This comprehensive approach to class selection was bolstered by input from agricultural experts, field surveys, and crop-specific considerations. It validated the relevance of the selected pest types, reaffirming the dataset's credibility and its applicability in practical pest management contexts. A detailed statistical description of the dataset is provided in Table 2, offering insights into the distribution of images across the 10 pest classes.

3.2. Data preprocessing

The performance and robustness of the proposed HPMA-based vision

Table 1
Taxonomy of pest classifications using CNN and Transformer.

Ref	CNN	Transformer	Field data	Public data	Performance metrics					Ablation Study
					Accuracy	Precision	Recall	F1-score	Confusion matrics	
[4]	✓		✓		✓	✓	✓	✓	✓	
[15]	✓			✓	✓					
[17]	✓			✓	✓					
[28]	✓	✓		✓	✓					
[29]	✓	✓		✓	✓	✓	✓	✓		
[32]		✓		✓		✓	✓	✓		✓
Ours		✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 2
Descriptions of collected pest dataset.

S.No	Pest name	Count	S.No	Pest name	Count
1.	Aphids	241	6.	Stem borer	279
2.	Beetle	239	7.	Saw fly	215
3.	Grashopper	261	8.	Weevils	196
4.	Army worm	230	9.	Thunder flies	254
5.	Field cricket	231	10.	Red ant	259

transformer were assessed using two open-source datasets. The first dataset, known as the Small Dataset (SD), consists of 2405 images depicting 10 targets. These images were specifically collected for this study, as illustrated in Fig. 1 and Table 2.

The second dataset [34], is the Medium Dataset (MD), comprising 4500 images spread across 40 classes (available at https://www.dlearningapp.com/web/DLFautoinsects.htm). Additionally, the effectiveness and versatility of the proposed HPMA model were further tested using the Large Dataset (LD) [35], which includes an extensive collection of 75000 images categorized into 102 classes (accessible at https://github.com/xpwu95/IP102/tree/master). These supplementary datasets were utilized to assess the performance of the HPMA model across a variety of pest datasets. The datasets, namely the newly built dataset, Xie dataset [36], and IP102, are consistently referred to as SD (Small Dataset), MD (Medium Dataset), and LD (Large Dataset) throughout the paper. To ensure consistency, all three datasets were resized and normalized. Furthermore, the SD and MD underwent image augmentation techniques, such as random rotation, random crop, zoom, horizontal flip, and vertical flip, to increase image quantity and improve model generalization. Augmentation was essential to rectify potential class imbalances in the dataset, ensuring comprehensive representation of pest populations. Additionally, normalization helped standardize pixel values, aiding in feature extraction and model convergence. Consequently, the augmented SD was increased from 2405 to 12025 images, and MD was increased from 4500 to 22500 images. The complete dataset, including the curated dataset and the augmented open-source datasets (SD and MD), was exclusively employed for the evaluation of the proposed HPMA model, ensuring comprehensive testing across diverse pest datasets.

3.3. Proposed method

The proposed model incorporates a novel attention mechanism called Hybrid Pooled Multihead Attention (HPMA) within a vision transformer framework to tackle pest classification tasks. An overview of the proposed model is depicted in Fig. 2. The overall architecture of the model is presented in Fig. 3, illustrating the workflow, which includes patch creation, embedding, the transformer with hybrid pooled multihead attention (HPMA), and an MLP used as the classifier head for pest classification. This will be discussed further below.

3.3.1. Patch creation

In this study, images of size 224 × 224 × 3 were utilized to create patches using the conv2d operation, drawing inspiration from the paper [37]. By applying conv2d with a stride and kernel size equal to the patch size of 16, the images were divided into non-overlapping patches of size 16 × 16. This resulted in a total of 196 patches arranged in a grid format with 14 rows and 14 columns. Consequently, the 3D input shape H × W × C, representing the height (H = 224), width (W = 224), and channel (C = 3) of the input image, was transformed into a 2D output shape $N_p \times (P^2 \times C)$, where N_p represents the number of patches. The number of patches, N_p, is defined as $(H \times W)/P^2$, and P^2 represents the patch resolution.

3.3.2. Patch and positional embedding

The patches are flattened to generate a 1D sequence of tokens. Each

Fig. 1. Pest image/count.

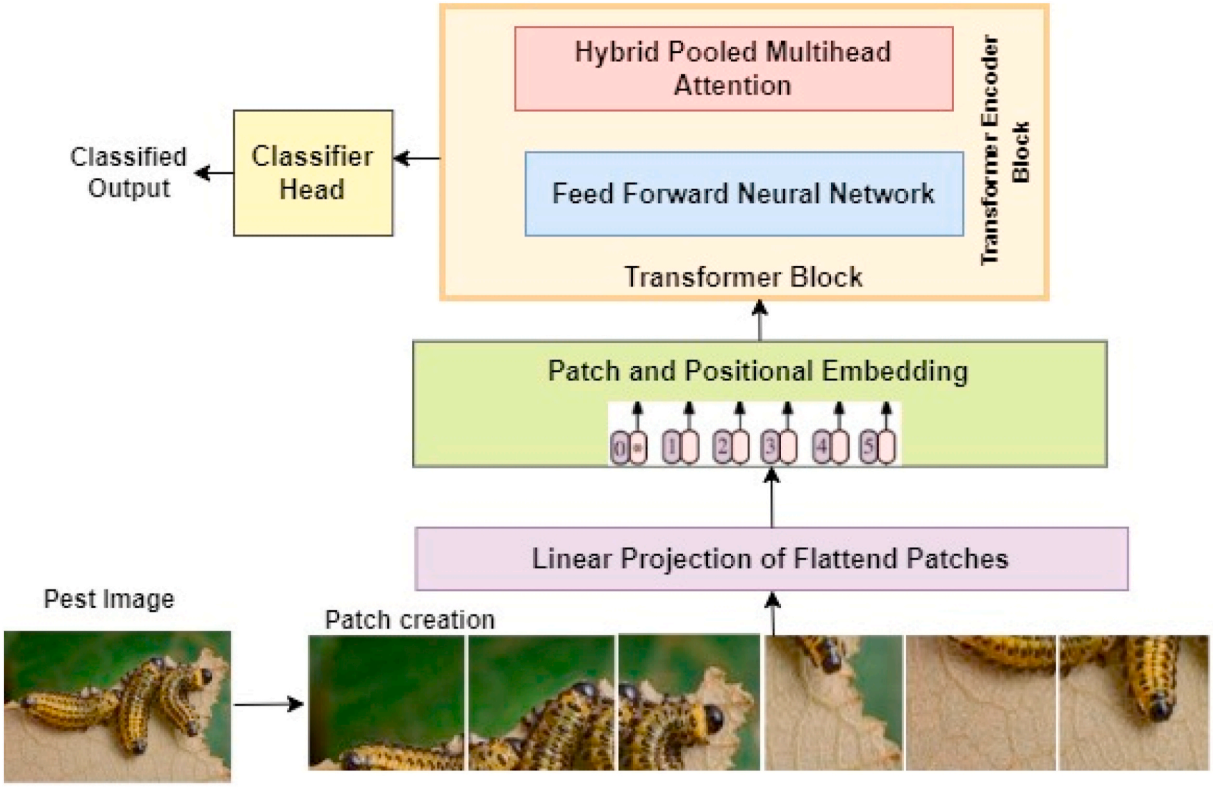

Fig. 2. Overview of the proposed model.

patch's pixels are represented by a certain number of tokens, which are used to represent each patch's pixels. The tokens are then embedded.

Into a higher-dimensional space. The patch embedding layer, a trainable linear layer, is used for this. Tokens are converted into a higher-dimensional space that captures the connections between various elements of the image by the patch embedding layer. Positional embeddings as well as patch embeddings are used by vision transformers. The spatial relationships between the patches are captured by these embeddings.

The final embedding for each patch is made up of the patch embeddings in addition to a learnable vector called the positional embedding that is assigned to each patch. The proposed hybrid pooled multihead attention (HPMA) is then fed the final embeddings of each patch. The following are the mathematical equations for patch and positional embeddings:

$$Z_{patch} = X_{patch} \cdot W_{patch} \quad \text{Eq.1}$$

$$Z_{pos} = E_{pos} \quad \text{Eq.2}$$

$$Z_{class} = X_{class} \cdot W_{class} \quad \text{Eq.3}$$

$$X = Z_{patch} + Z_{pos} + Z_{class} \quad \text{Eq.4}$$

Where, X_{patch} represents the flattened patches, W_{patch} represents the trainable linear projection weights for the patch embeddings, Epos represents the positional embedding matrix, X_{class} represents the class token input, W_{class} represents the trainable linear projection weights for the class token embedding, Z_{patch} represent the patch embeddings, Z_{pos} represents the positional embeddings, Z_{class} represents the class token embedding, X represents the final embeddings for each patch. This X will be the input for transformer encoder where the proposed HPMA finds out the attention score.

3.3.3. Hybrid pooled multihead attention (HPMA)

In the proposed model, a vision transformer architecture for pest classification is implemented with a novel attention mechanism called Hybrid Pooled Multihead Attention (HPMA). To capture spatial relationships and feature relevance in input images, this method creatively combines patch generation, embedding, and proposed HPMA methods. HPMA strategically integrates max pooling and min pooling techniques to create a hybrid pooled feature map, enabling adaptive feature weighting and enhancing classification accuracy. Significant progress in pest classification is anticipated as a result of this paradigm change.

The HPMA is the proposed modified attention where it uses the embedded feature map X as an input.

Lemma 1. *Let X be the embedded feature map, W_Q, W_K, and W_V be the learnable weight matrices specific to the queries Q, keys K, and values V, respectively, and dim_K be the dimension of the key K. Then, the HPMA model can be mathematically defined as follows*:

$$Q = X \times W_Q \quad \text{Eq.5}$$

$$Max_{pool} = Max_{pool}(X) \quad \text{Eq.6}$$

$$Min_{pool} = Min_{pool}(X) \quad \text{Eq.7}$$

$$Hybrid_{pooled} = concat\left(Max_{pool}, Min_{pool}\right) \quad \text{Eq.8}$$

$$K = Hybrid_{pooled} \times W_K \quad \text{Eq.9}$$

$$V = Hybrid_{pooled} \times W_V \quad \text{Eq.10}$$

$$Attention(Q, K, V) = softmax\left(QK^T \Big/ \sqrt{dim_K}\right)V \quad \text{Eq.11}$$

$$HPMA = Attention(Q, K, V) \quad \text{Eq.12}$$

Where X denotes the embedded feature map, W_Q, W_K, W_V are the learnable

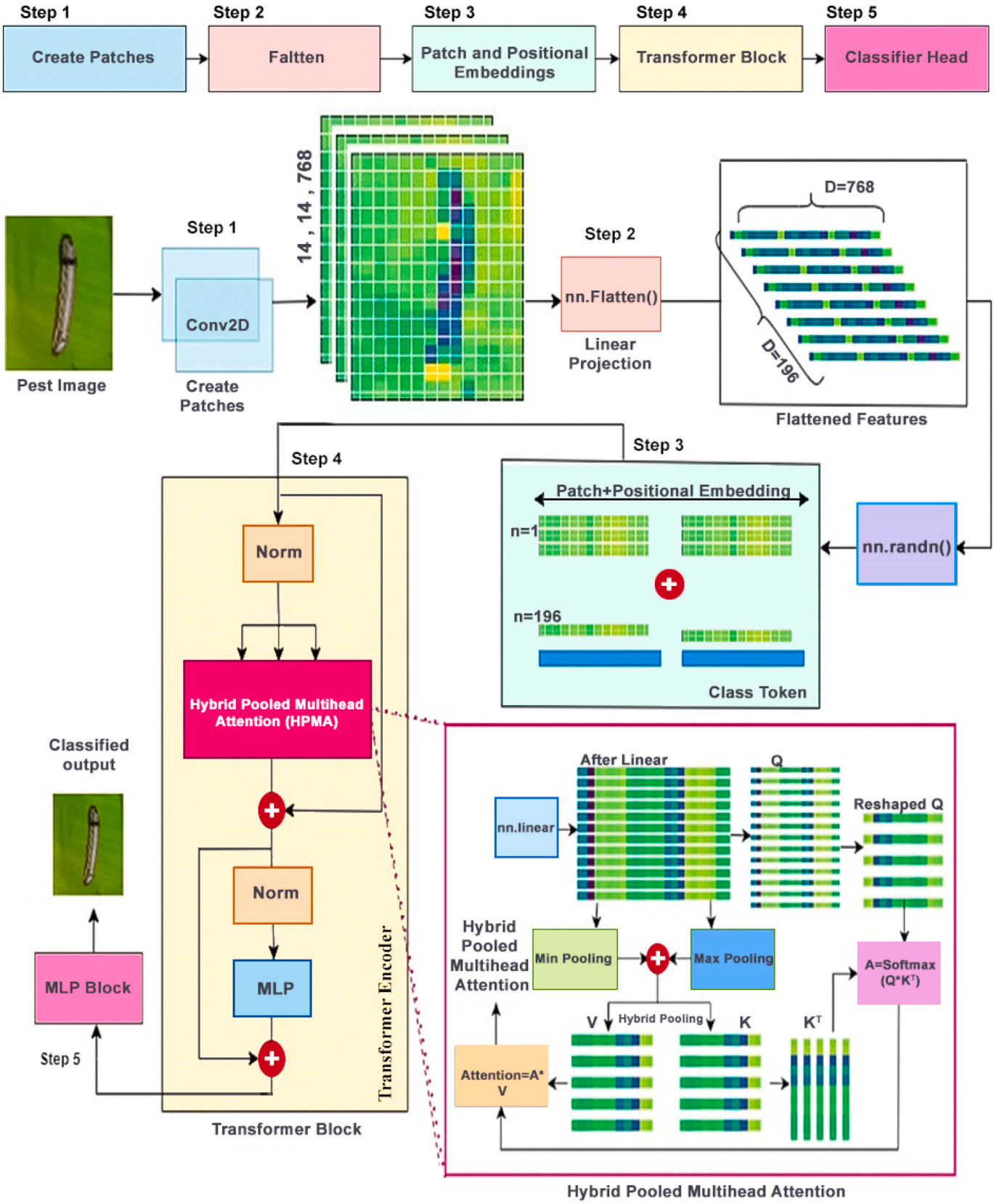

Fig. 3. Proposed architecture of HPMA model.. (Illustrate the overall architecture in five steps, with each step identified by a distinct block color. Clearly depict the flow of steps in the architecture, specifying the sequence in which they occur.)

weight matrices specific to the queries Q, keys K, and values V, respectively, × denotes the matrix multiplication, Max_{pool}& Min_{pool} represents the max pooling and min pooling, $Hybrid_{pooled}$ is the concatenated pooled attention map of max and min pooling, dim_K is the dimension of the key K, and HPMA is the hybrid pooled multihead attention.

***Proof of Lemma* 1**: *The mathematical definition of HPMA is derived by substituting Q, K, and V in HPMA equation as mentioned in the Lemma* 1. *Therefore Lemma* 1 *holds*:

3.3.4. Classifier head

The classifier head in a vision transformer is a key component that transforms learned visual features into class predictions. It typically includes fully connected layers followed by a softmax activation function. This combination allows the model to generate probability scores for different classes, enabling it to classify input images accurately. The input to the classifier head is the output of the proposed HPMA transformer encoder, which captures high-level visual features.

Algorithm 1. Hybrid Pooled Multihead Attention

```
Input:
    X: Embedded feature map
    W_Q: Learnable weight matrix for queries
    W_K: Learnable weight matrix for keys
    W_V: Learnable weight matrix for values
    dim_K: Dimension of keys
Output:
    Hybrid Pooled Multihead Attention (HPMA) Map.
Begins
    Hybrid Pooled Feature Map:
            Max_pool = max_pool(X)
            Min_pool = min_pool(X)
            Hybrid_pooled = concat(Max_pool, Min_pool)
    For i in range (6):
        For j in range (6):
            Extract the keys (K) and values (V) from Hybrid_pooled:
                K = Hybrid_pooled× W_K
                V = Hybrid_pooled× W_V
            Extract the query (Q) from X:
                Q = X × W_Q
            Compute the attention scores:
                Attention scores = softmax((QK^T) / √(dim_K))
            Hybrid pooled Multihead attention:
                HPMA = Attention_scores × V
        End For
    End For
End
```

The combined embeddings are passed through layer norm and then fed into the proposed hybrid pooled multihead attention block. The traditional self-attention uses the same value for query (Q), key (K), and value (V), which are extracted from the embedded feature map [38] *However, the proposed hybrid pooled multihead attention (HPMA) uses Q extracted from the feature map, similar to traditional self-attention, while V and K are extracted from the hybrid pooled feature map and Algorithm* 1*and* Fig. 4 *shows the work flow of proposed HPMA model.*

The HPMA model first applies max pooling and min pooling to the feature map X to obtain two pooled feature maps, Max_{pool} and Min_{pool}. The pooled feature maps are then concatenated to form a hybrid pooled feature map, $Hybrid_{pooled}$. The query (Q) is extracted from the original feature map X, and the keys (K) and values (V) are extracted from the hybrid pooled feature map $Hybrid_{pooled}$. The Q is reshaped based on the values of K and V. Then, the reshaped Q and the transposed K (K^T) are used in softmax operation to compute the attention score.

The attention score is then multiplied by V to obtain the hybrid pooled attention. The hybrid pooled attention is a weighted combination of the max pooled Max_{pool} and the min pooled Min_{pool} features. The most significant features in the feature map X are captured by the max pooled features, while the least significant features are captured by the min pooled features. The attention scores allow the algorithm to weight the max pooled features and the min pooled features according to their importance. Therefore, it is possible to make use of the HPMA model to extract a feature map's most crucial features. This method can capture both the most important and the least important features in the feature map due to the hybrid pooled attention, which is a weighted combination of the max pooled features and the min pooled features.

4. Results and discussion

In this section, details on experimental and hyperparameter settings, classification outcomes, and a comparative analysis of the proposed model with three datasets (SD, MD & LD) were presented. Additionally, a comparison with state-of-the-art algorithms and discussions on the findings were included.

4.1. Experimental and hyper parameter settings

The proposed work was trained and tested on a Google Colab machine with a GPU that supports CUDA. Versions 1.13.1 of Torch and 0.14.1 of Torchvision were used in the implementation. The pest classification involved splitting all three datasets into training, testing, and validation sets, with a ratio of 80 % for training, 10 % for testing, and 10 % for validation. The input size was defined as 224, 224, 3, and the patches were generated using conv2d [39]. These patches were then embedded with dimensions of 197, 768 [34]. Tables 3 and 4 below provides details on the parameters utilized in this study.

4.2. Classification outcome

4.2.1. Classification outcome of HPMA on collected dataset SD

The proposed Hybrid Pooled Multihead Attention (HPMA) model was initially evaluated on Dataset SD, focusing on training and testing metrics, including accuracy, loss, precision, recall, and F1 score. According to Fig. 5 and Table 5, the HPMA model for Dataset SD achieved a

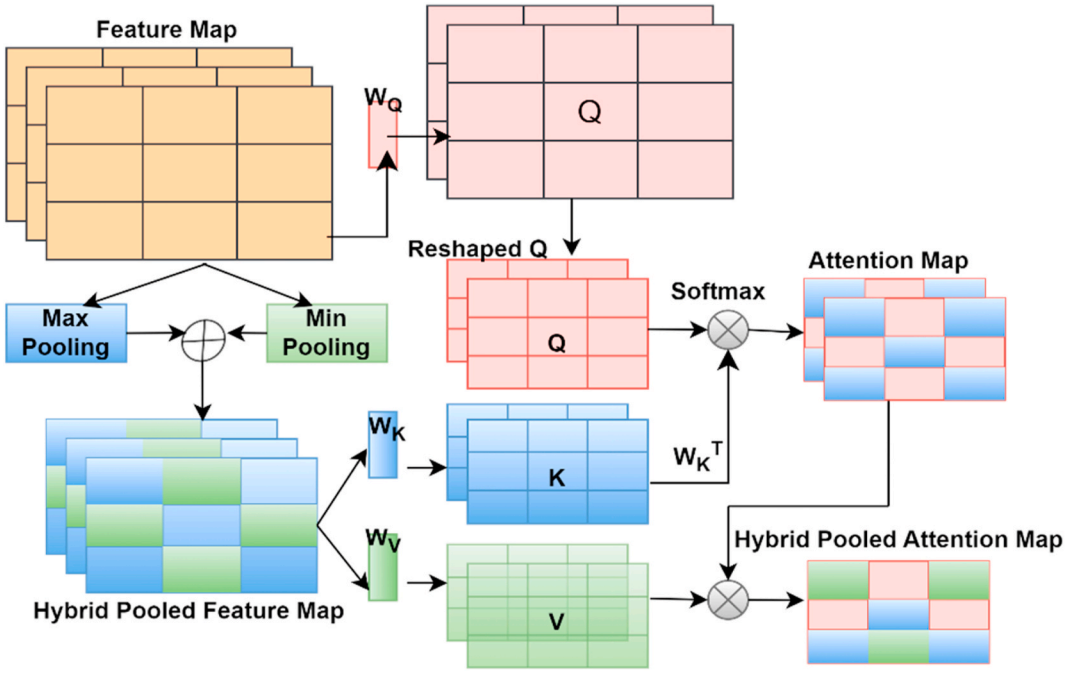

Fig. 4. Proposed hybrid pooled multihead attention.

training accuracy of 99.04 % and a testing accuracy of 98 % after 50 epochs. The training and testing losses were 0.3419 and 0.2011, respectively. Precision, recall, and F1 score were 0.9820, 0.9775, and 0.9751, respectively. These outcomes showcase the HPMA model's exceptional ability to accurately classify pests within Dataset SD.

The evaluation began with an examination of the Confusion Matrix from Fig. 6, shedding light on the classification performance of the HPMA model. This matrix revealed specific patterns of true positives, true negatives, false positives, and false negatives across different pest classes. Subsequently, a ROC curve for the proposed HPMA model for 10 classes in the collected dataset SD was analyzed. The ROC area for classes 0 to 9 ranged between 0.97 and 1.0, illustrating the model's effectiveness in distinguishing between various pest classes, as depicted in Fig. 7.

4.2.2. Classification outcome of HPMA on benchmark datasets MD and LD

The performance of the HPMA model was further evaluated on benchmark datasets MD and LD, highlighting its versatility and adaptability.

Classification Outcome of HPMA on Benchmark Dataset MD: As illustrated in Fig. 8 and detailed in Table 6, the HPMA model achieved an impressive training accuracy of 99.56 % and a testing accuracy of 98.02 % on Dataset MD after 50 epochs. The training and testing losses were 0.2101 and 0.3222, respectively, with precision, recall, and F1 score all at 0.9892, 0.9788, and 0.9888. These results underscore the HPMA model's robustness and generalizability when applied to benchmark Dataset MD.

Classification Outcome of HPMA on Benchmark Dataset LD: Dataset LD, a larger and more complex dataset, was also employed to assess the HPMA model's performance. As depicted in Fig. 9 and summarized in Table 6, the HPMA model exhibited commendable performance on Dataset LD.

After 50 epochs, it achieved a training accuracy of 97.23 % and a testing accuracy of 95.98 %. Both training and testing losses were 0.4268, and precision, recall, and F1 score were 0.9655, 0.9597, and 0.9595, respectively. These findings underscore the HPMA model's capacity to effectively categorize pests across diverse and challenging datasets, emphasizing its adaptability and versatility.

Comparisons of Benchmark dataset MD and LD with Existing Work: The comparison highlights the accuracy percentages of existing CNN-based methods and a proposed approach utilizing a Vision Transformer with Hybrid Pooled Multi-Head Attention (HPMA). Existing CNN-based methods achieve accuracies ranging from 89 % to 99 % on the MD dataset and from 67 % to 89 % on the LD dataset [2,19,40,41,42]. In contrast, the HPMA method demonstrates competitive performance, achieving 98 % accuracy on MD and notably good accuracy of 95 % on LD. These results emphasize the potential effectiveness of Vision Transformer-based approaches, particularly HPMA, in managing larger datasets for insect pest classification.

Table 3
Parameters for creating patches and embedding.

Parameter	Value
Input Size	224^2
Channel	3
Stride	16
Kernel	16
Patch size	16^2
Total patches	196 (14 × 14)
Embedded shape	197, 768

Table 4
Parameters for the proposed model.

Parameter	Value
Dimension	768
Head	6
Layer	6
Batch size	16
MLP size	1024
Activation	GeLu
MLP dropout	0.1
Attention dropout	0.1
Embedding dropout	0.1
Loss	Categorical cross entropy
Epochs	50
Optimizer	Adam

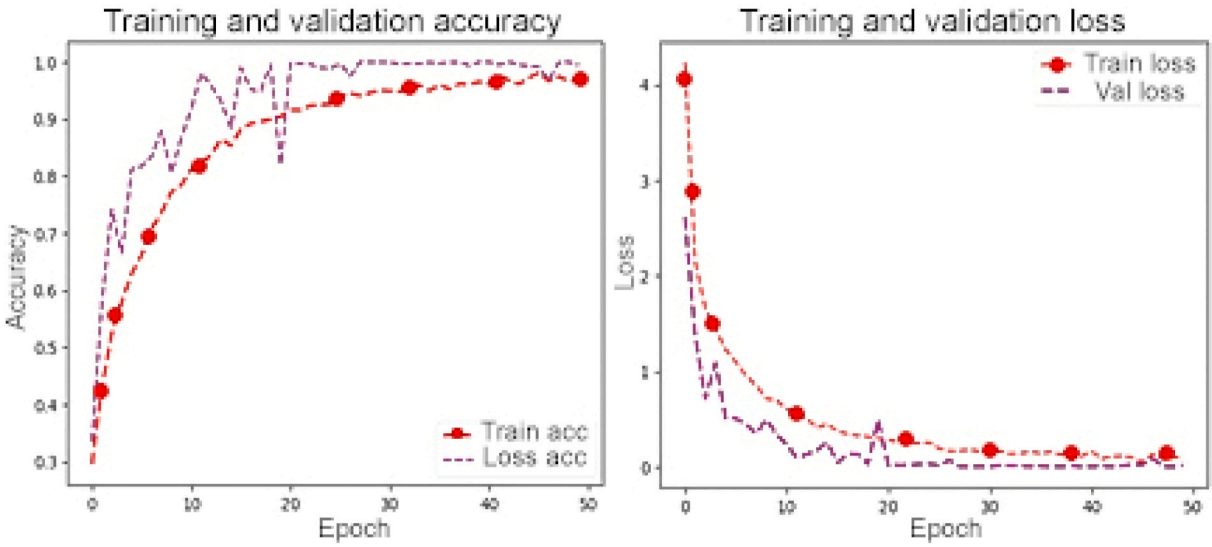

Fig. 5. Dataset SD accuracy and loss of HPMA.

Table 5
Classification report of HPMA on dataset SD.

Pest Classes	Precision	Recall	F1-Score	Support
Aphids	0.94	0.97	0.90	71
Army worm	0.97	0.97	0.97	63
Beetle	0.98	0.93	0.96	70
Field cricket	0.97	0.98	0.98	62
Grasshopper	0.98	0.97	0.97	72
Mites	0.98	0.93	0.96	60
Thunder fly	0.93	0.95	0.94	73
Sawfly	0.99	0.97	0.98	116
Stem borer	0.97	1.00	0.98	80
Weevils	1.00	0.95	0.97	62
Accuracy			0.98	729
Macro avg	0.97	0.97	0.97	729
Weighted avg	0.97	0.98	0.97	729

4.2.3. *Comparative analysis of datasets SD, MD, and LD*

It is clear from comparing the HPMA model's outcomes for the three datasets that the model performs consistently and robustly, as shown in Table 6 and Figs. 10 and 11. Dataset SD, which is a collected dataset, showed high accuracy, precision, recall, and F1 score, as shown in Table 5. For real pest classification applications, the model was able to achieve a high level of accuracy on the pests in this dataset. On benchmark Dataset MD, the HPMA model demonstrated remarkably flexible behavior. While working with a different dataset, it maintained good levels of accuracy, precision, and recall, indicating that the model's capabilities are not restricted to that dataset.

This flexibility is necessary to apply its use to different pest classifications. This versatility makes it essential to expand its use to other pest classification circumstances. The richness and diversity of Dataset LD further demonstrate the power of the HPMA model. The model nonetheless managed to attain outstanding accuracy, precision, and recall in spite of the difficulties provided by LD. Indicating that the HPMA model is well adapted to handle complicated, real-world circumstances where pest classification is frequently difficult, Figs. 10 and 11 gives an additional visual representation of these findings.

In summary, the comparative analysis unequivocally establishes that the HPMA model consistently achieves high-performance results across different datasets, showcasing its adaptability and reliability. While the model exhibits exceptional accuracy, precision, recall, and F1 scores in the collected small dataset (SD), it also demonstrates remarkable flexibility in maintaining commendable levels of performance across the benchmark medium dataset (MD) and the large dataset (LD). This versatility positions the HPMA model as an invaluable tool for precise and dependable pest classification, with the potential for widespread applications in diverse agricultural pest control.

4.3. *Comparative analysis with state of art algorithm*

The comparative analysis of the proposed Hybrid Pooled Multihead Attention (HPMA) model for pest classification was conducted alongside other state-of-the-art models. These models include convolution models, MLP models, hybrid transformer models, and transformer models [43, 40,44, 45]]. The analysis was performed across three datasets. Table 7 illustrates the varying parameter sizes, FLOPs, image sizes, and accuracy of these models. Notably, Transformer models exhibit the largest parameter sizes and FLOPs, yet they consistently achieve the highest accuracy [46,47].

In contrast, MLP models, being smaller and faster, sacrifice accuracy. Hybrid Transformer models strike a balance between precision and efficiency. Across all three datasets, the HPMA Transformer model, as proposed, attains cutting-edge accuracy. It manages to do so with parameter sizes and FLOPs comparable to other transformer models, demonstrating commendable efficiency. These results underscore the HPMA model's prowess in accurately categorizing pests while maintaining a favorable trade-off between model complexity, performance, and computational requirements.

4.4. *Discussion*

The exceptional performance of the HPMA model can be ascribed to its pioneering hybrid pooled multihead attention mechanism. This mechanism enables the model to effectively capture both local and global information, thereby augmenting its classification capabilities significantly. This adaptability plays a pivotal role in the model's consistent success across a wide array of datasets. The HPMA model's excellence lies in its capacity to distinguish between different pest classes. It achieves this by effectively filtering out irrelevant or noisy information, focusing on critical areas within the image, and understanding the connections between various image components. This adaptability bolsters the model's resilience and its capacity to generalize, rendering it an invaluable asset for pest classification.

Moreover, the HPMA model introduces a distinctive element with hybrid pooled "k" and "v" values. This sets it apart from conventional attention models, which typically rely on the same query ("q"), key ("k"), and value ("v") representations. This innovative approach substantially

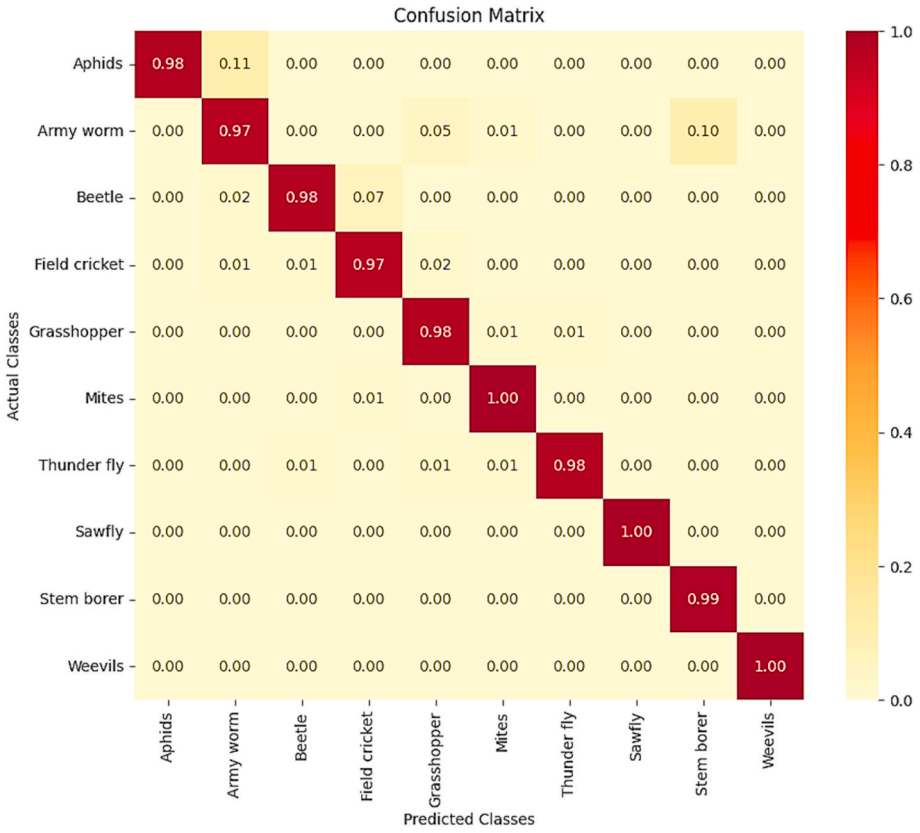

Fig. 6. Confusion matrix of HPMA model.

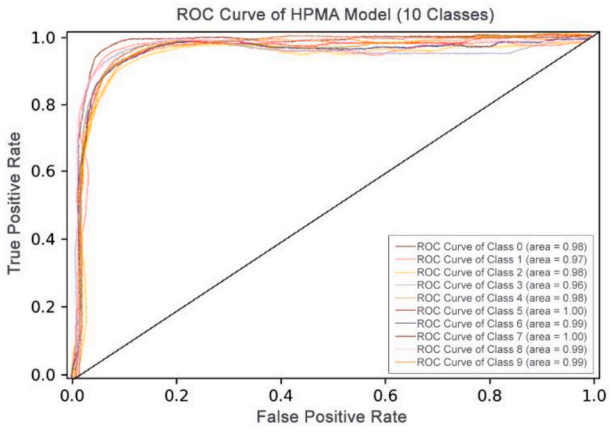

Fig. 7. ROC curve of HPMA model.

enhances the model's discriminative power, enabling it to extract more profound insights from the input data. The incorporation of these hybrid pooled "k" and "v" values significantly contributes to the model's remarkable performance on the assessed datasets.

The HPMA model introduces a hybrid pooling approach that considers both the most and least significant features in the attention calculation. This involves max and min pooling operations to create two pooled feature maps, which are then concatenated to form a hybrid pooled feature map. The model extracts the query from the original feature map and the keys/values from the hybrid pooled feature map, ensuring a comprehensive consideration of feature relationships. While the HPMA model showcases impressive discriminative power and robust accuracy, additional investigation is necessary to evaluate its performance on imbalanced datasets. Furthermore, optimizing its architecture for deployment on resource-constrained devices, commonly used in pest classification scenarios, is crucial. This will ensure its broader practical applicability in real-world pest management settings. The advantages of the HPMA model include improved feature discrimination, robustness to noise by focusing on informative parts of the feature map, and potential generalization to tasks beyond pest classification. The rationale for this design is to address the limitations of traditional self-attention. It aims to provide a nuanced understanding of feature distribution for improved performance in vision-based tasks requiring fine-grained discrimination. In summary, the HPMA model offers a novel approach to attention mechanisms, potentially leading to more accurate and robust results in vision tasks with subtle visual variations.

5. Ablation study

5.1. Investigating the proposed HPMA model

In order to gain a comprehensive understanding of the proposed Hybrid Pooled Multihead Attention (HPMA) model, an in-depth investigation was conducted through various analytical techniques. The insights obtained through feature map analysis, attention map analysis, mean distance calculation, and attention heatmap visualizations (depicted in Figs. 12–16) played a crucial role in unravelling the inner workings and classification procedures of the HPMA model [48].

5.1.1. Feature map analysis

Feature map analysis has enabled exploration into the hierarchical representation of input data within the model. Through a thorough examination of the feature maps, it has become feasible to discern how the model hierarchically extracts and processes information at various levels of abstraction. This process sheds light on the model's comprehension of intricate patterns and features within the input data. Fig. 12 visually encapsulates this analysis, providing a graphical representation of the hierarchical extraction and processing of information.

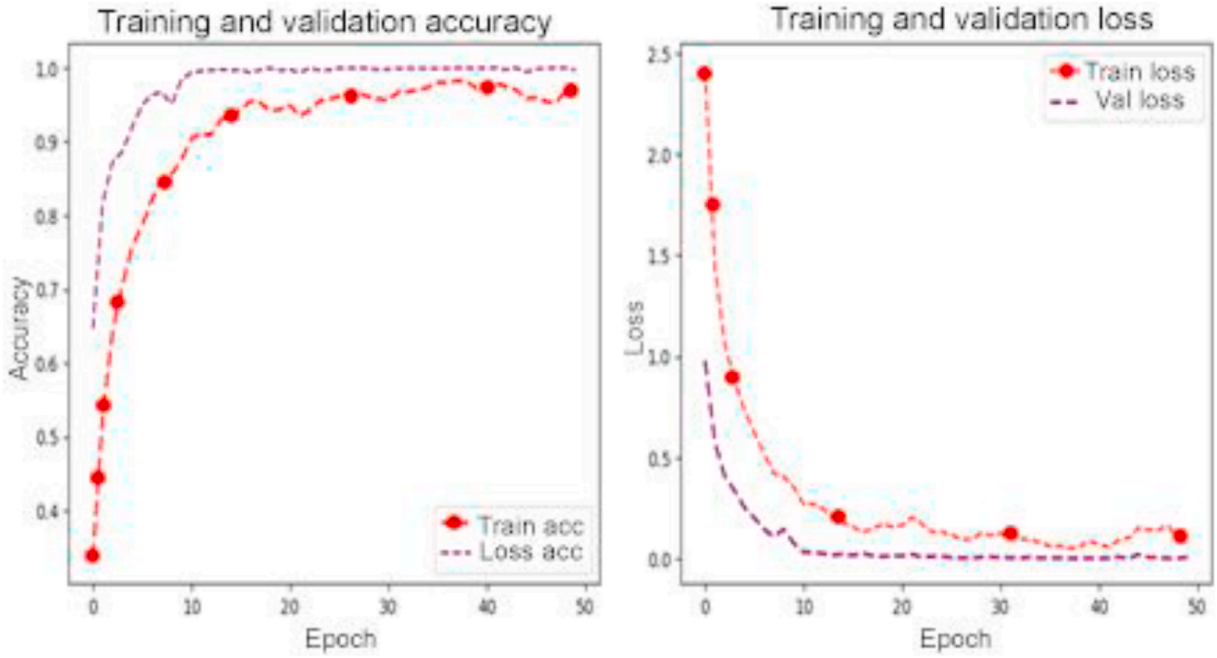

Fig. 8. Dataset MD accuracy and loss of HPMA.

Table 6
Numerical analysis of proposed HPMA model on three datasets.

Results	Dataset SD	Dataset MD	Dataset LD
Epochs	50	50	50
Train time	90 s	101 s	168 s
Train accuracy	0.9904	0.9956	0.9723
Test accuracy	0.9894	0.9802	0.9598
Train loss	0.3419	0.2101	0.5231
Test loss	0.2011	0.3222	0.4268
Precision	0.9820	0.9892	0.9655
Recall	0.9775	0.9788	0.9597
F1score	0.9751	0.9888	0.9595

5.1.2. *Attention map analysis*

The investigation included attention map analysis, a powerful technique for interpreting the model's predictions. Fig. 13 present attention maps that highlight the specific regions in the input images that the model deemed critical for accurate classification. This granular insight into the attention mechanism facilitated a precise understanding of the model's decision-making process and helped validate its reasoning.

5.1.3. *Mean distance calculation*

Mean distance calculation, as illustrated in Fig. 14, provided a quantitative measure of how information propagated and interacted between different layers (6 layers) of the model. This analysis was instrumental in comprehending how the model represented both local

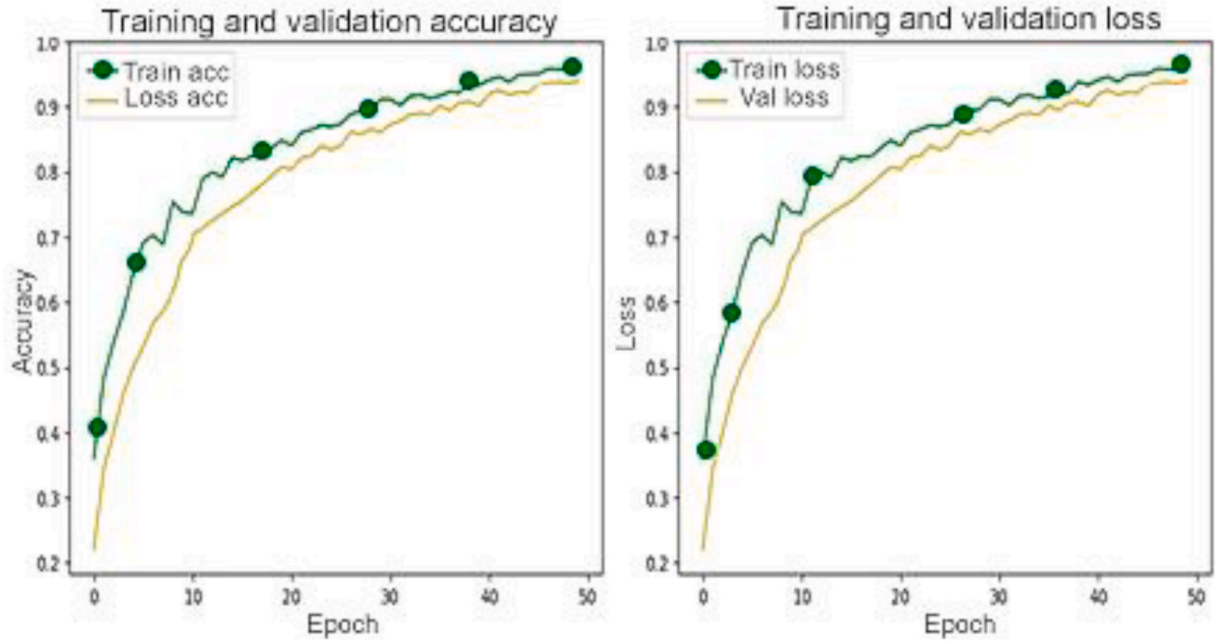

Fig. 9. Dataset LD accuracy and loss of HPMA.

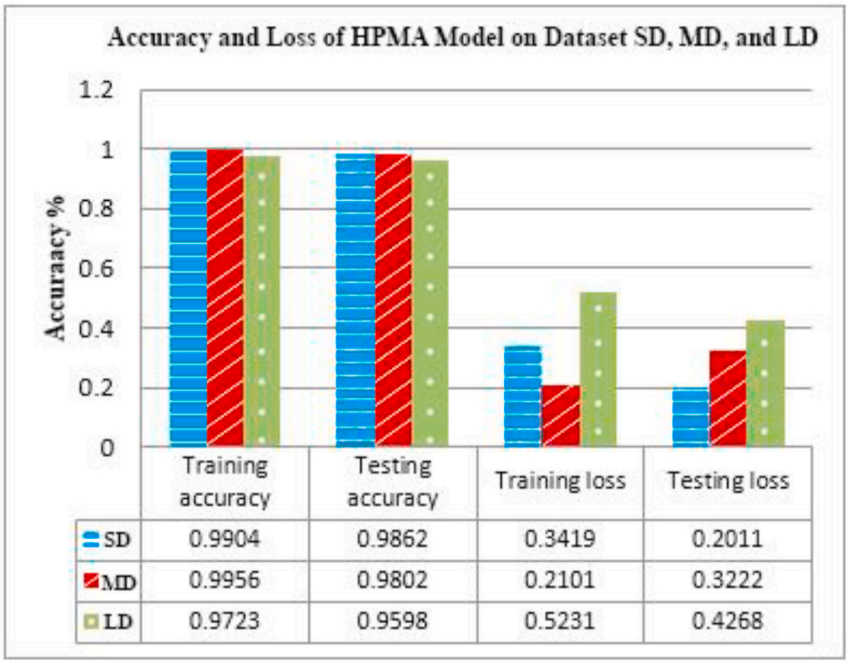

Fig. 10. Accuracy and loss of HPMA model.

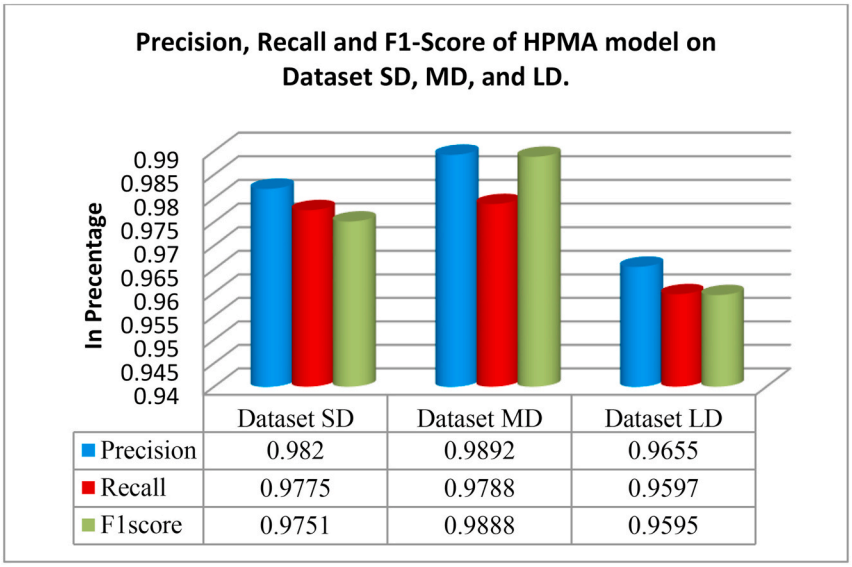

Fig. 11. Precision, Recall and F1 score of HPMA model.

and global dependencies within the image. It contributed to a nuanced understanding of the information flow and interaction dynamics within the HPMA model.

5.1.4. Multihead attention mechanism

To elucidate the multihead attention mechanism of the model, attention heatmaps for the six heads were examined. This analysis, as depicted in Fig. 15, provided insights into the individual contributions of each attention head in capturing various components of the image. Understanding the function of each head contributed to a holistic comprehension of the model's attention mechanism, enhancing transparency and interpretability.

5.1.5. Attention heatmap visualization

Fig. 16 presented attention heatmap visualizations for the largest army worm pest and the smallest tiny aphid pest. These visualizations offered additional insights into how the HPMA model assimilated both local and global information from the input. By scrutinizing these heatmaps, the HPMA model gained a deeper understanding of the specific image regions that played a pivotal role in the model's decision-making process, further enhancing interpretability.

In conclusion, the collective findings from the feature map, attention map, mean distance calculation, and attention heatmap analyses significantly visualize the overall understanding of the HPMA model's functionality. This enhanced understanding not only contributed to the interpretability of the model but also bolstered its performance and applicability in pest classification tasks. The insights gained from these analyses are crucial for refining the model and advancing its effectiveness in real-world applications.

Table 7
Comparative analysis of HPMA with SOTA

Model	Parameter (M)	Disk Size (MB)	GFLOPs	Image size	Accuracy (%)		
					SD	MD	LD
Convolution Net							
Inception v3	23	~102	5.2	224 × 224	**82**	77	89
EfficientNet	17	~78	2.3	224 × 224	**60**	73	76
MobileNetv2	2.3	~20.2	0.55	224 × 224	**87**	88	54
ResNet18	11.1	~58.4	2	224 × 224	**90**	76	80
ResNet 34	20	~89.2	3.1	224 × 224	**91**	89	79
VGG19	142	~559	15.56	224 × 224	**89**	88	64
VGG16	196	~589	24.1	224 × 224	**88**	91	62
MLP							
MLP-Mixer	115.2	~480	11	224 × 224	**57**	63	66
GMLP	25.6	~120	4	224 × 224	**76**	82	80
FNet	18.8	~83.2	3	224 × 224	**81**	85	79
Hybrid Transformer							
EfficinentNetB4 + ViT	23.2	~101.8	3.91	224 × 224	**91**	87	90
ResViT –S/16	22.1	~100.4	4.1	224 × 224	**90**	88	83
ResViT –B/16	86	~351	17.9	224 × 224	**88**	91	92
MobViT-S/16	13.6	~72.4	2.3	224 × 224	**63**	77	89
MobViT-B/16	27.9	~126.6	4.9	224 × 224	**71**	55	80
Transformer							
DeiT-S/16	22	~98	3.6	224 × 224	**81**	90	91
ViT-B/16	86	~354	17	224 × 224	**94**	95	93
Swin- B/16	88	~362	15.1	224 × 224	**95**	91	96
HPMA (our model)	**18.8**	**~83.2**	**3.1**	224 × 224	**98**	**98**	**95**

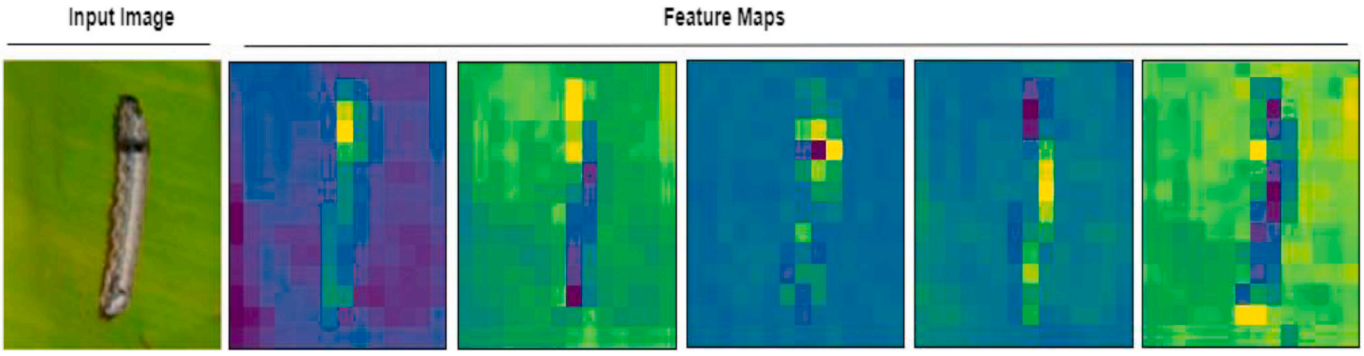

Fig. 12. Visualizing feature map.

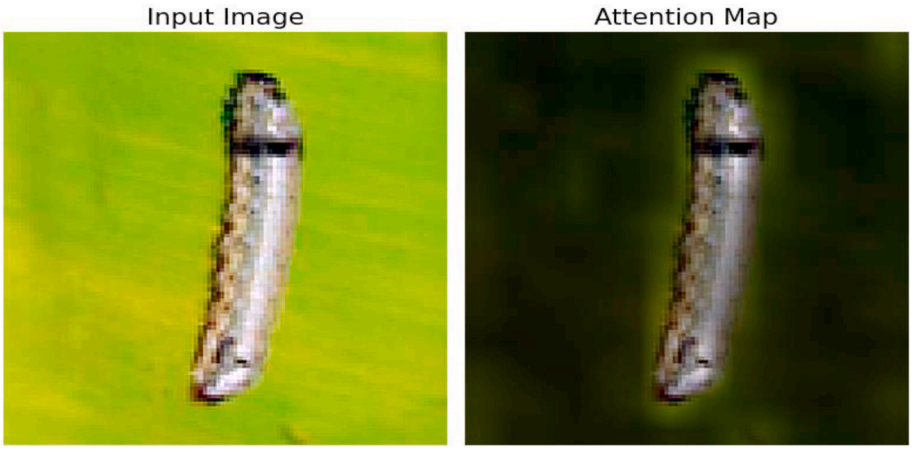

Fig. 13. Visualizing attention map.

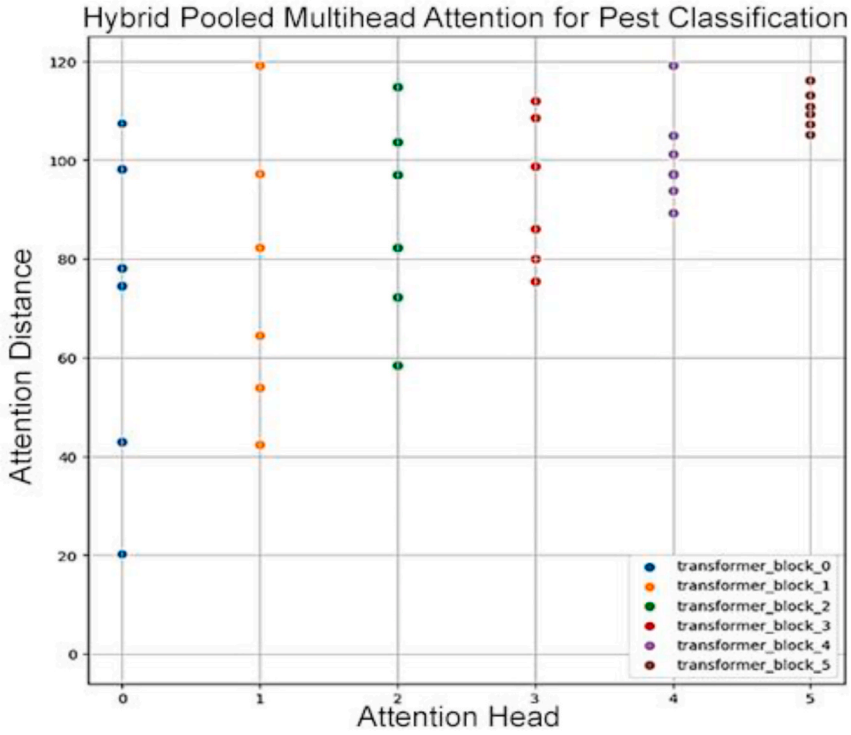

Fig. 14. Mean attention distance of 6 heads.

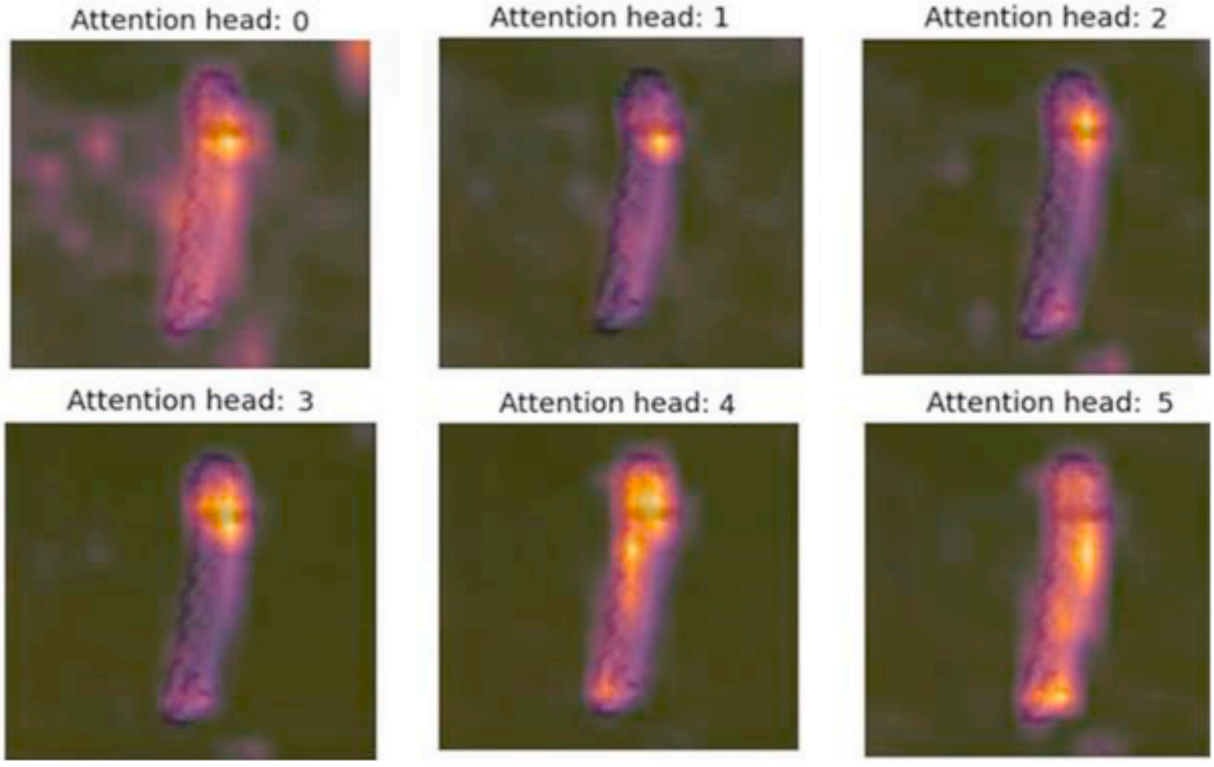

Fig. 15. Mean attention distance of 6 heads.

5.2. Effects of data split, pooling components and optimizers on HPMA model

The ablation study evaluated different pooling components in the HPMA model on three datasets (SD, MD, and LD). Results from Table 8 showed that the HPMA model with min + max pooling achieved the highest accuracy of 98 % on all datasets. This indicates its importance in capturing both local and global information precisely, with min pooling for fine-grained features and max pooling for coarse-grained features.

Other pooling combinations, such as max + product pooling and

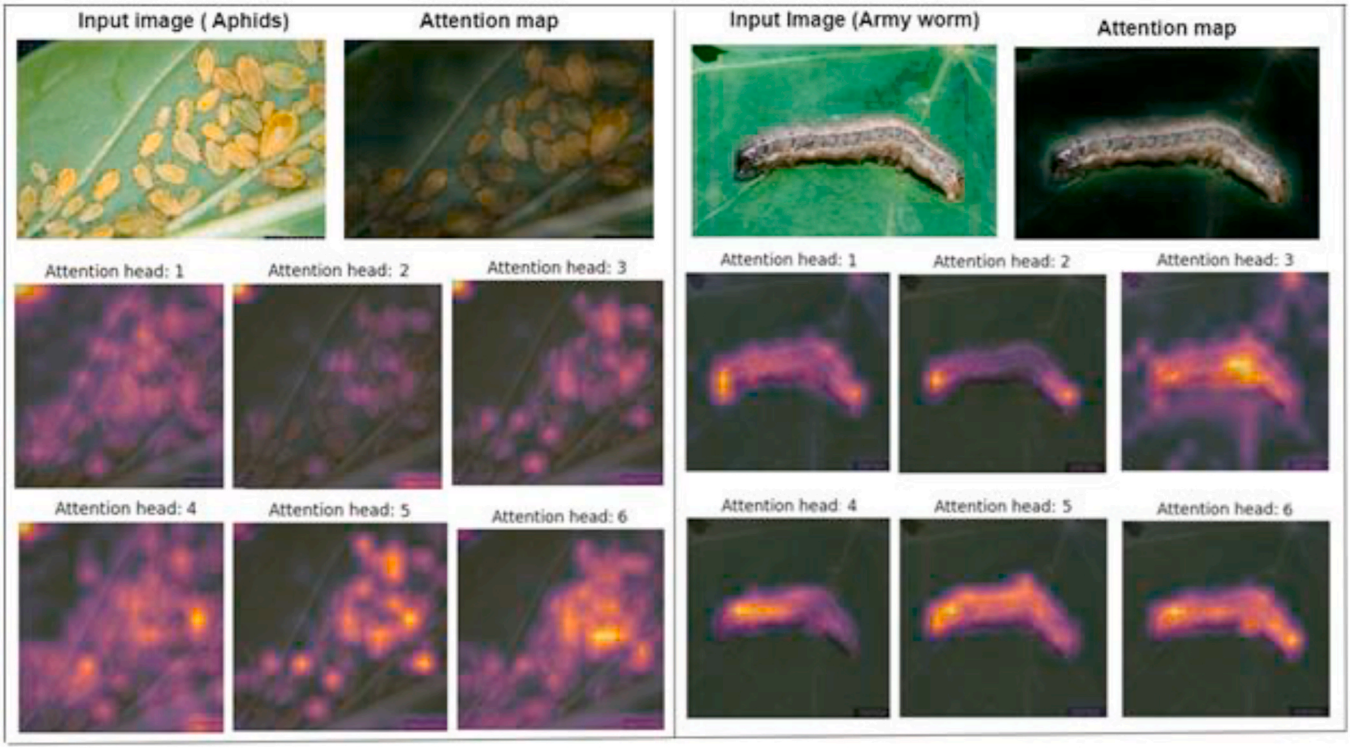

Fig. 16. Visualizing attention map and attention head of aphids and army worm of a HPMA model.

Table 8
Effects of pooling components.

Hybrid Pooling on Proposed HPMA	Accuracy in %		
	SD	MD	LD
Max + Product pooling	85	89	91
Max + Spatial pyramid pooling	82	71	90
Average + Spatial pyramid	66	89	81
Selective max + Average	92	90	73
HPMA (Min + Max pooling)	**98**	**98**	**95**

selective max + average pooling, also showed varying accuracies across datasets. The study provides insights into the significance of pooling methods, guiding future research for improved hybrid pooling architectures in the HPMA model and similar approaches.

By varying the batch size from 8 to 32 in the HPMA model, it was observed that a batch size of 16 achieved good accuracy across all three datasets. Similarly, when different optimizers were used, such as SGD, Adagrad, Rmsprop, and Adadelta, the accuracy of the HPMA model varied with increases of +1, +1.2, −0.5, and −1.5, respectively, on all three datasets [49]. However, when the Adam optimizer was used, the accuracy improved significantly by +1.8. These findings suggest that the HPMA model is well optimized with the Adam optimizer, yielding good accuracy on all three datasets.

To assess the impact of dataset splits, three configurations were evaluated 70/20/10 (train/validation/test), 60/20/20, and 80/10/10. While all splits resulted in satisfactory performance with accuracy variations of ±1 % or ±0.5 %, the 80/10/10 split yielded the highest accuracy. This suggests that the model is relatively insensitive to variations in dataset splits, and the 80/10/10 configuration might be most effective for this task. However, it's important to note that the data itself was not altered. Only the proportions used for training, validation, and testing were changed.

In summary, HPMA achieves remarkable accuracy in pest classification, with a training accuracy of 99.04 % and a testing accuracy of 98 % on a collected dataset (SD) consisting of 10 pest classes. Its effectiveness is also validated on two benchmark datasets (MD and LD), demonstrating its versatility and adaptability to diverse pest datasets. It achieves high accuracy on these benchmark datasets as well, showcasing its ability to handle different pest classification scenarios. HPMA is also computationally efficient, providing a favorable trade-off between model complexity, performance, and computational requirements. Additionally, its attention mechanisms and ablation studies help understand how it processes information and what components of the image are crucial for accurate classification. This interpretability enhances the model's transparency and trustworthiness. Overall, HPMA is a significant contribution to the field of pest classification. It is a novel, versatile, and efficient model that achieves remarkable accuracy on diverse pest datasets.

6. Conclusion and future work

The Hybrid Pooled Multihead Attention (HPMA) mechanism is presented in this study as a novel method for precise pest classification in agricultural applications. The HPMA model effectively captures both local and global features within images. It outperforms conventional CNN models and vision transformers by incorporating hybrid pooling techniques and altering the attention mechanism. The experimental outcomes show the HPMA model's superior accuracy on a newly built dataset, achieving an exceptional accuracy of 98 % across 10 pest classes. Additionally, benchmark datasets are used to validate the model's robustness and generalization abilities, yielding accuracy rates of 98 % and 95 % respectively. The ablation study confirms their critical roles in the model's exceptional performance in pest classification and advances the understanding of the fundamental components which make the entire HPMA model. Results of the study and ablation analysis show how well the proposed HPMA model performs in accurately classifying pests. Adopting such a model offers a promising strategy for addressing pest problems in agriculture, enabling quick pest control actions and reducing crop losses.

CRediT authorship contribution statement

T. Saranya: Writing – review & editing, Writing – original draft, Validation, Resources, Methodology, Funding acquisition, Data curation. **C. Deisy:** Visualization, Validation, Supervision, Investigation, Formal analysis. **S. Sridevi:** Formal analysis, Conceptualization, Funding acquisition, Investigation, Supervision.

Declaration of competing interest

My coauthors and I do not have any conflicts of interests to disclose.

Acknowledgement

The researchers would like to express their gratitude to the Thiagarajar College of Engineering (TCE) for their invaluable support in carrying out this research. Furthermore, they sincerely acknowledge the financial support received from the Thiagarajar Research Fellowship Scheme, affiliated with Thiagarajar College of Engineering in Madurai, Tamilnadu, India.

References

[1] T. Saranya, C. Deisy, S. Sridevi, K.S.M. Anbananthen, A comparative study of deep learning and Internet of Things for precision agriculture, Eng. Appl. Artif. Intell. 122 (2023) 106034.
[2] E. Ayan, H. Erbay, F. Varçın, Crop pest classification with a genetic algorithm-based weighted ensemble of deep convolutional neural networks, Comput. Electron. Agric. 179 (2020) 105809.
[3] Z. Su, J. Luo, Y. Wang, Q. Kong, B. Dai, Comparative study of ensemble models of deep convolutional neural networks for crop pests classification, Multimed. Tool. Appl. (2023) 1–20.
[4] M.T. Mallick, S. Biswas, A.K. Das, H.N. Saha, A. Chakrabarti, N. Deb, Deep learning based automated disease detection and pest classification in Indian mung bean, Multimed. Tool. Appl. 82 (8) (2023) 12017–12041.
[5] N. Dilshad, T. Khan, J. Song, Efficient deep learning framework for fire detection in complex surveillance environment, Comput. Syst. Sci. Eng. 46 (1) (2023) 749–764.
[6] H. Yar, Z.A. Khan, F.U.M. Ullah, W. Ullah, S.W. Baik, A modified YOLOv5 architecture for efficient fire detection in smart cities, Expert Syst. Appl. 231 (2023) 120465.
[7] S. Parez, N. Dilshad, T.M. Alanazi, J.W. Lee, Towards sustainable agricultural systems: a lightweight deep learning model for plant disease detection, Comput. Syst. Sci. Eng. 47 (1) (2023) 515–536.
[8] F. Qi, G. Chen, J. Liu, Z. Tang, End-to-end pest detection on an improved deformable DETR with multiheadcriss cross attention, Ecol. Inf. 72 (2022) 101902.
[9] Z. Niu, G. Zhong, H. Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.
[10] Vincent Christlein, Lukas Spranger, Mathias Seuret, Anguelos Nicolaou, Pavel Král, Andreas Maier, Deep generalized max pooling, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1090–1096.
[11] K. Rimal, K.B. Shah, A.K. Jha, Advanced multi-class deep learning convolution neural network approach for insect pest classification using TensorFlow, Int. J. Environ. Sci. Technol. 20 (4) (2023) 4003–4016.
[12] S. Yonbawi, S. Alahmari, R. Daniel, E.L. Lydia, M.K. Ishak, H.K. Alkahtani, A. Aljarbouh, S.M. Mostafa, Modified metaheuristics with transfer learning based insect pest classification for agricultural crops, Comput. Syst. Sci. Eng. 46 (3) (2023).
[13] A.A. Rani, K.L. Prasanna, M.S. Ashraf, A.K. Dey, M.A.A. Walid, D.R.K. Saikanth, Classification for crop pest on U-SegNet, in: 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2023, February, pp. 926–932.
[14] W. Li, T. Zheng, Z. Yang, M. Li, C. Sun, X. Yang, Classification and detection of insects from field images using deep learning for smart pest management: a systematic review, Ecol. Inf. 66 (2021) 101460.
[15] Y. Ai, C. Sun, J. Tie, X. Cai, Research on recognition model of crop diseases and insect pests based on deep learning in harsh environments, IEEE Access 8 (2020) 171686–171693.
[16] Z. Ünal, Smart farming becomes even smarter with deep learning—a bibliographical analysis, IEEE Access 8 (2020) 105587–105609.
[17] C.J. Chen, Y.Y. Huang, Y.S. Li, C.Y. Chang, Y.M. Huang, An AIoT based smart agricultural system for pests detection, IEEE Access 8 (2020) 180750–180761.
[18] R. Hadipour-Rokni, E.A. Asli-Ardeh, A. Jahanbakhshi, S. Sabzi, Intelligent detection of citrus fruit pests using machine vision system and convolutional neural network through transfer learning technique, Comput. Biol. Med. 155 (2023) 106611.
[19] W. Xia, D. Han, D. Li, Z. Wu, B. Han, J. Wang, An ensemble learning integration of multiple CNN with improved vision transformer models for pest classification, Ann. Appl. Biol. 182 (2) (2023) 144–158.
[20] M.T. Mallick, S. Biswas, A.K. Das, H.N. Saha, A. Chakrabarti, N. Deb, Deep learning based automated disease detection and pest classification in Indian mung bean, Multimed. Tool. Appl. 82 (8) (2023) 12017–12041.
[21] T. Zheng, X. Yang, J. Lv, M. Li, S. Wang, W. Li, An efficient mobile model for insect image classification in the field pest management, Engineering Science and Technology, an International Journal 39 (2023) 101335.
[22] A.A. Rani, K.L. Prasanna, M.S. Ashraf, A.K. Dey, M.A.A. Walid, D.R.K. Saikanth, Classification for crop pest on U-SegNet, in: 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), IEEE, 2023, February, pp. 926–932.
[23] L. Li, S. Zhang, B. Wang, Plant disease detection and classification by deep learning—a review, IEEE Access 9 (2021) 56683–56698.
[24] X. Lu, R. Yang, J. Zhou, J. Jiao, F. Liu, Y. Liu, B. Su, P. Gu, A hybrid model of ghost-convolution enlightened transformer for effective diagnosis of grape leaf disease and pest, Journal of King Saud University-Computer and Information Sciences 34 (5) (2022) 1755–1767.
[25] L. Jiao, C. Xie, P. Chen, J. Du, R. Li, J. Zhang, Adaptive feature fusion pyramid network for multi-classes agricultural pest detection, Comput. Electron. Agric. 195 (2022) 106827.
[26] A.I. Jajja, A. Abbas, H.A. Khattak, G. Niedbała, A. Khalid, H.T. Rauf, S. Kujawa, Compact convolutional transformer (CCT)-Based approach for whitefly attack detection in cotton crops, Agriculture 12 (10) (2022) 1529.
[27] H. Liu, Y. Zhan, H. Xia, Q. Mao, Y. Tan, Self-supervised transformer-based pre-training method using latent semantic masking auto-encoder for pest and disease classification, Comput. Electron. Agric. 203 (2022) 107448.
[28] Y. Peng, Y. Wang, CNN and transformer framework for insect pest classification, Ecol. Inf. 72 (2022) 101846.
[29] C. Wang, J. Zhang, J. He, W. Luo, X. Yuan, L. Gu, A two-stream network with complementary feature fusion for pest image classification, Eng. Appl. Artif. Intell. 124 (2023) 106563.
[30] S. Wang, Q. Zeng, W. Ni, C. Cheng, Y. Wang, ODP-Transformer: interpretation of pest classification results using image caption generation techniques, Comput. Electron. Agric. 209 (2023) 107863.
[31] X. Yang, Y. Luo, M. Li, Z. Yang, C. Sun, W. Li, Recognizing pests in field-based images by combining spatial and channel attention mechanism, IEEE Access 9 (2021) 162448–162458.
[32] Y. Zhang, L. Chen, Y. Yuan, Multimodal fine-grained transformer model for pest recognition, Electronics 12 (12) (2023) 2620.
[33] M.L. Huang, T.C. Chuang, A database of eight common tomato pest images, Mendeley Data 1 (2020).
[34] C. Xie, R. Wang, J. Zhang, P. Chen, W. Dong, R. Li, T. Chen, H. Chen, Multi-level learning features for automatic classification of field crop pests, Comput. Electron. Agric. 152 (2018) 233–241.
[35] X. Wu, C. Zhan, Y.K. Lai, M.M. Cheng, J. Yang, Ip102: a large-scale benchmark dataset for insect pest recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8787–8796.
[36] K. Thenmozhi, U.S. Reddy, Crop pest classification based on deep convolutional neural network and transfer learning, Comput. Electron. Agric. 164 (2019) 104906.
[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2020 *arXiv preprint arXiv:2010.11929*.
[38] Z. Pan, J. Cai, B. Zhuang, Fast vision transformers with hilo attention, Adv. Neural Inf. Process. Syst. 35 (2022) 14541–14554.
[39] Q. Dai, X. Cheng, Y. Qiao, Y. Zhang, Agricultural pest super-resolution and identification with attention enhanced residual and dense fusion generative and adversarial network, IEEE Access 8 (2020) 81943–81959.
[40] L. Nanni, A. Manfè, G. Maguolo, A. Lumini, S. Brahnam, High performing ensemble of convolutional neural networks for insect pest image detection, Ecol. Inf. 67 (2022) 101515.
[41] H.T. Ung, H.Q. Ung, B.T. Nguyen, An Efficient Insect Pest Classification Using Multiple Convolutional Neural Network Based Models, 2021 arXiv preprint arXiv: 2107.12189.
[42] N.E.M. Khalifa, M.O.H. Loey, M.H.N. Taha, Insect pests recognition based on deep transfer learning models, J. Theor. Appl. Inf. Technol. 98 (1) (2020) 60–68.
[43] Z. Su, J. Luo, Y. Wang, Q. Kong, B. Dai, Comparative study of ensemble models of deep convolutional neural networks for crop pests classification, Multimed. Tool. Appl. (2023) 1–20.
[44] I.O. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, Mlp-mixer: an all-mlp architecture for vision, Adv. Neural Inf. Process. Syst. 34 (2021) 24261–24272.
[45] J. Lee-Thorp, J. Ainslie, I. Eckstein, S. Ontanon, Fnet: Mixing Tokens with Fourier Transforms, 2021 arXiv preprint arXiv:2105.03824.
[46] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[47] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in transformer, Adv. Neural Inf. Process. Syst. 34 (2021) 15908–15919.
[48] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, A. Dosovitskiy, Do vision transformers see like convolutional neural networks? Adv. Neural Inf. Process. Syst. 34 (2021) 12116–12128.
[49] T. Saranya, C. Deisy, S. Sridevi, K.S. Muthu, M.A. Khan, Performance analysis of first order optimizers for plant pest detection using deep learning, in: International Conference on Machine Learning, Image Processing, Network Security and Data Sciences, Springer Nature Switzerland, Cham, 2022, December, pp. 37–52.