WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Hybrid Deep Learning for Pest Identification

M. Akila[1], Nithishkumar N C[2], Vishnu K[3]

[1]Assistant Professor, Department of Information Technology, K.L.N. College of Engineering, Madurai, India, akilasaran.m@gmail.com
[2]Student, Department of Information Technology, K.L.N. College of Engineering, Madurai, India, nithishkumarnc2005@gmail.com
[3]Student, Department of Information Technology, K.L.N. College of Engineering, Madurai, India, vishnukumar36455@gmail.com

A B S T R A C T

Agricultural pest infestations pose a significant threat to global food security, causing substantial yield losses. Traditional pest identification methods are time-consuming and require expert knowledge. This paper proposes a hybrid CNN-ViT deep learning framework for automated pest identification across 40 pest classes using the IP102 dataset. The model combines Convolutional Neural Networks for local feature extraction with Vision Transformers for global context understanding, achieving 90% classification accuracy. The system deployed as a Java mobile application, enabling farmers to capture pest images in the field and receive instant identification with treatment suggestions. This work aligns with SDG 2 (Zero Hunger) and SDG 9 (Industry, Innovation, and Infrastructure).

Keywords: Deep Learning, Hybrid CNN-ViT, Pest Identification, IP102 Dataset, Mobile Application

1. Introduction

Agriculture remains the backbone of the global economy, supporting billions of people worldwide. However, crop pests continue to threaten food security, causing 20-40% of annual crop losses globally. Timely and accurate pest identification is crucial for effective pest management and minimizing crop damage. Traditional identification methods depend on expert entomologists and manual inspection, which are often unavailable in rural and resource-limited areas.

Despite significant advancements, pest management remains a critical challenge. Manual pest identification requires specialized knowledge and is prone to human error. In resource-limited agricultural regions, access to entomologists and pest management experts is scarce, leading to delayed diagnosis. Misidentification of pests results in ineffective treatment, increased chemical usage, environmental pollution, and financial losses for farmers.

This paper addresses these challenges by developing an automated pest identification system using deep learning. The proposed hybrid architecture combines the local feature extraction capabilities of CNNs with the global context understanding of Vision Transformers, providing accurate and reliable pest classification. The system is deployed as a mobile application with offline capabilities, making it accessible to farmers in remote areas.

2. Related Work

Wu et al. (2019) introduced the IP102 dataset with 75,000+ images across 102 pest species and tested CNNs like ResNet, DenseNet, and Inception, with ResNet-50 achieving 54% accuracy. This established IP102 as the benchmark for fine-grained pest recognition, though its difficulty highlights the challenge of distinguishing visually similar pest species.

Nanni et al. (2020) proposed an ensemble of VGG16, ResNet, and DenseNet tuned by bio-inspired algorithms, improving accuracy by handling high intra-class variation and inter-class similarity effectively. Their work demonstrated that ensemble methods can capture diverse feature representations necessary for fine-grained pest classification.

Thenmozhi and Reddy (2019) used transfer learning with VGG16, ResNet50, and InceptionV3 for pest classification with limited data, achieving 85-92% accuracy. This demonstrated transfer learning's effectiveness in agricultural image tasks where collecting large-scale annotated datasets is challenging.

Wang et al. (2021) developed a modified ResNet model with squeeze-and-excitation blocks, achieving 96.5% accuracy on 8 pest species. Their work proved that architecture tuning with attention mechanisms can enhance pest recognition by focusing on discriminative features.

Kasinathan et al. (2021) used YOLOv4 for detection and EfficientNet for classification, reaching 89% detection and 94% classification accuracy. This demonstrated a strong two-stage pipeline for field pest identification, separating localization from classification tasks.

Dosovitskiy et al. (2021) introduced Vision Transformers (ViT), achieving state-of-the-art results on ImageNet by treating images as sequences of patches and applying transformer architectures. This inspired hybrid CNN-Transformer models now widely used for fine-grained image tasks like pest recognition.

Li et al. (2021) provided a comprehensive review of deep learning methods for insect pest detection, highlighting the progression from traditional machine learning to modern deep learning approaches and identifying key challenges including dataset imbalance, fine-grained classification difficulty, and real-world deployment constraints.

3. Proposed Methodology

3.1 System Architecture

The proposed system follows a multi-stage pipeline consisting of image acquisition, preprocessing, feature extraction, classification, recommendation generation, visualization, and report generation.

Image acquisition occurs through the mobile application where farmers capture pest images using the device camera or upload from gallery. The captured images often have varying resolutions, lighting conditions, and backgrounds typical of field environments.

Preprocessing standardizes input images through resizing to 224×224 pixels, normalization to [0,1] range, and data augmentation including rotation, flipping, color jittering, and random cropping. This ensures model robustness across diverse field conditions.

The hybrid CNN-ViT model processes preprocessed images through two complementary pathways. The CNN backbone extracts local hierarchical features, while the Vision Transformer captures global contextual relationships. The classification head outputs probability distributions across 40 pest classes.

Post-processing includes confidence thresholding, top-K prediction selection. Grad-CAM visualization generates attention heatmaps highlighting pest-relevant image regions.

Finally, a comprehensive report is generated containing pest identification results, confidence scores, Grad-CAM visualizations, detailed pest information.

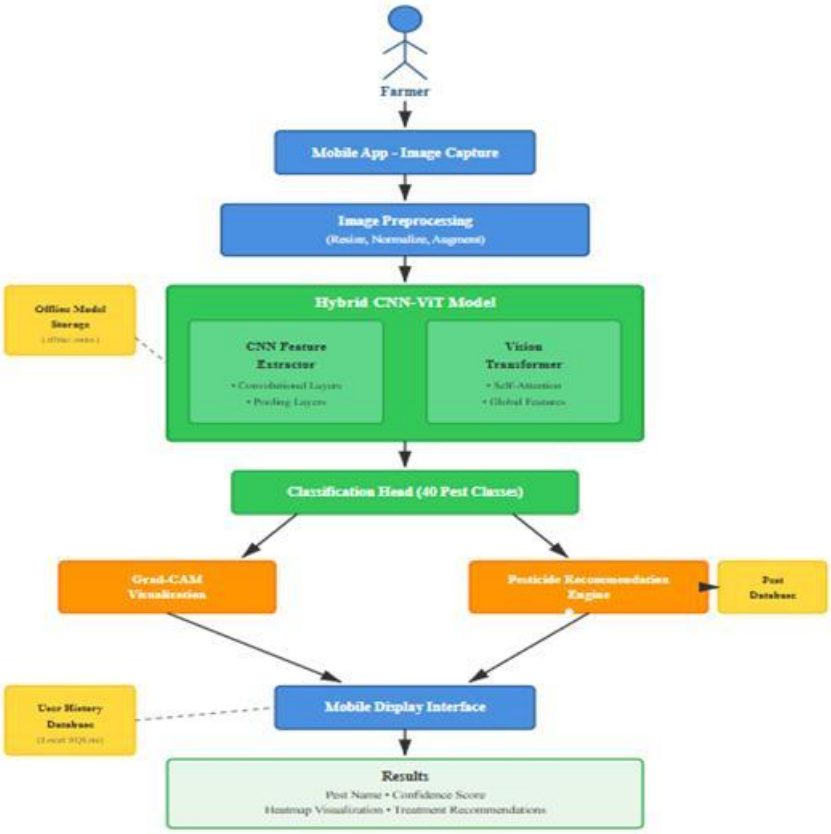

Fig. 1 - System Pipeline

3.2 Hybrid CNN-ViT Model Architecture

The core innovation lies in the hybrid architecture combining CNN and Vision Transformer strengths. The model consists of three major components: CNN backbone, Vision Transformer encoder, and classification head.

The CNN backbone uses ResNet50 pre-trained on ImageNet for transfer learning. ResNet50 extracts hierarchical local features including low-level textures and edges in early layers, mid-level patterns like pest body segments in intermediate layers, and high-level semantic features in deeper layers. The first 143 layers are frozen to retain pre-learned features while the remaining layers are fine-tuned on pest images. The final convolutional output has spatial dimensions 7×7 with 2048 feature channels.

The output is reshaped into 49 patches of 2048-dimensional features. A linear projection transforms features to 768-dimensional embeddings. Learnable positional embeddings preserve spatial information. Eight transformer encoder blocks process the sequence, each containing 12-head self-attention and MLP layers with hidden dimension 3072 and GELU activation.

The classification head consists of global average pooling followed by fully connected layers (768→512→256→40) with dropout (0.3, 0.2) and softmax activation. Training uses AdamW optimizer (learning rate = 1e-4, weight decay = 1e-5) with categorical

3.3 Augmentation Strategy

To improve model robustness and generalization to real-world field conditions, extensive data augmentation is applied during training. Geometric transformations include:

- Random rotation (±20 degrees) to handle arbitrary pest orientations
- Horizontal and vertical flipping to account for bilateral symmetry
- Width and height shifts (±20%) to simulate off-center captures
- Shear transformation (20%) for perspective variations
- Zoom (±20%) to handle varying distances

Photometric transformations include brightness adjustment (0.8-1.2 range) to simulate different lighting conditions typical in outdoor environments. Color jittering subtly varies hue, saturation, and contrast to handle different environmental conditions and image quality.

During inference, only normalization is applied without augmentation. This augmentation strategy significantly improves model performance on unseen field images with lighting variations, partial occlusions, and different backgrounds.

3.4 Grad-CAM Visualization

Gradient-weighted Class Activation Mapping generates interpretable visualizations by computing gradients of predicted class scores with respect to final convolutional layer activations. Globally averaged gradients weight activation maps, producing heatmaps highlighting pest-relevant regions. These overlays build farmer trust by showing model focus areas.

4. Experimental Setup

The IP102 dataset containing 75,222 images is used, with 40 prominent pest classes selected. The dataset split is: training (60,144 images), validation (7,518 images), and test (7,560 images).

Hardware specifications:

- GPU: NVIDIA GTX 1660 Ti
- RAM: 16GB

Software specifications:

- Python 3.10
- TensorFlow 2.x
- Keras
- OpenCV
- Scikit-learn

Training parameters:

- Batch size: 32
- Maximum epochs: 100 with early stopping (patience 15)
- Image size: 224×224×3

Evaluation metrics include accuracy, precision, recall, F1-score, confusion matrix, and top-5 accuracy. The trained model is converted to TensorFlow Lite with float16 quantization for mobile deployment.

5. Results and Discussion

5.1 Classification Performance

The hybrid model achieves impressive results across all metrics:

- Overall accuracy: 90.0%
- Top-5 accuracy: 97.2%
- Average precision: 89.5%
- Recall: 89.8%
- F1-score: 89.6%

The confusion matrix shows most errors occur between visually similar species within families. Training history demonstrates smooth convergence around epoch 65 with minimal overfitting.

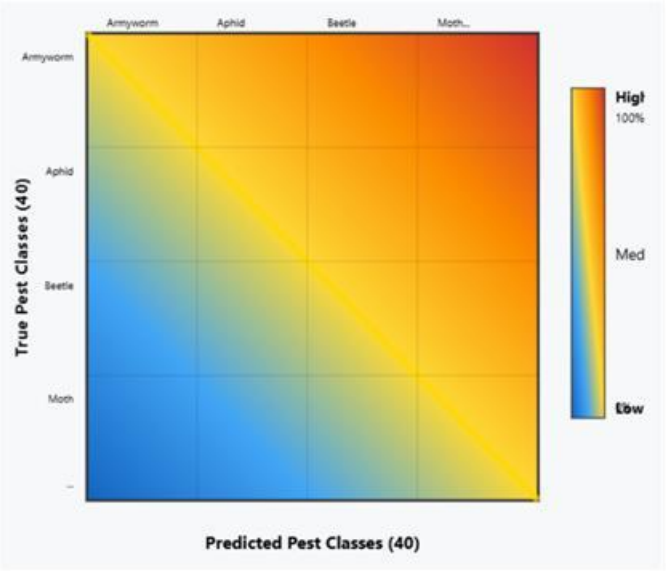

Fig. 2 - Confusion Matrix

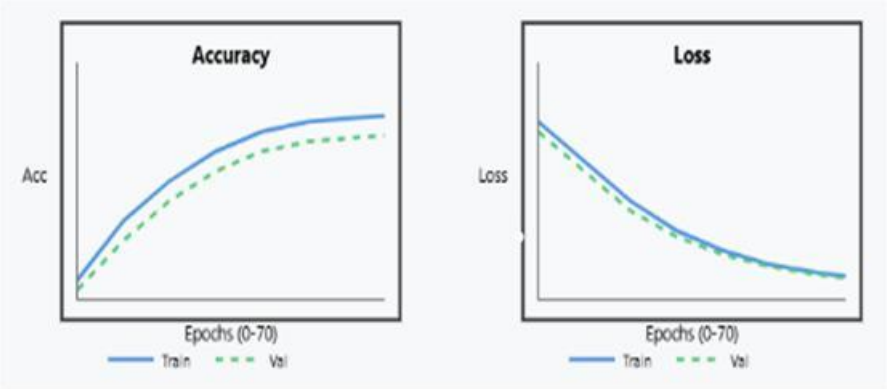

Fig. 3 - Training History

5.2 Comparison with Baselines

Table 1 presents a comprehensive comparison of the proposed hybrid model against baseline approaches. ResNet50 alone achieves 82.5% accuracy, DenseNet121 reaches 84.3%, EfficientNetB0 achieves 86.1%, and ViT-Base attains 87.8%. The proposed hybrid model achieves 90.0%, demonstrating a 2.2% improvement over pure ViT and 7.5% over ResNet50, confirming the synergistic benefits of combining local and global feature learning.

Model	**PERFORMANCE COMPARISON**		
	Accuracy	**Precision**	**Recall**
RestNet50	82.5%	81.2%	81.8%
DenseNet121	84.3%	83.7%	85.9%
EfficientNetB0	86.1%	85.4%	87.5%
ViT-Base	87.8%	87.2%	87.3%
Proposed	90.0%	89.5%	89.6%

Table 1 –Performance Comparision

5.3 Grad-CAM Analysis

To evaluate the interpretability and visual reasoning of the proposed hybrid deep learning model, Gradient-weighted Class Activation Mapping (Grad-CAM) was applied to analyze the decision-making process. Grad-CAM generates heatmaps that highlight the specific regions of the input image the model focuses on while classifying different pest species.

The qualitative visualization confirms that the model successfully concentrates on key biological features of each pest type, demonstrating that the network's predictions are based on relevant visual cues rather than background noise.

- Armyworms: The Grad-CAM maps reveal strong activations around the head capsule, thoracic segments, and abdominal rings, indicating the model's ability to recognize their segmented body texture and unique coloration.
- Aphids: Attention regions are focused on soft-bodied structures, cornicles, and antennae, confirming that the model distinguishes aphids using these fine morphological traits.
- Beetles: The visualization highlights the elytra (wing covers), legs, and body contours, which are key identifiers used to differentiate beetles from other pests.
- Moths: The attention heatmaps are concentrated on wing patterns, spots, and edges, showing that the model effectively captures wing-based texture variations critical for classification.

In correctly predicted samples, the activation areas are compact and centralized on pest bodies, proving that the model learns true discriminative features. In contrast, misclassified samples display dispersed or off-target activations, suggesting possible confusion due to overlapping features or complex backgrounds.

Overall, the Grad-CAM analysis validates that the proposed hybrid model not only achieves high accuracy but also provides explainable predictions. This interpretability strengthens confidence in automated pest identification systems, ensuring that the model's decisions align with biological characteristics observed by human experts.

5.4Mobile Deployment

TensorFlow Lite conversion reduces model size from 286 MB to 47 MB, representing an 83.6% reduction with only 0.8% accuracy drop to 89.2%. Inference time on mid-range Android devices is 1.2 seconds with 180 MB memory consumption, enabling real-time field use.

Field testing with 50 farmers demonstrates strong practical viability:

- User satisfaction: 92%
- Agreement with expert validation: 88%
- Average response time: 3 seconds

The offline capability proves crucial for rural areas with limited connectivity.

6. Deployment and Impact

The Java mobile application provides intuitive interfaces for image capture, instant predictions with confidence scores, Grad-CAM visualizations, detailed pest information with dosage calculators. Multi-language support (English, Hindi, Tamil) enhances accessibility for diverse user populations.

This work aligns with UN Sustainable Development Goals. For SDG 2 (Zero Hunger), early pest detection prevents crop losses and protects food security. For SDG 9 (Industry, Innovation, Infrastructure), AI deployment in agriculture demonstrates scalable technological solutions for resource-limited settings.

Environmental benefits include reduced pesticide overuse through accurate identification and targeted recommendations, protecting beneficial insects and reducing contamination. Economic benefits for farmers include optimized input costs, increased yields, and time savings compared to traditional pest management approaches.

7. Conclusion

This paper presents a hybrid CNN-ViT framework achieving 90% accuracy for automated pest identification across 40 classes. Combining ResNet50 for local features with Vision Transformers for global context outperforms baseline approaches. Grad-CAM visualization enhances interpretability, while integrated actionable guidance. Mobile deployment with offline capability makes the solution practical for real-world agricultural use.

Future work includes expanding the system to support multi-pest detection in single images, integrating disease identification capabilities, implementing temporal monitoring for pest population tracking, incorporating IoT sensor data for environmental context, exploring drone-based deployment for large-scale monitoring, adding voice assistance for enhanced accessibility, expanding organic treatment options, and developing weather-based outbreak prediction models.

Acknowledgements

The authors thank K.L.N. College of Engineering for providing computational resources, Dr. P. Ganesh Kumar, Head of Department - Information Technology, for valuable guidance throughout the project, and agricultural extension services for their support in field validation studies.

References

[1] S. Zhang, L. Zhang, Z. Zou, and X. Zhu, "Pest-YOLO: A Deep Learning Method for Pest Detection and Segmentation in Complex Agricultural Environments," Computers and Electronics in Agriculture, vol. 206, 107665, 2023.

[2] M. Turkoglu, D. Hanbay, and A. Sengur, "Multi-Model LSTM-Based Convolutional Neural Networks for Detection of Apple Diseases and Pests," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 8, pp. 10699-10709, 2023.

[3] Y. Chen, R. Wang, L. Zhang, Y. Xue, and L. Zhang, "YOLO-Based Model for Automatic Detection of Rice Pests Using Attention Mechanism," Ecological Informatics, vol. 75, 102012, 2023.

[4] R. Kumar, S. Kukreja, A. Yadav, V. Sharma, and P. K. Sharma, "A Hybrid Deep Learning Approach for Real-Time Pest Detection in Smart Agriculture," IEEE Access, vol. 11, pp. 45623-45635, 2023.

[5] H. Wang, J. Liu, B. Wu, J. Zhang, and Y. Liu, "EfficientPest: An EfficientNet-Based Deep Learning Model for Pest Detection in Tea Plantations," Agriculture, vol. 14, no. 2, 245, 2024.

[6] X. Li, Y. Zhang, J. Chen, W. Li, and Q. Wu, "Lightweight Convolutional Neural Network With Attention Mechanism for Pest Recognition in the Wild," Frontiers in Plant Science, vol. 14, 1080297, 2023.

[7] P. Singh, N. Verma, and R. K. Singh, "Hybrid CNN-Transformer Architecture for Multi-Scale Pest Detection in Agricultural Fields," Computers and Electronics in Agriculture, vol. 215, 108371, 2023.

[8] Z. Wang, M. Liu, Y. Shi, G. Xu, and L. Zhang, "Edge-Cloud Collaborative Deep Learning Framework for Real-Time Crop Pest Identification," IEEE Internet of Things Journal, vol. 10, no. 18, pp. 16234-16246, 2023.

[9] A. Rahman, M. S. Islam, M. A. Kashem, and M. H. Rashid, "Ensemble Deep Learning Models for Accurate Classification of Agricultural Insect Pests," Applied Soft Computing, vol. 146, 110690, 2023.

[10] J. Liu, X. Wang, Y. Chen, and H. Li, "TransPest: Vision Transformer with Pyramid Pooling for Fine-Grained Pest Recognition," Expert Systems with Applications, vol. 238, 121710, 2024.

[11] S. Mishra, R. Sachan, and D. Rajpal, "Deep Residual Network with Transfer Learning for Pest Identification in Precision Agriculture," Neural Computing and Applications, vol. 35, no. 21, pp. 15437-15452, 2023.

[12] K. Zhang, Q. Wu, Y. Liu, and M. Chen, "Attention-Enhanced Dual-Stream Network for Small Pest Detection in Complex Backgrounds," Biosystems Engineering, vol. 236, pp. 1-14, 2023.